APPROXIMATION METHODS FOR COVARIANCE MATRIX ESTIMATORS USED IN ANALYSIS OF DIARY AND INTERVIEW DATA FROM THE U.S. CONSUMER EXPENDITURE SURVEY

John L. Eltinge and Amang Sukasih
U.S. Bureau of Labor Statistics and Texas A&M University

Abstract: Multivariate analyses of complex survey data often make extensive use of estimators $\hat{V}$ of the variance-covariance matrix V of a random vector $\hat{y}$ where the variances and covariances are evaluated with respect to the sample design. For example, $\hat{V}$ is often used in computation of quadratic-form test statistics, and also may be used in computation of generalized least squares point estimators. However, these analyses can be problematic when $\hat{V}$ is based on a small or moderate number of degrees of freedom. This paper considers methods for approximation of V, and for computation of associated modified estimators of V. Principal emphasis is placed on exploratory evaluation of modeling assumptions. The methods discussed in this paper are motivated and illustrated by analyses of interview and diary data from the U.S. Consumer Expenditure Survey.

Keywords: Correlation matrix; Diagnostics; Eigenvalues and eigenvectors; Exploratory analysis; Generalized variance functions; Misspecification effect matrix.

1. Introduction: Multivariate Analysis of Interview and Diary Data in the U.S. Consumer Expenditure Survey

This paper describes some methods for approximation of covariance matrices used in the analysis of complex survey data. In the interest of space, we focus primarily on one specific application to data from the U.S. Consumer Expenditure Survey. A more general theoretical development will be considered elsewhere.

The U.S. Consumer Expenditure Survey (U.S. Bureau of Labor Statistics, 1997) is a household survey based on a stratified multistage sample of consumer units, which are roughly equivalent to households. Within each selected primary sample unit, some consumer units are asked to provide data through a quarterly interview process, while other consumer units are asked to complete expenditure diaries. Interviewed consumer units are asked to report their expenditures separately for each of the previous three complete methods. For example, in an interview conducted on November 14, the interviewer would ask the consumer unit to report expenditures separately for October (the proximate month), September (the middle month) and August (the most distant month). In addition, consumer units assigned to the diary group are asked to complete a weekly expenditure diary in each of two consecutive weeks. For the discussion below, we will use the labels $i = 1$ through 5 for, respectively, data from the proximate interview month, the middle interview month, the distant interview month, the first diary week and the second diary week.

Eltinge, Sukasih and Weber (2000) carried out a detailed analysis of the vectors $\hat{y}_p = (\hat{y}_{p1},...\hat{y}_{p5})'$ where each $\hat{y}_{pi}$ is an estimator of mean annual expenditures based on source i, using data collected for each month or week, respectively, in a specific year p. These estimators use standard weighting adjustments to account for unequal selection probabilities, unit nonresponse and differences in the lengths of (monthly or weekly) reference periods. Thus, if there were no problems with nonsampling errors, one would anticipate that each estimator $\hat{y}_{pi}$ would be approximately unbiased for the

true mean expenditure for year p. Conversely, differential patterns of underreporting or nonreporting of specific expenditures would cause one or more of the estimators $\hat{y}_{pi}$ to have different expectations.

Evaluation of these and related bias issues in Eltinge, Sukasih and Weber (2000) led to extensive use of a design-based covariance matrix estimator $\hat{V}_{eep}$ computed through standard balanced repeated replication methods using 44 sets of replicate weights. For many national level estimands in this application, $40 \times \hat{V}_{eep}$ would be distributed approximately as a Wishart random matrix with 40 degrees of freedom. However, for estimands involving relatively rare types of expenditures or subpopulations concentrated in a moderate number of geographical areas, $\hat{V}_{eep}$ may be considerably less stable than would be suggested by the nominal 40 degrees of freedom term. Consequently, it is of interest to consider approximation methods that may lead to variance estimators that are more stable than $\hat{V}_{eep}$. Section 2 develops a class of estimators based on approximations for the correlation matrix and univariate variances associated with the matrix $\hat{V}_{eep}$. Section 3 applies the proposed methods to one specific example from the Consumer Expenditure Survey, with special emphasis on diagnostics to assess the adequacy of approximations used in Section 2.

2. A Correlation-Based Approximation Method for Covariance Matrix Estimators

For the years $p = 1,..., P$, let V_{eep} equal the variance-covariance matrix of the approximate distribution of the vector $\hat{y}_p$, evaluated with respect to the sample design. Consider the decomposition,

$$V_{eep} = D_p R_p D_p \tag{2.1}$$

where $D_p = diag(V_{eep11}^{1/2},..., V_{eep55}^{1/2})$, $V_{eepii}^{1/2}$ is the square root of the i-th diagonal element of V_{eep}, and R_p is the 5×5 dimensional correlation matrix associated with V_{eep}.

The decomposition (2.1) suggests the following two-step method for approximation of V_{eep}, and for construction of an alternative estimator of V_{eep} that may be more stable than $\hat{V}_{eep}$. First, one may use univariate generalized variance function methods to compute alternative estimators of V_{eepii} based on the standard estimators $\hat{V}_{eep}$ and available auxiliary information. (For some background on generalized variance function methods, see, e.g., Wolter (1985, Chapter 5), Johnson and King (1987), Valliant (1987) and references cited therein.) Let V_{eepii}^* be the resulting modified univariate variance estimators and define $D_p^* = diag(\{V_{eep11}^*\}^{1/2},..., \{V_{eep55}^*\}^{1/2})$.

In addition, consider the assumption that the correlation matrices R_p are constant over all years p, so that $R_p = R$, say. This assumption may be reasonable in cases like the Consumer Expenditure Survey in which the basic sample design remains the same over multiple years. Under the assumption of a common correlation matrix R, a simple estimator of R is $R^* = P^{-1} \sum_{p=1}^{P} \hat{R}_p$, where $\hat{R}_p$ is the correlation matrix computed directly from $\hat{V}_{eep}$. We then may use D_p^* and R^* to define the modified estimator $V_p^* = D_p^* R^* D_p^*$. In general, the performance of V_p^* as an estimator of V_p will depend on the

relative magnitudes of the sampling variability of $\hat{V}_{eep}$, the approximation errors in the generalized variance functions used to construct D_p^*, and the approximation errors $R_p - R$.

3. Application to Consumer Expenditure Survey Data

We applied the general ideas of Section 2 to data from the Consumer Expenditure Survey for expenditures contained in one six-digit Universal Classification Code group, 360330 (men's accessories, e.g., hats, ties and belts) for $P = 11$ years, 1987 through 1997. First, to develop appropriate estimators D_p^*, we explored the univariate variance-function patterns for each of the sources 1 through 5. Figure 1 displays a time plot of the sample coefficient of variation, $se(\hat{y}_{pi})/\hat{y}_{pi}$, with the plotting symbol set equal to the source label i. Due to differences in the numbers of respondents and the lengths of reference periods, the coefficients of variation for the diary sources ($i = 4$ and 5) are larger than those for the interview sources ($i =$ 1, 2 and 3). In addition, note that none of the five sources display any pronounced pattern of increase or decrease of the coefficients of variation over time. Figure 2 displays a plot of $se(\hat{y}_{pi})$ against $\hat{y}_{pi}$, with the plotting symbol again set equal to i. Taken across all five sources, this plot is roughly consistent with the simple linear regression model,

$$se(\hat{y}_{pi}) = \gamma_0 + \gamma_1 y_{pi} + \text{error} \tag{3.1}$$

where y_{pi} is defined to equal the expectation of $\hat{y}_{pi}$ and γ_0 and γ_1 are fixed coefficients. An ordinary least squares regression fit of the 55 points (5 sources across 11 years) displayed in Figure 2 resulted in a sample R^2 value equal to 0.83, suggesting a relatively good fit. Figure 2 also displays a fairly pronounced clustering of the $\hat{y}_{pi}$ for the interview sources (1, 2 and 3) and diary sources (4 and 5), respectively. This is consistent with point estimation bias issues raised in Section 1 and discussed further in Eltinge, Sukasih and Weber (2000). Initial exploratory work with expansion of model (3.1) to include additional covariates did not lead to substantial improvements, and thus will not be considered further here.

Second, recall that the multivariate decomposition (2.1) and the modified variance-covariance matrix estimator V_p^* were based on the assumption that the matrix R_p is constant across $p = 1,...,P$. To explore the consistency of this assumption with our data, Figures 3 and 4 present time plots of the sample correlation values $\hat{R}_{ijp}$, say, computed directly from the initial sample variance-covariance matrix $\hat{V}_{eep}$. Figure 3 displays results for R_{12}, the correlation between the interview sources 1 and 2. The symbol E represents the point estimate of this correlation. The symbols L and U represent lower and upper 95% pointwise confidence bounds for this correlation, based on a standard Fisher Z transformation (Snedecor and Cochran, 1967, p. 185) approach, under the assumption that $40 \times \hat{V}_{eep}$ is approximately distributed as a Wishart random matrix with 40 degrees of freedom. Note that these confidence bounds are relatively wide, reflecting the relatively large amount of sampling variability encountered in the standard estimators $\hat{V}_{eep}$. Also, Figure 3 does not display any pronounced patterns of increase or decrease of R_{12} over time; and with the exception of 1988, 1993 and 1994, the confidence intervals for R_{12} contain the value zero. Figure 4 displays the corresponding results for R_{34}, the correlation between the interview source 3 and the diary source 4. Note especially that for all eleven years, the 95% confidence interval for R_{34} contains zero.

Third, we considered the estimated misspecification effect matrix $M_p = (V_p^*)^{-1/2} \hat{V}_{eep} (V_p^*)^{-1/2}$

where $(V_p^*)^{-1/2}$ equals the inverse of the symmetric square root of V_p^*. For some general background on misspecification effects, see, e.g., Skinner (1989) and references cited therein. For the current discussion, it suffices to note that under correctly specified variance function models, the approximation $R_p = R$, and additional regularity conditions, the matrix M_p converges to the $k \times k$ dimensional identity matrix I_k, and each of its eigenvalues converge to one. Consequently, we can obtain a partial indication of the adequacy of our estimator V_p^* through examination of the eigenvalues of M_p. Figure 5 displays a time plot of the eigenvalues of M_p for the years 1987 through 1997. The plotting symbols 1 through 5, respectively, correspond to the largest through the smallest eigenvalues for a given year p; and the symbol m corresponds to the arithmetic average of the five eigenvalues. Figure 6 presents a similar plot for the case in which V_p^* is computed under the additional constraint $R^* = I_k$; cf. the relatively weak evidence of nonzero correlation displayed in Figures 3 and 4. Note especially that in Figure 6, the eigenvalues tend to be distributed more tightly around one, compared to the eigenvalues in Figure 5. To some degree, this may reflect the greater stability of V_p^* induced by the constraint $R^* = I_k$. Finally, note that deviations of the observed eigenvalues from the value one reflect the combined effects of the sampling variability of $\hat{V}_{eep}$ and V_p^*, and the lack of fit in our approximations for D_p and R_p. To identify deviations that are not attributable to sampling variability alone (and thus indicate problems with lack of fit), it is useful to compare the observed eigenvalues with quantiles of the appropriate reference distributions. In general, these reference distributions may be obtained through simulation work, as outlined in Lee and Eltinge (2001).

4. References

Eltinge, J.L., Sukasih, A. and Weber, W. (2000). Feasibility of constructing combined estimators using consumer expenditure interview and diary data. Paper presented to the Bureau of Labor Statistics Conference on Issues in measuring Price Change and Consumption, June 8, 2000.

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

Lee, S.R. and Eltinge, J.L. (2001). Exploratory analysis of estimated design effect matrices computed from complex survey data. Submitted for publication.

Skinner, C.J. (1989). Introduction to Part A. In C.J. Skinner, D. Holt and T.M.F. Smith (eds.) *Analysis of Complex Surveys.* New York: Wiley.

Snedecor, G.W. and Cochran, W.G. (1967). *Statistical Methods, Sixth Edition.* Ames, Iowa: Iowa State University Press.

U.S. Bureau of Labor Statistics (1997). Chapter 16: Consumer expenditures and income. *BLS Handbook of Methods.* U.S. Department of Labor, Bureau of Labor Statistics Bulletin 2490, April, 1997. Washington, DC: U.S. Government Printing Office.

Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association* **82**, 499-508.

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