COVARIANCE ESTIMATORS FOR THE CURRENT POPULATION SURVEY

Abdoulaye Adam and Wayne A. Fuller
Abdoulaye Adam, Iowa State University, Ames, IA 50010

KEY WORDS: **Rotation Sampling, Autocorrelation, Time series**

1. Introduction

The Current Population Survey (CPS) is a monthly household survey conducted by the Bureau of Census for the Bureau of Labor Statistics. The objective of the survey is to provide estimates of characteristics of the U.S. Labor force. The survey is described in Hanson (1978). The sample is divided into subgroups called rotation groups. The sample for a month consists of eight rotation groups, divided in such a way that 1/8 of the sample is interviewed for the first time, 1/8 for the second time, ..., and 1/8 for the eighth time. The first interview is defined as the first time—in—sample, the second as the second time—in—sample, etc. A particular rotation group is in the sample for four months, rotates out for eight months, and then re—enters the sample for four more months. This rotation pattern is called a 4—8—4 rotation scheme.

In the Current Population Survey, the direct survey estimate at time t can be modeled as

$$X_t = x_t + u_t , \tag{1.1}$$

where X_t is the direct estimate, x_t is the population characteristic of interest and u_t is an error due to sampling. The sampling error structure has been investigated by several authors. Using the 1975 CPS data, Train et al. (1978) estimated the total sampling variance as the sum of the within self—representing primary sampling units variance and the total variance of non—self—representing primary sampling units. Breau and Ernst (1983) assumed a covariance structure based on the 4—8—4 rotation scheme of the Current Population Survey to estimate the covariances between rotation groups. Lent (1991) gave a description of the method used by the Bureau of Labor Statistics to compute variance and correlation estimates for the CPS state data.

In this paper, we present a components of variance model for the sampling error $\{u_t\}$. Three components are identified and esitmated. These are a replicate component that is due to variation between primary sampling units, a permanent component associated with rotation groups within primary sampling units, and a transient component associated with rotation groups within primary sampling units.

2. Components of Variance Model for the Sampling Error

The replication procedure developed by Fay (1989) was used by the Census Bureau to assign replicate factors to rotation groups. The data used to estimate the sampling error covariance function consists of 48 replicates and the full sample estimates for eight time—in—sample observations in twelve months created from the 1987 CPS data. The sum of the eight time—in—sample observations for a replicate is an unbiased estimator of the population total. The weighted linear combination forming a replicate is such that the expected value of the squared deviation between the replicate estimate and the full sample estimate is one fourth of the variance of the full sample estimate.

The 1987 data set can be arranged in a data matrix $\mathbf{M}$ whose columns consist of observations on a set of groups of individuals. The groups are called "rotation groups," and the columns of $\mathbf{M}$ are called "streams." In the organization, a rotation group appears in only one stream. In each stream, the rotation groups follow the 4—8—4 rotation scheme of the CPS. Either three or four rotation groups appear in a stream during the 12 months.

For the 1987 data set, we consider the analysis of variance decomposition,

$$y_{tjk} = \mu + v_j + \alpha_t + \tau_k + \gamma_\ell + \zeta_{tk} + \epsilon_{tjk} , \tag{2.1}$$

$$\Sigma_j v_j = \Sigma_t \alpha_t = \Sigma_k \tau_k = \Sigma_\ell \gamma_\ell = 0 ,$$

$$\Sigma_t \zeta_{tk} = 0 \text{ for all } k , \text{ and } \Sigma_k \zeta_{tk} = 0 \text{ for all } t ,$$

where y_{tjk} is an estimate of a characteristic, such as total employed, obtained from the j—th replicate for the k—th time—in—sample at time t, μ represents the overall mean, v_j is the replicate

effect, α_t represents the time effect, τ_k is the time—in—sample effect, γ_ℓ is the effect of rotation groups, ζ_{tk} represents interactions among time—in—sample, time and group effects, and ϵ_{tjk} is the error term. The subscript $\ell = \ell(t, k)$ is completely determined by the subscripts t and k , i.e., once a month and a time—in—sample are given, the rotation group is uniquely determined by the configuration of the data matrix **M** . The time effect is sometimes called a month effect when dealing with the year 1987. For the 1987 CPS data used in the analysis, $t = 1, 2, ..., 12$, $k = 1, 2, ..., 8$, and $j = 1, 2, ..., 48$.

The matrix **M** has 96 entries representing combinations of three factors, month, time—in—sample, and group. There are 12 months, 8 times—in—sample and 27 groups. The three factors are not orthogonal. Therefore, the sums of squares and degrees of freedom can be partitioned in different ways. One possible partition is shown in Table 1. The sum of squares for months and time—in—sample were first computed. The main effect for groups has been adjusted for the month and time—in—sample main effects and has 25 degrees of freedom. The remaining 52 degrees of freedom in Table 1 correspond to interactions among the main effects.

The three characteristics, Civilian Labor Force, total employed, and total unemployed, were coded by dividing each entry by 100,000. The analyses of variance in Table 1 reveal that the mean squares for all factors in model (2.1) are much bigger than the residual mean square for all characteristics. The ratio of the mean square of a factor to the residual mean square does not have an F—distribution because the replicates are not independent.

The 4—8—4 rotation scheme of the CPS requires a sixteen—month period to complete the

Table 1. Analysis of variance for employed, unemployed, and Civilian Labor Force, 1987.

Source	d.f.	Mean Squares			
		Employed	Unemployed	CLF	Employment Rate (%)
Replicates	47	1.2134	0.1785	1.0762	0.0788
Months	11	3553.2435	268.4974	2377.7240	149.8368
Time–in–Sample	7	458.1149	75.4759	891.2340	20.3164
Groups[1]	25	113.4492	20.7709	91.5742	9.0534
Interactions	52	12.9719	6.7783	11.7550	2.8615
Residual	4465	0.2458	0.0554	0.2112	0.0247

[1]The groups mean square is adjusted for month and time–in–sample.

observations on a rotation group. Because we have data only for the twelve months of 1987, some of the possible correlations cannot be computed directly. Let y_{tjk} be the j—th replicate estimate for total employed in month t and k—th time—in—sample and let y_{t0k} be the corresponding full sample estimate for month t and k—th time—in—sample. Lent (1991) used the following formulae to estimate correlations,

$$\hat{\mathrm{Var}}(y_{t0k}) = 12^{-1} \sum_{j=1}^{48} (y_{tjk} - y_{t0k})^2 \tag{2.2}$$

$$\hat{\mathrm{Cov}}(y_{t0k}, y_{t+h,0,k+h})$$

$$= 12^{-1} \sum_{j=1}^{48} (y_{tjk} - y_{t0k})(y_{t+h,j,k+h} - y_{t+h,0,k+h}) \tag{2.3}$$

$$\hat{\rho}_{t,t+h} = \frac{\hat{\mathrm{Cov}}(y_{t0k}, y_{t+h,0,k+h})}{[\hat{\mathrm{Var}}(y_{t0k})\hat{\mathrm{Var}}(y_{t+h,0,k+h})]^{1/2}}, \tag{2.4}$$

where $\hat{\rho}_{t,t+h}$ is the correlation between time t and time t+h for the direct estimate.

For our computations, the data were arranged in the data matrix **M** . When the replicate effects are removed, the streams are assumed to be independent. Thus, when the replicate effects are removed, the covariance matrix of the vec of the data matrix **M** is block diagonal, where the vec of a matrix **M** is the column vector formed by the columns of **M** arranged chronologically.

The form of the covariance matrix when the streams are assumed independent is block diagonal. Since no rotation group belongs to more than one stream, the block diagonal form of the covariance matrix was used to estimate the correlations between observations made on the same rotation group at different points in time.

The correlations at lag 1, 2, 3, 9, 10, and 11 are obtained by taking the averages of all correlations at the respective lags. The average correlations are given in Table 2. The autocorrelations are similar to those estimated by Breau and Ernst (1983) using CPS data from September 1976 through December 1977. The survey was redesigned after the 1980 Census, but the correlations for the two periods are similar.

Table 2. Average autocorrelations within a rotation group for 1987 CPS

Lag	No. Obs.	Employed	Unemployed	Civilian Labor Force	Unemployment Rate
1	66	0.8068 (0.0062)	0.4979 (0.0136)	0.7876 (0.0068)	0.5187 (0.0132)
2	40	0.7332 (0.0106)	0.3788 (0.0199)	0.7197 (0.0111)	0.4019 (0.0195)
3	18	0.6856 (0.0182)	0.3230 (0.0312)	0.6668 (0.0192)	0.3484 (0.0306)
9	3	0.6732 (0.0461)	0.1566 (0.0832)	0.6377 (0.0504)	0.2034 (0.0818)
10	4	0.7191 (0.0354)	0.2691 (0.0686)	0.6187 (0.0452)	0.3159 (0.0665)
11	3	0.6038 (0.0536)	0.1401 (0.0830)	0.4910 (0.0614)	0.2138 (0.0814)
Ave. 9–11	10	0.6708 (0.0252)	0.1966 (0.0450)	0.5861 (0.0298)	0.2443 (0.0439)

The estimates for 1987 are slightly lower for the first few lags and slightly larger for long lags than those for 1976–1977. Estimates of the standard errors of the average correlations are given in parentheses below the estimates. At each lag, the estimated standard errors of the average correlations were computed by inverting an approximate 95% confidence interval for the mean of the z–scores. The z–scores were obtained by the transformation,

$$z_i = 0.5 \, \mathrm{Log}\{(1 + \hat{\rho}_i)(1 - \hat{\rho}_i)^{-1}\},$$

where $\hat{\rho}_i$ is the estimate of the correlation ρ_i. The standard errors are only approximations. The correlations are not independent because more than one correlation is computed from some rotation groups and the computed correlations are autocorrelations.

Chi–square values for the tests that, at each lag, the different values estimated the same correlations were computed. The computation of the chi–square values is described in Snedecor and Cochran (1980, p. 187). These chi–square values were computed as if the correlations were independent. The correlations are not independent for the reasons given above. Therefore, the chi–square values do not have the tabulated chi–square distributions. However, except for unemployed and unemployment rate at lag one, the values of the chi–square give no reason to reject the hypothesis of equal correlations. An analysis of variance shows that the large chi–square values for unemployed and unemployment rate at lag one are due to a large within group variation where groups are formed on the basis of time–in–sample pairs. Because the variation is within group variation, we retain the model in which the correlation is constant at each lag.

Direct computation using formula (2.4) does not provide estimates of all possible correlations. Therefore, a model was developed and used to construct estimates of the remaining correlations. Consider the sum of the replicate and error effects,

$$r_{tjk} = v_j + \epsilon_{tjk}, \tag{2.5}$$

where v_j and ϵ_{tjk} are defined in model (2.1). Let an estimator of r_{tjk} be

$$\hat{r}_{tjk} = y_{tjk} - \hat{\mu} - \hat{\alpha}_t - \hat{\tau}_k - \hat{\gamma}_\ell - \hat{\zeta}_{tk}, \tag{2.6}$$

where $\hat{\mu}$, $\hat{\alpha}_t$, $\hat{\gamma}_k$, $\hat{\tau}_k$, $\hat{\gamma}_\ell$ and $\hat{\zeta}_{tk}$ are the ordinary least squares estimates of the corresponding parameters of model (2.1). That is, the estimated effects correspond to estimating the means of the 96 entries of the data matrix $\mathbf{M}$. Thus, $\hat{r}_{tjk}$ is the original observation with estimated month, time–in–sample, their interactions effects removed. In model (2.1), observations were indexed by month, replicate, and time–in–sample. On the basis of the data matrix $\mathbf{M}$, we can identify particular rotation groups. If we know the month and the time–in–sample of an observation, we know the rotation group of the observation. With a slight abuse of notation, let r_{gjk} be the value of r_{tjk} obtained when we use the rotation group index in place of the time index. Thus, r_{gjk} is the original observation for the k–th time–in–sample of the g–th rotation group in the j–th replicate when the effects of factors of model (2.1) except replicate are removed. We assume

$$r_{gjk} = v_j + e_{gj} + a_{gjk}, \tag{2.7}$$

$$a_{gjk} = \sum_{\ell=1}^{3} \xi_\ell \, a_{g,j,k-\ell} + b_{gjk},$$

where v_j is the replicate effect, e_{gj} is the permanent effect of rotation group g within replicate j, and a_{gjk} is a transient effect

associated with rotation group g. It is assumed that the transient rotation group effect is a stationary third order autoregressive process. It is assumed that $\{v_j\}$, $\{e_{gj}\}$, and $\{a_{gjk}\}$ are independent sequences. While the method of constructing the replicates is complicated, we assume that the construction is such that the v_j are independent. It is assumed that

$$(v_j, e_{gj}, b_{gjk}) \sim \text{Ind.}[0, \text{diag}(\sigma_v^2, \sigma_e^2, \sigma_b^2)] .$$

It follows that

$$\gamma_r(h) = E\{r_{gjk} r_{g,j,k+h}\}$$

$$= \sigma_v^2 + \sigma_e^2 + \rho_a(h)\sigma_a^2 , \tag{2.8}$$

$$\rho_r(h) = \frac{\sigma_u^2 + \sigma_e^2 + \rho_a(h)\sigma_a^2}{\sigma_u^2 + \sigma_e^2 + \sigma_a^2} , \tag{2.9}$$

$$V\{r_{gjh}\} = \sigma_u^2 + \sigma_e^2 + \sigma_a^2 , \tag{2.10}$$

where $\rho_r(h)$ is the autocorrelation function of r_{gjk}, $\gamma_r(h)$ is the autocovariance function of r_{gjk}, and $\rho_a(h)$ is the autocorrelation function of a_{gjk}. Thus $\rho_r(h)$ is the correlation between r_{gjk} and $r_{g,j,k+h}$, where r_{gjk} is the observation on a single rotation group which is in the sample for the k—th time.

An estimate of σ_v^2 was obtained from an analysis of variance table constructed by taking the four observations from each of the two rotation groups in each stream for which four observations are available. Estimates of σ_v^2 are $\hat{\sigma}_v^2 = 0.00552$ for employed, $\hat{\sigma}_v^2 = 0.00065$ for unemployed, $\hat{\sigma}_v^2 = 0.00499$ for Civilian Labor Force, and $\hat{\sigma}^2 = 2.81\ 10^{-8}$ for unemployment rate.

Remember that replicates are weighted linear combinations of observations on the original sampling units. Assume the original sample to be a simple random sample of primary sampling units of size n, and assume an equal number of observations were made on each of eight rotation groups in each primary sampling unit. Then the variance σ_v^2 of the replicate effects is an estimate of $0.25n^{-1}\sigma_{psu}^2$, where σ_{psu}^2 is the variance of the primary sampling unit effect, and the estimated population total is the mean of n primary sampling unit estimated totals.

Because the ϵ_{gjk} of model (2.1) is the sum of e_{gj} and a_{gjk} of model (2.7), the residual mean squares of Table 1 provide estimates of $\sigma_e^2 + \sigma_a^2$ for total employed, total unemployed, Civilian Labor Force, and unemployment rate, respectively.

To estimate the autoregressive coefficients, ξ, of (2.7), we use the correlations of Table 2. An iterative estimation scheme was used which utilized the fact that the correlations are nearly constant for lags greater than eight. Therefore, to construct initial estimates of the ξ_j, we assume that the autocorrelations of a_{gjk} can be treated as zero for $h > 9$. Then,

$$\gamma_a(h) \doteq \gamma_r(h) - \gamma_r(\text{ave}) \qquad \text{for } h = 0, 1, 2, 3 . \tag{2.11}$$

where

$$\gamma_r(\text{ave}) = 3^{-1}[\gamma_r(9) + \gamma_r(10) + \gamma_r(11)]$$

Using the estimated $\hat{\gamma}_a(h)$, the initial estimates of ξ_j are obtained using the Yule—Walker equations. In the Yule—Walker equations, the estimated autocovariances were defined by

$$\hat{\gamma}_a(h) = [\hat{\rho}_r(h) - \hat{\rho}_r(\text{ave})]\hat{V}(r_{gjk})$$

for $h = 0, 1, 2, 3$,

$$\hat{V}(r_{gjk}) = \hat{\sigma}_v^2 + \hat{\sigma}_e^2 + \hat{\sigma}_a^2 ,$$

$$\hat{\rho}_r(\text{ave}) = 3^{-1}[\hat{\rho}_r(9) + \hat{\rho}_r(10) + \hat{\rho}_r(11)] ,$$

where $\hat{\rho}_r(\text{ave})$ and $\hat{\rho}_r(h)$ are taken from Table 2. Using the estimates from the Yule—Walker equations, second round estimates of the ξ_j were then computed. The average of the estimated autocovariances for lags 9, 10, and 11 of the first round model were computed and improved estimates of the first three covariances of a were computed as

$$\hat{\gamma}_a(h) = \hat{\gamma}_r(h) - \hat{\gamma}_r(\text{ave}) + 3^{-1}[\tilde{\gamma}_a(9) + \tilde{\gamma}_a(10) + \tilde{\gamma}_a(11)] ,$$

where $\tilde{\gamma}_a(h)$ is the covariance computed from the first round estimated parameters. A third iteration was conducted, and because the estimate changed little from the second to the third iteration, the third—round estimates were accepted as the final estimates. The estimated parameters are given in Table 3.

The estimates of the three variance components are given in Table 4 for the four characteristics. These variances must be multiplied by four when used to construct the estimates of the standard errors of the CPS estimates of totals.

From Table 4, we see that the permanent effect of a rotation group is smaller relative to the transient effect for unemployed than for Civilian Labor Force. The contributions of the three sources, replicate effect, permanent and transient rotation group effects to the variance of the direct estimate are about the same for unemployed and unemployment rate, and the percentage contribution of replicate variance for employed and Civilian Labor Force are similar. The transient effect of rotation groups is responsible for 74% of the variance of unemployed and unemployment rate, but is responsible for only 36% of the variance of the estimate of Civilian Labor Force.

Using equation (2.9), the estimates of the variance components can be used to compute estimated autocovariances for observations on a single rotation group of a replicate for any given lag.

Using the variance component for replicate effect, the autocovariances of rotation groups and the number of overlapped rotation groups by lag, one can obtain the estimated autocovariance functions of the direct estimates of the characteristics of the Current Population Survey. The direct estimate is the simple sum of the eight estimates associated with the eight rotation groups. The direct estimate is not the published estimate. The published estimator is a composite estimator. The direct estimates for January 1987 are 1084.70, 86.35, 1171.05 and 7.37 for employed, unemployed, Civilian Labor Force, and unemployment rate, respectively, where the first three estimates are in 100,000's and unemployment rate is in percent. The estimated standard errors for these estimates, including the replicate effect are 3.046, 1.393, 2.835, and 0.00916 for employed, unemployed, Civilian Labor Force, and unemployment rate, respectively.

Table 3. Estimates of parameters of transient processes a_{gjk}.

Model $a_{gjk} = \xi_1 a_{gj,k-1} + \xi_2 a_{gj,k-2} + \xi_3 a_{gj,k-3} + b_{gjk}$					
Characteristic	$\hat{\xi}_1$	$\hat{\xi}_2$	$\hat{\xi}_3$	$\hat{\sigma}_b^2$	$\hat{\sigma}_a^2$
Employed	0.40481	0.04270	−0.04945	0.06841	0.08263
Unemployed	0.33422	0.08452	0.05267	0.03831	0.04508
CLF	0.43415	0.11345	0.00318	0.06756	0.08962
Unemp. rate (%)	0.32855	0.07524	0.04382	0.01579	0.01934

Table 4. Estimates of σ_u^2, σ_e^2, and σ_a^2.

	Variance component			
Characteristic	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_a^2$	Total
Employed	0.00552	0.16319	0.08263	0.25134
Unemployed	0.00065	0.01037	0.04508	0.05610
Civilian labor force	0.00499	0.12159	0.08962	0.21620
Unemployment rate (%)	0.00028	0.00466	0.01934	0.02428

Acknowledgements

This research was partly supported by Joint Statistical Agreement 91—21 with the U.S. Bureau of the Census.

References

Adam, A. (1992), *Covariance estimation for characteristics of the Current Population Survey.* Ph.D. dissertation, Iowa State University, Ames, IA.

Breau, P., and Ernst, L. R. (1983), Alternative estimators to the current composite estimator. *Proceedings of the American Statistical Association, Section on Survey Research Methods,* 397—402.

Fay, R. E. (1984), Some properties of estimates of variance based on replication methods. *Proceedings of the American Statistical Association, Section on Survey Research Method,* Washington, D.C.

Fay, Robert E. (1989), Theory and application of replicate weighting for variance calculations. *Proceedings of the American Statistical Association, Section on Survey Research Methods,* 212–217.

Fuller, Wayne A. (1976), *Introduction to Statistical Time Series.* John Wiley and Sons, New York.

Hanson, R. H. (1978), *The Current Population Survey – Design and Methodology.* U.S. Bureau of the Census, Washington, D.C.

Huang, E. T., and Ernst, L. R. (1981), Comparison of an alternative estimator to the current composite estimator in CPS. *Proceedings of the American Statistical Association, Section on Survey Research Methods,* 303–308.

Lent, J. (1991), Variance estimation for Current Population Survey Small Area Labor Force Estimate. *Proceedings of American Statistical Association, Section on Survey Research Methods, forthcoming.*

Snedecor, G. W., and Cochran, W. G. (1980), *Statistical Methods,* Iowa State UniversityPress, Ames, Iowa.

Train, G. et al. (1978), The Current Population Survey variances, inter–relationships and design effects. *Proceedings of the American Statistical Association, Section on Survey Research Methods,* 443–448.