VARIANCE ESTIMATION FOR ESTIMATES OF EMPLOYMENT CHANGE IN THE CURRENT EMPLOYMENT STATISTICS SURVEY

Steve Woodruff, Bureau of Labor Statistics,
Room 4985, 2 Massachusetts Ave N.E., Washington D.C. 20212

KEY WORDS: Generalized Variance, Model Based

1. INTRODUCTION

The Bureau of Labor Statistics' (BLS) Current Employment Statistics (CES) Survey gathers data monthly from over 380,000 nonagricultural business establishments for the purpose of estimating total employment, women and production workers, hours, and earnings. Estimates are made for over 1,500 industry cells, complimenting the demographic detail provided by estimates of employment from the Current Population Survey (CPS). Monthly estimates of level and month-to-month change in employment are of primary importance to the users of these data. In addition to the CES survey, each state conducts a complete count of the employment of its business population every quarter following the guidelines of the Unemployment Insurance (UI) system. Except for a few industries exempt from UI coverage, this complete count is used by the CES as a benchmark to which survey estimates are revised and to which they are compared to derive a measure of total error.

This total error includes: bias due to changes in the population caused by births and deaths, sample nonresponse, response bias, and variance under an extremely well documented regression model. This model variance is to be estimated in this paper.

For this paper we restrict ourselves to estimating the variance of the estimate of month-to-month change in the all employment variable. Later, this variance estimator will be used to construct an estimator for the variance of the total employment estimates.

The purpose of this initial study is to answer the following questions.

1) The month to month link is the ratio of the total matched sample employment for the current month over the total matched sample employment for the previous month, where the "matched sample" consists of those units which reported employment for both months. Thus the month to month link is an estimate of employment change between adjacent months. How much variance should one expect in the month to month link?

The CES sample data for a given month arrives at BLS in a piecemeal fashion over an interval of several months duration. All the data that has arrived in time for release on the first Friday of the first month following the reference month is called first closing data, first friday of the second month following the reference month is called second closing data, and so on. Under this CES data flow, what is the variance of first, second, and third closing links? These variances will permit us to place confidence intervals about these estimated links for each closing. If such intervals contain one, then the month to month link is at best a weak indicator of the direction in which employment is moving.

2) What size of closing revisions (difference between two different closing links for the same month) should be expected from this type of variability alone? What size of closing revision should be reason to suspect that in addition to the variability described in 1) above, there is a regular and estimable bias component in the closing revision? That is, what size of closing revision would indicate that there is a significant difference between the underlying month to month link of first closing responders and the underlying month to month link of later responders? How can we estimate such bias and reduce the closing revisions?

Royall and Cumberland (1978,1981), and Royall and Eberhardt (1975) looked at the general problem of estimating the variances of ratio and regression estimators. Their findings lead to some of the estimators tested here. The specific problem of estimation of CES variances has been studied by West (1984), Royall (1981), and Madow & Madow (1978). The estimators considered here include variations on their suggestions, with the emphasis on computational simplicity (a generalized variance estimator).

Section two describes the estimates of change, their stochastic properties, and some different ways to estimate their variance. Section three contains a simulation study where the proposed variance estimators can be compared to the true variance. Section four describes a study of the best of the proposed estimators in a CES estimation cell.

2. SOLUTIONS

2.1 Definitions, Models, and Variances

For a given pair of adjacent months, let a_1 be the total employment in the matched sample for the current month at first closing. The matched sample is the set of units that have data for both the current month and the previous month at first closing (the matched sample at any other closing is defined similarly). Let b_1 be the total employment in the matched sample for the previous month at first closing.

Define a_2, as the total employment in the current month for those units that were added to the matched sample between first and second closing and define b_2 similarily. a_3 is total employment for the current month for those units that were added to the matched sample between second and third closing and similarly for b_3. Then, for

example, the total employment in the matched sample for the current month at third closing is $a_1+a_2+a_3$.

The first closing month to month link is $\hat{\beta}_1 = a_1/b_1$. The second closing month to month link is $\hat{\beta}_2 = (a_1+a_2)/(b_1+b_2)$ and so on for third and later closings.

We will be estimating the variance of these links as well as the variance of closing revisions of these links. For example, the first to second closing revision of the month to month link estimate is:

$$R_2 = (a_1/b_1)-[(a_1+a_2)/(b_1+b_2)] = \hat{\beta}_1 - \hat{\beta}_2.$$

Other between closing link revisions are defined similarly and estimating their variance will be exactly analogous to estimating the variance of R_2. For the remainder of this document we consider only first closing links, second closing links, and the revision, R_2, between these links. From now on drop the subscript 2 on R_2 and refer to it as R.

Under the model that describes establishment "all employment" in the Current Employment Statistics (CES) survey we have:

$$a_1 = \beta_1 b_1 + \varepsilon_1 \text{ and } a_2 = \beta_2 b_2 + \varepsilon_2 \quad (2.1.1)$$

where $E(\varepsilon_1) = E(\varepsilon_2) = 0$,
β_1 and β_2 are unknown constants,
$Var(\varepsilon_1|b_1)=Kb_1(\alpha)$, & $Var(\varepsilon_2|b_2)=Kb_2(\alpha)$
the εs are independent,
K and α are unknown constants and the $b_i(\alpha)$ are given as follows:

$$b_i = \sum_{k \varepsilon s_i} x_k \quad \& \quad b_i(\alpha) = \sum_{k \varepsilon s_i} x_k^{\alpha},$$

s_i is the set of i^{th} closing units (the matched sample for the reference month and its immediate predecessor at i^{th} closing) and x_k is "all employment" in the kth sample establishment for the past month.Let y_k denote "all employment" in the kth sample establishment for the current (reference) month. Then $a_i = \sum_{k \varepsilon s_i} y_k$, for example. .

This model implies that the conditional variance of the first and second closing links given the $\{b_i \& b_i(\alpha)\}$ are respectively:

$$V(a_1/b_1) = Kb_1(\alpha)/b_1^2 \text{ and} \quad (2.1.2)$$

$$V([a_1+a_2]/[b_1+b_2]) = K[b_1(\alpha)+b_2(\alpha)]/(b_1+b_2)^2$$

The expected value over the distribution of the $\{b_i \& b_i(\alpha)\}$ (both with respect to the sampling distributionand the model 2.1.1) of these conditional variance terms are the unconditional variances of these closing links and thus these variance expressions (2.1.2) are unbiased estimators of the unconditional variances ot the closing links (given that we know K and α).

We will say that there is no systematic difference between first and second closing responders for estimating month to month change (first and second closing links have the same expected value) by equating the ßs ($\beta_1 = \beta_2$). In this case E(R)=0 and the conditional variance of R given b_1 and b_2 is:

(2.1.3)

$$V_R = [Kb_1(\alpha)(b_1+b_2)^2 + b_i^2(b_1(\alpha)+ b_2(\alpha))K - 2Kb_1(\alpha)b_1(b_1+b_2)]/b_i^2(b_1+b_2)^2$$

Note that this variance expression does not involve the ßs. The conditional expected value of R given the $\{b_i \& b_i(\alpha)\}$ is $A = (\beta_1-\beta_2)[b_2/(b_1+b_2)]$ and this is zero when $\beta_1=\beta_2$. Thus, letting δ be an error term, we can write:

$$R = A + \delta \text{ where } E(\delta)=0 \text{ and } Var(\delta)=E(V_R)$$

(since $E(\delta|b)=0$ and $Var(\delta|b)=V_R$).

Note that the unconditional variance of R is the expected value over the $\{b_i \& b_i(\alpha)\}$ of V_R and thus V_R is an unbiased estimate of this unconditional variance of R (given the constants α and K). Similarily, $V(\hat{\beta}_1)$ & $V(\hat{\beta}_2)$ are unbiased estimates of the unconditional variance of $\hat{\beta}_1$ and $\hat{\beta}_2$.

With this structure we can test the hypothesis: A=0. That is, the closing revision is caused by the natural variability in the month to month links [as described in question 1) above] and not due to some underlying difference with respect to employment change between first and second closing reporters.

2.2 Parameter Estimation

At unit level the model (2.1.1) is:

$$y_i = \beta x_i + \varepsilon'_i \text{ where } E(\varepsilon'_i|x_i)=0 \text{ and } V(\varepsilon'_i|x_i)=Kx_i^{\alpha}$$

(2.2.0)

where y_i is the i^{th} establishment's employment for the current (reference) month and x_i is that establishment's

employment for the previous month. ß is either $ß_1$ or $ß_2$ depending on whether the unit reported for first closing or second closing. The ε'_i for two different sample units (values of i) are independent.

This model implies that the conditional expected value of $(y_i - ßx_i)^2$ given x_i is Kx_i^α or $E_{xi}(y_i - ßx_i)^2 = Kx_i^\alpha$ where E_{xi} denotes conditional expectation given x_i. Using historical data, ß can be estimated with $\hat{\beta}$ where $\hat{\beta}$ is at least a third closing link. Then we have approximately:

$$(y_i - \hat{\beta} x_i)^2 = f(x_i) + \delta_i = Kx_i^\alpha + \delta_i, \quad (2.2.1)$$

where $E(\delta_i)=0$ and $Var(\delta_i)$ is finite for each sample unit with data for both the current month and the previous month.

Six cases are studied here; two linear approximations to the function, $f(x)=Kx^\alpha$, by three models for the variance of the $\{\delta_i\}$, $Var(\delta_i)$. The six cases are summarized (and named) in the table below.

Table 2.2.1

Var(δ_i) / f(x)	σ^2	$\sigma^2 x_i$	$\sigma^2 x_i^2$
Cx	V_2	V_{srs}	V_{wls}
c + dx	V_3	V_5	V_4

For example, we get V_{srs} when we approximate (2.2.1) with:

$(y_i - \hat{\beta} x_i)^2 = f(x_i) + \delta_i = Cx_i + \delta_i$, where $Var(\delta_i) = \sigma^2 x_i$, and under this linear model, compute the BLUE of C, $\hat{C}$. Thus

$$V_{srs} = \hat{C} / \sum x_i .$$

For notational simplicity, we restrict ourselves to estimating the variance of $\hat{\beta}_1$ in the remainder of this section. The other cases, $\hat{\beta}_2$ and R, are analogous. Thus, in the rest of this section, the summations are over s_1 and n is the number of sample units in s_1.

We obtain V_2 by substituting $\hat{\alpha}=1$ and $\hat{C}=(X^tX)^{-1}X^tZ$ into (2.1.2) and (2.1.3) where X is the column vector of x_j j=1,2,...,n and Z is the column vector of $(y_j - \hat{\beta} x_j)^2$ j=1,2,....,n. If we let $\hat{V}_j = (y_j - \hat{\beta} x_j)^2$ then V_2 for the variance of $\hat{\beta}_1$ is given by: $\frac{1}{\Sigma x_j \Sigma x_j^2} \Sigma x_j \hat{V}_j$ where all the sums are over s_1.

If we substitute the variance function c + dx for Kx^α, in (2.1.2) and (2.1.3) we get V_3, where c and d are estimated as the vector $(\hat{c}, \hat{d})^t = (X^tX)^{-1}X^tZ$. Here X is the nx2 matrix, the j^{th} row of which is, (1 , x_j) for j=1,2,....,n, and Z the column vector of $(y_j - \hat{\beta} x_j)^2$ for j=1,2,.....,n. For example, V_3 for the first closing link, $\hat{\beta}_1$, is $(n\hat{c} + \hat{d}\, b_1)/b_1^2$.

Note that both V_2 and V_3 are BLUE under the questionable assumption that the variance of the $\{(y_j - \hat{\beta} x_j)^2\}$, is the same across all sample units. Modelling hetroscadasticity into these unit level variances we get V_{srs}, V_{wls}, V_5, and V_4 by computing their corresponding estimators from the least squares estimates of C (or c & d) under the hetroscadastic structure given in the table above. V_{wls} is the weighted least squares variance estimator, Royall (1981), and V_{srs} is the variance of a link assuming simple random sampling.

When $\hat{V}_j = Cx_j + \delta_j$, where $E(\delta_j)=0$, Table 2.2.1 defines variance estimators for three cases,

$E(\delta_j^2)=$

1) σ^2 2) $\sigma^2 x_i$ 3) $\sigma^2 x_i^2$

For each of these three cases the BLUE, $\hat{V}(\hat{\beta}_1)$, is

1) $V_2 = \frac{1}{\Sigma x_j \Sigma x_j^2} \Sigma x_j \hat{V}_j$ 2) $V_{srs} = \frac{1}{(\Sigma x_j)^2} \Sigma \hat{V}_j$

3) $V_{wls} = \frac{1}{n \Sigma x_j} \Sigma \hat{V}_j / x_j$

Now suppose, $\hat{V}_j = Cx_j^q + \delta_j$, then $V(\hat{\beta}_1) = \frac{C}{(\Sigma x_j)^2} \Sigma x_j^q$ and the expected values of the three estimators, V_2, V_{srs}, and V_{wls}, are respectively:

1) $\frac{C}{\Sigma x_j^2 \Sigma x_j} \Sigma x_j^{1+q}$ 2) $\frac{C}{(\Sigma x_j)^2} \Sigma x_j^q$

3) $\frac{C}{n\Sigma x_j}\Sigma x_j^{q-1}$

By Jensen's inequality we have for all q≥1:

1) ≥ 2) ≥ 3) and for all q≤1:

3) ≥ 2) ≥ 1)

Thus V_{srs} remains unbiased for all q and V_2 & V_{wls} are positively or negatively biased depending on whether q≥1 or q≤1.

3. AN EVALUATION

The purpose of this section is to establish the existance and strength of the relationship between the variance estimators proposed here and the variances that they are supposed to be estimating. When the data behaves according to the model given in Section 2, how do the six variance estimators defined in Table 2.2.1 perform?

Tables 3.1 and 3.2 contain the results of a simulation study where the data follow the model given in Section 2. The April 1988 CES sample micro data for all employment for SIC 1531 and establishments with less than 100 employees were used as the "previous" month's data. The May data was generated directly from this April data following the model in Section 2.

These tables are based on 50 replications of this data generation. Actual CES closing codes for the months of April/May 1988 as well as actual CES data for April was used in order to keep as close as practically possible to the data flow of this survey. The all employment data for May is generated from the April data following the model (2.1.1) with K=.0004 and α= the value given in the left-most column for each row of the tables below.

The entries in the following tables are the estimated variances times 10^7. For example, in Table 3.1, for α=1.2, the average of the estimated variances of $\hat{\beta}_1$ using V_2 is 9.7×10^{-7}.

The Target column contains, for each α:

$$V(\beta) = (1/49)\sum_{i=1}^{50}(\hat{\beta}_{1i} - \overline{\beta}_1)^2$$, where the sum is over the 50 replications of the data generation described above and $\hat{\beta}_{1i}$ is the i^{th} replicate link for closing 1, $\overline{\beta}_1$ is the average over the 50 replications of these 1^{st} closing replicate links. A good variance estimator should tend to the value in the Target column.

Table 3.2 contains estimated variances of the variance estimates whose means are contained in Table 3.1. They are computed exactly analogously to V(β) with $\hat{\beta}_{1i}$ replaced with, in the case of V_2, with V_{2i} to give estimates of the variance of the $\{V_2\}$ in Table 3.1.

50 replications are not enough to get more than a rough idea of what works (and what doesn't). The relative error of the variance estimates is rather substantial in this case study (20% to 40%). For example, the relative variance of V_{srs} for α = 1.4 is $18.2/(18.4)^2$ = .054 and the rel error is .23.

Note in Table 3.1 that for the values of $1< \alpha \leq 2$ we have $V_2 \geq V_{srs} \geq V_{wls}$, exactly as predicted in the last paragraph of section 2.

Table 3.1 First Closing Variance Estimates: $Vx10^{-7}$

α	V_2 V_3	V_{srs} V_5	V_{wls} V_4	Target
1.2	9.7 9.2	9.5 9.2	9.3 9.2	8.5
1.4	19.2 17.4	18.4 17.4	17.4 17.5	22.9
1.6	44.9 37.3	40.0 37.3	34.9 36.6	57.9
1.8	89.5 76.0	78.6 76.0	67.1 74.2	57.4
2.0	171 140	147 140	121 136	127

We should mention that V_3 and V_5 are algebraically identical. This anomoly is due to estimating the parameters c & d with the same data used to estimate β_1. If these parameters are estimated from historical data (preferable) then V_3 and V_5 will only be similar.

Royall and Cumberland (1978,1981) note that V_{wls} has a negative bias in all the situations they consider. As we have seen in section 2, it can also have a positive bias but V_{wls} can also have a much smaller variance than the alternatives considered above. The net result is that the MSE of V_{wls} is smaller than that of V_{srs}, although V_{srs} is "uniformly" unbiased (see last paragraph of section two).

Royall, Cumberland, and Eberhardt suggest as a robust (against bias) alternative, the Jacknife variance estimator. Unfortunately, because of the vast amount of

computation already required just to produce the basic estimates each month, a computationally simple variance estimator is required if variances are to be produced monthly along with the estimates.

Table 3.2 Estimated Variances of First Closing Variance Estimates : $Vx10^{-14}$

α	V_2 V_3	V_{srs} V_5	V_{wls} V_4
1.2	7.4 6.0	5.8 6.0	6.1 6.1
1.4	27.0 16.8	18.2 16.8	14.6 16.9
1.6	310 134	142 134	64.6 104
1.8	1421 767	770 767	370 653
2.0	5330 2347	2566 2347	1111 1925

4. SOME ESTIMATES OF VARIANCE IN SELECTED SICs

In this section we summarize the results of computing variance estimates for links and revisions in several CES estimation cells. To be conservative (overestimate) we used V_2 and V_3.

Variance estimates were computed from historical data for April 1988 through March 1989 in SICs, 2921, 5251, 1531, and 8631. These SICs were further divided into various size classes to give a total of ten industry by size estimation cells.

The parameter estimates, $\hat{C}$ for V_2 and $(\hat{c},\hat{d})$ for V_3, vary substantially between different SICs and size classes. Within an SIC/size cell they can vary a great deal over time. The values of the parameters, $\hat{C}$ or $(\hat{c},\hat{d})$, that are used to estimate the variance of the links and revisions come from averaging nine estimates for these parameters in each estimation cell. These nine estimates are derived from estimates computed using the data for the eleven pairs of adjacent months, after deleting the largest and smallest of the eleven estimates for each parameter.

Table 4.1 contains the first and second closing link estimates and their difference, the first to second closing revision for each of the eleven pairs of months in SIC 1531 for units with less than 100 employees. The other 9 CES estimation cells were fairly similar and are not given here because of space constraints. The first row in Table 4.1 contains the estimate of 2σ (using V_2) for each link or revision directly beneath. This 2σ estimate is obtained by averaging eleven estimates of 2σ obtained from each of the eleven pairs of adjacent months. This average sigma seemed appropriate because there can be considerable variability across months for these variance estimates.

In Table 4.1., the Jul - Aug first closing link is 0.992 and the 2σ for these first closing links is .03. The * is used to denote that the particular entry is beyond two standard errors of unity for the links or two standard errors of zero for the revisions. Thus the * indicates that the trend being measured by the link may be statistically significant.

Table 4.1. seems to say that for SIC 1531 and those units with less than 100 employees, the closing revisions are largely of a magnitude that does not indicate much underlying difference between first and second closing links. Another way to say this is that these first and second closing links appear to be measuring the same thing and the estimation process is "under control".

By second closing when more sample units have reported, the standard error estimate is smaller.

Table 4.1 Estimates, Revisions, and Estimated Error (V_2) in SIC 1531, Employment between zero and 100

	β_1	β_2	R
Estimated 2-sigma	.03	.022	.019
Months			
Apr-May	1.028	1.017	-.001
May-Jun	1.024	1.033*	0.009
Jun-Jul	1.013	1.006	-.007
Jul-Aug	0.992	0.998	-.004
Aug-Sep	0.982	0.994	0.012
Sep-Oct	1.008	1.006	-.002
Oct-Nov	0.973	0.975*	0.002
Nov-Dec	0.975	0.984	0.009
Dec-Jan	0.983	0.980	-.003
Jan-Feb	0.960*	0.980	-.020*
Feb-Mar	0.998	0.998	0.000

5. CONCLUSIONS

We studied a fairly diverse set of quick and dirty estimators. Hopefully, at least one of the six variance estimators will fit any situation that we might come accross in CES estimation. If MSE is the most important criterion, then V_{wls} is the estimator of choice. If we wish to be conservative (overestimate variance) then V_2 would be a likely candidate in the smaller employment size cells and V_3 in the larger cells. In cases where we know virtually nothing about the micro variances then V_{srs} will be unbiased under a wide range of alternative models.

Once we estimate these model coefficients with the historical micro data, all six estimators reduce to little more than generalized variance estimators.

In section four, we estimated variances and constructed confidence intervals for the first and second closing estimates and for their difference (the first to second closing revision). It appears that the main difference between closing estimators is variance and that bias is not a major factor in most SIC/size cells. That is, there is no systematic difference other than variance between first and second closing estimates.

We will soon have the actual universe data for California and several other states with which to test the variance estimators suggested here.

My thanks to George Stamas, Chief, Current Employment Statistics for his many helpful suggestions.

REFERENCES

1) Madow, L and Madow, W (1978), "On Link Relative Estimators", ASA Proceedings of the Section on Survey Research Methods, 534-539.

2) Royall, Richard and Cumberland, W.G. (1978), Variance Estimation in Finite Population Sampling, Journal of The American Statistical Association, 351-358.

3) Royall, Richard and Cumberland, W.G. (1981), An Empirical Study of the Ratio Estimator and Estimators of its Variance, Journal of the American Statistical Association, 66-77

4) Royall, Richard and Cumberland, W.G. (1981), Reply to Comments on, "An Empirical Study of the Ratio Estimator and Estimators of its Variance", Journal of the American Statistical Association, 87-88.

5) Royall, Richard. (1981), "Study of the Role of Probability Models in 790 Survey Design and Estimation" , Bureau of Labor Statistics contract report #80-98.

6) Royall, R.M. and Eberhardt, K.R. (1975), "Variance Estimators for the Ratio Estimator", Sankhya, Ser C,37, 43-52.

7) West, S. (1984), "A Comparison of Estimates for the Variance of Regression-Type Estimators in a Finite Population", ASA Proceedings of the Section on Survey Research Methods,