THE INFLUENCE OF IMPLICIT METHODOLOGIES IN THE ANALYSIS OF THE CCO AND THE CPS

Yves Thibaudeau, Donna Kostanich, Randall Parmer
U.S Bureau of the Census, Washington DC 20233

Introduction

The Current Population Survey (CPS) and the CATI-CAPI Overlap Survey (CCO) both yield estimates of the unemployment rate and other labor force characteristics. The two surveys exhibit differences in collection methodology and in estimation methodology. The differences in collection methodology are well documented. Polivka (1994) reports on the overall effect of these differences. Kostanich and Cahoon (1994) investigate a variety of design differences. Thompson (1994) specifically looks at mode of collection. The objective of this paper is to delineate the effects of the difference in estimation methodology.

Background

The CPS samples approximately 59,000 households each month. Prior to January 1994 centralized interviewing was used for 9% of the households in the survey. A centralized interview is conducted using computer assisted interviewing and the standard questionnaire. All the other CPS interviews are conducted using the standard questionnaire on paper. From September 1992 to December 1993, the CCO was being conducted in parallel to the CPS. The motivation for conducting the CCO was to obtain a benchmark of the new methodology. The new methodology was later implemented in the CPS. The CCO has approximately 12,000 housing units in sample each month and 18% of the interviews are centralized. All the CCO interviews are computer assisted. In addition, a new questionnaire is used for all the CCO interviews. There are several other important differences between the collection methodologies of the CPS and the CCO. The reader is referred to Kostanich (1994) and Polivka (1994).

The CCO was designed to produce benchmark estimates of labor force characteristics, in order to measure differences in collection methodologies between the CCO and the CPS. Accordingly, the estimation process used for the CCO should be the same process that is used for the CPS. However, because of sample size differences between the two surveys, the estimation methodologies differ slightly.

Stratification and Post-Stratification Differences Between the CCO and the CPS.

The CPS was designed to produce estimates at the state level. Hence the primary strata for the CPS are defined by the states boundaries. Unfortunately, the restricted size of the CCO does not allow for the luxury of accurate estimates at the state level. For the CCO, the primary strata are defined by the boundaries of the regions.

Associated with a stratification at the state level is a set of adjustments, also at the state level, to reconciliate CPS population estimates with independent population estimates derived from the Decennial Census. For the CCO, these adjustments are bypassed. This is one of the differences in estimation methodology. The effects of this difference on the national unemployment estimate is one of the concerns addressed by the paper.

Another discrepancy between the estimation processes of the two surveys may be observed in the post-stratification patterns. Indeed, the demographic post-stratification is defined after observing the size of various demographic subpopulations. For example, for the CPS, it may be decided to create post-strata defined by three subpopulations: Black males age 18-19, black males age 20-24, and black males age 25-29. Then adjustments to reconciliate the population estimates for those categories with Census based population estimates are carried through. However, for the CCO, the size of the sample subpopulations may be too small, sometimes zero, to warrant post-stratification on these subpopulations. The three post-strata may have to be merged in one bigger post-stratum: Black male age 18-29. Then the reconciliation with the Census based figures is carried through only at the level of this larger post-stratum. This estimation difference may affect the estimates of the labor force characteristics, not only for the categories involved in the merging of post-strata, but for the entire population as well. This paper explores this issue.

Providing a Basis to Assess the Effects of the Estimation Differences

The preceding section explains why the CPS estimation process cannot be readily implemented with CCO data. The bottom line is that the difference in estimation methodology between the two surveys may be contributing to the overall difference between the CCO and CPS, in terms of estimates. Since the estimation methodology couldn't be controlled for, it was decided to test for it separately. Note that the CCO estimation process can be made to function with CPS data: The state adjustment in the CPS processing can be bypassed, as it is done for the CCO. Moreover, the demographic post-stratification pattern of the CCO is expected to be inclusive with respect to the post-stratification pattern of the CPS. It is assumed that, since the CPS is larger, the demographic post-strata are finer and selected CPS post-strata can be merged together to recreate any CCO post-stratum.

Recall that the objective of this research is to assess the differences in expected values due to differences in estimation methodologies between the CPS and CCO. Since it is possible to replicate the CCO estimation methodology, but using CPS observations, the following experiment can be set-up:

- Process the CCO as usual. Record the post-stratification pattern.
- Process the CPS as usual. Keep the final weights.
- Process the CPS with modifications: Bypass the state level adjustments and force the post-stratification pattern of the CCO on the CPS. This is the Modified CPS process (MCPS). Keep the final weights of the MCPS.
- Examine the relationship between the set of final weights corresponding to the CPS and the set of final weights corresponding to the MCPS. Establish a relation between the two sets.

Analyzing The Results

The remainder of the paper reports on experimental results obtained from CPS data for the month of January 1993. For simplicity, the paper focus on measuring the unemployment rate. Accordingly, attention is confined to the part of the CPS and the MCPS representing the population in the labor force. The remainder of the population is ignored. The labor force portion of the CPS and the MCPS is made of the 109,108 individuals. Figure 1 shows the MCPS weights versus the CPS weights for the 109108 individuals in the labor force. The graph suggests a linear relation between the two sets of weights. The following general model is formulated:

$$E\{W_{MCPS}\} = A * W_{CPS} + B \qquad (1)$$

Model 1 express a simple linear transformation from the CPS weights to the weights of the MCPS, plus a random noise centered at zero. The parameters of the model are easily estimated:

$$\hat{A} = .9536$$

$$\hat{B} = 103.7$$

$$\hat{\sigma} = 251.28$$

$$\overline{W}_{MCPS} = 1787.41$$

$$R^2 = .9295$$

The estimated relationship between the weights of the MCPS and those of the CPS is represented by a line with a positive slope less than 1, and with a positive value at the origin. The following interpretation is deduced: The high (above average) weights of the CPS are mapped to lower weights for the MCPS, while the low (below average) weights of the CPS are mapped to higher weights for the MCPS.

The validity of model 1 was also assessed: Figure 1 suggests using variance stabilizing transformations. The most appropriate appears to be the log-log transformation. Figure 2 is the representation of the relation in the log-log scale. The data set seems more suited for ordinary least squares regression in this scale. However, the value of R^2 is raised to .9634 from .9295, which is marginally better. The independence assumption, underlying ordinary least squares, is violated, since several constraints are imposed on the weights, through the post-stratification adjustments. Clustering and other design artifacts also introduce dependencies between weights. However, the very large size of the data set offsets these dependencies. The offshoot is that overall, model 1 is deemed reasonable for purpose of estimation.

Testing for $H_0: A = 1, B = 0$

One particular realization of model 1 has a special meaning. It is the case where $A = 1$, and $B = 0$. This corresponds to the situation where

the MCPS is the original CPS, plus a random noise centered at zero. An assessment of this proposition is given through a formal hypothesis test.

Under the null hypothesis, no parameters are estimated and therefore no degrees of freedom are lost. The residual sum of squares has 109,108 degrees of freedom. Under the general model (model 1) two parameters are estimated and thus the residual sum of squares has 109,106 degrees of freedom, two less than the number of observations. The Analysis is summarized in Table 1.

Consider the following F statistic:

$$F = \frac{SS_{M2} - SS_{M1}}{SS_{M1}} \times \frac{d.f.^1}{d.f.^2 - d.f.^1}$$

The actual value of this F statistic is 2112.4. Under the null hypothesis, this F statistic behaves as an F random variable with 2 degrees of freedom on the numerator and 109,106 degrees of freedom on the denominator. Accordingly, the results are very significant and the null hypothesis is rejected. Therefore, the expectations of the MCPS weights are not CPS weights, and the effect of the estimation methodology is significant.

Effect on the Unemployment Rate

The preceding section shows that the final weights are distorted in a very particular way when the MCPS process is substituted for the original CPS process. This fact does not entail any foregone conclusion regarding the difference in unemployment rate between the CPS and MCPS. Recall that the unemployment rate is a ratio and thus linear transformations applied to the denominator and the numerator may partially cancel each other. The unemployment rates of the CPS and MCPS are presented in table 2, along with their difference and its standard deviation, for selected subpopulations. These results were obtained using the replication features of the VPLX software, monitored and designed by Robert Fay (1990). The difference between the CPS and the MCPS unemployment rate is significant for the male and white male categories. However, the estimated differences between the two rates are small. For the entire population, the difference in unemployment rate is not significant. Subsequently, annual average unemployment rate were also investigated. No significant effect from the estimation methodology was detected for any subpopulation (see Kostanich and Cahoon, 1994).

CONCLUSIONS

Keeping the collection methodology constant, it is clear that estimation methodology used for the CCO produces weights different, in expected value, from those generated by the CPS estimation methodology. So far, these differences do not appear to be causing significant distortions of the unemployment rate. The interference caused by the discrepancies of the estimation methodologies, in terms of the unemployment rate, when evaluating the differences in collection methodologies between the CPS and CCO, can be ignored.

REFERENCES

Fay, R.E., "Variance Estimates for Complex Samples", Proceedings of the Section on Survey Research Methods, 1990, ASA, pp. 266-271.

Kostanich, D.L., Cahoon, L.S., "CPS Bridge Team Technical Report 3", 1994, Census-BLS.

Polivka, A.E., "CPS Overlap Analysis Team Technical Report 1", 1994, Census-BLS.

Thompson, J., "CPS Overlap Analysis Team Technical Report 3", 1994, Census-BLS.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Table 1. Summary of Hypothesis Testing for H_0: $A = 1$, $B = 0$

	DEGREES OF FREEDOM MODEL 1	SUM OF SQUARES MODEL 1	DEGREES OF FREEDOM MODEL 2	SUM OF SQUARES MODEL 2
MODEL	2	4.39378	0	4.39111
RESIDUAL	109,106 ($d.f.^1$)	.0688921 (SS_{M1})	109,108 ($d.f.^2$)	.0715597 (SS_{M2})
TOTAL	109,108	4.46267	109,108	4.46267

TABLE 2. CPS and MCPS Unemployment Rates

CATEGORY	CPS UNEMPL. RATE	MCPS UNEMPL. RATE	DIFFERENCE MCPS-CPS	STD. DEV. DIFFERENCE
ALL	8.060	8.016	-.043	.0287
FEMALE	7.297	7.273	-.024	.0460
MALE	8.696	8.638	-.059	.0302
BLACK	15.019	14.930	-.090	.1314
WHITE	7.081	7.040	-.040	.0282
BLACK FEM.	13.973	13.848	-.125	.2271
BLACK MALE	16.084	16.029	-.056	.1803
WHITE FEM.	6.228	6.220	-.008	.0400
WHITE MALE	7.773	7.707	-.066	.0293

Figure 1: MCPS Weights (Y) vs CPS Weights (X)

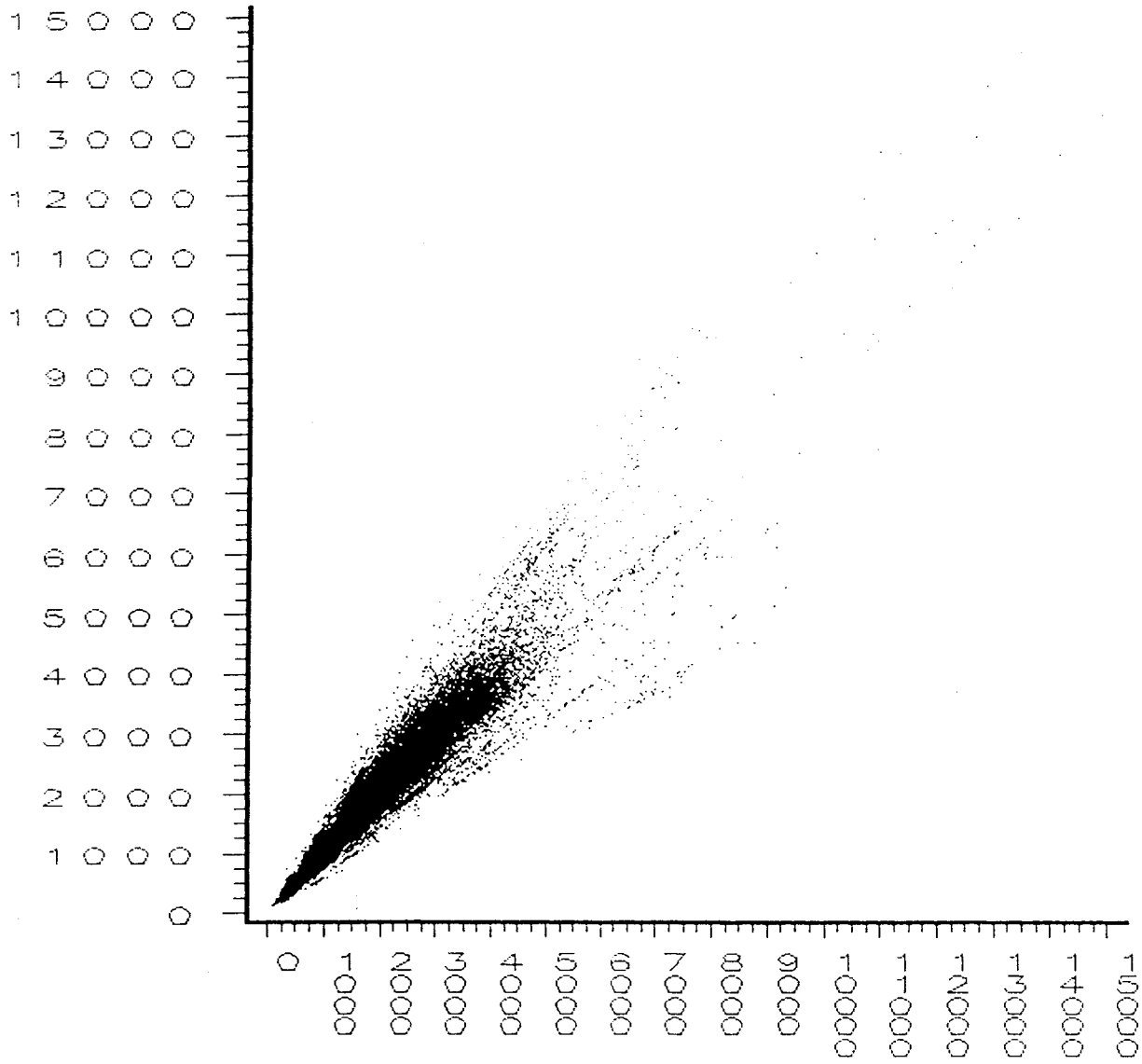

Figure 2: MCPS (X) vs CPS (Y); log−log Scale