LEAST SQUARES AND RELATED ANALYSES FOR COMPLEX SURVEY DESIGNS

Wayne A. Fuller[1]

1. INTRODUCTION AND MODEL

Assume that a sample of clusters of elemental units is selected from a finite population divided into L strata. The total sample of n clusters (primary sampling units) is given by

$$n = \sum_{h=1}^{L} n_h . \tag{1}$$

where $n_h \geq 2$ is the number of clusters selected in the h-th stratum. A column vector of characteristics

$$\underset{\sim}{Y}_{hij} = (Y_{hij1}, Y_{hij2}, \cdots, Y_{hijp})' \tag{2}$$

is observed for the j-th elemental unit in the i-th cluster of the h-th stratum. The vector $\underset{\sim}{Y}_{hij}$ is quite general. For example, some elements of the vector can be the powers of products of other entries. Also, one element can be, and often will be, identically equal to one. The cluster totals for the vector are defined by

$$\underset{\sim}{Y}_{hi.} = \sum_{j=1}^{m_{hi}} \underset{\sim}{Y}_{hij} . \tag{3}$$

where m_{hi} is the number of elements in the hi-th cluster.

We shall be interested in the behavior of locally continuous functions of a linear function of the vector of cluster means

[1] Wayne A. Fuller, Department of Statistics, Iowa State University.

$$\hat{\underset{\sim}{\theta}} = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{i=1}^{n_h} \underset{\sim}{Y}_{hi}., \tag{4}$$

where W_h are fixed weights. Often the weights are

$$W_h = N_h N^{-1}, \tag{5}$$

where N_h is the number of clusters in the h-th stratum and N is the total number of clusters in the population. For the weights (5) the linear function in (4) is the usual unbiased estimator of the finite population mean per cluster. Another set of weights that often is of interest is the set of unit weights

$$W_h = n^{-1} n_h. \tag{6}$$

Our model permits us to consider functions of the mean per element. The usual estimator of the mean per element for a particular Y-variable is the ratio of the mean per cluster for the Y-variable to the mean per cluster of the number of elements. The mean number of elements per cluster is the cluster mean of a Y-variable that is identically one.

Our discussion can be easily expanded to include various forms of subsampling within clusters. Because such expansions add little to the generality of the discussion and add considerable notational complexity, we restrict our attention to single stage sampling within strata.

Our discussion rests heavily on the following central limit theorem for samples from a finite population.

Theorem 1. Let $\{\xi_r : r = 1, 2, \ldots\}$ be a sequence of stratified finite populations. Let the population in the h-th stratum of the r-th population be a random sample of size $N_{rh} \geq N_{r-1,h}$ selected from a p dimensional infinite population with absolute $2 + \delta$, where $\delta > 0$, moments bounded by $M_\delta < \infty$. Let the covariance matrix for the rh-th infinite population be $\underset{\sim}{\Sigma}_{rh}$. Let $L_r \geq L_{r-1}$ be the number of strata in the finite population and let a simple random

sample of n_{rh} ($n_{rh} \geq 2$ and $n_{rh} \geq n_{r-1,h}$) units be selected in the h-th stratum. Let $f_{rh} = N_{rh}^{-1} n_{rh}$ be a triangular array such that

$$0 \leq f_{rh} < M_{fu} < 1,$$

where M_{fu} is a fixed number. Let $\underset{\sim}{Y}_{rhi.}$ be the total for the i-th cluster selected in the h-th stratum for the r-th population and let

$$\hat{\underset{\sim}{\theta}}_r = \sum_{h=1}^{L_r} W_{rh} n_{rh}^{-1} \sum_{i=1}^{n_{rh}} \underset{\sim}{Y}_{rhi.},$$

$$\underset{\sim}{\theta}_{rf} = \sum_{h=1}^{L_r} W_{rh} N_{rh}^{-1} \sum_{i=1}^{N_{rh}} \underset{\sim}{Y}_{rhi.},$$

$$\underset{\sim}{\theta}_r = \sum_{h=1}^{L_r} W_{rh} \underset{\sim}{\mu}_{.h..},$$

where $\underset{\sim}{\theta}_{rf}$ is the finite population parameter and $\underset{\sim}{\mu}_{.h..}$ is the mean of the infinite population used to generate the h-th stratum of the finite population. Assume

$$0 < M_{SL} < \left| n_r \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1} \underset{\sim}{\Sigma}_{rh} \right| < M_{SU} < \infty,$$

where the M's are fixed numbers and assume that

$$n_r = \sum_{h=1}^{L_r} n_{rh} \longrightarrow \infty,$$

$$\sup_h \left[\sum_{t=1}^{L_r} W_{rt}^2 n_{rt}^{-1} \right]^{-1} W_{rh}^2 n_{rh}^{-2} \longrightarrow 0,$$

as $r \to \infty$, where W_{rh} is a triangular array of weights. Then

$$[\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}\}]^{-\frac{1}{2}}(\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}) \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}),$$

$$[\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\}]^{-\frac{1}{2}}(\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r) \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}),$$

where

$$\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}\} = \sum_{h=1}^{L_r} W_{rh}^2(1 - f_{rh})n_{rh}^{-1}\,\hat{\underset{\sim}{\Sigma}}_{rh},$$

$$\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\} = \sum_{h=1}^{L_r} W_{rh}^2 n_{rh}^{-1}\,\hat{\underset{\sim}{\Sigma}}_{rh},$$

$$\hat{\underset{\sim}{\Sigma}}_{rh} = (n_{rh} - 1)^{-1} \sum_{i=1}^{n_{rh}} (\underset{\sim}{Y}_{rhi.} - \bar{\underset{\sim}{Y}}_{rh..})(\underset{\sim}{Y}_{rhi.} - \bar{\underset{\sim}{Y}}_{rh..})'.$$

$$\bar{\underset{\sim}{Y}}_{rhi.} = n_{rh}^{-1} \sum_{i=1}^{n_{rh}} \underset{\sim}{Y}_{rhi.}.$$

The proof of this theorem follows from Theorems 1 and 2 of Fuller (1975) and can be extended to multistage samples. Also see Krewski and Rao (1981) and Isaki and Fuller (1982).

Most of our applications are to continuous functions of $\hat{\underset{\sim}{\theta}}$.

Corollary 1. Let the assumptions of Theorem 1 hold. Let $\underset{\sim}{g}(\underset{\sim}{\theta})$ be a vector valued function of $\underset{\sim}{\theta}$, where $\underset{\sim}{g}(\underset{\sim}{\theta})$ is continuous with continuous first derivatives for $\underset{\sim}{\theta}$ in the sphere $|\underset{\sim}{\theta} - \underset{\sim}{\theta}_r| \leq \delta$ for all r, where $\delta > 0$ is fixed. Let $\underset{\sim}{G}(\underset{\sim}{\theta})$ be the nonsingular matrix of first derivatives of $\underset{\sim}{g}(\underset{\sim}{\theta})$, where the i,j-th element of $\underset{\sim}{G}(\underset{\sim}{\theta})$ is

$$\frac{\partial g_i(\underset{\sim}{\theta})}{\partial \theta_j},$$

$g_i(\underset{\sim}{\theta})$ is the i-th element of $\underset{\sim}{g}(\underset{\sim}{\theta})$ and θ_j is the j-th element of $\underset{\sim}{\theta}$. Then

$$[\underset{\sim}{G}(\hat{\underset{\sim}{\theta}}_r)\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_{rf}\}\underset{\sim}{G}'(\hat{\underset{\sim}{\theta}}_r)]^{-\frac{1}{2}}[\underset{\sim}{g}(\hat{\underset{\sim}{\theta}}_r) - \underset{\sim}{g}(\underset{\sim}{\theta}_{rf})] \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}),$$

$$[\underset{\sim}{G}(\hat{\underset{\sim}{\theta}}_r)\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\}\underset{\sim}{G}'(\hat{\underset{\sim}{\theta}}_r)]^{-\frac{1}{2}}[\underset{\sim}{g}(\hat{\underset{\sim}{\theta}}_r) - \underset{\sim}{g}(\underset{\sim}{\theta}_r)] \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}).$$

Corollary 1 is stated for the Taylor estimator of the variance of the approximate distribution of $\underset{\sim}{g}(\hat{\underset{\sim}{\theta}}_r) - \underset{\sim}{g}(\underset{\sim}{\theta}_r)$. Suitably defined replication estimators of the variance can also be used. Replication methods include balanced replication methods (see McCarthy (1969)), jackknife methods (See Miller (1974)) and bootstrap methods (see Efron (1979, 1981)). While these methods can be adapted to the sampling situation, the adaptation is not always immediate (see Rao and Wu (1983)).

One class of continuous functions of $\hat{\underset{\sim}{\theta}}$ that deserves special attention is that obtained by using $\hat{\underset{\sim}{\theta}}$ as the dependent variable in a generalized least squares fit.

Corollary 2. Let the assumptions of Theorem 1 hold. Let $\underset{\sim}{\theta}$ satisfy

$$\underset{\sim}{\theta} = \underset{\sim}{h}(\underset{\sim}{\alpha}).$$

where $\underset{\sim}{\alpha}$ is a k-dimensional vector ($k \leq p$), $\underset{\sim}{h}(\underset{\sim}{\alpha})$ is a continuous function of $\underset{\sim}{\alpha}$, with continuous first and second derivatives for all $\underset{\sim}{\alpha}$ in an open sphere containing the true $\underset{\sim}{\alpha}_r$ for all r. Let the parameter space for $\underset{\sim}{\alpha}$ be an open bounded subset of k-dimensional Euclidean space. Let $\hat{\underset{\sim}{\alpha}}_r$ be the vector that minimizes

$$[\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{h}(\underset{\sim}{\alpha}_r)]'\hat{\underset{\sim}{V}}^{-1}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\}[\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{h}(\underset{\sim}{\alpha}_r)].$$

Then

$$[\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\alpha}}_r\}]^{-\frac{1}{2}}(\hat{\underset{\sim}{\alpha}}_r - \underset{\sim}{\alpha}_r) \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{I}),$$

where

$$\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\alpha}}_r\} = [\underset{\sim}{H}(\hat{\underset{\sim}{\alpha}}_r)\hat{\underset{\sim}{V}}^{-1}\{\hat{\underset{\sim}{\theta}}_r - \underset{\sim}{\theta}_r\}\underset{\sim}{H}'(\hat{\underset{\sim}{\alpha}}_r)]^{-1},$$

and $\underset{\sim}{H}(\hat{\underset{\sim}{\alpha}}_T)$ is the matrix of first derivatives of $\underset{\sim}{h}(\underset{\sim}{\alpha})$ with respect to $\underset{\sim}{\alpha}$ evaluated at $\hat{\underset{\sim}{\alpha}}$.

2. MEANS, RATIOS AND REGRESSIONS

An elementary application of Theorem 1 is the estimation of the mean per cluster and the setting of approximate confidence limits for the mean per cluster. Often the parameter of interest for the mean estimator is the finite population mean per cluster, in which case the finite population correction $(1 - f_h)$ would be included in the variance estimator.

A slightly more complex application is the estimation of the difference between the means per cluster for two domains. If we let

Y_{hij1} = observation on characteristic of interest if element hij is in domain 1
= 0 otherwise,

Y_{hij2} = observation on characteristic of interest if element hij is in domain 2
= 0 otherwise,

Y_{hij3} = 1 if element hij is in domain 1
= 0 otherwise,

Y_{hij4} = 1 if element hij is in domain 2
= 0 otherwise.

the estimated difference between the mean per element in the two domains is

$$\underset{\sim}{g}(\hat{\theta}) = g(\bar{\underset{\sim}{Y}}_{...}) = \bar{Y}^{-1}_{...3}\bar{Y}_{...1} - \bar{Y}^{-1}_{...4}\bar{Y}_{...2}. \tag{7}$$

Two methods of computing the Taylor estimator of variance are often used. The first method computes the estimator of Corollary 1 directly from the matrices $\underset{\sim}{G}(\hat{\underset{\sim}{\theta}}_T)$ and $\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_T - \underset{\sim}{\theta}_T\}$ or $\hat{\underset{\sim}{V}}\{\hat{\underset{\sim}{\theta}}_T - \underset{\sim}{\theta}_{Tf}\}$. An algebraically identical computational procedure is to define the observations

$$\underset{\sim}{Z}(\underset{\sim}{Y}_{hi.}, \hat{\underset{\sim}{\theta}}) = \hat{\underset{\sim}{Z}}_{hi} = \underset{\sim}{G}(\hat{\underset{\sim}{\theta}})(\underset{\sim}{Y}_{hi.} - \bar{\underset{\sim}{Y}}_{h..}) \tag{8}$$

and to compute the ordinary stratified estimator of the variance of the mean per cluster for $\hat{\underset{\sim}{Z}}_{hi}$.

$$\hat{\underset{\sim}{V}}\{\hat{\bar{\underset{\sim}{Z}}}_{..}\} = \hat{\underset{\sim}{V}}\{\underset{\sim}{g}(\bar{\underset{\sim}{Y}}_{...})\}$$

$$= \sum_{h=1}^{L} W_h^2(1 - f_h)n_h^{-1}(n_h - 1)^{-1} \sum_{j=1}^{n_h}(\hat{\underset{\sim}{Z}}_{hi} - \hat{\bar{\underset{\sim}{Z}}}_{h.})(\hat{\underset{\sim}{Z}}_{hi} - \hat{\bar{\underset{\sim}{Z}}}_{h.})', \tag{9}$$

where

$$\hat{\bar{\underset{\sim}{Z}}}_{..} = \sum_{h=1}^{L} W_h \hat{\bar{\underset{\sim}{Z}}}_{h.},$$

$$\hat{\bar{\underset{\sim}{Z}}}_{h.} = n_h^{-1} \sum_{i=1}^{n_h} \hat{\underset{\sim}{Z}}_{hi}.$$

For example, the computational form (9) is used in Super Carp. See Hidiroglou et al. (1980, p. 32).

The analyst may be interested in inferences for the particular finite population sampled or for the superpopulation when working with quantities such as differences of means.

One of the more frequent analytic uses of survey data is the computation of regression equations. In fact, the difference between domain means can be expressed as a regression coefficient. Although the vector of regression coefficients is of the form $\underset{\sim}{g}(\hat{\underset{\sim}{\theta}})$ described in the previous section. it may be advantageous to partition the $\underset{\sim}{Y}$-vector of Section 1 into several parts and to give the regression coefficients explicit expressions. The regression equation can be written as

$$Y_{hij} = \underset{\sim}{X}'_{hij}\underset{\sim}{\beta} + e_{hij}, \tag{10}$$

where Y_{hij} is the dependent variable, the vector $\underset{\sim}{X}_{hij}$ is a k-dimensional

vector of explanatory variables. The weighted least squares estimator of $\underset{\sim}{\beta}$ is

$$\hat{\underset{\sim}{\beta}}_W = \left[\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \underset{\sim}{X}'_{hij}\right]^{-1} \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} Y_{hij}. \tag{11}$$

The weights W_{hij} are permitted to be a function of hij, but we will assume that the weights are fixed in the sense that they depend only on the elemental identification. This precludes from consideration (except as an approximation) the use of weights that are a function of other elements entering the sample.

Under mild assumptions on the moments of the superpopulation generating the finite population, Theorem 1 is applicable to the estimator defined in (11). If the selection probabilities are denoted by π_{hij}, then the estimator $\hat{\underset{\sim}{\beta}}_W$ is a consistent estimator of the finite population vector

$$\underset{\sim}{B}_f = \left[\sum_{h=1}^{L}\sum_{i=1}^{N_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \pi_{hij} \underset{\sim}{X}'_{hij}\right]^{-1} \sum_{h=1}^{L}\sum_{i=1}^{N_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \pi_{hij} Y_{hij}. \tag{12}$$

It follows from (12) that the estimator (11) is a consistent estimator of the finite population regression coefficient when W_{hij} is proportional to the inverse of the selection probabilities. The error in $\hat{\underset{\sim}{\beta}}_W$ as an estimator of $\underset{\sim}{B}_f$ is

$$\hat{\underset{\sim}{\beta}}_W - \underset{\sim}{B}_f = \left[\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \underset{\sim}{X}'_{hij}\right]^{-1} \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} v_{hij}, \tag{13}$$

where

$$v_{hij} = Y_{hij} - \underset{\sim}{X}'_{hij}\underset{\sim}{B}_f.$$

By Theorem 1 and Corollary 1 a consistent estimator of the variance of the approximate distribution of $\hat{\underset{\sim}{\beta}}_W - \underset{\sim}{\beta}$ is

$$\hat{V}\{\hat{\underset{\sim}{\beta}}_W - \underset{\sim}{\beta}\} = \hat{\underset{\sim}{A}}^{-1}\hat{\underset{\sim}{G}}\hat{\underset{\sim}{A}}^{-1}. \tag{14}$$

where

$$\hat{\underset{\sim}{A}} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} W_{hij} \underset{\sim}{X}'_{hij},$$

$$\hat{\underset{\sim}{G}} = (n - 1)(n - k)^{-1} \sum_{h=1}^{L} n_h (n_h - 1)^{-1} \sum_{i=1}^{n_h} \hat{\underset{\sim}{d}}_{hi.} \hat{\underset{\sim}{d}}'_{hi.},$$

$$\hat{\underset{\sim}{d}}_{hi.} = \sum_{j=1}^{m_{hi}} \hat{\underset{\sim}{d}}_{hij},$$

$$\hat{d}_{hij} = W_{hij} \underset{\sim}{X}_{hij} \hat{v}_{hij},$$

$$n = \sum_{h=1}^{L} \sum_{i=1}^{n_h} m_{hi},$$

$$\hat{v}_{hij} = Y_{hij} - \underset{\sim}{X}'_{hij} \hat{\underset{\sim}{\beta}}_W,$$

and $\underset{\sim}{\beta}$ is the superpopulation analog of $\underset{\sim}{B}_f$. This particular form of the estimator of variance was suggested by Fuller (1975) and is used in Super Carp.

One of the frequently asked questions faced by survey statisticians is: "In computing the regression equation, should I use the sampling weights?" As with most such questions, the answer is "It depends." The fact that the question is asked generally means that the questionner has in mind inference for a population beyond the finite population sampled. This does not mean that the particular superpopulation is completely defined or definable. It does suggest that the questionner is postulating that the finite population is generated by a superpopulation in which some type of linear model holds. One quantification of the hypothesis that weights are not required is the superpopulation hypothesis

$$H_0: \underset{\sim}{\theta}_\pi = \underset{\sim}{\theta}_{(1)}. \tag{15}$$

where the $\underset{\sim}{\theta}$'s are superpopulation analogs of (12),

$$\underset{\sim}{\theta}_\pi = \left[\sum_{h=1}^{L} \sum_{i=1}^{N_h} E_\xi \sum_{j=1}^{m_{hi}} \{\underset{\sim}{X}_{hij} \pi_{hij} \underset{\sim}{X}'_{hij}\}\right]^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} E_\xi \left\{\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij} \pi_{hij} Y_{hij}\right\},$$

$$\underset{\sim}{\theta}_{(1)} = \left[\sum_{h=1}^{L}\sum_{i=1}^{N_h} E_\xi\left\{\sum_{j=1}^{m_{hi}} \underset{\sim}{X}_{hij}\underset{\sim}{X}'_{hij}\right\}\right]^{-1} \sum_{h=1}^{L}\sum_{i=1}^{N_h} E_\xi\left\{\sum_{j=1}^{m_h} \underset{\sim}{X}_{hij}Y_{hij}\right\}, \tag{16}$$

and E_ξ denotes expectation with respect to the superpopulation. This is a testable hypothesis. It seems that, at a minimum, a test of this hypothesis should be constructed if one performs an unweighted analysis of a sample with unequal selection probabilities.

If the null hypothesis also includes the hypothesis that the estimator with unit weights is the minimum variance estimator, then the test of the hypothesis is given by the statistic

$$F^k_{n-L-2k} = k^{-1}\,\hat{\underset{\sim}{\delta}}'_2\hat{\underset{\sim}{V}}^{-1}_{22}\hat{\underset{\sim}{\delta}}_2. \tag{17}$$

where

$$(\hat{\delta}'_1,\ \hat{\underset{\sim}{\delta}}'_2)' = \left[\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{Z}_{hij}\underset{\sim}{Z}'_{hij}\right]^{-1} \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \underset{\sim}{Z}_{hij}Y_{hij},$$

$$\underset{\sim}{Z}'_{hij} = (\underset{\sim}{X}'_{hij},\ \underset{\sim}{X}'_{hij}W_{hij}),$$

and

$$\hat{\underset{\sim}{V}} = \begin{pmatrix} \hat{\underset{\sim}{V}}_{11} & \hat{\underset{\sim}{V}}_{12} \\ \hat{\underset{\sim}{V}}_{21} & \hat{\underset{\sim}{V}}_{22} \end{pmatrix} \tag{18}$$

is defined by (14) with $\underset{\sim}{Z}_{hij}$ replacing $\underset{\sim}{X}_{hij}$. As the notation suggests, the statistic is approximately distributed as Snedecor's F with k and n - L - 2k degrees of freedom.

Example 1. Table 1 contains observations on 37 area segments collected by the Statistical Reporting Service, U.S. Department of Agriculture in northcentral Iowa in 1978. Two determinations on the hectares of soybeans are reported. The first is obtained by personal interview in the June Enumerative Survey. The second is obtained from a classification of Landsat data based upon a classifier developed by the Statistical Reporting Service. The original objective of the study was to use the Landsat data to construct a regression

estimator of the total acres. We use the data to illustrate the computation of regression statistics from survey data. The sample most nearly approximates a stratified sample with strata identified in the column headed "county". The inverse of the sampling rates is given in the weight column. The estimated regression equation for the regression of interview hectares on satellite hectares defined by estimator (11) is

$$\hat{Y} = \underset{(8.332)}{-11.845} + \underset{(0.0922)}{1.1602X},$$

where the numbers in parentheses are the standard errors obtained from the estimated covariance matrix calculated by equation (14).

Calculations were performed using Super Carp. If the equation and standard errors are calculated using unit weights in equations (11) and (14), respectively, we have

$$\hat{Y} = \underset{(9.282)}{-3.927} + \underset{(0.0963)}{1.0850X}.$$

If we calculate the F-test suggested in equation (17), we obtain

$$F_{23}^{2} = 2.81.$$

At first glance, this test is large enough to cause to suspicion about the equality of the two coefficients. Because this sample is very small and because of the structure of the weights, the test is nearly a test between two lines, the line for county one, and the average line for the remaining counties. In this small sample the deviations from the line in county one are small. Hence, the estimated standard errors of the coefficients for the two added variables are small. This phenomenon is discussed further in Section 3. If one uses the ordinary regression F-test that assumes homogeneous error variances and ignores the stratification, one obtains

$$F_{33}^{2} = 0.68.$$

While this statistic is not distributed as Snedecor's F, it does make one feel more comfortable with the assumption that the two weighting procedures are estimating the same equation.

Table 2 contains the standard errors of regression coefficients estimated under alternative assumptions. The estimated standard errors for the intercept behave much as one might anticipate. The stratified weighted sample procedure has the smallest estimated standard error followed by the stratified unit weight procedure and the ordinary least squares procedure. Do not forget these are estimated standard errors. The two stratified procedures are consistent under the stratified model. The weighted estimator has smaller variance because the observations for stratum 1, the stratum with the largest weight, lie closer to the estimated line than do the points in other strata. The ordinary least squares estimated standard error is not consistent under the stratified model. If the sample is treated as a cluster sample of counties, the estimated standard errors for the intercept are about 30 to 40 percent larger than the corresponding values for the stratified sample.

The estimated standard errors for the slope display a different behavior. The smallest estimated standard error is associated with the unit weight cluster estimation, and the largest estimated standard error is associated with ordinary least squares. Roughly speaking, the variation of slopes among clusters is small relative to the within cluster variation. Because the weights are inversely correlated with the observed variability, the weighted estimators have smaller estimated variances. This is a small sample, but it is sufficient to demonstrate that unit weights do not always produce smaller variances than sample weights and that stratification and clustering can have rather complex effects on the estimated variances of the regression coefficients.

3. WHAT IS A LARGE SAMPLE?

Our discussion has rested on the large sample properties of estimators and of estimators of variance. If the limiting normal distribution is being used to establish confidence intervals, the size of the sample required for a good approximation depends upon the nature of the original population. For

example, if the characteristic is a rare zero-one item (probability less than 0.05, say), a very large sample (more than 1,400 for a simple random sample (Cochran, 1977, p. 58)) will be required for the normal approximation. The binomial with small p is only one example of the very skewed populations often encountered in sampling practice. Measures of size such as gross sales of firms, number of employees of firms, number of animals per farm, and family income are examples of skewed populations for which large samples are required before the distribution of the mean approaches normality. On the other hand, the distribution of the mean for items such as family size may approximate the normal distribution for small (less than 100) sample sizes.

*The use of the Taylor expansion is semi-nonparametric in that the approxi*mation holds, in large samples, under very mild assumptions on the population. The large sample requirements are met if we have no isolated points in our sample space. The method may perform poorly in situations where the generating distribution and sample size are such that an observation or observations are isolated from the remaining cluster of points. We consider the problem of estimating the variance of the vector of regression coefficients used to test the effect of weighting on the coefficients in the soybean example. The original vector is

$$(1,\ X,\ XW,\ W),$$

and the hypothesis to be tested is the hypothesis that the coefficients for XW and W are zero. To illustrate the problems associated with variance estimation for the vector of coefficients for the soybean data set, we create a vector that is orthogonal in the unit weight metric. The matrix of observations on the transformed independent variables is composed of the residuals obtained in the regression of each variable, except the first, on the elements preceeding it in the original vector. Table 3 contains the transformed regression variables $(X - \bar{X},\ RWX,\ RW)$. Only a few digits have been retained to make it easier to read the table.

When we regress Y on $(1,\ X - \bar{X},\ RWX,\ RW)$ we obtain

$$\hat{Y} = \underset{(2.24)}{95.34} + \underset{(0.093)}{1.085}(X - \bar{X}) + \underset{(0.044)}{0.093 \times 10^{-2}}RWX - \underset{(0.023)}{0.015}RW,$$

where the estimated standard errors were computed for a stratified sample with unit weights using expression (14). If the regression and standard errors are computed by ordinary least squares, we obtain

$$\hat{Y} = \underset{(3.37)}{95.34} + \underset{(0.113)}{1.085}(X - \bar{X}) + \underset{(0.086)}{0.093 \times 10^{-2}}RWX - \underset{(0.034)}{0.015}RW.$$

The estimated standard error for the coefficient of RWX obtained by Taylor methods is about one half of that obtained by ordinary least squares methods. This can be explained by the data configuration.

The first observation on RWX is much larger in absolute value than any other observation. Of the total sum of squares for RWX, 67 percent is due to this observation. The Taylor approximation to the variance uses the sample variance of deviates called $\hat{d}_{hij}$ in (14) to estimate the variance of the statistic. The deviations from regression, denoted by $\hat{v}$, are given in the last column of Table 3. The $\hat{v}$ value for observation one is among the smaller values. The mean square for the residuals is 421. The product $(RWX)(\hat{v})$ for the first observation is -1113. This product is of the same order of magnitude as the product for observations 3, 33 and 36. Therefore, while the first observation is responsible for about 67 percent of the sum of squares of RWX, it is responsible for only about 15 percent of the sum of squares of $(RWX)(\hat{v})$. This is because $\hat{v}^2$ for the first observation is less than one tenth of the average of the squares of the other observations. Furthermore, the squared deviation for the first observation is biased downward because the method of least squares will cause the estimated plane to pass close to an observation that is separated from the other observations. Thus, if all of the observations have the same error variance, the Taylor method will produce an estimate of the variance of the coefficient for RWX that is biased downward.

Did the procedure underestimate the variance for this sample? We do not know. If we use the parametric procedure of ordinary least squares, we assign the pooled estimate of error variance to the separated observation. It is not possible to determine if this procedure is correct because our estimate of variance for the separated observation is a one degree of freedom estimator.

In this situation most people will feel more comfortable assuming that the variance for the separated point is the same as the variance of the other points rather than taking the small observed variance of the single point.

In the nonparametric world a single observation contains little information about the variability of the population that generated the observation. Furthermore, an observation separated from other observations is essentially a single observation. In the full parametric world the separated observation is in the fold because the separated observation is specified to have been created by the same generating mechanism that created the other observations. For data of the type displayed in Table 3, the answer obtained by parametric methods rests very heavily on assumptions about the error variance.

In the estimation of variances, one measure of the numerical size of the sample is the number of cluster degrees of freedom. Thus, for example, the estimated covariance matrix for a k-dimensional vector random variable is singular unless

$$\sum_{h=1}^{L} (n_h - 1) > k.$$

In setting approximate confidence intervals it seems reasonable to use Student's t distribution with degrees of freedom no greater than $\Sigma (n_h - 1)$. Because the variance of an estimated variance is a function of the fourth moments of the population, estimated variances are notoriously unreliable. The coefficient of variation for the squares is $2^{\frac{1}{2}}$ for the normal and considerably larger for many other common distributions.

If the error variances in the strata are unequal or if unequal weights are applied to the estimates of different strata, the variance of the variance estimator can be considerably different from that suggested by a simple calculation of error degrees of freedom. Table 4 has been constructed using the data configurations of Table 1 to illustrate these effects on the estimated variance. In the first column we assume that stratification is ineffective in that we assume each stratum variance is equal to the variance of the population. We assume the parent population to be normal so that we can give an explicit expression for the variance of the variance. In this situation stratification produces an estimated error variance for a mean with a variance

that is proportional to $(26.6)^{-1}$ while a simple random sample produces a variance of the estimated variance that is proportional to 36^{-1}. The effective degrees of freedom for the stratified sample is slightly less than 27 because of the unequal sample sizes within strata. If we use the sample weights of Table 1 and the usual stratified variance estimator, the variance of the estimated variance is proportional to $(4.6)^{-1}$. This large reduction is due to the large weight for the first stratum. If the variance in the first stratum is one half of the variance in other strata, then the effective degrees of freedom for the variance estimator is 12.4. In the last column we give the effective degrees of freedom for the simple random sample if the variance of the simple random sample is twice that of the stratified sample. This illustrates the fact that stratification can reduce both the variance of the estimated mean and the variance of the estimated variance of the mean.

While we are unable to specify the number of error degrees of freedom required for our approximations, it is clear that we shall be uncomfortable with a small number of degrees of freedom, particularly with unequal weights.

The theory of Corollary 1 uses a linear approximation to the nonlinear function of the sample means to approximate the behavior of the nonlinear function. If this approximation is to perform well, the curvature of the function must be small relative to the standard error of the sample means. For example, if the function is quadratic

$$g(\bar{Y}) = \alpha_1\bar{Y} + \alpha_2\bar{Y}^2,$$

the linear approximation is

$$g(\bar{Y}) \doteq \alpha_1\mu + \alpha_2\mu^2 + (\alpha_1 + 2\alpha_2\mu)(\bar{Y} - \mu).$$

The expected value of $g(\bar{Y})$ is

$$E\{g(\bar{Y})\} = \alpha_1\mu + \alpha_2[\mu^2 + V\{\bar{Y}\}].$$

For the linear approximation to perform well we must have small $V\{\bar{Y}\}$ and/or

small α_2.

In summary, to be comfortable with the use of large sample theory we require:

1. A reasonable number of observations in the sense that no observations are widely separated from the main clusters of observations. This is another way of saying that the Taylor deviates are such that the mean of the deviates is nearly normally distributed.
2. A reasonable number of effective error degrees of freedom for the estimator of variance.
3. The curvature of the nonlinear function of sample means to be small relative to the standard error of the sample means.

ACKNOWLEDGEMENTS

This research was partly supported by Research Agreement 58-319T-1-00054X with the Statistical Reporting Service of the U.S. Department of Agriculture. I thank Nancy Hasabelnaby for computations and Carol Francisco for comments.

REFERENCES

[1] Cochran, W.G. (1977). Sampling Techniques 3rd Ed. Wiley, New York.

[2] Efron, B. (1979). Bootstrap method: Another look at the jackknife. Ann. Statist. 7, pp. 1-26.

[3] Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika 68, pp. 589-599.

[4] Fuller, W.A. (1975). Regression analysis for sample survey. Sankhyā Series C 37, pp. 117-132.

[5] Fuller, W.A. and Hidiroglou, M.A. (1978). Regression estimation after correcting for attenuation. J. Amer. Statist. Assoc. 73, pp. 99-104.

[6] Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1980). Super Carp, Department of Statistics, Iowa State University, Ames, Iowa.

[7] Isaki, C. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. J. Amer. Statist. Assoc. 77, pp. 89-96.

[8] Kish, L. and Frankel, M.R. (1974). Inference from complex samples. J. Roy. Statist. Soc. B 36, pp. 1-22.

[9] Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. Ann. Statist. 9, pp. 1010-1019.

[10] McCarthy, P.J. (1965). Stratified sampling and distribution-free confidence intervals for the median. J. Amer. Statist. Assoc. 60, pp. 772-783.

[11] McCarthy, P.J. (1969). Pseudo-replication: Half-samples. Rev. Int. Statist. Inst. 37, pp. 239-264.

[12] Miller, R.G., Jr. (1974). The jackknife - a review. Biometrika 61, pp. 1-15.

[13] Rao, J.N.K. and Wu, C.F.J. (1984). Bootstrap with stratified samples. Technical Report No. 19 of the Laboratory for Research in Statistics and Probability. Carleton University, Ottawa, Canada.

Table 1: Soybean Area Determined by Two Methods

County	Segment	Weight	Soybean Hectares	
			Interview (Y)	Satellite (X)
1	1	502	8.09	24.75
1	2		106.03	98.10
1	3		103.60	112.50
2	1	212	6.47	43.20
2	2		63.82	80.10
3	1	188	43.50	61.65
3	2		71.43	92.70
3	3		42.49	74.25
4	1	190	105.26	98.10
4	2		76.49	99.45
4	3		174.34	152.10
5	1	134	95.67	57.60
5	2		76.57	66.15
5	3		93.48	91.80
6	1	189	37.84	34.65
6	2		131.12	97.65
6	3		124.44	116.10
7	1	172	144.15	136.35
7	2		103.60	99.45
7	3		88.59	99.90
7	4		115.58	123.30
8	1	114	99.15	85.50
8	2		124.56	121.50
8	3		110.88	77.40
8	4		109.14	102.60
8	5		143.66	133.65
9	1	193	91.05	75.15
9	2		132.33	85.95
9	3		143.14	112.05
9	4		104.13	81.90
9	5		118.57	80.55
10	1	93	102.59	117.90
10	2		29.46	39.15
10	3		69.28	72.00
10	4		99.15	99.45
10	5		143.66	155.25
10	6		94.49	85.50

Table 2: Estimated Standard Errors of Regression Coefficients Calculated by Alternative Procedures

Procedure	Estimated standard Error	
	$\hat{\beta}_0$	$\hat{\beta}_1$
Ordinary least squares	10.747	0.1116
Stratified; sample weights	8.332	0.0922
Cluster; sample weights	11.121	0.0823
Stratified; unit weights	9.282	0.0963
Cluster; unit weights	13.256	0.1071

Table 3: Data for Transformed Regression Problem

Stratum	Cluster	Weight	$X - \bar{X}$	$10^{-2}RWX$	RW	$\hat{v}$
1		502	-67	-195	167	6
1		502	7	25	336	6
1		502	21	68	369	-15
2		212	-48	1	1	-37
2		212	-11	4	24	-19
3		188	-30	10	-7	-20
3		188	1	5	7	-26
3		188	-17	8	-1	-35
4		190	7	4	12	3
4		190	8	4	13	-28
4		190	61	-3	38	14
5		134	-34	28	-53	34
5		134	-25	23	-51	6
5		134	0	5	-47	-3
6		189	-57	13	-20	3
6		189	6	4	11	29
6		189	25	2	20	3
7		172	45	-9	8	1
7		172	8	3	-6	-1
7		172	8	2	-6	-16
7		172	32	-5	3	-14
8		114	-6	10	-67	8
8		114	30	-22	-66	-2
8		114	-14	18	-68	28
8		114	11	-5	-67	1
8		114	42	-32	-65	5
9		193	-16	7	4	13
9		193	-6	6	9	43
9		193	21	3	22	26
9		193	-10	6	7	19
9		193	-11	6	6	35
10		114	26	-24	-90	-21
10		114	-52	63	-84	-16
10		114	-19	26	-87	-9
10		114	8	-4	-89	-6
10		114	64	65	-93	-16
10		114	-6	12	-88	3

Table 4: Efficiency of Estimated Variance under Alternative Assumptions

Procedure	Equivalent degrees of freedom	
	$V_{SRS} = V_{st}$	$V_{SRS} = 2V_{st}$
Simple random sampling	36	9
Strat. Sa., unit weights, equal var.	26.6	26.6
Strat. Sa., unequal weights, equal var.	4.8	4.8
Strat. Sa., unequal weights, $\sigma_1^2 = 0.5\sigma^2$	13.9	13.9