“Challenging Research Issues in Statistics and Survey Methodology at the BLS”

Topic Statement: Stability of Linearization-Based Variance Estimators Computed from Potentially Unstable Estimators of First Derivatives

Key words: Asymptotics; Balanced repeated replication; Degrees of freedom; Inference; Nonlinear function of means; Replication-based variance estimation; t distribution approximation; Wishart distribution approximation.

Contact for further discussion:
John L. Eltinge
Office of Survey Methods Research, PSB 1950
Bureau of Labor Statistics
2 Massachusetts Avenue NE
Washington, DC 20212
Telephone: (202) 691-7404
Fax: (202) 691-7426
E-mail: Eltinge.John@bls.gov

Background, Definitions and Notation:

In the analysis of complex survey data, we often need to estimate the variance of the approximate distribution of a random vector $f(\hat{\bar{Y}})$, where $\hat{\bar{Y}}$ is an estimator of a k-dimensional population mean $\bar{Y}$ computed from complex survey data involving n sample elements, and $f(y)$ is a continuously differentiable m-dimensional real function of the k-dimensional real argument y.

In the complex-survey literature, regularity conditions on the sample design and population lead to results on the consistency of $\hat{\bar{Y}}$ for $\bar{Y}$ and the convergence in law of $n^{1/2}(\hat{\bar{Y}}-\bar{Y})$ to a normal distribution with mean 0 and $k \times k$-dimensional variance-covariance matrix V_{lim}. Additional regularity conditions then lead to development of a consistent estimator $\hat{V}_{\text{lim}}$ of V_{lim}, and to results on the limiting multivariate standard normal distribution of $n^{1/2}\hat{V}_{\text{lim}}^{-1/2}(\hat{\bar{Y}}-\bar{Y})$ where $\hat{V}_{\text{lim}}^{-1/2}$ is the inverse of the symmetric square root of $\hat{V}_{\text{lim}}$.

Furthermore, under additional regularity conditions (e.g., Korn and Graubard, 1990), $d\hat{V}_{\text{lim}}$ is distributed approximately as a Wishart(V_{lim}, d) random matrix, where d is a known “degrees of freedom” term computed from the number of primary sample units and the number of strata. Korn and Graubard (1990) also consider extensions of this Wishart approximation for cases in which $(\hat{\bar{Y}}-\bar{Y})$ is replaced by a corresponding difference between an estimator and true value for a vector of regression coefficients.

Now consider variance estimator for $f(\hat{\bar{Y}})$. In formal terms, we wish to estimate $V_{\lim f}$, defined to be the $m \times m$-dimensional variance-covariance matrix of the limiting distribution of $n^{1/2}\{f(\hat{\bar{Y}}) - f(\bar{Y})\}$. Under a standard linearization approach, we define the $m \times k$ matrices

$F(y) = \partial f(y)/\partial y$, $\bar{F} = \{\partial f(y)/\partial y\}|_{y=\bar{Y}}$ and $\hat{F} = \{\partial f(y)/\partial y\}|_{y=\hat{\bar{Y}}}$. In practical applications, these matrices often are functions of additional variables, the presence of which is suppressed in the current notation.

Under regularity conditions, one can show that $V_{\lim f} = \bar{F} V_{\lim} \bar{F}'$, and one commonly defines the corresponding random matrix $\hat{V}_{\lim f} = \hat{F}\hat{V}_{\lim}\hat{F}'$ and uses it as an estimator of $V_{\lim f}$. Furthermore, one generally attributes to $\hat{V}_{\lim f}$ the same degrees of freedom term, d, that was previously attributed to $\hat{V}_{\lim}$, and thus treat $d\hat{V}_{\lim f}$ as if it were distributed approximately as a Wishart$(V_{\lim f}, d)$ random matrix. In addition, under conditions on the function $f(\cdot)$ and its derivatives, and additional regularity conditions, one can establish that $n^{1/2}\hat{V}_{\lim f}^{-1/2}\{f(\hat{\bar{Y}}) - f(\bar{Y})\}$ converges in law to a multivariate standard normal distribution. For some general background on such asymptotic approaches, see, e.g., Krewski and Rao (1981), Binder (1983), Francisco and Fuller (1991), Binder and Patak (1994) and Shao (1996).

Issue: In samples of moderate size, the estimated matrix of first derivatives, $\hat{F}$, may itself demonstrate nontrivial random variability.

Questions on Properties of Standard Variance Estimators for Nonlinear Functions of Estimated Means, and Modifications of Said Variance Estimators:

1. Assume that $\hat{F}$ has a nontrivial amount of random variability, relative to the random variability of $\hat{V}_{\lim}$. In formal terms, assume that the differences $\hat{F} - F$ and $\hat{V}_{\lim} - V_{\lim}$ are of the same order in probability. In a standard asymptotic setting for complex surveys, this generally would occur when the degrees-of-freedom term d is increasing at the same rate as n. Thus, we are excluding from consideration the case in which d is fixed, as would occur with a fixed number of strata and primary sample units, and increasing numbers of sample elements within each primary sample unit. In addition, although we are using an asymptotic framework, we are implicitly excluding from consideration the cases in which d and n are so large that the errors differences $\hat{F} - F$ and $\hat{V}_{\lim} - V_{\lim}$ have a trivial effect on inference for $f(\bar{Y})$ based on $n^{1/2}\hat{V}_{\lim f}^{-1/2}\{f(\hat{\bar{Y}}) - f(\bar{Y})\}$.

Under what additional conditions can one establish that for some positive real number d_F , $d_F \hat{V}_{\text{lim}\, f}$ is distributed approximately as a Wishart$(V_{\text{lim}\, f}, d_F)$ random matrix?

2. Under the conditions of question (1), what is an appropriate estimator of d_F ?

3. Under the conditions of question (1), one might wish to produce an estimator of $\hat{F}$ that is more stable. This occurs, for example, in quantile estimation when one uses smoothed density estimators in the computation of related variance estimators in some cases.

   For general classes of smooth functions $f(\cdot)$, what are appropriate procedures for computation of stabilized versions of $\hat{F}$ and use in the resulting variance estimators $\hat{V}_{\text{lim}\, f} = \hat{F}\hat{V}_{\text{lim}}\hat{F}'$? To what extent, if at all, would such a "smoothing" approach be related to methods of variance estimation derived from estimating equations? Also, to what extent can results on variance-estimator stability in the literature on semi-parametric estimation (e.g., Ritov, 1991; and Bickel et al., 1993) be extended to work with complex survey data?

4. To what extent do the issues (and prospective solutions) in (1)-(3) extend to replication-based variance estimators (in which the replication procedure produces, in an informal sense, a nonparametric difference-based estimator of the derivative matrix F)?

5. Other authors have identified cases in which customary standard normal or t distribution approximations are problematic for quantities like $n^{1/2}\hat{V}_{\text{lim}}^{-1/2}(\hat{\bar{Y}} - \bar{Y})$ or related univariate t statistics, largely related to the correlation of the "numerator" term $\hat{\bar{Y}}$ with the "denominator" term $\hat{V}_{\text{lim}}$ for some cases involving, e.g., proportions of rare events . See, e.g., Casady, Dorfman and Wang (1998).

   To what extent are these distributional problems exacerbated for $n^{1/2}\hat{V}_{\text{lim}\, f}^{-1/2}\{f(\hat{\bar{Y}}) - f(\bar{Y})\}$ due to the possible correlation of the derivative matrix $\hat{F}$ with the "numerator" term $f(\hat{\bar{Y}})$?

6. Finally, some authors consider inference for $f(\bar{Y})$ based on bootstrap procedures, e.g., confidence intervals based on bootstrap-*t* methods. To what extent do the issues and solutions from (1)-(5) carry over to bootstrap-based inference methods?

Acknowledgements

The author thanks Moon Jung Cho, Alan Dorfman, Partha Lahiri and Michael Sverchkov for helpful comments on earlier versions of this topic statement.

References:

Bickel, P. J., C.A.J. Klaassen, Y. Ritov, Y. and J.A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore: John Hopkins University Press.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279-292.

Binder, D.A. and Z. Patak (1994). Use of estimating functions for estimation from complex surveys, *Journal of the American Statistical Association* **89** 1035-1043.

Casady, R.J., A.H. Dorfman and S. Wang (1998). Confidence intervals for domain parameters when the domain sample size is random, *Survey Methodology* **24**, 57-67.

Francisco, C.A. and W.A. Fuller (1991). Quantile estimation with a complex survey design, *The Annals of Statistics* **19**, 454-469.

Korn, E.L. and Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics, *The American Statistician* **44**, 270-276

Krewski, D. and J.N.K. Rao (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods, *The Annals of Statistics* **19** , 1010-1019.

Ritov, Y. (1991). Estimating functions in semi-parametric models. Pp. 319-336 in *Estimating Functions* (V.P. Godambe, ed.), Oxford: Clarendon Press.

Shao, J. (1996). Resampling methods in sample surveys (with discussion), *Statistics* **27**, 203-254.