Analysis of the Current Population Survey State Level Variance Estimates

Khandaker A. Mansur and Richard R. Griffiths , U.S. Census Bureau*
Khandaker A. Mansur, U.S. Census Bureau, Washington, D.C. 20233

Key Words: Variance; Current Population Survey; autocorrelation; time series.

1. Introduction

The Current Population Survey (CPS) is a monthly household survey conducted by the Bureau of the Census for the Bureau of Labor Statistics. Its primary purpose is to provide official labor force estimates for the nation as a whole. The CPS sample design is a two-stage stratified cluster design. It has two types of strata: self representing (SR) consisting of only one primary sampling unit (PSU) and non-self representing (NSR) consisting of more than one PSU. In SR strata estimators have only within-PSU variance and in NSR strata estimators have both within-PSU and between-PSU variance. To estimate the between-PSU variance, NSR strata are collapsed into groups of two or three pseudostrata because we select only 1 sample PSU in each NSR stratum. If the population means or variances of strata in the same pseudostrata are different, the collapsing method induces a bias in the variance estimator.

The problem with the current method of calculating CPS state level variances for labor force characteristics is that it does not give accurate state level variance estimates because of small state sample sizes and bias due to collapsing of the NSR strata. To address these issues, we are investigating modeling the variance estimates.

Before fitting any variance models, we conducted a preliminary analysis of several years of monthly CPS variance estimates. In this paper we report on several analyses designed to identify the autoregressive moving average (ARMA) models that best fit the state level variance time series and to determine if differences in sampling error variances exist by states and by months.

The paper is organized as follows: Section 2 discusses the methodology currently used to estimate state level variances; Section 3 discusses the analysis based on time series plots of state level standard errors; Section 4 describes an analysis of variance for state and month effects in the variance estimates; Section 5 discusses an analysis for the identification of ARMA models; and Section 6 provides a summary and describes areas of future research.

2. The Current Variance Estimation Methodology

The CPS total variance is composed of two types of variance, the variance due to sampling of housing units within PSUs (within-PSU variance) and the variance due to the selection of a subset of all NSR PSUs (between- PSU variance). Between-PSU variance estimates can not be calculated directly. Rather, they are calculated as the difference between the estimates of the total variance and within-PSU variance.

The CPS currently uses two methods to compute the monthly state level variance estimates, a modified balanced half-sample approach used to compute the total variance estimates and a successive difference replication method used to estimate within-PSU variance. The theoretical basis for the successive difference method was discussed by Wolter (1985) and extended by Fay and Train (1995) to produce the successive difference replication method. For detailed information about the variance estimation, see the U.S. Bureau of the Census, Bureau of Labor Statistics (2000).

3. Time Series Plots for Standard Errors

We first analyze the data using plots of standard errors of two labor force characteristics, civilian labor force (CLF) and number unemployed (UE). We produced plots of the estimated total standard errors of the estimators of these two characteristics for all states: Figure 1 displays these plots for Alabama (AL), Alaska (AK) and California (CA) over the period January 1996 to November 1999.

We can see from Figure 1 that plots for UE are more oscillatory than those of CLF. This suggests that standard errors of unemployed are not as highly correlated as those for CLF from month-to-month. This has important implications for a state level variance model. It suggests that a different model may be needed for each characteristic to account for the different correlation structures.

Another way to look at the month-to-month correlation between standard error estimates is to examine plots of standard error estimates from one month versus those of the previous month. Such

plots appear in Figure 2. These plots exhibit a linear relationship if there is a nonzero correlation between standard error estimates one month apart (lag of 1). We see that plots for CLF exhibit such a linear relationship whereas for UE, the linear trend is less evident. Again, this is an indication that the standard error estimates time series is more oscillatory for UE than it is for CLF. This further indicates less autocorrelation at lag 1 for UE standard error estimates.

4. State and Month Effects in the Variance Estimates.

4.1. Analysis of Variance (ANOVA)

State effects exist if there is a difference in the level of variance estimates from one state to another. For example, the plots in Figure 1 show that the standard error estimates in Alabama are larger than those in Alaska. This is clearly an indication of a state effect. A month effect is similarly defined as a difference in standard error estimates from one month to another.

In order to see whether state and month effects exist in the variance data, we performed an analysis of variance (ANOVA). The ANOVA indicates the amount of variation in the standard error estimates explained by state and month effects, thereby indicating the importance of these effects for inclusion in a state level variance model. The ANOVA model we used is

$$\log(n_{st}\ v_{st}) = \log(\sigma^2) + S_s + M_t + e_{st} \quad (1)$$

where n_{st} is the CPS sample size in state s and month t; v_{st} is the estimated variance in state s, month t for the characteristic of interest (UE or CLF); $\log(\sigma^2) + S_s + M_t$ is the mean of $\log(n_{st}\ v_{st})$; S_s is the effect for state s; M_t is the effect for month t; and e_{st} is the random error term in state s and month t.

The typical ANOVA assumptions don't hold for this model, because the $\log(n_{st}\ v_{st})$ variables are dependent over time. Because of this we didn't perform any standard statistical tests. Instead, we examined the sums of squares attributable to the different factors. In particular, we looked at the ratio of the sum of squares of the factor (SS_{factor}) to the sum of squares of total (SS_{total}). Table 4.1 summarizes these results.

Table 4.1: $\mathbf{SS}_{factor}$ / $\mathbf{SS}_{total}$

Characteristic	State	Month	Error
UE	97.0%	0.9%	2.0%
CLF	100.0%	0%	0%

The table displays the percent of total variation explained by each factor for UE and CLF. The table shows that state effects are explaining almost all of the variation in the standard error estimates. Month effects are very small; they are not as important as state effects. This analysis indicates the need for state effect terms, but no month effect term, in the model.

It is important to note the fact that month effects are not important does not mean that lag effects are unimportant. Indeed, we plan to include a time series component in our model, acknowledging the importance of lag effects. Instead, unimportance of month effects may indicate that the time series is stationary in the weak sense.

4.2. Analysis of Covariance (ANCOVA)

We extended our analysis of the previous section to account for the relationship variance has with the characteristic estimate. We attempted to determine if the state effect is still apparent after accounting for this relationship. To do this, we fit the following ANCOVA model:

$$\log(n_{st}\ v_{st}) = \beta \log(Y_{st}^2) + \log(\sigma^2) + S_s + M_t + e_{st}$$

where the terms common to (1) have the same interpretation; Y_{st} is the estimate for the characteristic Y in state s, month t; and β is a regression coefficient. Results of the ANCOVA are presented in Table 4.2.

Table 4.2: $\mathbf{SS}_{factor}$ /$\mathbf{SS}_{total}$

Characteristic	Estimate	State	Month	Error
UE	0.9%	99.0%	0.1%	0%
CLF	0.6%	99.3%	0.1%	0%

The results confirm those given in Table 4.1. We conclude that state effects dominate while the month effect is inconsequential.

5. Identification of Time Series Process

In this section we discuss an analysis aimed at identifying the ARMA models that best fit the state level variance time series.

A strong autocorrelation in the sampling error arises from the use of a 4-8-4 rotating panel design that generates complex patterns of sample overlap over time. In addition, when a cluster of housing units is permanently dropped from the sample, it is replaced by nearby units, resulting in correlations from non-identical households in the same rotation panel (Train, Cahoon, and Makens, 1978). To investigate this, we attempt to identify ARMA models that fit the CPS state level variance estimators time series.

We used Akaike's Information Criterion (AIC) to determine the best fitting ARMA models. AIC is calculated as $-2\ln(L) + 2k$, where L is the likelihood function and k is the number of parameters in the model (Akaike, 1974). The model with the smallest AIC is the best-fitting model.

Tables 5.1a and 5.1b display the AIC for the fit of several ARMA processes to the natural logarithm of standard error estimates for the CLF and UE estimates of several states. The smallest AIC for each state in a table is given in bold. We see from these tables that the model that most consistently fits best is the AR(1) model: for CLF, the AR(1) and ARMA(2,2) models resulted in the smallest AIC values and, for UE, the AR(1) model resulted in the smallest AIC values for most of the states examined.

Table 5.1a: **AIC Values for CLF of Several ARMA Models**

State	AR(1)	AR(2)	MA(1)	MA(2)	ARMA (1,1)	ARMA [1](1,2)	ARMA(2,2)	ARMA[2](1,1)(1,0) $_{12}$
AL	995	**982**	989	987	988	988	**982**	988
AK	**782**	784	**782**	784	784	787	788	784
AZ	**947**	949	956	954	949	951	954	949
CA	1026	1028	1036	1031	1030	1029	**1024**	1028
GA	**984**	991	990	991	991	992	985	991
HI	819	820	825	826	820	822	**816**	820
IL	**968**	970	968	967	**968**	972	969	970
NY	**965**	966	972	969	966	967	**965**	966
UT	**840**	842	842	842	842	844	845	842

Table 5.1b: **AIC Values for UE of Several ARMA Models**

State	AR(1)	AR(2)	MA(1)	MA(2)	ARMA (1,1)	ARMA (1,2)	ARMA (2,2)	ARMA(1,1)(1,0) $_{12}$
AL	903	**901**	902	903	904	**901**	**901**	904
AK	**750**	752	751	951	752	753	753	752
AZ	**886**	889	887	889	889	890	888	889
CA	**953**	956	954	956	956	957	960	956
GA	961	958	961	956	960	**954**	956	959
HI	**793**	796	794	796	796	797	800	796
IL	**933**	935	934	935	935	937	939	935
NY	**907**	907	908	908	907	801	801	907
UT	798	797	798	798	799	909	908	792

[1] ARMA(p,q) is an autoregressive moving average model with p being the number of autoregressive parameters and q the number of moving average parameters.

[2] ARMA(p,q)(l,m)$_s$ is a seasonal autoregressive moving average model with l the order of the seasonal autoregressive process, m the order of the seasonal moving average process and s the span of the seasonality.

We also examined some autocorrelation plots which provided us some indication that the AR(1) model provided the best fit. These plots generally showed that only the autocorrelation at lag 1 was different from zero.

Note: We did not account for nonstationarity in these analyses. Some of the state level variance time series may be nonstationary and it might be appropriate to first difference them before trying to fit an ARMA model to them.

6. Summary and Future Research

The main goal of the research is to see what the analyses tell us about the structure of CPS state level standard error estimates, and what they mean for us in trying to develop a state level variance model. What we have learned from the analyses is the following:

- Section 3 indicates that the monthly standard error estimates for UE and CLF exhibit some month-to-month correlation. CLF standard error estimates seem to have a higher level of autocorrelation than do UE standard error estimates. Thus, it seems reasonable that a variance model would include a component to account for the autocorrelation.

- Section 4 tells us that the state level standard error estimates are subject to a state effect, even after accounting for their relationship with the characteristic estimate. It thus appears that accounting for this state effect in a variance model will be important.

- Section 5 gives us an idea of the type of ARMA process that best describes the state level variance time series. It first indicates that an AR(1) model might be adequate for describing both the UE and CLF standard error estimates time series; but it also cautions that a higher order ARMA process might better describe some of the time series.

In the future, we plan to continue our investigation in several areas. These include the following:

- We will continue our research to find an appropriate ARMA process which will ultimately be used in the state level variance model.

- We will continue our investigation to see whether the same time series model of the variance estimates works for all the states or separate models are needed for groups of states.

Acknowledgments

The authors would like to thank Harland H. Shoemaker, Douglas Bond and Mahdi Sundukchi for reviewing the paper and their helpful comments and suggestions.

*This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

References

Akaike, H . (1974), " A New Look at the Statistical Model Identification,*" IEEE Transactions on Automatic Control, AC-19, 716-723.*

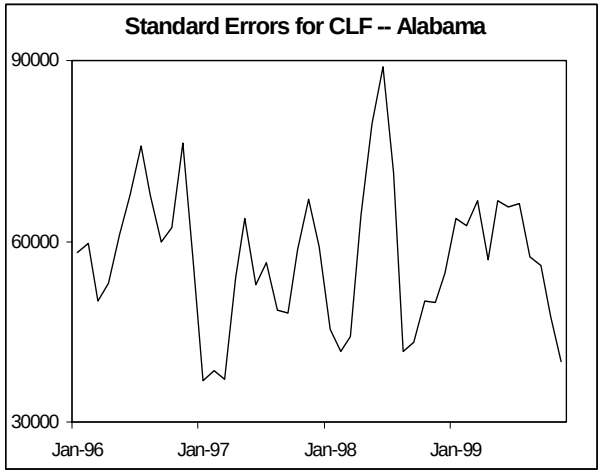
Fay, R. and Train, G. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties, " *Proceedings of the Section on Government Statistics, American Statistical Association, 154-159.*

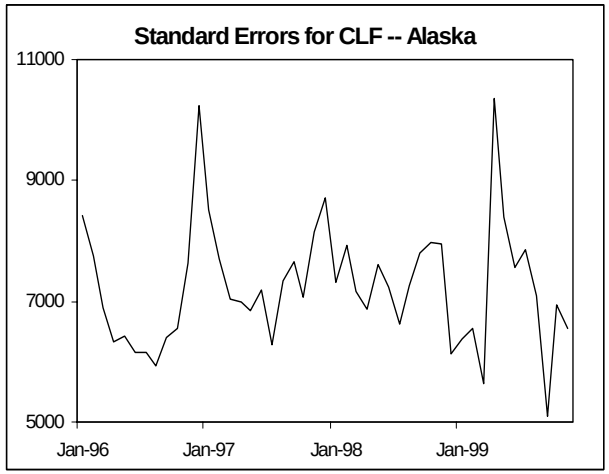
Train, G., Cahoon, L. and Makens, P. (1978), " The Current Population Survey Variances, Inter Relationships, and Design Effects, *" Proceedings of the Section on Survey Research Methods, American Statistical Association, 443-448.*

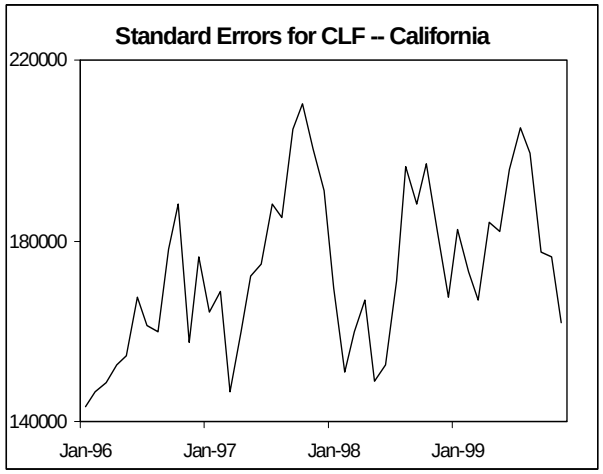
U.S. Bureau of the Census, Bureau of Labor Statistics. (2000), *Current Population Survey: Design and Methodology, Technical Paper 63*, Washington, D.C.

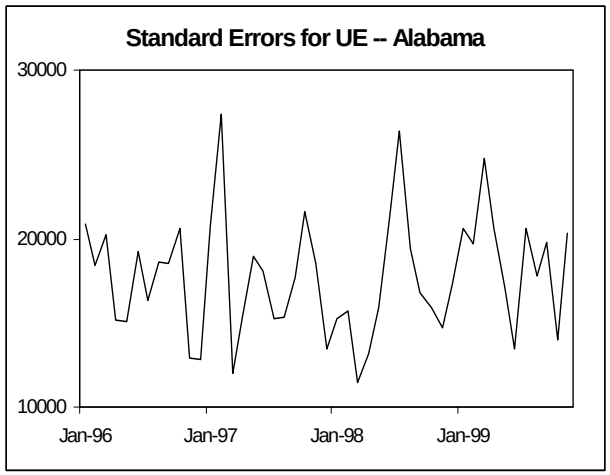
Wolter, K. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

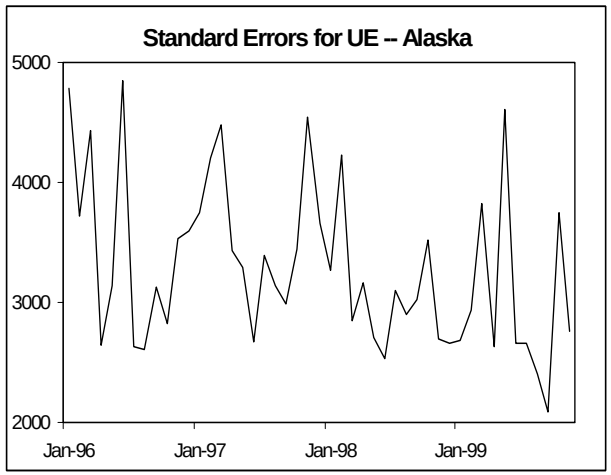
Figure 1. **Time Series Plots of Standard Errors for CLF and UE**











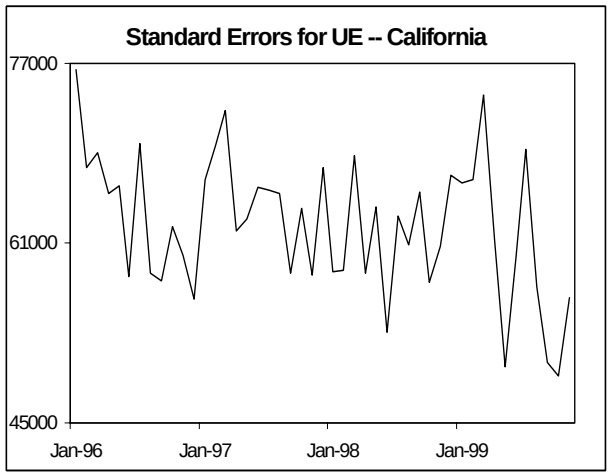


Figure 2. Plots of Current Month Standard Error (Y-Axis) Versus Previous Month Standard Error (X-Axis) for CLF and UE