EVALUATION OF NATIONAL CPS LABOR FORCE VARIANCES

Tamara Sue Zimmerman, Bureau of Labor Statistics, Room 4985, 2 Massachusetts Ave., N.E., Washington, DC 20212

The Current Population Survey (CPS), a nationwide household survey conducted by the Bureau of the Census for BLS, provides official labor force estimates for the noninstitutional population of working-age. Variances for the labor force estimators are based on a modified half sample replication method. Since the variances are based on a sample the variance estimates have a considerable amount of error. One method to remove some of the sampling error in the variances is to group similar labor force items and fit a generalized variance function to the variance estimates. This paper evaluates the current variance estimation procedure and suggests another procedure for generalizing variances for national labor force statistics.

KEY WORDS: Variance estimation, Balanced Half Sample Replication, Generalized Variance Functions, Current Population Survey

Introduction

The Current Population Survey (CPS) is a nationwide household survey conducted by the Bureau of the Census for the Bureau of Labor Statistics (BLS). Using a multistage stratified sample of about 48,000 households, the CPS provides official labor force estimates for the working-age noninstitutional population.

Variance estimates for the labor force estimators are based on a modified half sample replication method. For publication purposes, labor force items with similar mean to variance relationship are grouped together and a generalized variance function (GVF) is fit to the data. Labor force statistics, along with the parameter estimates for the GVF are published monthly in the publication Employment and Earnings.

BLS is researching a model-based approach to seasonal adjustment for national labor force series using a signal-plus-noise approach. This method depends upon “good” estimates of the variances. If the variances are too high, then too much noise is removed from the signal, resulting in a signal that is too smooth. If the variances are too small, not enough of the noise is removed from the signal and the signal is too noisy.

This paper evaluates the current variance estimation procedure and suggests another procedure for generalizing variances for national labor force statistics.

Section 1. Sample Design

The CPS sample consists of independent samples in each state and the District of Columbia. State sample sizes are determined by a specified coefficient of variation (CV). The first stage of sampling involves dividing the United States into primary sampling units (PSUs), which are metropolitan areas, a large county, or a group of smaller counties. Grouping similar PSUs based on information obtained from the Census Bureau and other information forms strata. Large PSUs are placed in strata by themselves and are sampled with certainty. These are defined as self-representing (SR) PSUs. The remaining PSUs grouped in a strata are called non-self representing (NSR) PSUs since the selected PSU represents the other PSUs in the strata.

In the second stage of sampling, a systematic sample of clusters of 4 housing units (ultimate sampling units, USUs) is selected within each sampled PSU. To account for new construction, a sample of building permits supplements the sampled USUs.

Section 2. Labor Force Estimation Procedure

Once a sample is selected, labor force estimates are computed by applying a series of weights to each person selected in the sample. The weights include:

- Baseweight and special weights derived from CPS sampling probabilities,
- Nonresponse adjustment,
- First-stage ratio adjustment,
- Second-stage ratio adjustment,

- Composite estimation.

Section 3. Current Variance Estimation Procedure

CPS variance estimates take into account sampling and non-sampling error. Obtaining an unbiased variance estimate is nearly impossible to obtain due to the CPS complex sample design and estimation process. Selecting 1 PSU per stratum and systematic sampling of clusters of households within the sampled PSU creates problems. Furthermore the ratio adjustments in the estimation procedure exacerbates the problem.

Section 3.1 Replication

One common way of computing variance estimates in complex sampling designs and estimation procedures is through replication. The basic idea is to repeatedly subsample the full sample using the same principles of selection as used for the full sample. Estimates from the subsamples are obtained just like those prepared from the full sample.

A successive difference replication method is currently used for estimating variances for the CPS labor force estimates. Fay and Train (1995) extended the basic theory of replication to be applicable to the CPS.

In general, replication methods require the selection of 2 PSUs per stratum. However in the CPS, only one 1 PSU is selected per stratum. The solution applied to the CPS sampling scheme is creating pseudo-strata and pseudo-PSUs.

SR-PSUs are considered pseudo-stratum. These are divided into several pieces with about the same population size. Each piece serves as a pseudo-PSU. NSR-PSUs are paired together to form pseudo-stratum where each NSR-PSU is considered a pseudo-PSU. In states where there is an odd number of NSR-PSUs, 3 NSR-PSUs are grouped together. The grouping of the NSR-PSUs introduces an upward bias in the replicate variances; a “between-stratum” component is artificially added which is not present in the CPS variance. However, efforts are made to group the NSR-PSUs to minimize the upward bias.

Since January 1996, 160 replicate estimates are produced each month. Throughout the rest of this document, I will use the following notation. The estimated variance for a labor force item is computed as follows:

$$\hat{V}\left(\hat{Y}_o\right) = \frac{4}{160}\sum_{r=1}^{160}\left(\hat{Y}_r - \hat{Y}_o\right)^2$$

where $\hat{Y}_o$ is the CPS full, original, sample estimate of the labor force characteristic and $\hat{Y}_r$ is the replicate estimate.

Section 3.2 Bias Estimation

In 1996, the Census Bureau conducted research on estimating the bias on the replicate variances at the state level. Using this information, I computed the bias of the national estimates

$$Bias\left\{\hat{V}\left(\hat{Y}_o\right)\right\} = 100\times\left\{\hat{V}\left(\hat{Y}_o\right) - V\left(\hat{Y}_o\right)\right\}\Big/V\left(\hat{Y}_o\right).$$

$\hat{V}\left(\hat{Y}_o\right)$ is the replicate variance estimate based on grouping the NSR-PSUs while $V\left(\hat{Y}_o\right)$ is the true variance computed without any grouping of the PSUs. (Rothass, 1996).

Census research only provided estimates of the bias for unemployment and the working-age civilian labor force (CLF). I use the bias of the CLF as a proxy for that of the employment since most of the CLF is comprised of employed people. For unemployment the replicate variances overestimate the true variance by 4.1% while for the CLF or employment, the replicate variance overestimates by 6.0%. The bias is larger for CLF than the unemployment since the CLF is sum of both unemployment and employment.

3.2 Reliability Estimation

In order to compute the reliability of the replicate variance estimates I used two methods. The first method (Method 1) defines $d_r = 4\left(\hat{Y}_r - \hat{Y}_o\right)^2$. The replicate variance is written as the average of the d_r's

$$\hat{V}\left(\hat{Y}_o\right) = \left(160\right)^{-1}\sum_{r=1}^{160} d_r = \bar{d}$$

The basic textbook formula for the variance of a sample mean was used to estimate the variance for the replicate variance estimate.

$$\tilde{V}\left(\hat{V}\right) = \left(159\right)^{-1}\left(160\right)^{-1}\sum_{r=1}^{160}\left(d_r - \bar{d}\right)^2$$

The second method (Method 2) I used involved a Monte Carlo simulation. I generated 500 simple random samples with replacement of size 160 of the replicate estimates, computed the replicate variance for each sample and then took the standard deviation across the 500 samples.

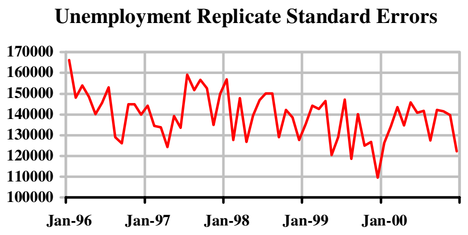
The table below provides the coefficient of variation based on these two methods. Both methods provide a CV of about 11%. The CV for unemployment is slightly higher than that for employment. However, the problem with both of these methods is that both violate an independence assumption. Therefore these estimates may be incorrect. We need a method compute the variances without the independence assumption.

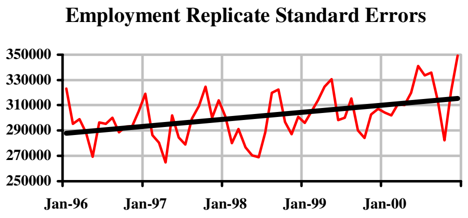
$\%CV = 100\sqrt{\tilde{V}\left(\hat{V}\right)}\big/\left(\hat{V}\right)$		
	Unemployment	**Employment**
Method 1	11.26%	11.09%
Method 2	11.24%	11.03%

Section 4. Generalizing Replicate Variances

Below are the standard errors for unemployment and employment based on the replicate variances. There are several things to notice:

1. The replicate standard errors are very noisy, especially for the unemployment estimates.





2. The employment replicate standard errors have an upward trend.

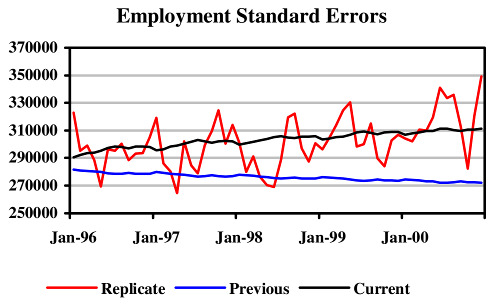
In order to remove some of the noise in the replicate variances, BLS generalizes the replicate variances. Previous generalization research was only based on 12 months of replicate estimates. Labor force items assumed to have the same mean to expected value were grouped to borrow strength across time and labor force items. The model used to generalize the variances was expressed in terms of the relative variance.

$$\hat{R}\left(\hat{Y}_o\right) = a + b\big/\hat{Y}_o$$

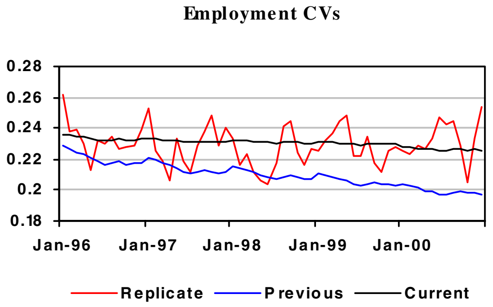
where $\hat{R}\left(\hat{Y}_o\right) = \hat{V}\left(\hat{Y}_o\right)\big/\hat{Y}_o^2$

An iterative reweighted least squares procedure was used to estimate the model parameters.

In the current research 60 months of replicate estimates were available; however none of the labor force items were grouped. That is, each item was individually generalized across time. Several models where tested but the same model used in the previous research proved to provide the best results in terms of diagnostics. Furthermore, it can be shown theoretically that the model above should be the model of choice. Below is a graph illustrating the standard errors for employment based on previous and current research. Notice that the standard errors based on previous research has a downward trend while those based on current research follow the upward trend of the replicate standard errors. This is most likely due to the grouping of the labor force items in the previous research.



The CPS sample size is determined by a reliability requirement based on a specified CV. For the nation, the specified CV for the monthly level of unemployment is 1.8 percent assuming a 6% unemployment rate. This roughly translates to a CV of about 0.2% for employment. I calculated the CVs using the standard errors for employment from the previous and current research. As can be seen the CVs are a little higher than expected. This is due to the upward bias in the replicate variances caused by the grouping of the NSR-PSUs.



Section 5.0 Future Research

There is a great deal of work that needs to be addressed. Further research needs to be conducted on estimating the bias and reliability of the replicate variance estimates. Other means of generalizing the replicate variances needs to be further investigated.

Another issue not addressed in this paper is the production of historical variance estimates to be used in the model-based seasonal adjustment procedure. This procedure depends on both historical and current variance estimates. Since the CPS replicate variance estimation system was only implemented in 1996, research much be conducted on how to get historical variance estimates. I have conducted some research on this but the historical variance estimates I produced did not seem reasonable. Therefore, I plan to address this issue further.

References

Fay, R., and Train, G. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," Proceedings of the Section on Government Statistics, American Statistical Association, 66. 154-159.

Rothass, R. (1996), "Collapsing Strata to Calculate 1990 Design Variances for CPS (VAR90-17)," Internal U.S. Census Bureau Memorandum.