EXPLORATORY ANALYSIS OF GENERALIZED VARIANCE FUNCTION MODELS FOR THE U.S. CURRENT EMPLOYMENT SURVEY

Larry L. Huff, John L. Eltinge, and Julie Gershunskaya, BLS
Larry L. Huff, Statistical Methods Division, OEUS, PSB 4985, U.S. Bureau of Labor Statistics,
2 Massachusetts Ave., N.E., Washington, DC 20212-0001, Huff_L@bls.gov

Key Words: **Chi-square distribution approximations; Design-based inference; Lack of fit; Residual plot; Lognormal approximations;**

1. Introduction

For the Current Employment Statistics Program, approximately unbiased and stable variance estimators are important for the empirical evaluation of standard design-based point estimators, and for production of related small domain estimators. In some cases, standard design-based variance estimators can be relatively unstable, which may lead to consideration of alternative variance estimators based on generalized variance functions. This paper presents an exploratory analysis of generalized variance function models for estimates of total monthly employment within domains determined by the intersection of metropolitan statistical area and major industrial division. Three topics receive principal attention: (a) a detailed description of features of the underlying sample design that are important in variance estimation; (b) graphical evaluation of potential biases in generalized variance function estimators; and (c) omnibus measures of the relative magnitudes of the fixed and random components of model lack of fit.

2. Survey Background

The Current Employment Statistics Program is conducted by the U.S. Bureau of Labor Statistics as a Federal-State cooperative program. The Bureau specifies the design of the survey and operational procedures in close coordination with the States. On a monthly basis, the Bureau produces national estimates while the States produce State and local area estimates. The data collected for the survey includes all employees, production workers, production workers hours paid, and production workers payroll. This data is collected each month for a sample of approximately 220,000 State Unemployment Insurance (UI) accounts from each of the 50 States and the District of Columbia. The primary estimate made from the survey is the monthly total "all employee" estimate which is published approximately 3 weeks after the reference period of the collected data. Because of the importance of the payroll employment estimates produced and the timeliness of the estimates, the CES estimates are recognized as a leading economic indicator. They provide one of the first available signs of the state of the economy each month. The estimates are also used as input into many other major economic indicators.

3. Sample Allocation and Selection

The sample for the survey is selected each year from a sample frame compiled from State Unemployment Insurance accounts. These UI accounts are compiled by BLS as part of another Federal-State cooperative program known as the Covered Employment and Wages (ES-202) Program. The data is collected for this program by the States under contract with the BLS and consists of over 7,000,000 individual establishment records representing virtually every employer in each of the 50 States and the District of Columbia. These UI account and establishment records include information on total employment, Standard Industry Code (SIC), and county or area which is used to code the Metropolitan Statistical Area code (MSA). The individual establishment records in each UI account have the same types of information coded as that collected and coded for the UI account. Each of these establishment records may operate in a slightly different SIC and area than that coded for the "parent" UI account record. The UI account parent records (the sample unit) on the sample frame are stratified by State into 11 major industrial divisions (MID), and 8 employment size classes (Size) for a total of 88 strata for each State. The largest Size units (Size 8 – 1,000 or more employees) are selected with absolute certainty. The sample collection resources are fixed for each State through an administrative process. The resources required to collect the certainty units are removed from each State's total. The remainder of the sample for each State is allocated to the remaining 77 non-certainty strata using a program that optimizes the allocation to provide the best estimate (smallest sampling error) of State total employment. The input into this process includes the estimated resources required to collect data from each unit and the over-the-month coefficient of variation for employment as calculated from the sample frame. Thus, the sample is truly a State based design. The sample strata are defined by

State, MID, and Size. Before sample selection, the units within each stratum are sorted by MSA to ensure that MSAs have sample units selected from them in direct proportion to the number of units in the MSA. The sample is selected from each State/MID/Size stratum, after sorting by MSA, by taking one random start for the stratum and then selecting the remainder of the sample units by taking every N/n th unit in the stratum. This does not ensure that all MSAs within the stratum will have sample units selected, however, it does ensure that if units are selected from a stratum with a probability of ¼ and an MSA in the stratum has 4 units, then 1 unit from the MSA will be selected. Within a given State x MID x Size stratum, units are sorted only according to MSA and an uninformative permanent random number. This provides a degree of randomness for the MSA sample selected within each stratum. This also provides a degree of independence between the MSA samples that are used in making MSA/MID estimates of all employees. After sample selection, each sample unit is given a sample weight which is equal to the inverse of the probability of selection. For additional background in the CES sample design, see Butani et al, (1997), Werking (1997) and references cited therein.

4. Point Estimation

We will limit our discussion of estimation to all employment (AE) estimates since that is the principal statistic estimated from the survey and the one estimate where the need for reliable and stable sampling error information is the strongest. For a given month, the individual establishment data is collected for all responding establishments within each selected UI account. This provides us the actual MID and MSA where the employees in the establishment are working. When making estimates, the individual establishment records are used to ensure that we place the employees in the appropriate industry and area. The form of the monthly estimate of AE is referred to as a weighted link relative estimator:

$$\hat{Y}_t = \frac{\sum_{i \in M_t} w_i y_{i,t}}{\sum_{i \in M_t} w_i y_{i,t-1}} \hat{Y}_{t-1} \qquad (4.0)$$

where $\hat{Y}_t$ = total employment estimate for month t, $\hat{Y}_{t-1}$= total employment estimate for month t-1, $\sum_{i \in M_t} w_i y_{i,t}$ = summed weighted employment total in month t for matched sample units at time t and t-1, i.e., matched sample units at time t reporting non-zero data for month t and t-1, and $\sum_{i \in M_t} w_i y_{i,t-1}$ = summed weighted employment total in month t-1 for matched sample units at time t.

Once each year the estimates are benchmarked or adjusted to the true population employment values from the Covered Employment and Wages Program. For t=0, the estimator shown above is started with Y_0 in the place of $\hat{Y}_{t-1}$, Y_0 being the true population value at the benchmark month or month 0.

5. Variance Estimation

Variance estimation is accomplished using balanced half-sample (BHS) methodology. The BHS method addresses all of the CES design features including stratification, allowances for imputation variance and for the finite population correction. Details of the procedure are provided in Wolter et al, (1998). The basic form for the variance estimator is:

$$\hat{v}_k(\hat{\theta}) = \frac{1}{\gamma^2 k} \sum_{\alpha=1}^{k} (\hat{\theta}_\alpha^+ - \hat{\theta})^2$$

where $\hat{\theta}$ = the full sample weighted link relative estimator (4.0) for total employment as described above; k is the number of half samples (both the half sample and its complement half sample are used); γ = a mixing parameter used to weight the half samples $(1+\gamma)$, and the complement of the half samples $(1-\gamma)$, with γ set = 0.5; and $\hat{\theta}_\alpha^+$ = the half sample weighted link relative estimator for the α th half sample (using the half sample and its complement). The weights used for these half sample estimates are adjusted for the half sample, imputation, and the finite population correction factor.

The set of half samples used for calculating variances for Statewide/All Industry estimates are constructed by employing the use of a Hadamard matrix with columns representing different strata and rows designating different half samples. The number of sample strata in each State is 66 since there are 11 MIDs and 6 size classes. (For purposes of variance estimation, the largest 3 size classes are collapsed together.) A Hadamard Matrix of order 68 is used to designate the half samples. This results in 68 half sample replicates used in calculation of Statewide/All

Industry estimates, which in turn produce variance estimates for the aggregate estimates.

The estimates that are of interest for our study are estimates of total employment in a given MID within a specified MSA. The variances needed for these estimates have two purposes. The first is for use by the States in analyzing their small area estimates made using the weighted link relative estimator (1) described above. The second is for use in weighting the weighted link relative estimate in a weighted least squares small area estimator. In calculating the variances for MSA/MID estimates, the only remaining stratification uses the 6 combined size classes within the MID. A new Hadamard matrix of order 8 is used to define 8 half samples for each of the MSA/MID variance estimates. The columns represent the six size strata (the first and last colunms are omitted) and the rows designate the 8 half samples. The variance estimates calculated have only 6 degrees of freedom and display a substantial degree of variability.

6. Finding a Generalized Variance Function

Due to the above mentioned stability problems for standard design-based estimators, we explored the possibility of using generalized variance functions (GVFs) for small domains defined by the intersection of MSA and MID. For some general background on GVFs, see Johnson and King (1987), Valliant (1987) and references cited therein. Woodruff (1992, 1993) considered generalized variance functions for high-level point estimators from the CES under its previous quota-sample design. The present paper restricts attention to results under the CES probability design.

We consider a linear regression (GVF) model with $\ln(\hat{V}_{mt})$ as a dependent variable, where $\hat{V}_{mt}$ is the BHS estimate of variance for $\hat{Y}_{mt}$, the employment estimator for domain m in month t. After exploring many alternative GVF models, the search was narrowed to the model:

$$\ln(\hat{V}_{mt}) = \gamma_0 + \gamma_1 \ln(x_{m0}) + \gamma_2 t + \gamma_3 \ln(n_{mt}) + e_i \qquad (6.0)$$

where $x_{m0} =$ the true employment in area m for the benchmark period 0; t = month label for number of months from benchmark period; $n_{mt} =$ number of responding sample UI Accounts in domain m at time t; and, e_{mt} = a random error term with expectation equal to zero and variance equal to σ_e^2.

If $e_{mt} \sim N\left(0, \sigma_e^2\right)$, an approximately unbiased estimator of the design expectation of $\hat{V}_{mt}$ is:

$$V_{mt}^* \equiv \exp\{\hat{\sigma}_e^2 / 2 + \hat{\gamma}_0 + \hat{\gamma}_1 \ln(x_{m0}) + \hat{\gamma}_2 t + \hat{\gamma}_3 (n_{mt})\} \quad (6.1)$$

In many applications, a variance estimator follows approximately a chi-square or lognormal distribution. To evaluate the adequacy of these approximations, for the CES, we produced the quantile-quantile plots displayed in Figures 1 and 2 for data collected for the Wholesale Trade industry in M = 100 MSAs and T = 12 months. Figure 1 displays a plot of the quantiles of the relative remainder terms

$$d_{mt} = (V_{mt}^*)^{-1} (\hat{V}_{mt} - V_{mt}^*) \qquad (6.2)$$

(vertical axis) against the corresponding quantiles of a standardized chi-square distribution on six degrees of freedom. Note especially that the upper tail of the distribution of d_{mt} is much more extreme than would be anticipated under a standardized chi-square distribution on six degrees of freedom.

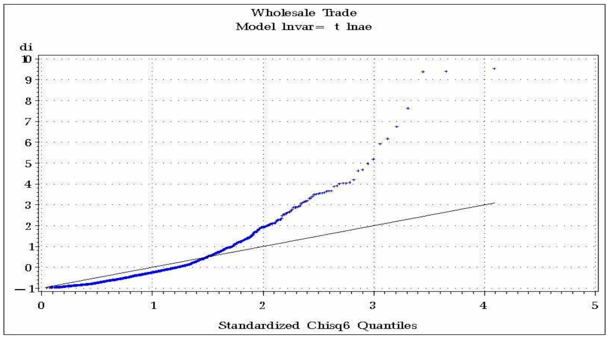

Figure 1

Figure 2 presents the corresponding lognormal plots. Note that under model (6.0), the lognormal distribution provides a better approximation to the upper tail of the distribution of d_{mt}.

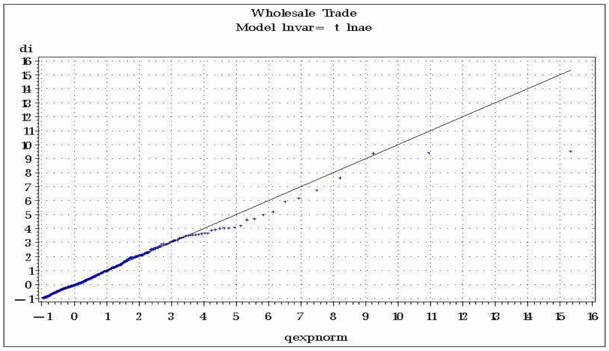

Figure 2

7. Diagnostics to Assess the Adequacy of Approximation (6.0): Direct Evaluation of Remainder Terms

Define the relative remainder terms

$$d_{mt} = (V_{mt}^{*})^{-1}(\hat{V}_{mt} - V_{mt}^{*}) \qquad (7.1)$$

and their average

$$\bar{d}_{.t} = M^{-1}\sum_{m=1}^{M} d_{mt} \qquad (7.2)$$

and variance

$$S_{.t}^{2} = (M-1)^{-1}\sum_{m=1}^{M}(d_{mt} - \bar{d}_{t})^{2} \qquad (7.3)$$

for a given month t. Routine arguments then show that $S_{.t}^{2}/M$ is an approximately unbiased estimator of $V(\bar{d}_{t})$. Thus, under additional regularity conditions, the random variables

$$\{M^{-1}S_{.t}^{2}\}^{-1/2}\bar{d}_{.t} \qquad (7.4)$$

should follow approximately a t distribution on M-1 degrees of freedom, provided $\sum_{m=1}^{M} E\left(\frac{\hat{V}_{mt} - V_{mt}^{*}}{V_{mt}^{*}}\right) = 0$.

In addition, define

$$\bar{d} = T^{-1}\sum_{t=1}^{T}\bar{d}_{.t} = (MT)^{-1}\sum_{t=1}^{T}\sum_{m=1}^{M} d_{mt} = M^{-1}\sum_{m=1}^{M}\bar{d}_{m.}$$

where $\bar{d}_{m.} = T^{-1}\sum_{t=1}^{T} d_{mt}$

Thus, $\bar{d}$ is the average of M independent random variables $\bar{d}_{m.}$. Therefore, an approximately unbiased estimator of $V(\bar{d})$ is $(M^{-1}S_{..}^{2})$ where $S_{..}^{2} = (M-1)^{-1}\sum_{m=1}^{M}(\bar{d}_{m.} - \bar{d})^{2}$. This in turn means that under additional regularity conditions,

$$(M^{-1}S_{..}^{2})^{-1/2}\bar{d}_{..} \qquad (7.5)$$

should follow approximately a t distribution on M-1 degrees of freedom, provided

$$\sum_{m=1}^{M}\sum_{t=1}^{T} E\left(\frac{\hat{V}_{mt} - V_{mt}^{*}}{V_{mt}^{*}}\right) = 0 \qquad (7.6)$$

Thus, (7.5) provides a summary indicator of the overall relative bias, if any, of V_{mt}^{*} as an estimator of $E(\hat{V}_{mt})$. Similarly, expression (7.4) provides month specific indications (averaging over metropolitan areas) of the overall relative bias of V_{mt}^{*}.

We applied the ideas leading to expressions (7.4) and (7.5) to data from T=12 months (January through December, 2000) for five industries. Figure 3 displays the values $\bar{d}_{t}$ and the corresponding approximate pointwise 95% confidence intervals

$$\bar{d}_{t} \pm t_{M-1,.975}(M^{-1}S_{.t}^{2})^{1/2}$$

for the month-specific average relative bias terms

$M^{-1}\sum_{m=1}^{M} E\left(\frac{\hat{V}_{mt} - V_{mt}^{*}}{V_{mt}^{*}}\right)$. The plotting symbols A through E correspond, respectively, to five industries, construction, combined construction and mining, durables manufacturing, nondurables manufacturing and wholesale trade. For these five industries, M was equal to 61, 36, 131, 100, and 100 respectively. The values of M vary across industry because we omitted from consideration any MSA x MID combinations that had less than 12 responding sample UI accounts in any month between March 1999 and December 2000. Note especially that in all cases, the confidence intervals in Figure 3 include the value zero, which would be consistent with the unbiasedness condition (7.6).

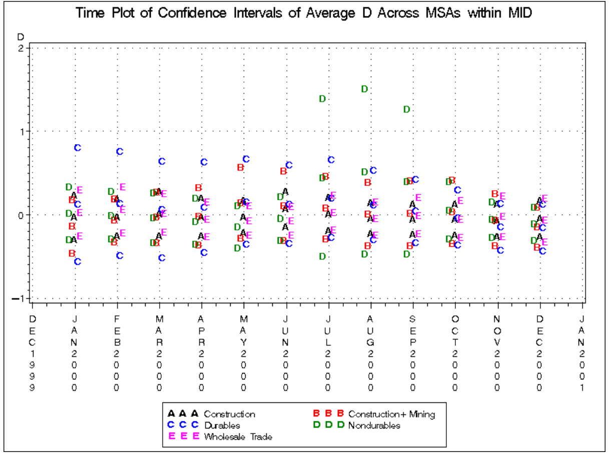

Figure 3

8. Diagnostics for Chi-Square Approximations

Note that (7.4) and (7.5) depend on distributional assumptions only in a limited way, e.g., through approximate normality of a mean of M independent random variables. If we also assume that

$$c\hat{V}_{mt} / V_{mt}^{*}$$

follows a chi-square distribution on c degrees of freedom for some c>0, then for a given month t the terms (d_{mt}^{2}) , m=1, …, M are independent and identically distributed with expectation equal to $V(d_{mt}) = 2/c$. In particular, if c=6, then $V(d_{mt}) = 1/3$. Define

$$D_t = M^{-1} \sum_{m=1}^{M} d_{mt}^{2} .$$

If our estimator ratio $\hat{V}_{mt} / V_{mt}^{*}$ satisfies the chi-square distributional approximation, then the terms

$$\hat{R}_{LFt} = D_t - 1/3$$

have a mean equal to zero, and an approximately unbiased variance estimator is

$$\hat{V}(\hat{R}_{LFt}) = M^{-1}(M-1)^{-1} \sum_{m=1}^{M} [(d_{mt})^2 - (D_t)]^2 \qquad (8.1)$$

with associated confidence intervals

$$\hat{R}_{LFt} \pm t_{M-1,.975} \sqrt{\hat{V}(\hat{R}_{LFt})} \qquad (8.2)$$

A confidence interval (8.2) that falls entirely above zero would correspond to relative differences d_{mt} that are more variable than anticipated under a X_6^2 approximation. This might be attributable to $\hat{V}_{mt}$ being associated with fewer degrees of freedom than the nominal c=6. On the other hand, this phenomenon might also arise from a lack of fit of the values $\ln(\hat{V}_{mt})$ to model (6.0). Note especially that the diagnostics $\bar{d}_t$ and $\bar{d}$ are sensitive to systematic deviations of V_{mt}^{*} from $E(\hat{V}_{mt})$, across all areas in a given month t, or all months. In contrast with this, $\hat{R}_{LFt}$ will reflect local deviations $(V_{mt}^{*})^{-1}(\hat{V}_{mt} - V_{mt}^{*})$ that may not necessarily all have the same sign.

Similarly, define the aggregate goodness-of-fit measure

$$\hat{R}_{LF} = D_0 - 1/3$$

with associated variance estimator

$$\hat{V}(\hat{R}_{LF}) = M^{-1}(M-1)^{-1} \sum_{m=1}^{M} \left[T^{-1} \sum_{t=1}^{T} (d_{mt})^2 - D_0 \right]^2 \qquad (8.3)$$

where

$$D_0 = (MT)^{-1} \sum_{m=1}^{M} \sum_{t=1}^{T} (d_{mt})^2$$

and approximate 95% confidence interval

$$\hat{R}_{LF} \pm t_{M-1,.975} \sqrt{\hat{V}(\hat{R}_{LF})} \qquad (8.4)$$

Figure 4 displays the confidence intervals (8.2) for January through December of 2000 for the same MSAs and the same five industries considered in Figure 3, with the same industry labels A through E.

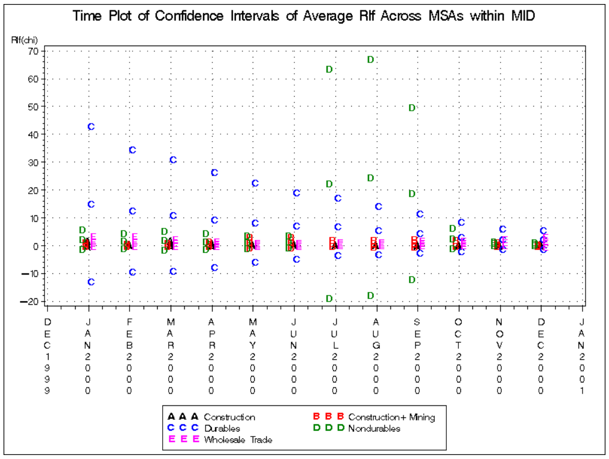

Figure 4

9. Diagnostics for Lognormal Approximations

The preceding subsection presented diagnostics intended to identify cases in which the relative errors d_{mt} deviated substantially from their expectations under an idealized chi-square approximation. If we instead had $\ln(\hat{V}_{mt} / V_{mt}^{*})$ following a normal distribution with mean zero and variance σ_e^2, the corresponding diagnostics would be the same as in Section 8, but with $\hat{R}_{LFt}$ replaced by

$$\tilde{R}_{LFt} = D_t - \left[\exp(\hat{\sigma}_e^2) - 1\right]$$

and associated approximate 95% confidence interval

$$\tilde{R}_{LFt} \pm t_{M-1,.975}\sqrt{\hat{V}(\hat{R}_{LFt})} \qquad (9.1)$$

Note especially that relative errors d_{mt} that display a greater degree of dispersion (heavier tails) than would be observed under a lognormal model will tend to produce confidence intervals (9.1) that fall entirely above zero. Similarly, relative errors that display less dispersion (lighter tails) than would be observed under a lognormal model will tend to produce confidence intervals that fall entirely below zero. (Conversely, confidence intervals that include zero are consistent with a lognormal model for the relative errors d_{mt} .

Similar comments apply to the quantity averaged over time,

$$\tilde{R}_{LF} = D_0 - \left[\exp(\hat{\sigma}_e^2) - 1\right]$$

with associated confidence interval equal to (8.4) with $\hat{R}_{LF}$ replaced by $\tilde{R}_{LF}$.

Figure 5 displays the confidence intervals (9.1), with labeling similar to that for Figure 2.

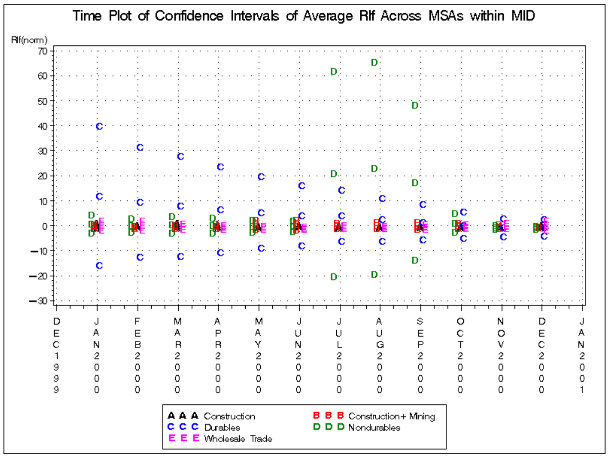

Figure 5

10. Acknowledgements

The authors thank Moon Jung Cho for helpful discussions of the issues covered in this paper. Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

11. References

Butani, S., Harter, R., and Wolter, K. (1997). Estimation Procedures for the Bureau of Labor Statistics Current Employment Statistics Program. *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 523-528.

Butani, S., Stamas, G. and Brick, M. (1997). Sample Redesign for the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 517-522.

Cochran, W. G. (1977). Sampling Techniques, 3rd ed. New York: John Wiley.

Johnson, E.G., and King, B.F. (1987). Generalized variance functions for a complex sample survey. *Journal of Official Statistics*, 3, 235-250.

Valliant, R. (1987). Generalized Variance Functions in Stratified Two-Stage Sampling. *Journal of the American Statistical Association* **82** 499-508.

Werking, G. (1997). Overview of the CES Redesign. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 512-516.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer Verlag.

Wolter, K., Huff, L., and Shao, J. (1998) Variance Estimation for the Current Employment Statistics Survey. Presented at the Joint Statistical Meetings, Dallas, August 13, 1998.

Woodruff, S. (1992). Variance Estimation for Estimates of Employment Change in the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 626-631.

Woodruff, S. (1993). Generalized Variance Functions for the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 860-865.