A Method of Reconciling Discrepant National and Sub-National Employment and Unemployment Estimates*

DRAFT, May 19, 2004, Revised August 16, 2004

P.A.V.B. Swamy [a], Jatinder S. Mehta and I-Lok Chang

[a] Bureau of Labor Statistics, Statistical Methods Staff, 2 Massachusetts Ave, N.E., Washington, DC 20212

Key Words: Sample survey, Small area estimation, Errors in the variables model; Heteroscedastic nonlinear regression; Stochastic coefficient estimation; Iteratively rescaled generalized least squares estimation.

Considering two discrepant estimates of employment or unemployment for each of several geographical areas, this paper finds two models of their conditional variations across areas both at a point in time and through time. One of these models is shown to predict the nation's employment or unemployment better than the other model. It improves one of the two estimates and corrects for nonresponse and measurement-error biases and sampling and non-sampling errors in the other estimate for each area. The improved estimate is equal to the corrected estimate.

1. Introduction

In this paper, we consider a situation where we have at least two different estimates of employment or unemployment made from different data sets for each of several small geographical areas. The sources of two of these data sets are the Current Population Survey (CPS) and the American Community Survey (ACS). The problem with CPS data is that for many sub-state areas, they are either unavailable or too sparse. For such areas, Local Area Unemployment Statistics (LAUS) estimates produced by the LAUS Program within the Bureau of Labor Statistics (BLS) will be used in place of the CPS estimates.

Section 2 derives the first two moments of two conditional distributions from a joint distribution of the CPS and ACS estimators across states at a point in time. The mean of one of these two distributions is shown to predict the nation's employment or unemployment better than that of the other distribution. An empirical example is given in Section 3. Section 4 concludes.

2. A Method of Simultaneously Improving One and Correcting Another of Two Estimates

Let Y be the finite population value of a population characteristic for a geographical area at a period. Such a value may not be equal to the value of the population characteristic obtained through a census (100% sample) because (i) while conducting the census, some of the units in the population may not be measured (omissions) or may be measured more than once (duplication) and (ii) measurement, response, editing, coding, and tabulating errors may enter into the census data. For these reasons, we treat Y as an unknown "true value" and try to learn about it from the available data that are not perfect. This concept of a "true value" is not different from Cochran's (1977, p. 377) idea of a "correct value." The population characteristics that are of interest in this paper are employment and unemployment. The geographical areas that are of interest in this paper are states.

Let Y_{it} denote the "true value" of Y for state i at time t. (1)

Let the number of states in the nation be n. Then $Y_t = \sum_{i=1}^{n} Y_{it}$ is the "true value" of national employment or unemployment. The estimators of Y_{it} based on data for time t only from the sample units within area i are called the direct survey estimators. Let $\hat{Y}_{it}^{ACS}$ and $\hat{Y}_{it}^{CPS}$ denote the direct ACS and CPS estimators of Y_{it}, respectively.[1] The ACS is not designed to make monthly estimates, so we can only get ACS estimates of the annual averages of Y_{it}. Let t index years.

2.1 Choosing Between the Means of Two Conditional Distributions Derived from a Joint Distribution of Two Estimators

*Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics. Thanks are due to J.N.K. Rao, Tamara Zimmerman and Edwin Robison for helpful comments.

[1] The ACS is not fully implemented yet. Only smaller-scale versions of the ACS were conducted in 2000, 2001, and 2002. In this paper, these smaller-scale ACS data for 2000 are used to evaluate $\hat{Y}_{it}^{ACS}$.

We write $\hat{Y}_{it}^{CPS} = Y_{it} + \varepsilon_{it}^{CPS}$ and $\hat{Y}_{it}^{ACS} = Y_{it} + \varepsilon_{it}^{ACS}$, where ε_{it}^{CPS} and ε_{it}^{ACS} denote the errors of $\hat{Y}_{it}^{CPS}$ and $\hat{Y}_{it}^{ACS}$, respectively. Here inferential disasters can be avoided by making assumptions about ε_{it}^{CPS} and $\varepsilon_{\text{it}}^{\text{ACS}}$ that are attentive to the CPS- and ACS-design features, respectively. It is true that $\hat{Y}_{it}^{CPS}$ and $\hat{Y}_{it}^{ACS}$ are unbiased in probability sampling. It is also true that adjustments have been made in these estimators for nonresponse and, in addition, measures have been taken to control the presence of various sources of nonsampling errors in the CPS and ACS. However, the approximations involved in these adjustments and measures (see Cochran (1977, Chapter 13)) may not have permitted the complete elimination of all the biases nonresponse and measurement errors have produced in the estimates that are computed from the CPS and ACS data. Because of the difficulty of ensuring that no such approximations are present, we define ε_{it}^{CPS} and ε_{it}^{ACS} broadly to include both sampling and non-sampling errors and assume that these errors may not have zero means. The amounts of biases that are still remaining in $\hat{Y}_{it}^{CPS}$ and $\hat{Y}_{it}^{ACS}$ after these estimators have been adjusted for bias are given by $E\varepsilon_{it}^{CPS}$ and $E\varepsilon_{it}^{ACS}$, respectively, which may not be zero.

The estimator, $\hat{Y}^{CPS}$, is the BLS standard for measuring *Y* at the national level but not at the sub-national level, where the CPS sample size may be too small to yield direct survey estimates with meaningful accuracy. This standard suggests that a *necessary* condition for an estimate y_{it} of Y_{it} to be accurate is that $\sum_{i=1}^{n} y_{it}$ is equal to the CPS estimate of Y_t.

The common component Y_{it} of $\hat{Y}_{it}^{CPS}$ and $\hat{Y}_{it}^{ACS}$ shows that there is a relationship between $\hat{Y}_{it}^{CPS}$ and $\hat{Y}_{it}^{ACS}$. One form of this relationship is obtained by replacing the left-hand side of the identity

$$\mathrm{Y_{it}} = \mathrm{Y_{it}} \qquad (2)$$

by $\hat{\mathrm{Y}}_{\text{it}}^{\text{CPS}} - \varepsilon_{\text{it}}^{\text{CPS}}$ and its right-hand side by $\hat{\mathrm{Y}}_{\text{it}}^{\text{ACS}} - \varepsilon_{\text{it}}^{\text{ACS}}$. Doing so gives

$$\hat{\mathrm{Y}}_{\text{it}}^{\text{CPS}} = \alpha_{0\text{it}}^{\text{CPS}} + \alpha_{1\text{it}}^{\text{ACS}} \hat{\mathrm{Y}}_{\text{it}}^{\text{ACS}} \qquad (3)$$

where $\alpha_{0\text{it}}^{\text{CPS}} = \varepsilon_{\text{it}}^{\text{CPS}}$ and $\alpha_{1\text{it}}^{\text{ACS}} = 1 - (\varepsilon_{\text{it}}^{\text{ACS}} / \hat{\mathrm{Y}}_{\text{it}}^{\text{ACS}})$.

If $\hat{Y}_{it}^{CPS}$ is based on a very small sample, then it is not a very precise estimator. To improve its precision, its relationship with $\hat{Y}_{it}^{ACS}$ in (3) may be used. This relationship can "borrow strength" cross-sectionally, over time, or both, more effectively than the relationship, $\hat{Y}_{it}^{CPS} = Y_{it} + \varepsilon_{it}^{CPS}$ with Y_{it} replaced by a function of some auxiliary variables and random error, when relevant variables are omitted from the latter relationship, when the included auxiliary variables are measured with error, or when the function's true functional form is misspecified. Ignoring the effects of these misspecifications leads to incorrect inferences, as Freedman and Navidi (1986) and Swamy, et al. (2003a, b) show.

A big virtue of model (3) is that by eliminating Y_{it} from (2) it avoids all the misspecification problems associated with the regressions of Y_{it} on auxiliary variables. For example, consider the usual assumption that the auxiliary variables, which are used as some of the determinants of Y_{it} are independent of 'the' excluded determinants themselves. This assumption shown to be either meaningless or false by Pratt and Schlaifer (1984, 1988) is avoided in (3). What (3) cannot avoid, however, is the errors-in-the-variables problem: both its regressand $\hat{\mathrm{Y}}_{\text{it}}^{\text{CPS}}$ and its regressor $\hat{\mathrm{Y}}_{\text{it}}^{\text{ACS}}$ are subject to error. Here, it should be pointed out that to estimate (3), the econometric method of instrumental variables cannot be used because these variables do not exist in the usual situations, where (i) relevant explanatory variables are omitted, (ii) the included explanatory variables are measured with error, and (iii) the unknown functional forms of relevant relationships are misspecified, as shown by Chang, et al. (2000).

The roles of $\hat{\mathrm{Y}}_{\text{it}}^{\text{CPS}}$ and $\hat{\mathrm{Y}}_{\text{it}}^{\text{ACS}}$ in (3) can be interchanged by replacing the left- and right-hand sides of (2) by $\hat{Y}_{it}^{ACS} - \varepsilon_{it}^{ACS}$ and $\hat{Y}_{it}^{CPS} - \varepsilon_{it}^{CPS}$, respectively. Doing so gives

$$\hat{Y}_{it}^{ACS} = \alpha_{0it}^{ACS} + \alpha_{1it}^{CPS} \hat{Y}_{it}^{CPS} \qquad (4)$$

where $\alpha_{0it}^{ACS} = \varepsilon_{it}^{ACS}$ and $\alpha_{1it}^{CPS} = 1 - (\varepsilon_{it}^{CPS} / \hat{Y}_{it}^{CPS})$.

We now show that (3) is preferable to (4). In these models, the regressors are not independent of their respective coefficients. To account for the correlation between α_{1it}^{ACS} and $\hat{Y}_{it}^{ACS}$ in (3), we assume that given $\hat{\mathrm{Y}}_{\text{it}}^{\text{ACS}} = y_{it}^{ACS}$,

$$\alpha_{0it}^{CPS} = \pi_{00}^{CPS} + \pi_{01}^{CPS}\left(\frac{1}{y_{it}^{ACS}}\right) + \zeta_{0it}^{CPS} \quad (5)$$

$$\alpha_{1it}^{ACS} = \pi_{10}^{ACS} + \pi_{11}^{ACS}\left(\frac{1}{y_{it}^{ACS}}\right) + \zeta_{1it}^{ACS} \quad (6)$$

where the vectors, $(\zeta_{0it}^{CPS}, \zeta_{1it}^{ACS})'$, are independent of $\hat{Y}_{it}^{ACS}$ and are independently distributed with mean vector zero and constant covariance matrix $\sigma_1^2\Delta_1$ as i varies for fixed t.[2]

Assumption (5) says that the mean function, $E(\alpha_{0it}^{CPS} | \hat{Y}_{it}^{ACS} = y_{it}^{ACS}) = \pi_{00}^{CPS} + \pi_{01}^{CPS}(1/y_{it}^{ACS})$, with one constant and one variable term, is a proxy for the bias, $E\,\varepsilon_{it}^{CPS}$, and ζ_{0it}^{CPS}, representing the fluctuating component of ε_{it}^{CPS}, is independent of $\hat{Y}_{it}^{ACS}$. It is analogous to one of Cochran's (1977, pp. 377-379) assumptions. The mean function is a good proxy for $E\,\varepsilon_{it}^{CPS} = 0$ because $\pi_{01}^{CPS}(1/y_{it}^{ACS})$ takes only tiny values and zero is a convenient value for π_{00}^{CPS}, as we show in Section 3 below. Good proxies for $E\,\varepsilon_{it}^{CPS} \neq 0$ can be obtained by adding additional regressors to (5). Quite possibly, some of these regressors are the proportions of the population in state i falling in black and Hispanic racial groups with different unemployment rates. Thus, using various proxies for $E\,\varepsilon_{it}^{CPS}$ including the one for $E\,\varepsilon_{it}^{CPS} = 0$ in (5), we can investigate the statistical consequences of various departures from the assumption that $E\,\varepsilon_{it}^{CPS} = 0$. Assumption (6) says that the correlation between α_{1it}^{ACS} and $\hat{Y}_{it}^{ACS}$ is due to the function, $\pi_{10}^{ACS} + \pi_{11}^{ACS}(1/y_{it}^{ACS})$, but once this function is subtracted from α_{1it}^{ACS}, the remainder ζ_{1it}^{ACS} is independent of $\hat{Y}_{it}^{ACS}$. Given that α_{1it}^{ACS} is a function of $1/\hat{Y}_{it}^{ACS}$ and is correlated with $\hat{Y}_{it}^{ACS}$, assumption (6) is reasonable. This assumption may justify assumption (5) because (i) α_{0it}^{CPS} and α_{1it}^{ACS}, being the coefficients of the same equation, may be affected by the same set of variables and (ii) if α_{1it}^{ACS} is affected by $1/\hat{Y}_{it}^{ACS}$, then α_{0it}^{CPS} may also be affected by the same variable.

To account for the correlation between α_{1it}^{CPS} and $\hat{Y}_{it}^{CPS}$ in (4), we assume that given $\hat{Y}_{it}^{CPS} = y_{it}^{CPS}$,

$$\alpha_{0it}^{ACS} = \pi_{00}^{ACS} + \pi_{01}^{ACS}\left(\frac{1}{y_{it}^{CPS}}\right) + \zeta_{0it}^{ACS} \quad (7)$$

$$\alpha_{1it}^{CPS} = \pi_{10}^{CPS} + \pi_{11}^{CPS}\left(\frac{1}{y_{it}^{CPS}}\right) + \zeta_{1it}^{CPS} \quad (8)$$

where the vectors, $(\zeta_{0it}^{ACS}, \zeta_{1it}^{CPS})'$, are independent of $\hat{Y}_{it}^{CPS}$ and are independently distributed with mean vector zero and constant covariance matrix $\sigma_2^2\Delta_2$ as i varies for fixed t.

Inserting (5) and (6) into (3) gives

$$\hat{Y}_{it}^{CPS} = \pi_{00}^{CPS} + \pi_{01}^{CPS}\left(\frac{1}{y_{it}^{ACS}}\right) + \pi_{10}^{ACS} y_{it}^{ACS} + \pi_{11}^{ACS} + \zeta_{0it}^{CPS} + \zeta_{1it}^{ACS} y_{it}^{ACS} \quad (9)$$

where only the sum of π_{00}^{CPS} and π_{11}^{ACS} is identifiable. Equation (9) is a nonlinear regression model with heteroscedastic disturbances. Under (5) and (6), it implies that the conditional distribution of $\hat{Y}_{it}^{CPS}$ given $\hat{Y}_{it}^{ACS} = y_{it}^{ACS}$, has mean equal to the right-hand side of (9) with the last two terms suppressed and variance equal to $\mathrm{var}(\zeta_{0it}^{CPS}) + \mathrm{var}(\zeta_{1it}^{ACS})(y_{it}^{ACS})^2 + 2\mathrm{cov}(\zeta_{0it}^{CPS}, \zeta_{1it}^{ACS})y_{it}^{ACS}$, where the variances and covariance are given by the elements of $\sigma_1^2\Delta_1$.

It follows from (3) that

$$\varepsilon_{it}^{CPS} = \alpha_{0it}^{CPS} \quad (10)$$

$$\varepsilon_{it}^{ACS} = (1 - \alpha_{1it}^{ACS})\hat{Y}_{it}^{ACS} \quad (11)$$

$$Y_{it} = \hat{Y}_{it}^{CPS} - \alpha_{0it}^{CPS} = \alpha_{1it}^{ACS}\hat{Y}_{it}^{ACS} \quad (12)$$

Equation (12) gives the true value, Y_{it}, by eliminating the differences between $\hat{Y}_{it}^{CPS}$ and $\hat{Y}_{it}^{ACS}$. Summing it over i gives

$$\sum_{i=1}^{n}\left(\hat{Y}_{it}^{CPS} - \alpha_{0it}^{CPS}\right) = \sum_{i=1}^{n}\alpha_{1it}^{ACS}\hat{Y}_{it}^{ACS} \quad (13)$$

The left- and right-hand sides of this equation give two estimators of Y_t that are equal to each other with probability (w.p.) 1. If, in addition, $\sum_{i=1}^{n}\alpha_{0it}^{CPS} = 0$ w.p. 1, then these two estimators are equal to the CPS estimator, $\sum_{i=1}^{n}\hat{Y}_{it}^{CPS}$. It is desirable to obtain such

[2] The time subscript t is fixed at a value because currently, the ACS data are available only for the years 2000-2002.

estimators if $|\sum_{i=1}^{n}\hat{Y}_{it}^{CPS} - Y_t| < |\sum_{i=1}^{n}\hat{Y}_{it}^{ACS} - Y_t|$ w.p. 1. The BLS belief in this condition is very high. Thus, under the condition, $\sum_{i=1}^{n}\alpha_{0it}^{CPS} = 0$ w.p. 1, the estimators on both sides of the second equality sign in (12) satisfy both the BLS standard and the implied necessary condition (stated at the end of the second paragraph of Section 2.1) for them to be precise. The condition, $\sum_{i=1}^{n}\alpha_{0it}^{CPS} = 0$ w.p. 1, is nearly satisfied if the sums of $\pi_{00}^{CPS} + \pi_{01}^{CPS}(1/y_{it}^{ACS})$ and ζ_{0it}^{CPS} over i are near zero w.p. 1.

If we consider the equation set, {(4), (7), (8)}, then (13) changes to

$$\sum_{i=1}^{n}\left(\hat{Y}_{it}^{ACS} - \alpha_{0it}^{ACS}\right) = \sum_{i=1}^{n}\alpha_{1it}^{CPS}\hat{Y}_{it}^{CPS} \quad (14)$$

which is equal to the BLS preferred estimator $\sum_{i=1}^{n}\hat{Y}_{it}^{CPS}$ if $\alpha_{1it}^{CPS} = 1$ w.p. 1 for all i and t. This condition is very strong and is unlikely to be satisfied in practice. Therefore, we can conclude that equation (4) usually gives an estimate of Y_t that is different from its BLS preferred (CPS) estimate. Since it is easier to satisfy the condition, $\sum_{i=1}^{n}\alpha_{0it}^{CPS} = 0$ w.p. 1, than to satisfy the condition, $\alpha_{1it}^{CPS} = 1$ w.p. 1 for all i and t, we can easily obtain an estimate of Y_t equal to its CPS estimate by using the set, {(3), (5), (6)}, which is equivalent to (9), instead of the set, {(4), (7), (8)}. This is the reason why we prefer (9) to the latter set.

Under (7) and (8), (4) gives the conditional distribution of $\hat{Y}_{it}^{ACS}$ given $\hat{Y}_{it}^{CPS} = y_{it}^{CPS}$, whose mean is equal to

$$E(\hat{Y}_{it}^{ACS} \mid \hat{Y}_{it}^{CPS} = y_{it}^{CPS}) = \pi_{00}^{\mathrm{ACS}} + \pi_{01}^{ACS}\left(\frac{1}{y_{it}^{CPS}}\right) + \pi_{10}^{CPS}y_{it}^{CPS} + \pi_{11}^{CPS} \quad (15)$$

and whose variance is equal to $\mathrm{var}(\zeta_{0it}^{ACS}) + \mathrm{var}(\zeta_{1it}^{CPS})(y_{it}^{CPS})^2 + 2\mathrm{cov}(\zeta_{0it}^{ACS}, \zeta_{1it}^{CPS})y_{it}^{CPS}$, where the variances and covariance are given by the elements of $\sigma_2^2\Delta_2$.

Given the data, $(y_{it}^{CPS}, y_{it}^{ACS})$ for i = 1, 2, ..., n and fixed t, an Iteratively Re-Scaled Generalized Least Squares (IRSGLS) method is used to obtain good approximations to the minimum variance linear unbiased estimators of the coefficients, π_{01}^{CPS}, π_{10}^{ACS}, $(\pi_{00}^{CPS} + \pi_{11}^{ACS})$, and the best linear unbiased predictors of the errors, ζ_{0it}^{CPS} and ζ_{1it}^{ACS}, of equation (9).[3] These approximations are denoted by $\hat{\pi}_{01}^{CPS}$, $\hat{\pi}_{10}^{ACS}$, $\widehat{(\pi_{00}^{CPS} + \pi_{11}^{ACS})}$, $\hat{\zeta}_{0it}^{CPS}$, and $\hat{\zeta}_{1it}^{ACS}$, respectively. The corresponding estimate of $\sigma_1^2\Delta_1$ is denoted by $\hat{\sigma}_1^2\hat{\Delta}_1$. Arbitrary prior values of the unknown parameters have little or no influence on these estimates. Without (5) and (6) these IRSGLS estimators are inconsistent because a *necessary* condition for their consistency and asymptotic normality is that ζ_{0it}^{CPS} and ζ_{1it}^{ACS} are independent of $\hat{Y}_{it}^{ACS}$ and other variables included on the right-hand side of (5) and (6). Sufficient conditions for their consistency and asymptotic normality have been worked out in the econometrics literature.

2.2 Bias- and Error-Corrected Version of the CPS Estimator

Equations (5), (10), and (12) can be used to show that given $\hat{Y}_{it}^{ACS} = y_{it}^{ACS}$ and $\pi_{00}^{CPS} = 0$,

$$E(\hat{Y}_{it}^{CPS} - Y_{it})^2 = \left(\pi_{01}^{CPS}\frac{1}{y_{it}^{ACS}}\right)^2 + \mathrm{var}(\zeta_{0it}^{CPS}) \quad (16)$$

where the first term on the right-hand side is the square of the bias, $E\varepsilon_{it}^{CPS}$. Adding additional regressors on the right-hand side of (5) may push the magnitude of this bias above that of $\pi_{01}^{CPS}\frac{1}{y_{it}^{ACS}}$ and decrease the magnitude of the fluctuating component of ε_{it}^{CPS}, below that of ζ_{0it}^{CPS}. An estimate of $E\varepsilon_{it}^{CPS}$ is

$$\widehat{Bias}^{CPS} = \hat{\pi}_{01}^{CPS}\frac{1}{y_{it}^{ACS}} \quad (17)$$

The standard error of (17) is given by the square root of

$$\mathrm{var}(\hat{\pi}_{01}^{CPS})\left(\frac{1}{y_{it}^{ACS}}\right)^2 \quad (18)$$

From equations (5) and (12) we have

$$\hat{Y}_{it}^{BECCPS} = \hat{Y}_{it}^{CPS} - \hat{\alpha}_{0it}^{CPS} = \hat{Y}_{it}^{CPS} - \hat{\pi}_{01}^{CPS}\left(\frac{1}{y_{it}^{ACS}}\right) - \hat{\zeta}_{0it}^{CPS}, \quad (19)$$

if $\pi_{00}^{CPS} = 0$. This is what we call bias- and error-corrected (BEC) version of the CPS

[3] Chang, et al. (2000) developed software for the IRSGLS method.

estimator, $\hat{Y}_{it}^{CPS}$. It is close to $\hat{Y}_{it}^{CPS}$ if $\hat{\alpha}_{0it}^{CPS}$ is close to 0 w.p. 1. Estimator (19) can be more precise than $\hat{Y}_{it}^{CPS}$.

2.3 Improved ACS Estimator

An estimator of Y_{it} that gives a good approximation to (12) is

$$\hat{Y}_{it}^{IACS} = \hat{\alpha}_{1it}^{ACS} y_{it}^{ACS} = \left[\hat{\pi}_{10}^{ACS} + \hat{\pi}_{11}^{ACS} \left(\frac{1}{y_{it}^{ACS}} \right) + \hat{\zeta}_{1it}^{ACS} \right] y_{it}^{ACS} \qquad (20)$$

where $\hat{Y}_{it}^{IACS}$ denotes the improved ACS (IACS) estimator and $\hat{\pi}_{11}^{ACS}$ is the IRSGLS estimator of π_{11}^{ACS} when π_{00}^{CPS} is restricted to be zero. The standard error of (20), denoted by se($\hat{Y}_{it}^{IACS}$), is given by the square root of the conditional variance of $\hat{Y}_{it}^{IACS}$ given $\hat{Y}_{it}^{ACS} = y_{it}^{ACS}$, which is

$$\begin{aligned} &\mathrm{var}(\hat{\pi}_{10}^{ACS})(y_{it}^{ACS})^2 + \mathrm{var}(\hat{\pi}_{11}^{ACS}) + \mathrm{var}(\hat{\zeta}_{1it}^{ACS}) \times \\ &(y_{it}^{ACS})^2 + 2\mathrm{cov}(\hat{\pi}_{10}^{ACS}, \hat{\pi}_{11}^{ACS}) y_{it}^{ACS} + \\ &2\mathrm{cov}(\hat{\pi}_{10}^{ACS}, \hat{\zeta}_{1it}^{ACS})(y_{it}^{ACS})^2 \\ &+ 2\mathrm{cov}(\hat{\pi}_{11}^{ACS}, \hat{\zeta}_{1it}^{ACS}) y_{it}^{ACS} \end{aligned} \qquad (21)$$

2.4 Comparison of CPS, ACS, Improved ACS, and Bias- and Error-Corrected CPS Estimators

From (11) it follows that the amount of bias in $\hat{Y}_{it}^{ACS}$ is $E\varepsilon_{it}^{ACS} = E(1 - \alpha_{1it}^{ACS})\hat{Y}_{it}^{ACS}$. An idea of the magnitudes of ε_{it}^{ACS} and $E\varepsilon_{it}^{ACS}$ can be obtained by examining the estimate, $\hat{\varepsilon}_{it}^{ACS} = (1 - \hat{\alpha}_{1it}^{ACS}) y_{it}^{ACS}$, of (11) and the estimate of $E(1 - \alpha_{1it}^{ACS})\hat{Y}_{it}^{ACS}$ equal to $y_{it}^{ACS} - \hat{\pi}_{10}^{ACS} y_{it}^{ACS} - \hat{\pi}_{11}^{ACS}$ when $\pi_{00}^{CPS} = 0$.

The second equality in (12) is preserved under the IRSGLS estimation of (9). Consequently, (19) and (20) have the same conditional distribution given $\hat{Y}_{it}^{ACS} = y_{it}^{ACS}$ and yield the same estimate of Y_{it} when they are evaluated at the IRSGLS estimates of π's and ζ's in (9). They both are close to $\hat{Y}_{it}^{CPS}$ if $\hat{\alpha}_{0it}^{CPS}$ is close to 0 w.p. 1. Thus, (20) resolves the difficult problem of choosing between the two discrepant estimates of Y_{it} given by the CPS and the ACS.

The previous studies found that for the year 2000, the amount of bias in $\hat{Y}_{it}^{ACS}$ exceeds that in $\hat{Y}_{it}^{CPS}$ for a majority of states. Consequently, $\hat{Y}_{it}^{ACS}$ without the improvements considered in Section 2.3 is not preferable to $\hat{Y}_{it}^{CPS}$. Now we need to compare (20) with $\hat{Y}_{it}^{CPS}$. The conditional distributions of (20) and $\hat{Y}_{it}^{CPS}$ given $\hat{Y}_{it}^{ACS} = y_{it}^{ACS}$ across i for fixed t are called the conditional cross-sectional distributions. Suppose that F_{it}^{IACS}, μ_{it}^{IACS}, and ω_{it}^{IACS} (= (21)) denote the conditional cross-sectional distribution function of (20), its mean and variance, respectively. Let F_{it}, μ_{it}, and ω_{it} denote the conditional cross-sectional distribution function of $\hat{Y}_{it}^{CPS}$, its mean and variance, respectively. Under (5), $\mu_{it} = \pi_{00}^{CPS} + \pi_{01}^{CPS}(1/y_{it}^{ACS}) + Y_{it}$ and $\omega_{it} = \mathrm{var}(\zeta_{0it}^{CPS})$ if Y_{it} is fixed. Unlike ω_{it}, the CPS design variance of $\hat{Y}_{it}^{CPS}$ does not take into account the variation in $\hat{Y}_{it}^{CPS}$ across i and hence is not equal to ω_{it}.

We say that (20) is better than $\hat{Y}_{it}^{CPS}$ if

$$\begin{aligned} &F_{it}^{IACS}(y_{it} + \mu_{it}^{IACS}) - F_{it}^{IACS}(-y_{it} + \mu_{it}^{IACS}) \geq \\ &F_{it}(y_{it} + \mu_{it}) - F_{it}(-y_{it} + \mu_{it}) \end{aligned} \qquad (22)$$

for each y_{it}. A *necessary* condition for (22) to hold is that $\omega_{it}^{IACS} \leq \omega_{it}$ (see Rao (1973, p. 96)). It is satisfied for fixed Y_{it} if (21) $\leq$ var(ζ_{0it}^{CPS}). The difficulty here is that the population values of ω_{it}^{IACS} and ω_{it} are always unknown and the condition, $\omega_{it}^{IACS} \leq \omega_{it}$, may not be false when the sample estimates of ω_{it}^{IACS} and ω_{it} violate it. Furthermore, to verify the condition, $\omega_{\mathrm{it}}^{\mathrm{IACS}} \leq \omega_{\mathrm{it}}$, if we use the same data that we have used to estimate the unknown quantities of (9), then we will be judging the validity of our estimated model (9) by the data from which we have derived it (see Friedman and Schwartz (1991, pp. 46-48))! This is not the right thing to do. Therefore, the practical verification of the condition, $\sum_{i=1}^{n} \alpha_{0it}^{CPS} = 0$, given below (13) gives more useful information about the relative accuracies of the estimates given by (20) and $\hat{Y}_{it}^{CPS}$ than that of the inequality, $\omega_{it}^{IACS} \leq \omega_{it}$.

A *necessary* condition for (20) to *stochastically dominate* $\hat{Y}_{it}^{CPS}$ is that ω_{it}^{IACS} +

$(\mu_{it}^{IACS} - Y_{it})^2 \leq \omega_{it} + (\mu_{it} - Y_{it})^2$ (see Rao (1973, p. 315, (5a.1.2)) and Lehmann and Casella (1998, p. 400, Problem 4.14)). It is possible that $(\mu_{it}^{IACS} - Y_{it})^2 \leq (\mu_{it} - Y_{it})^2$ because (20) is equal to (19), which is corrected for bias.

Changing the set of regressors in (5) and (6) changes the values of μ_{it}^{IACS}, μ_{it}, ω_{it}^{IACS}, and ω_{it}. One way of limiting the effects of misspecifications in (5) and (6) is to restrict attention to those regressors, which when included in (5) and (6), yield close approximations to design consistent estimates. This can be accomplished by using in (5) and (6) those regressors that make (20) lie between $\hat{Y}_{it}^{CPS}$ - 2se($\hat{Y}_{it}^{CPS}$) and $\hat{Y}_{it}^{CPS}$ + 2se($\hat{Y}_{it}^{CPS}$), since the estimator, $\sum_{i=1}^{n} \hat{Y}_{it}^{CPS}$, is design consistent.

3. Example

Let Y_{it} denote the annual average unemployment for t = 2000 and let i index the 50 states and the District of Columbia. Following the IRSGLS method and using the CPS and ACS estimates, $(y_{it}^{CPS}, y_{it}^{ACS})$, of 51 Y_{it}'s given in columns (B) and (D) of Table 1, we obtain the following estimates for model (9):

$$y_{it}^{CPS} = \underset{(0.12511)}{-0.064101}\left(\frac{1}{y_{it}^{ACS}}\right) + \underset{(0.03034)}{0.76659}\, y_{it}^{ACS} - \underset{(4199.8)}{2151.8} + \hat{\zeta}_{0it}^{CPS} + \hat{\zeta}_{1it}^{ACS} y_{it}^{ACS} \qquad (23)$$

where the figures appearing in parentheses below the coefficient estimates are asymptotic (large sample) standard errors and $\pi_{00}^{CPS} = 0$. The realized value of the design coefficients of variation (CV) of CPS estimates are given in column (C) of Table 1.

The estimate of the coefficient on $(1/y_{it}^{ACS})$ in equation (23), although has the right sign, is insignificant. This shows that the estimate of the bias, $E\varepsilon_{it}^{CPS}$, of the CPS estimate, y_{it}^{CPS}, based on (17) is insignificant. The estimated values of $E\varepsilon_{it}^{CPS}$ lie between -5.8E-06 and -6.5E-08 and those of $E\varepsilon_{it}^{ACS}$ lie between 4714.6 and 232926.6; for the formula used to estimate $E\varepsilon_{it}^{ACS}$, see Section 2.4. Thus, under (5) and (6), the estimated absolute biases of $\hat{Y}_{it}^{ACS}$ are much larger than those of $\hat{Y}_{it}^{CPS}$, which are close to zero.

In (23), $-37760 \leq \hat{\zeta}_{0it}^{CPS} \leq 34198$ and $-0.15115 \leq \hat{\zeta}_{1it}^{ACS} \leq 0.16419$. When (5) and (6) are evaluated at the estimates in (23), we obtain $-37760 \leq \hat{\alpha}_{0it}^{CPS} \leq 34198$ and $0.55068 \leq \hat{\alpha}_{1it}^{ACS} \leq 0.90853$. All the estimates of α_{1it}^{ACS} have the right sign. Inserting the estimates of α_{0it}^{CPS} and α_{1it}^{ACS} into (10) and (11) gives $-37760 \leq \hat{\varepsilon}_{it}^{CPS} \leq 34198$ and $4933.534 \leq \hat{\varepsilon}_{it}^{ACS} \leq 139309.2$. The estimate of the error, $\hat{\varepsilon}_{it}^{CPS} = \hat{\alpha}_{0it}^{CPS}$, of the CPS estimate, y_{it}^{CPS}, is different from that of $E\varepsilon_{it}^{CPS}$ because the IRSGLS estimate of ζ_{0it}^{CPS} is large. Also, $|\hat{\varepsilon}_{it}^{ACS}| > |\hat{\varepsilon}_{it}^{CPS}|$ for 45 states. For Hawaii and North Carolina, $|\hat{\varepsilon}_{it}^{ACS}|$ is more than 170 and 269 times $|\hat{\varepsilon}_{it}^{CPS}|$, respectively. Also, for some states, such as DE, KS, MT, RI, VT, and WY, $\hat{\varepsilon}_{it}^{CPS}$ appears to be large in absolute value even though the CPS and ACS estimates are very close to each other. However, such large values do not arise when the Blacks/Population and Hispanics/Population ratios for state i are included in (5) and (6) as additional regressors (see column (I) of Table 1).

The formulas (20) and (21) were used to obtain the estimates, IACS1, in column (E) and their standard errors in column (F) of Table 1, respectively. For each state, the IACS1 estimate is equal to the bias- and error-corrected CPS estimate, as shown in Section 2.4, but is smaller than the ACS estimate, y_{it}^{ACS}. The latter result is obtained because the ACS data classified far more people as unemployed than the CPS data did for 2000.

All these results arise as a direct consequence of (5) and (6). Including in (5) and (6) the Blacks/Population and Hispanics/Population ratios for state i as additional regressors generally brings the estimates based on $\hat{Y}_{it}^{CPS} - \hat{\alpha}_{0it}^{CPS}$ much closer to y_{it}^{CPS} by reducing substantially the fluctuating component of $\hat{\alpha}_{0it}^{CPS}$ without substantially increasing the bias of $\hat{Y}_{it}^{CPS}$ (see columns (E), (G), and (I) of Table 1).

The estimate of $-\sum_{i=1}^{51} \hat{\alpha}_{0it}^{CPS}$ given at the bottom of column (E) of Table 1 is -5 showing

that the sum, $\sum_{i=1}^{51}\left(y_{it}^{CPS} - \hat{\alpha}_{0it}^{CPS}\right)$, of the state BECCPS estimates based on (19) is equal to the sum, $\sum_{i=1}^{51} y_{it}^{CPS}$, of the state CPS estimates. This means that under (5) and (6), both the IACS1 and the BECCPS estimates of Y_{it}'s cohere with the national CPS estimate (of $\sum_{i=1}^{51} Y_{it}$), which is regarded as the BLS standard. This coherency property, however, does not exactly hold if the ratios of blacks and Hispanics to the total population in state i are included in (5) and (6) as additional regressors (see columns (G) and (I) of Table 1).

The CV of IACS1 estimates in column (F) of Table 1 are very high for many states. However, they get reduced for all states except DC if the proportion of state i's population who were black is included in (5) and (6) as a third regressor (see column (H) of Table 1). The range of the CV of IACS1 or IACS2 estimates is reduced when the proportion of state i's population who were Hispanic is included in (5) and (6) as a fourth regressor (see column (J) of Table 1). For each state, the CV in columns (F), (H), and (J) are larger than the CV in column (C). This result should not be regarded as a demonstration that the CPS estimates are more accurate than the IACS3 estimates because the CV of the CPS estimates ignore the variability of $\hat{Y}_{it}^{CPS}$ across i and those of IACS3 take into account the variability of the cross-sectional estimates of coefficients and error terms in (23). The IACS3 estimates can be more accurate than the other estimates presented in Table 1 because for all states except one, they reside between the lower and upper 95% confidence limits of the corresponding CPS estimate.

4. Conclusions

Even with estimators that are unbiased in probability sampling, nonresponse and measurement errors may produce biases in the estimates that we are able to compute from the data. Therefore, it is important that appropriate adjustments for bias are made in the estimators and, in addition, measures are taken to control the presence of various sources of nonsampling errors in the surveys. If these adjustments and measures are inaccurate, then the estimates that we are able to compute from the data may still be biased and may contain nonsampling errors besides sampling errors. This paper develops good methods of removing these biases and errors. The method can be used to remove the differences between any pair of discrepant national and sub-national employment or unemployment estimates. The method's practical behavior is demonstrated on a real dataset.

References

Chang, I., Swamy, P.A.V.B., Hallahan, C. and Tavlas, G.S. (2000), "A Computational Approach to Finding Causal Economic Laws," *Computational Economics*, 16, 105-136.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd edition, New York: John Wiley & Sons.

Freedman, D.A. and Navidi, W.C. (1986), "Regression Models for Adjusting the 1980 Census (with discussion)," *Statistical Science*, 1, 3-39.

Friedman, M. and Schwartz, A.J. (1991), "Alternative Approaches to Analyzing Economic Data," *The American Economic Review*, 81, 39-49.

Lehmann, E.L. and Casella, G. (1998), *Theory of Point Estimation*, 2nd edition, New York: Springer.

Rao, C.R. (1973), *Linear Statistical Inference and its Applications*, 2nd edition, New York: John Wiley & Sons.

Swamy, P.A.V.B., Chang, I., Mehta, J.S. and Tavlas, G.S. (2003a), "Correcting for Omitted-Variable and Measurement-Error Bias in Autoregressive Model Estimation with Panel Data," *Computational Economics*, 22, 225-253.

Swamy, P.A.V.B., Tavlas, G.S. and Chang, I. (2003b), "How Stable are Monetary Policy Rules: Estimating the Time-Varying Coefficients in Monetary Policy Reaction Function for the U.S." *Computational Statistics & Data Analysis*, forthcoming.

Table 1: Direct and Improved Estimates of Unemployment for the 50 States and the D.C. During 2000

StAbbr	CPS	se(CPS)% / CPS	ACS	IACS1	se(IACS1)% / IACS1	IACS2	se(IACS2)% / IACS2	IACS3	se(IACS3)% / IACS3
(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)
AL	96978	8.10	134383	103061	19.08	97635	11.67	96229	13.09
AK	21373	6.76	24674	16059	31.48	17295	18.77	21871	11.29
AZ	98085	7.98	150382	114869	19.15	111060	9.77	107889	13.70

AR	55278	8.13	80232	61855	19.26	59559	10.79	56548	12.58
CA	835331	3.10	988710	849401	17.23	839316	8.59	843676	12.17
CO	65042	9.46	94290	73814	18.82	72539	9.47	71309	12.22
CT	40064	12.27	74750	63682	17.51	57199	9.96	46463	14.23
DE	16428	8.68	16947	10274	43.73	10604	32.75	15578	14.05
DC	17653	7.26	22875	15065	32.57	13070	41.56	16549	25.15
FL	280793	4.50	396620	269741	21.64	279558	10.19	286078	12.49
GA	155576	7.56	193889	158667	17.89	148949	11.23	153768	12.08
HI	25267	8.99	35702	25205	24.39	26479	13.30	28458	11.76
ID	31969	7.31	37291	25082	25.22	27459	12.95	33259	10.53
IL	280926	4.65	347670	291055	17.56	280375	8.92	282583	10.86
IN	99817	9.05	123002	86517	20.82	95087	9.26	99634	10.85
IA	41326	10.21	63321	48732	19.69	48883	10.02	44555	12.78
KS	51761	8.83	48841	29737	25.92	35227	11.84	48875	8.96
KY	81043	8.35	109211	81980	19.55	83431	9.43	81721	11.83
LA	110592	7.26	146198	109751	19.49	106226	12.90	108616	13.47
ME	23772	9.54	27220	17853	29.59	19466	16.62	23679	11.09
MD	106682	8.67	139970	104481	19.60	102609	11.97	105612	12.68
MA	87701	7.45	113099	82078	20.21	87170	9.33	89387	11.18
MI	182848	5.55	263720	169488	22.82	185245	10.21	184208	12.58
MN	90958	8.89	95724	56761	24.84	74155	9.41	88462	9.69
MS	74423	7.25	99271	73591	19.84	69673	15.14	72102	14.72
MO	101860	8.85	136955	103334	19.39	103627	9.49	102270	11.76
MT	23648	7.48	23631	14845	33.47	16473	19.18	22366	10.60
NE	28120	9.43	29627	19141	28.80	20975	16.09	27625	10.30
NV	42457	7.86	60214	44682	20.55	45513	10.52	48483	11.93
NH	19183	10.77	21918	14043	34.40	15263	20.59	19473	11.58
NJ	159974	5.46	220139	155354	20.76	161976	9.75	163761	11.91
NM	42331	7.52	63475	48445	19.85	48945	10.03	54963	14.42
NY	418855	3.55	507074	431383	17.33	420638	8.71	420826	10.77
NC	149668	6.48	198593	149485	19.45	148484	10.35	149459	11.97
ND	10520	9.61	14495	8826	49.42	9501	32.46	12475	14.60
OH	232986	5.15	279842	248530	16.52	230668	8.62	231245	10.70
OK	50532	9.51	96741	87892	16.20	73528	9.56	56527	15.05
OR	88361	7.66	112952	81358	20.36	87991	9.35	90663	11.24
PA	251237	4.79	316102	258742	17.95	250625	8.96	250817	11.14
RI	22249	8.71	23709	15114	32.91	16394	19.81	22670	10.93
SC	75102	8.90	98133	71562	20.18	70365	13.12	73496	13.08
SD	9447	10.48	13000	7689	55.87	8308	37.14	11480	15.36
TN	110409	8.46	164329	118360	20.31	118070	10.26	111752	12.98
TX	440736	4.14	572323	442125	19.10	440248	9.29	450006	13.15
UT	37152	8.80	66133	54103	18.44	51719	9.81	43610	13.79
VT	9695	10.49	10980	6046	69.85	6641	46.65	10523	16.05
VA	79094	10.49	150989	116853	18.90	101365	11.43	83787	16.07
WA	159026	7.78	161811	133407	17.74	139153	8.39	156877	9.25
WV	44562	7.12	51111	34347	23.28	38074	11.10	43615	10.55
WI	105419	8.67	123719	83455	21.71	96977	9.13	104644	10.47
WY	10360	8.39	11909	6742	63.10	7384	41.85	11857	14.99
Total	5694669		7357896	5694664		5677173		5782378	
			1663227	-5		-17496		87709	

Note: se = standard error. The IACS1-3 are the improved ACS estimates obtained from equation (3). The IACS1 were obtained by subjecting the coefficients of (3) to equations (5) and (6), respectively. The IACS2 were obtained by subjecting the coefficients of (3) to (5) and (6) with Blacks/Population as an additional regressor, respectively. The IACS3 were obtained by subjecting the coefficients of (3) to (5) and (6) with Blacks/Population and Hispanics/Population as additional regressors, respectively. The figures appearing below the totals are the differences between the totals and the CPS total.