BIOMETRICS 56, 645–646
June 2000

## READER REACTION

# A Note on Robust Variance Estimation for Cluster-Correlated Data

**Rick L. Williams**

Research Triangle Institute, P.O. Box 12194, Research Triangle Park, North Carolina 27709-2194, U.S.A.
*email:* willy@rti.org

SUMMARY. There is a simple robust variance estimator for cluster-correlated data. While this estimator is well known, it is poorly documented, and its wide range of applicability is often not understood. The estimator is widely used in sample survey research, but the results in the sample survey literature are not easily applied because of complications due to unequal probability sampling. This brief note presents a general proof that the estimator is unbiased for cluster-correlated data regardless of the setting. The result is not new, but a simple and general reference is not readily available. The use of the method will benefit from a general explanation of its wide applicability.

KEY WORDS: Between-cluster variance estimator.

There are many situations where data are observed in clusters such that observations within a cluster are correlated while observations between clusters are uncorrelated, so-called cluster-correlated data. For example, the typical teratology screening experiment involves administration of a compound to pregnant dams of a rodent species, followed by evaluation of the fetuses in a litter for various types of malformations. In this situation, the fetuses within a particular litter are correlated while any two fetuses from different litters are independent. Similarly, dental studies often collect data on each tooth surface for each of several teeth from a set of patients. Again, observations from the same patient are correlated while any two observations from different patients are independent. Another example is repeated measurements or recurrent events observed on the same person. As before, observations at different time points from the same person are correlated while any two observations from different patients are independent. As a final illustration, sample surveys often use multistage sample designs. For example, a sample of hospital patients might start out with a sample of geographic areas (such as counties), followed by a sample of hospitals within the selected geographics areas, ending with a sample of hospital discharges abstracted from the selected hospitals. Here we have a three-stage design consisting of geographic areas, hospitals, and hospital discharges. If the geographic areas were selected with replacement, then selected discharges from two geographic areas would be uncorrelated while two discharges from the same geographic area would be correlated.

A major statistical problem with cluster-correlated data arises from intracluster correlation, or the potential for clustermates to respond similarly. This phenomenon is often referred to as overdispersion or extra variation in an estimated statistic beyond what would be expected under independence. Analyses that assume independence of the observations will generally underestimate the true variance and lead to test statistics with inflated Type I errors.

The following presents an unbiased variance estimator for a linear statistic from cluster-correlated data. The approach uses the well-known, but not well-documented, robust between-cluster variance estimator for cluster-correlated data. This approach is used extensively in sample survey research where clustered data are commonly encountered. See, e.g., Hansen, Hurwitz, and Madow (1953, Section 6.7) or Särndal, Swensson, and Wretman (1992, Section 4.5). These two references from the sample survey literature justify the variance estimator under the assumptions that the primary clusters are sampled with replacement, while any sampling plan that allows unbiased estimation of the primary cluster totals can be used within a cluster. In the sample survey situation, with-replacement sampling of the primary clusters implies that observations between primary clusters are uncorrelated. In the general situation, the critical assumption is that the observations between clusters are uncorrelated.

The following notation describes the general cluster-correlated data situation. Let $z_{jk}$ be the $k$th observation ($k = 1, 2, \ldots, n_j$) from the $j$th cluster ($j = 1, 2, \ldots, m$). Assume, without loss of generality, that $\mathrm{E}[z_{jk}] = 0$. Further assume that $\mathrm{cov}(z_{jk}, z_{jk'}) = \sigma_{jkk'}$ and that $\mathrm{cov}(z_{jk}, z_{j'k'}) = 0$ when $j \neq j'$. These assumptions are very general and allow the variance to be heteroscedastic, both between and within clusters, and allow for an arbitrary dependence structure among observations within a cluster. For example, there could be three or more levels of nesting, as in the dental example above (tooth surfaces nested within teeth nested within patients) or an autoregressive process for repeated measurements over time on the same person.

First, consider the simple linear statistic $z = \Sigma_j \Sigma_k z_{jk}$ and note that

$$\mathrm{var}[z] = \sum_j \mathrm{var}\left[\sum_k z_{jk}\right] = \sum_j \sum_k \sum_{k'} \sigma_{jkk'}.$$

Letting $z_j = \Sigma_k z_{jk}$ and $\bar{z} = \Sigma_j z_j/m$, the between-cluster variance estimator is then given by

$$S^2 = \frac{m}{m-1}\sum_j (z_j - \bar{z})^2 = \frac{m}{m-1}\left[\sum_j z_j^2 - m\bar{z}^2\right].$$

We want to show that $\mathrm{E}[S^2] = \Sigma_j \Sigma_k \Sigma_{k'} \sigma_{jkk'} = \mathrm{var}[z]$. First, note that

$$\mathrm{E}\left[z_j^2\right] = \sum_k \sum_{k'} \mathrm{E}[z_{jk} z_{jk'}] = \sum_k \sum_{k'} \sigma_{jkk'}.$$

Also,

$$\mathrm{E}\left[\bar{z}^2\right] = \frac{1}{m^2}\sum_j \sum_{j'} \mathrm{E}[z_j z_{j'}] = \frac{1}{m^2}\sum_j \mathrm{E}\left[z_j^2\right]$$
$$= \frac{1}{m^2}\sum_j \sum_k \sum_{k'} \sigma_{jkk'}$$

because observations from different clusters are uncorrelated. Thus,

$$\mathrm{E}\left[S^2\right] = \frac{m}{m-1}\left[\sum_j \mathrm{E}\left[z_j^2\right] - m\mathrm{E}\left[\bar{z}^2\right]\right]$$
$$= \sum_j \sum_k \sum_{k'} \sigma_{jkk'} = \mathrm{var}[z].$$

Hence, we have the desired result that the between-cluster variance estimator, $S^2$, is an unbiased estimator of the variance of a linear statistic. Notice that we only need to know to which cluster each observation belongs without regard to the dependence structure of observations within a cluster.

The above is not a new result, but it is poorly documented. It has been available in the sample survey literature since at least 1953 (Hansen et al., 1953, Section 6.7). However, we are not aware of a general proof that the between-cluster variance estimator is unbiased for cluster-correlated data. The proofs in the sample survey literature are not easily applied because of the complications due to unequal probability sampling. The wide applicability of the results is often not well recognized because of the lack of a clear reference.

On a final note, the between-cluster variance estimator can be combined with a Taylor series linearization approach (Woodruff, 1971; Binder, 1983) to yield, as the number of clusters grows large, consistent variance estimates of nonlinear statistics. This approach replaces the original data with a linear approximation which can then be used as shown above. For example, Taylor series linearization with the between-cluster variance estimator was used by Rao and Colin (1991) for the proportion of malformed fetuses for teratology studies, by Fuller (1975) for linear regression coefficients in complex sample surveys, by Bieler and Williams (1995) for logistic regression in teratology studies, and by Williams (1995) for Kaplan–Meier survival functions. The Taylor series linearization approach with the between-cluster variance estimator is closely related to the generalized estimating equation (GEE) approach of Liang and Zeger (1986) and, in some situations, the two approaches are the same when assuming working independence. The Taylor series linearization approach is much older, with its roots in sample survey research reaching back to the early 1950s. The GEE approach attempts to improve estimation by including assumptions about the within-cluster correlation structure in the estimating equations.

## Résumé

Il existe un estimateur simple et robuste de la variance pour des données corrélées par groupe. Alors que cet estimateur est bien connu, la documentation le concernant est limitée et son large champs d'application est souvent mal compris. Il est largement utilisé dans la recherche d'enquête par échantillon, mais dans la littérature sur les enquêtes par échantillon les résultats ne sont pas facilement appliqués à cause des complications dues aux inégales probabilités d'échantillonnage. Cette courte note présente la preuve générale que l'estimateur est non biaisé pour des données corrélées par groupe quelle que soit la composition. Bien que le résultat ne soit pas nouveau, aucune référence simple et générale n'est facilement disponible. L'utilisation de la méthode pourra bénéficier d'une explication générale de son large domaine d'application.

## References

Bieler, G. S. and Williams, R. L. (1995). Cluster sampling techniques in quantal response teratology and developmental toxicity studies. *Biometrics* **51,** 764–776.

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

Fuller, W. A. (1975). Regression analysis for sample surveys. *Sankhya C* **37,** 117–132.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sample Survey Methods and Theory,* Volume I, *Methods and Applications.* New York: Wiley.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Rao, J. and Colin, D. (1991). Fitting dose–response models and hypothesis testing in teratological studies. In *Statistics in Toxicology,* D. Krewski and C. Franklin (eds). New York: Gordon and Breach.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Williams, R. L. (1995). Product-limit survival functions with correlated survival times. *Lifetime Data Analysis* **1,** 171–186.

Woodruff, R. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* **66,** 411–414.

*Received June* 1999. *Revised October* 1999.
*Accepted November* 1999.