Journal of Official Statistics
Vol. 6. No. 3, 1990, pp. 295–310


# Estimation in Finite Population Under a Calibration Model

*Raj S. Chhikara and James J. McKeon*[1]

**Abstract:** In this paper, we address the estimation of a finite population mean under a calibration model. We describe a general class of regression estimators that yield the standard regression estimator and the alternative classical estimator as special cases. Formulas are derived for the asymptotic bias and variance of the general regression estimator. Simulations were carried out to compare three special estimators on the basis of bias, relative efficiency, and robustness to nonnormality of the model error. It is shown that the standard regression estimator is the most efficient and robust of all three. Also, estimation of the bias and variance of this regression estimator is examined, and several variance estimators are compared.

**Key words:** LANDSAT; regression type estimators; relative efficiency; variance estimation.

## 1. Introduction

The present study of regression estimation in finite population sampling is motivated by a practical problem of crop acreage estimation using satellite data as auxiliary information. The U.S. Department of Agriculture (USDA) conducts an annual survey, called the June Enumerative Survey (JES), to collect land use and crop acreage data which are used to make crop acreage estimates, among others, at the state and national levels. The survey methodology is based on probability sampling with data collected for a sample of area segments and crop acreage estimates made using a Horvitz–Thompson type estimator (U.S. Department of Agriculture 1975.) The sampling error at the national level is designed to be about 2% for the major crop estimates. At the state and crop reporting district levels, the sampling error is considerably larger. To reduce the crop acreage estimation error at these levels, the USDA in the mid-seventies proposed using satellite (LANDSAT) data as auxiliary information and to estimate crop acreages using a regression estimator (Sigman, Hanuschak, Craig, Cook, and Cardenas 1978).

LANDSAT is a near-earth orbiting satellite equipped with a multispectral scanner (MSS) that measures the energy emitted from the ground in various wavelength bands. The spectral responses depend upon the characteristic of ground vegetation and can be utilized to estimate acreages of

[1] Mathematical Sciences Program, University of Houston-Clear Lake, Houston, Texas 77058, U.S.A.

**Acknowledgement**: This research is supported under a cooperative research grant from the National Agricultural Statistical Service of the U.S. Department of Agriculture, Washington D.C.
The authors wish to thank the two referees and the Editor for their many helpful comments and suggestions made on the earlier version of the paper.

different crops in each area sample unit. As the satellite data provide a complete coverage of an area sampling frame, the satellite based estimates made across all units can be used to obtain more efficient crop acreage estimates for an area of interest.

The basic approach of USDA is to acquire LANDSAT data over a stratum, called an analysis district, that contains a number of JES sample units. The LANDSAT data are classified into different possible crop types using a discriminant analysis technique and crop acreage estimates are obtained for each sample unit as well as for the entire analysis district. The actual acreages for the JES sample units are regressed onto the corresponding satellite derived estimates, and the standard regression estimate of the mean crop acreage in the analysis district is determined for a crop. This approach to crop acreage estimation has already been used by the USDA for several major crops in several states (Holko and Sigman 1984).

Chhikara and Houston (1984) studied the USDA crop acreage estimation procedure and discussed the characteristics of the model that relates to the actual and the estimated crop acreage for the area sample units. They showed that the actual crop acreage and the corresponding satellite derived estimate are linearly related in an area segment, provided the classification procedure is fixed and known. Since a sample of satellite data is used to estimate the discriminant function, a certain amount of variability is introduced in the classification rule. This, in turn, introduces the variability in satellite derived crop acreages for a segment. Furthermore, the linear relation expected between the two acreages is likely to vary across segments. Considering these two kinds of variability, the authors modelled the relationship between the actual and estimated crop acreage. It was shown that in order to achieve a linear model in this application, the estimated crop acreage (auxiliary variable) should be regarded as a dependent variable and the actual crop acreage as an independent variable. If, however, the roles of these two variables are reversed, the resulting model will have the error term correlated with the regressor variable and hence, it will not be a linear model. Thus, the USDA crop acreage estimation procedure needs to be evaluated when the model characteristics are different from the usual regression model. The model that utilizes both the ground observed and the satellite derived estimates of crop acreages for the sample units is briefly stated in the following.

Suppose the area of interest (population) consists of $N$ area sample units. For the crop of interest let $y_i$ be the actual crop acreage and $x_i$ be its estimate obtained from LANDSAT data for the $i$th unit, $i = 1, 2, \ldots, N$. Then the estimated crop acreage ($x$) is related to the actual crop acreage ($y$) in an area sample unit via a linear model

$$x_i = \alpha + \beta y_i + e_i \tag{1.1}$$

where $\alpha$ and $\beta$ are functions of the classification errors associated with the discriminant function and $e_i$ denotes the difference between the estimate $x_i$ and its expected value, and given $y_i$

$$\mathscr{E}(e_i) = 0$$
$$\mathscr{V}(e_i) = \sigma^2 \tag{1.2}$$

and the $e_i$ are independently distributed random variables. Chhikara and Houston (1984) describe the salient characteristics of this model which they investigated in detail.

In this paper the expectation with respect to the model error distribution will be shown by script letters as in (1.2).

From a theoretical viewpoint, the USDA crop acreage estimation as described above can be treated as a general problem of estimation of the finite population mean $\bar{Y} = \Sigma y_i/N$, where unit $i$ is associated with two numbers $(x_i, y_i)$ with population mean $\bar{X} = \Sigma x_i/N$ known and the $y_i$ fixed but unknown, and the pairs $(x_i, y_i)$ are observed for a set of $n$ sample units, and the model in (1.1) – (1.2) holds. Here the values $x_1, x_2, \ldots, x_N$ are realized observations of independent random variables, say $X_1, X_2, \ldots, X_N$.

Clearly, this formulation represents the superpopulation case where $\alpha$ and $\beta$ are model parameters, and the $y_i$ values are fixed in repeated realizations of the finite population sampled from the superpopulation. This treatment of a finite population as an independent sample of size $N$ from an infinite superpopulation has been well argued in the literature. The estimation of the finite population mean or total has been studied extensively by considering both the theory of finite population sampling (which is purely an estimation approach) and the classical linear model theory (which is mainly a prediction approach). In the usual linear regression case where the regressor is an independent variable, the standard regression estimator of population mean $\bar{Y}$ is biased if viewed solely in terms of finite population sampling, but it is an unbiased estimator under the model-based theory (Cochran 1977). Fuller (1975) and Hartley and Sielken (1975) further elaborate on the superpopulation viewpoint for the regression estimation of the finite population parameters. Fuller provides the asymptotic properties of the estimators as $N$ and $n$ become large, whereas Hartley and Sielken develop mainly the finite sample theory results. The prediction approach has been favored over the conventional estimation approach by Royall (1970), who also advocates the use of purposive sampling to achieve "good" predictors. In their study of model-robustness, Royall and Herson (1973) show that the "balanced" sample designs are preferable to random sampling when a postulated linear model breaks down in the sense that the higher order polynomial regression terms are omitted. An excellent treatment on the topic is given in Cassel, Särndal, and Wretman (1977).

Presently, we consider the estimation of $\bar{Y}$ assuming a realistic superpopulation model as argued and stated in (1.1) – (1.2). This is a calibration model since the response is the independent variable and the auxiliary variable is the dependent variable. Clearly, the calibration model differs from the usual linear regression model where the response corresponds to the dependent variable and the auxiliary variable is the independent one. It is in this way that the present study is different from those cited above.

A number of papers have been devoted to the choice of an estimator in calibration problems. For example, Krutchkoff (1967) advocated the use of the regression estimator whereas Berkson (1969) favored the classical estimator (see Section 2 for the definitions of these estimators). Williams (1969), among others, shows that the classical estimator is consistent but has an infinite variance. The regression estimator is shown to be inconsistent; it is biased toward the mean by an amount proportional to the distance between the mean and the point value to be estimated. There is, however, an important difference between the calibration problem treated in these papers and the present one. In the case of finite population sampling, the choice of sample units directly affects the population quantity to be estimated, whereas in the standard calibration case, the quantity estimated does not depend on the sample. This and the other differences that exist between the two cases parallel those discussed in Royall and Herson (1973) for the estimation of finite population total assuming a linear regression model.

In Section 2 we introduce a general regression estimator which has the standard regression and classical estimator as special cases. The asymptotic bias and variance of the general estimator are discussed in Section 3. Because their analytical expressions are complex, a simulation study was conducted to evaluate numerically the bias and relative efficiency for the three estimators given in Section 2. The results of this study are presented in Section 4.

## 2. Estimators

An estimator of the form

$$\hat{\bar{Y}} = \bar{y} + \gamma(\bar{X} - \bar{x}) \tag{2.1}$$

is an unbiased estimator of $\bar{Y}$ if $\gamma$ does not depend upon $\bar{x}$ since $E(\bar{y}) = \bar{Y}$ and $E(\bar{x}) = \bar{X}$, where $E$ denotes the expectation with respect to the random sampling design. The variance of $\hat{\bar{Y}}$ is minimized if

$$\gamma = \text{Cov}(\bar{y}, \bar{x})/\text{Var}(\bar{x}). \tag{2.2}$$

Replacing $\gamma$ in (2.2) by its estimate

$$t = \sum_1^n (x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum_1^n (x_i - \bar{x})^2 \tag{2.3}$$

yields the standard regression estimator

$$\hat{\bar{Y}}_R = \bar{y} + t(\bar{X} - \bar{x}). \tag{2.4}$$

The large sample properties of $\hat{\bar{Y}}_R$ are well known, provided the $x_i$, $i = 1, 2, \ldots, N$, are assumed fixed (Cochran 1977).

Hung and Fuller (1987) address a problem similar to the present one where the auxiliary variable values are estimated. They, however, ignore the modeling aspect and simply use the regression estimator given in (2.4) and investigate its asymptotic properties.

Taking the model (1.1) into consideration, the classical least-squares estimation approach yields another estimator of $\bar{Y}$ given by

$$\hat{\bar{Y}}_C = \bar{y} + (\bar{X} - \bar{x})/b \tag{2.5}$$

where

$$b = \sum_1^n (x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum_1^n (y_i - \bar{y})^2. \tag{2.6}$$

The classical estimator in (2.5) minimizes the model error in (1.1), whereas the basis for the regression estimator given in (2.4) is simply to estimate $\bar{Y}$ as efficiently as possible. Chhikara and Deng (1988) discuss the conditional inference of the two estimators, given sample units with $y_1, \ldots, y_n$ fixed. It is seen easily that $\hat{\bar{Y}}_C$ has a conditional bias of order $1/n$. Moreover, $\hat{\bar{Y}}_R$ is conditionally a biased estimator of $\bar{Y}$.

The two estimators given in (2.4) and (2.5) are special cases of a general regression-type estimator defined by

$$\hat{\bar{Y}}_g = \bar{y} + a(\bar{X} - \bar{x}) \tag{2.7}$$

where

$$a = bs_y^2/(b^2 s_y^2 + cs^2) \tag{2.8}$$

with $0 \leqslant c < \infty$, $b$ as in (2.6)

$$s_y^2 = \sum_1^n (y_i - \bar{y})^2/(n - 1)$$

and

$$s^2 = \sum_1^n [x_i - \bar{x} - b(y_i - \bar{y})]^2/(n - 1). \tag{2.9}$$

For $c = 0$ in (2.8), the classical estimator $\hat{\bar{Y}}_C$ is obtained, and for $c = 1$, the regression estimator $\hat{\bar{Y}}_R$ is obtained. Moreover, other estimators of $\bar{Y}$ can be constructed by choosing $c$ differently. One estimator that merits consideration is obtained by choosing $c = 1/(n - 1)$ and is denoted by $\hat{\bar{Y}}_U$. This choice of $c = 1/(n - 1)$ modifies the classical estimator only slightly, yet it corrects the major drawback of the classical estimator, namely, an infinite variance. This follows because of a finite range of values that $\hat{\hat{Y}}_U$ would take.

One suggestion was to choose $c = 1/2$, but this arbitrary choice was of no special interest since it would lead to intermediate results between $\hat{\hat{Y}}_C$ and $\hat{\hat{Y}}_R$. It should be noted that constant $c$ in (2.8) is introduced in order to combine the different estimators into one as given in (2.7). This allows us to investigate various estimators simultaneously once we derive the results for $\hat{\hat{Y}}_g$. Based on similar ideas, Tin (1965) constructed and investigated several ratio estimators.

In the next section, we first outline the properties of $\hat{\hat{Y}}_g$ by evaluating its asymptotic bias and variance. This is followed by a detailed investigation of the three estimators $\hat{\hat{Y}}_C$, $\hat{\hat{Y}}_U$, and $\hat{\hat{Y}}_R$ for their asymptotic bias and variance. The analytical results are numerically evaluated using simulations and the bias and variances are studied. It is seen that the regression estimator $\hat{\hat{Y}}_R$ has the smallest variance. Although its bias is of order $1/n$, this bias is numerically smaller than that of $\hat{\hat{Y}}_C$. The estimator $\hat{\hat{Y}}_U$ has a bias of order $1/n^2$, but it is still inferior to $\hat{\hat{Y}}_R$ in terms of mean square error.

## 3. Asymptotic Bias and Variance

First, we need to introduce notation for certain population quantities. For the model error distribution in (1.1), we denote the third and fourth moments by $\mu_3$ and $\mu_4$, respectively, and the fourth cumulant by $K_4 = \mu_4 - 3\sigma^4$. We assume that these quantities are finite. Next, for the finite population of $y$'s, we denote the variance by $S_y^2$, the third central moment by $S_{3y}$ and the fourth cumulant, $(S_{4y} - 3S_y^4)$ by $K_{4y}$. The corresponding quantities for a sample of $y$'s are similarly denoted but with a lower case $s$.

The bias and variance of $\hat{\hat{Y}}_g$ are derived in Appendix A. After proper substitutions in (A.25) and (A.30) and simplification of the resulting expressions, it follows that the bias of $\hat{\hat{Y}}_g$ to $0(1/n)$ is

$$\text{Bias}\,(\hat{\hat{Y}}_g) = (1 - n/N)cA[\mu_3 - \beta\sigma^2 S_{3y}/S_y^2]/nV \quad (3.1)$$

and the mean square error of $\hat{\hat{Y}}_g$ to $0(1/n^2)$ is

$$\begin{aligned}\text{MSE}\,(\hat{\hat{Y}}_g) = {} & \{[(1 - n/N)/n]\,[A^2\sigma^2 + c^2\sigma^4 S_y^2/V^2] \\ & + 2(1 - c)\sigma^2 A^2 P + (\beta^2 S_y^2 + \sigma^2)Q \\ & + A\beta c^2\sigma^4(6A\beta S_{3y}^2/S_y^2 - K_{4y})/nV^2 S_y^2 \\ & + 2Ac[(1 - 8A\beta + 6A^2\beta^2)\mu_3 S_{3y}/S_y^2 \\ & - AK_4]/nV + A^2c^2[(\beta^2 S_y^2 \\ & + (3 - 2c)\sigma^2)K_4 + 6\mu_3^2]/nV^2\} \quad (3.2)\end{aligned}$$

where

$$P = [\sigma^2(\beta^2 S_y^2 - 3c^2\sigma^2) + 2c^2\sigma^4]/[(n - 1)V^2] \quad (3.3)$$

and

$$Q = \sigma^2 S_y^2[(c\sigma^2 - \beta^2 S_y^2)^2 + 4c^2 A\beta\sigma^2 V]/[(n - 1)V^4] \quad (3.4)$$

with

$$V = \beta^2 S_y^2 + c\sigma^2 \quad \text{and} \quad A = \beta S_y^2/V. \quad (3.5)$$

Next, the variance of $\hat{\hat{Y}}_g$ can be expressed to $0(1/n^2)$ by making use of the expressions in (3.1) and (3.2).

For an arbitrary $c$ value, the above expressions are fairly complex. One may, however, notice that when $c = 0$, one of the terms of $0(1/n)$ and the last three terms of $0(1/n^2)$ given in (3.2) vanish. On the other hand, one of the terms of $0(1/n^2)$ involves $(1 - c)$ which becomes zero when $c = 1$. Hence, there is no clear indication which of the possible estimators has a smaller mean square error. Since the approximation of variance or mean square error involving terms up to $0(1/n)$ may not be adequate for small samples, it would be beneficial to consider the terms of $0(1/n^2)$ in evaluating these estimators. We now consider the three special cases of $\hat{\bar{Y}}_g$ as discussed in the last section.

### 3.1. *Classical estimator* ($\hat{\bar{Y}}_C$)

This estimator given in (2.5) follows by specifying $c = 0$ in (2.7) – (2.8). For $c = 0$, the expression in (3.1) becomes zero. Thus, $\hat{\bar{Y}}_C$ is unbiased to $0(1/n)$. On the other hand, it can be verified from (A.2) that the conditional bias of $\hat{\bar{Y}}_C$ to $0(1/n)$, given sample units with $y$'s fixed, is

$$(\sigma^2/[(n - 1)\beta^2 s_y^2])(\bar{Y} - \bar{y}). \qquad (3.6)$$

Next, the variance of $\hat{\bar{Y}}_C$ to $0(1/n^2)$ obtained from (3.1) and (3.2) is given by

$$\text{Var}(\hat{\bar{Y}}_C) = [(1 - n/N)/n](\sigma^2/\beta^2) \times [1 + 3\sigma^2/\beta^2 S_y^2)/(n - 1)]. \qquad (3.7)$$

### 3.2. *Modified classical estimator* ($\hat{\bar{Y}}_U$)

The conditional bias given in (3.6) for the classical estimator is directly a result of the bias of $1/b$ used for an estimator of $1/\beta$. If we specify $c = 1/(n - 1)$ in (2.8) and use the resulting expression

$$a = b/[b^2 + s^2/(n - 1)s_y^2] \qquad (3.8)$$

for an estimation of $1/\beta$, then $a$ in (3.8) is conditionally an unbiased estimator of $1/\beta$ up to $0(1/n)$. Accordingly, the modified classical estimator $\hat{\bar{Y}}_U$ obtained by taking $c = 1/(n - 1)$ is conditionally unbiased to $0(1/n)$.

Next, it follows from (3.1) that the bias is of $0(1/n^2)$ when $c = 1/(n - 1)$ and thus $\hat{\bar{Y}}_U$ is unconditionally unbiased to $0(1/n)$, as is the case with the classical estimator $\hat{\bar{Y}}_C$. Its variance to $0(1/n^2)$ is easily obtained from (3.1) – (3.2) and is given by

$$\text{Var}(\hat{\bar{Y}}_U) = [(1 - n/N)/n](\sigma^2/\beta^2) \times [1 + (1 + \sigma^2/\beta^2 S_y^2)/(n - 1)]. \qquad (3.9)$$

It may be noticed that the two classical estimators $\hat{\bar{Y}}_C$ and $\hat{\bar{Y}}_U$ have the same variance to $0(1/n)$, and they differ only by a fixed multiple in the term of $0(1/n^2)$. It is obvious that the modified estimator is expected to have smaller mean square error and is preferable over the other estimator.

### 3.3. *Regression estimator* ($\hat{\bar{Y}}_R$)

This estimator corresponds to the case of $c = 1$ in (2.7) – (2.8). It can be easily deduced from (3.1) that $\hat{\bar{Y}}_R$ has a bias to $0(1/n)$ given by

$$\text{Bias}(\hat{\bar{Y}}_R) = [(1 - n/N)/n] \times [\beta S_y^2 \mu_3/(\beta^2 S_y^2 + \sigma^2)^2 - \rho^2(1 - \rho^2) S_{3y}/S_y^2] \qquad (3.10)$$

where $\rho$ is the population correlation coefficient between $x$ and $y$, which can be expressed in terms of other parameters as follows

$$\rho^2 = \beta^2 S_y^2/(\beta^2 S_y^2 + \sigma^2). \qquad (3.11)$$

It may be noted that the bias of $\hat{\bar{Y}}_R$ directly depends upon the population skewness; the first term in (3.10) involves the skewness of the model error distribution and the second term involves the skewness of $y$-values.

Next, the variance of $\hat{\bar{Y}}_R$ to $0(1/n^2)$ is obtained from (3.2) and (3.10) as follows

$$\text{Var}(\hat{\bar{Y}}_R) = [(1 - n/N)/n] \times \{[1 + 1/(n - 1)](1 - \rho^2) S_y^2$$

$$+ \rho^2(1 - \rho^2)^2[6\rho^2 S_{3y}^2/\sigma^2 - K_{4y}]/nS_y^2$$
$$+ 2\beta(1 - 2\rho^2 - 6\rho^2(1 - \rho^2))\mu_3 S_{3y}/$$
$$[nS_y^2(\beta^2 S_y^2 + \sigma^2)^2]$$
$$+ \beta^2 S_y^4[6\mu_3^2/(\beta^2 S_y^2 + \sigma^2) - K_4]/$$
$$n(\beta^2 S_y + \sigma^2)^4\} - [\text{Bias}\ (\hat{\bar{Y}}_R)]^2. \quad (3.12)$$

Note that the variance of a classical estimator is a function of $\sigma^2/\beta^2 = (1/\rho^2 - 1)S_y^2$ which becomes large as $\beta^2$ or $\rho^2$ becomes small. On the other hand, the variance of the regression estimator is bounded. In Var $(\hat{\bar{Y}}_R)$, the third and fourth moments of the model error and also those of $y$-values are of opposite signs in the terms that involve these higher moments and thus they may make these terms negligible.

It may be mentioned here that the expression for $\rho^2$ in (3.11) involves the finite population parameter $S_y^2$ and the model parameters $\beta$ and $\sigma^2$, but this presents no problem because the linear model (1.1) is postulated for the superpopulation with respect to variable $x$ only, and not the finite target population of $y$-values.

## 4 Numerical Evaluations

Simulation studies were made to evaluate the performance of the three estimators discussed above. A finite population consisting of 500 units was considered, where their $y$-values were generated according to a right skewed distribution so that $S_{3y}$ and $K_{4y}$ were nonzero. For the sake of convenience in simulating random numbers, we chose for the right skewed distribution a noncentral chi-square variable obtained by $W = (1/2)(Z + \sqrt{3})^2$, where $Z$ is the standard normal variate. The random variable $W$ has mean 2.0 and variance 3.5. Again, for the sake of convenience, we let $\alpha = 0$, $\beta = 1$ in model (1.1) so that the $x$-values were obtained according to the relation, $x_i = y_i + e_i$. The error $e_i$'s were generated according to three different distributions, one normal and two nonnormal. Of the two nonnormal error distributions, one was the random variable $W$ and the other was a double exponential. All error distributions were adjusted to have a mean of zero. The values generated using these nonnormal distributions allowed us to evaluate the effect of skewness and kurtosis on the bias and variances of the three estimators discussed earlier.

For the error variance $\sigma^2$, three different values were chosen, and these choices were made so that the population parameter $\rho^2 =$ 0.25, 0.70, and 0.90. Equation (3.11) describes $\sigma^2$ in terms of $\rho^2$ and thus $\sigma^2$ can be easily obtained since $\beta = 1$ and $S_y^2$ is computed from the set of 500 generated $y$-values. Thus, nine independent sets were considered for generating errors, and all errors in a set were generated independently of one another as well as separately for each replication. The simulation evaluations parallel exactly the theoretical discussion where both the model error and sampling error components were taken into account in obtaining the estimated bias and variances for the three estimators.

To a certain extent, the choice of these inputs for the simulation study was motivated by the crop acreage estimation problem described earlier. The correlation between the LANDSAT derived crop acreage estimates and the actual acreages would depend upon the crop growth stage at the time of LANDSAT data acquisition, competing crops and other ground covers, and the crop size, etc. If the data acquisition corresponds to an early growth stage for a crop or if its spectral signatures are similar to those of other ground cover types, it would lead to a low to moderate correlation, whereas highly distinct spectral signatures for different ground covers would provide high correlation between the actual acreages and their

LANDSAT derived estimates for a crop. Next, the crop size distribution for segments in an area often involves higher frequencies for the smaller crop acreages and fewer counts for the larger acreages; so this suggests the appropriateness of using a right skewed distribution for generating a finite population of $y$-values.

Samples of size $n = 4$, 10, and 25 were drawn from the population and all three estimates were computed from each of the samples. Again, these sample sizes were chosen keeping in view the above crop acreage estimation problem. In agricultural surveys conducted by USDA, the number of sample units varies across strata, ranging from only a few to many area segments allocated to a stratum.

The relative efficiency, bias, and variance of each estimator based on 1000 replications were determined. The relative efficiency is obtained as a ratio of the observed variance of sample mean $\bar{y}$ to that of an estimator.

Table 1 gives the relative efficiencies for the three estimators in each case of the error distributions. Three error cases listed as (i)–(iii) are indicated in terms of their skewness, $\gamma_1 = \mu_3/\sigma^{3/2}$, and kurtosis, $\gamma_2 = \mu_4/\sigma^4 - 3$, computed from the $e_i$ generated across all replications using error distributions: (i) normal errors ($\gamma_1 = 0$, $\gamma_2 = 0$), (ii) right skewed errors ($\gamma_1 = 1.1$, $\gamma_2 = 1.1$), (iii) double exponential errors ($\gamma_1 = 0.4$, $\gamma_2 = 3.4$). The numerical results show that the regression estimator $\hat{\bar{Y}}_R$ is the most efficient of all three estimators in all three cases (i)–(iii). Unless both the sample size and correlation are small (i.e., $n = 4$ and $\rho^2 = 0.25$), $\hat{\bar{Y}}_R$ has a relative efficiency that exceeds 1 and it approaches almost 10 when $\rho^2 = 0.90$ and $n = 25$. The classical estimator $\hat{\bar{Y}}_C$ is highly inefficient compared to the sample mean $\bar{y}$ when $n = 4$ or $\rho^2 = 0.25$, whereas the modified classical estimator $\hat{\bar{Y}}_U$ is almost as efficient as $\hat{\bar{Y}}_R$ with the exception of $\rho^2 = 0.25$, in which case it is also consistently inefficient. In any case, relative efficiencies of the estimators are highly influenced by the correlation, as expected. All three estimators seem to be fairly robust with respect to any departure from normality of

*Table 1. Relative efficiencies of estimators of population mean*

| $\rho^2$ | Sample size $n$ | Estimator $\hat{\bar{Y}}_R$ | $\hat{\bar{Y}}_U$ | $\hat{\bar{Y}}_C$ |
|---|---|---|---|---|
| (i) Normal error distribution ($\gamma_1 = 0$, $\gamma_2 = 0$) | | | | |
| 0.25 | 4 | 0.805 | 0.640 | 0 |
| | 10 | 1.207 | 0.612 | 0.004 |
| | 25 | 1.276 | 0.321 | 0.020 |
| 0.70 | 4 | 1.891 | 1.616 | 0.012 |
| | 10 | 2.782 | 2.088 | 1.095 |
| | 25 | 2.989 | 2.166 | 2.031 |
| 0.90 | 4 | 4.690 | 4.413 | 0.062 |
| | 10 | 7.575 | 7.216 | 6.527 |
| | 25 | 9.761 | 9.030 | 8.933 |
| (ii) Right skewed error distribution ($\gamma_1 = 1.1$, $\gamma_2 = 1.1$) | | | | |
| 0.25 | 4 | 0.644 | 0.501 | 0 |
| | 10 | 1.122 | 0.599 | 0.010 |
| | 25 | 1.280 | 0.374 | 0.054 |
| 0.70 | 4 | 2.055 | 1.835 | 0.090 |
| | 10 | 2.792 | 1.943 | 1.416 |
| | 25 | 3.357 | 2.425 | 2.284 |
| 0.90 | 4 | 3.743 | 3.189 | 0.070 |
| | 10 | 8.651 | 8.204 | 2.554 |
| | 25 | 9.070 | 8.420 | 8.321 |
| (iii) Double exponential error distribution ($\gamma_1 = 0.4$, $\gamma_2 = 3.4$) | | | | |
| 0.25 | 4 | 0.755 | 0.570 | 0 |
| | 10 | 1.286 | 0.656 | 0 |
| | 25 | 1.203 | 0.382 | 0.160 |
| 0.70 | 4 | 1.910 | 1.631 | 0.022 |
| | 10 | 2.994 | 2.184 | 0.025 |
| | 25 | 3.313 | 2.441 | 2.313 |
| 0.90 | 4 | 4.811 | 4.160 | 0.468 |
| | 10 | 8.137 | 6.827 | 4.912 |
| | 25 | 8.454 | 7.738 | 7.623 |

*Table 2. Estimated vs actual bias of the regression estimator $\hat{\bar{Y}}_R$ when $\bar{Y} = 2.0$*

| $\rho^2$ | Sample size $n$ | Bias: Estimated | Bias: Actual |
|---|---|---|---|
| (i) Normal error distribution ($\gamma_1 = 0, \gamma_2 = 0$) | | | |
| 0.25 | 4 | −0.152 | −0.077 |
| | 10 | −0.037 | −0.050 |
| | 25 | −0.014 | −0.018 |
| 0.70 | 4 | −0.110 | −0.107 |
| | 10 | −0.040 | −0.048 |
| | 25 | −0.014 | −0.019 |
| 0.90 | 4 | −0.054 | −0.070 |
| | 10 | −0.017 | −0.033 |
| | 25 | −0.007 | −0.012 |
| (ii) Right skewed error distribution ($\gamma_1 = 1.1, \gamma_2 = 1.1$) | | | |
| 0.25 | 4 | (−0.070) | 0.110 |
| | 10 | 0.009 | 0.016 |
| | 25 | 0.010 | 0.049 |
| 0.70 | 4 | −0.069 | −0.028 |
| | 10 | −0.010 | −0.000 |
| | 25 | −0.004 | −0.001 |
| 0.90 | 4 | −0.034 | −0.055 |
| | 10 | −0.010 | −0.023 |
| | 25 | −0.004 | −0.004 |
| (iii) Double exponential error distribution ($\gamma_1 = 0.4, \gamma_2 = 3.4$) | | | |
| 0.25 | 4 | −0.160 | −0.124 |
| | 10 | −0.044 | −0.084 |
| | 25 | −0.019 | −0.020 |
| 0.70 | 4 | −0.160 | −0.154 |
| | 10 | −0.039 | −0.043 |
| | 25 | −0.017 | −0.017 |
| 0.90 | 4 | −0.052 | −0.070 |
| | 10 | −0.019 | −0.027 |
| | 25 | −0.008 | −0.002 |

the error distribution; this can be seen from the results in case (i) versus cases (ii) and (iii).

The observed bias for the two classical estimators was found insignificant. Although the regression estimator is biased, its observed bias was small. In Table 2 the observed (actual) and estimated values of the bias of $\hat{\bar{Y}}_R$ are given. The estimated bias was obtained as an average of the estimates resulting from (3.10) with population parameters replaced by the corresponding sample quantities and $1/(n-1)$ replaced by $1/(n-3)$; the latter substitution is justified because $E(1/s_y^2) = (n-1)/(n-3)\,S_y^2$, and the difference is of $0(1/n^2)$. Except in one instance, marked in the table with parenthesis, the sample estimates of bias are fairly in agreement with the actual biases with respect to both sign and magnitude. The exception occurs in the case of nonnormal errors with $\rho^2 = 0.25$ and $n = 4$ where the expected bias is $-0.070$ and the observed bias is 0.110. This is attributed to poor sample estimates of the third moments.

Among the three estimators discussed, $\hat{\bar{Y}}_R$ is clearly the best choice for the estimation of $\bar{Y}$. In the next section we discuss its variance estimation.

## 5. Variance Estimation for $\hat{\bar{Y}}_R$

The sample analogue of the mean square error in (3.12) provides an estimator of MSE ($\hat{\bar{Y}}_R$). This, however, requires computation of second, third, and fourth sample moments. Since some terms involving the third and fourth moments have opposite signs, one wonders whether their inclusion in the expression of the variance estimator leads to a better estimate of Var ($\hat{\bar{Y}}_R$). Thus, we compare the variance estimator obtained as a sample analogue from (3.10) and (3.12) with three other variance estimators which are both special cases and easily computed.

The large sample variance obtained from (3.10) and (3.12) by retaining only terms of $0(1/n)$ yields the usual large sample variance estimator given by Cochran (1977, p. 195), namely

$$v_L = (1 - n/N)\, s_e^2/n \qquad (5.1)$$

where $s_e^2$ is the residual mean squared error obtained by linearly regressing the $y_i$ on the $x_i$, that is

$$s_e^2 = \sum_1^n [y_i - \bar{y} - t(x_i - \bar{x})]^2/(n - 2). \quad (5.2)$$

We use the divisor $(n - 2)$ instead of $(n - 1)$ in (5.2) since $s_e^2$ is an unbiased estimator of $(1 - \rho^2) S_y^2$ whenever the regression of $y$ on $x$ is linear. Next, if we retain terms of $0(1/n^2)$, but ignore the third and fourth moments of the $y$'s and those of the error distribution as reflected in (3.10) and (3.12), the estimator in (5.1) can be improved. This variance estimator is given by

$$v_N = (1 - n/N)[1 + 1/(n - 3)](s_e^2/n). \quad (5.3)$$

Clearly, $v_N$ is most appropriate if the $y_i$ and the $e_i$ are each normally distributed.

Another estimator is obtained by the sample analogue of the expected variance of $\hat{\bar{Y}}_R$ as derived by Cochran (1942); the expression is also given in Cochran (1977, p. 197). Its derivation, however, is based on the assumption of a linear regression of $y$ on $x$ and that the model errors are normally distributed. Although this is not applicable to our problem, we consider it simply for the sake of comparison. The corresponding variance estimator is given by

$$v_M = (1 - n/N) \times [1 + 1/(n - 3) + 2G_1^2/n^2]s_e^2/n \quad (5.4)$$

where $G_1 = S_{3x}/S_x^3$, the relative skewness of the $x$'s in a realized finite population.

Let $v_A$ be the variance estimator obtained from (3.10) and (3.12) with population parameters replaced by their sample quantities and $(n - 1)$ by $(n - 3)$. Again the term $1/(n - 3)$ in replacement of $1/(n - 1)$ in (3.12) is justified because the conditional variance of $\hat{\bar{Y}}_R$, given fixed $y$'s, involves $1/s_y^2$ and $E(1/s_y^2) = (n - 1)/(n - 3) S_y^2$. The difference is of $0(1/n^2)$. This reasoning also applies in obtaining $v_N$ in (5.3).

Each of these variance estimates were computed and averaged over the 1000 replications in our simulations discussed in

*Table 3. Ratio of an estimated to actual variance of $\hat{\bar{Y}}_R$*

| $\rho^2$ | Sample size $n$ | Variance estimator $v_L$ | $v_N$ | $v_A$ | $v_M$ |
|---|---|---|---|---|---|
| (i) Normal error distribution ($\gamma_1 = 0, \gamma_2 = 0$) | | | | | |
| 0.25 | 4 | 0.588 | 1.175 | 1.245 | 1.244 |
| | 10 | 0.881 | 1.007 | 1.048 | 1.016 |
| | 25 | 1.015 | 1.061 | 1.073 | 1.062 |
| 0.70 | 4 | 0.504 | 1.007 | 1.074 | 1.073 |
| | 10 | 0.832 | 0.951 | 1.012 | 0.962 |
| | 25 | 0.878 | 0.918 | 0.943 | 0.920 |
| 0.90 | 4 | 0.434 | 0.868 | 0.913 | 0.930 |
| | 10 | 0.766 | 0.875 | 0.907 | 0.898 |
| | 25 | 0.915 | 0.957 | 0.972 | 0.960 |
| (ii) Right skewed error distribution ($\gamma_1 = 1.1, \gamma_2 = 1.1$) | | | | | |
| 0.25 | 4 | 0.400 | 0.800 | 0.849 | 0.851 |
| | 10 | 0.799 | 0.913 | 0.998 | 0.928 |
| | 25 | 0.896 | 0.937 | 0.989 | 0.940 |
| 0.70 | 4 | 0.504 | 1.009 | 1.076 | 1.069 |
| | 10 | 0.799 | 0.913 | 0.998 | 0.928 |
| | 25 | 0.896 | 0.937 | 0.989 | 0.940 |
| 0.90 | 4 | 0.345 | 0.690 | 0.724 | 0.735 |
| | 10 | 0.779 | 0.890 | 0.912 | 0.905 |
| | 25 | 0.858 | 0.897 | 0.896 | 0.900 |
| (iii) Double exponential error distribution ($\gamma_1 = 0.4, \gamma_2 = 3.4$) | | | | | |
| 0.25 | 4 | 0.476 | 0.952 | 1.004 | 1.011 |
| | 10 | 0.905 | 1.034 | 1.084 | 1.049 |
| | 25 | 0.910 | 0.951 | 0.966 | 0.953 |
| 0.70 | 4 | 0.457 | 0.915 | 0.973 | 0.971 |
| | 10 | 0.779 | 0.891 | 0.950 | 0.903 |
| | 25 | 0.953 | 0.997 | 1.024 | 0.999 |
| 0.90 | 4 | 0.415 | 0.830 | 0.874 | 0.887 |
| | 10 | 0.809 | 0.924 | 0.963 | 0.941 |
| | 25 | 0.859 | 0.898 | 0.911 | 0.901 |

Section 4. The ratio of an averaged variance to the observed variance of $\hat{\bar{Y}}_R$ was computed in each case. Table 3 lists these variance ratios for the four variance estimators.

The numerical results in Table 3 show that the large sample variance estimator $v_L$ underestimates Var $(\hat{\bar{Y}}_R)$ and the underestimation is substantial, particularly in the case of $n = 4$. The other three variance estimators do not always underestimate the actual variance for this case, especially when $\rho^2 = 0.25$ and the errors are normally distributed. In the case of normal errors, $v_A$ and $v_M$ show no significant improvement over $v_N$ even though these two estimators include terms which involve the third and fourth sample moments of $y$'s.

In the case of nonnormal errors, $v_A$ and $v_M$ show a significant difference based on a 10% significance level sign-test, where $v_A$ is preferred over $v_M$. Although $v_A$ shows a slight improvement over $v_N$ as an estimator of Var $(\hat{\bar{Y}}_R)$, the small magnitude of their difference may not justify the use of $v_A$ due to the complexity of its computation.

In conclusion, the variance estimator $v_N$ performs fairly well compared to the various alternatives and it is easy to compute. Hence, it is the best variance estimator among the four discussed here.

## Appendix A

### Derivation of Bias $(\hat{\bar{Y}}_g)$ and Var $(\hat{\bar{Y}}_g)$

First we derive the conditional mean and variance of $\hat{\bar{Y}}_g$, given sample units with $y_1$, $y_2, \ldots, y_n$ fixed, by taking the expectation with respect to the error distribution of the linear model in (1.1).

Using the identity

$$\mathscr{E}(UW) = \mathscr{E}(U)\mathscr{E}(W) + \mathscr{C}ov\,(U, W) \tag{A.1}$$

it follows that the *conditional* mean and variance of $\hat{\bar{Y}}_g$, given the $y$'s, are

$$\mathscr{E}(\hat{\bar{Y}}_g) = \bar{y} + \mathscr{E}(a)\mathscr{E}(\bar{X} - \bar{x}) + \mathscr{C}ov\,(a, \bar{X} - \bar{x}) \tag{A.2}$$

and

$$\mathscr{V}(\hat{\bar{Y}}_g) = [(\mathscr{E}(a))^2 + \mathscr{V}(a)] \times [(\mathscr{E}(\bar{X} - \bar{x}))^2 + \mathscr{V}(\bar{X} - \bar{x})] + \mathscr{C}ov\,(a^2, (\bar{X} - \bar{x})^2) - [\mathscr{E}(a)\mathscr{E}(\bar{X} - \bar{x}) + \mathscr{C}ov\,(a, \bar{X} - \bar{x})]^2 \tag{A.3}$$

where $a$ is defined in (2.8),

$$\mathscr{E}(\bar{X} - \bar{x}) = \beta(\bar{Y} - \bar{y}) \tag{A.4}$$

*Table 4. Conditional expectations and covariances for differential terms*

| $Q$ | $d$ | $f$ | $d^2$ | $f^2$ | $d \times f$ |
|---|---|---|---|---|---|
| $\mathscr{E}(Q)$ | 0 | 0 | $\dfrac{\sigma^2}{(n-1)s_y^2}$ | $\dfrac{2\sigma^4}{n-1} + \dfrac{K_4}{n}$ | 0 |
| $\mathscr{C}ov\,(Q, \bar{e})$ | 0 | $\dfrac{\mu_3}{n}$ | $\dfrac{\mu_3}{n^2 s_y^2}$ | $\dfrac{4\mu_3\sigma^2}{n(n-1)} + \dfrac{K_5}{n^2}$ | 0 |
| $\mathscr{C}ov\,(Q, \bar{e}^2)$ | 0 | $\dfrac{K_4}{n^2}$ | $0(1/n^3)$ | $\dfrac{2\mu_3^2}{n^2} + 0(1/n^3)$ | 0 |

and

$$\mathscr{V}(\bar{X} - \bar{x}) = (1 - n/N)\sigma^2/n. \quad (A.5)$$

To simplify the expressions in (A.2) and (A.3), we now evaluate the conditional mean, variance and covariances involving $a$. Letting

$$v = b^2 s_y^2 + cs^2 \quad (A.6)$$

the coefficient

$$a = s_y^2(b/v) \quad (A.7)$$

can be expanded in a series by the finite difference method as follows:

Suppose

$$b = \beta + d \quad (A.8)$$

$$s^2 = \sigma^2 + f. \quad (A.9)$$

Then

$$v = v_0 + D = v_0(1 + D/v_0) \quad (A.10)$$

where

$$v_0 = \beta^2 s_y^2 + c\sigma^2 \quad (A.11)$$

$$D = (2\beta d + d^2) s_y^2 + cf. \quad (A.12)$$

Next, $a$ in (A.7) can be represented by the series

$$a = s_y^2[(\beta + d)/v_0] \times [1 - D/v_0 + D^2/v_0^2 + \ldots] \quad (A.13)$$

which is convergent since $|D| < v_0$ implies $0 < v < 2v_0$. It follows that the expansion (A.13) converges in probability to $a_0$, where

$$a_0 = s_y^2\beta/v_0.$$

Retaining terms in (A.13) up to the quadratic, we can write

$$a = a_0 + a_0(1/\beta - 2a_0)d - (a_0^2/\beta)cf/s_y^2 + a_0^2(4a_0 - 3/\beta)d^2 + (a_0^3/\beta)c^2f^2/s_y^4 - (a_0^2/\beta^2)(1 - 4a_0\beta)cdf/s_y^2. \quad (A.14)$$

Taking expectations and covariances, conditional on given $y$'s, and making use of the results given in Table 4, it follows that, ignoring the finite population correction for the time being

$$\mathscr{E}(a) = a_0 + a_0^2(4a_0 - 3/\beta)\sigma^2/(n - 1)s_y^2 + c^2(a_0^3/\beta^2)(2\sigma^4 + K_4)/(n - 1)s_y^2 \quad (A.15)$$

$$\mathscr{V}(a) = a_0^2(1/\beta - 2a_0)^2\sigma^2/(n - 1)s_y^2 + c^2(a_0^4/\beta^2)(2\sigma^4 + K_4)/(n - 1)s_y^4 \quad (A.16)$$

$$\mathscr{Cov}\,(a, \bar{X} - \bar{x}) = c(a_0^2/\beta)\mu_3/ns_y^2 - a_0^2(4a_0 - 3/\beta)\mu_3/n^2s_y^2 - 4c^2\sigma^2\mu_3/n^2s_y^4 \quad (A.17)$$

and

$$\mathscr{Cov}\,(a^2, (\bar{X} - \bar{x})^2) = (1/n^2)\{-2c(a_0^3/\beta ns_y^2) \times [K_4/n - 2\beta(\bar{Y} - \bar{y})\mu_3] + 6c^2a_0^4/(n^2\beta^2s_y^4)[\mu_3^2 - 4\beta\sigma^2\mu_3(\bar{Y} - \bar{y})] - 2a_0^2\mu_3(\bar{Y} - \bar{y})/(\beta n^2s_y^2)[(1 - 2a_0\beta)^2 - 2a_0\beta(4a_0\beta - 3)]\}. \quad (A.18)$$

Making substitutions from (A.4) – (A.5) and (A.15) – (A.18) into (A.2) and (A.3), the conditional mean and variance of $\hat{\bar{Y}}_g$ to $0(1/n^2)$ are obtained.

Next, we derive the *unconditional* bias and variance of $\hat{\bar{Y}}_g$ by taking further expectations with respect to the random sampling of $y$'s. It follows that the unconditional mean of $\hat{\bar{Y}}_g$ to $0(1/n)$ is given by

$$E\mathscr{E}(\hat{\bar{Y}}_g) = \bar{Y} + [(1 - n/N)/n]c\mu_3 E(a_0/v_0) + \beta \text{ Cov } [\mathscr{E}(a), (\bar{Y} - \bar{y})]. \quad (A.19)$$

Since $\mathscr{E}(a) = a_0 + 0(1/n)$ as shown in (A.15), $\mathscr{E}(a)$ may be replaced by $a_0$ in evaluating the covariance term in (A.19). Next, an expansion for $a_0$ as a function of $s_y^2$ can be found in a manner similar to the expansion in (A.13). Let

$$s_y^2 = S_y^2 + u$$

where $S_y^2$ is the finite population variance of $y$'s. Then

$$v_0 = V + \beta^2 u$$
$$= V[1 + (A\beta/S_y^2)u] \qquad (A.20)$$

where

$$V = \beta^2 S_y^2 + c\sigma^2 \qquad (A.21)$$
$$A = \beta S_y^2/V. \qquad (A.22)$$

Now expanding $v_0^{-1}$, it follows that

$$a_0 = \beta S_y^2(1 + u/S_y^2)v_0^{-1}$$
$$= A[1 + (1 - A\beta)u/S_y^2 - \beta(1 - A\beta)Au^2/S_y^4 + \ldots] \qquad (A.23)$$

which converges because $\beta^2|u| < V$ implies that $0 < v_0 < 2V$.

Making use of the expansion in (A.23), it follows that the covariance term in (A.19) to $0(1/n)$, ignoring higher than fourth moment terms, is given by

$$\text{Cov}\,(\mathscr{E}(a), (\bar{Y} - \bar{y}))$$
$$= -(1 - n/N)A(1 - A\beta)S_{3y}/nS_y^2 \qquad (A.24)$$

where $S_{3y}$ is the third population central moment of $y$'s. Recognizing that $E(a_0/v_0) = A/V$ to $0(1/n)$, it follows from (A.19) that the bias of $\hat{\bar{Y}}_g$ to $0(1/n)$ is

$$\text{Bias}\,(\hat{\bar{Y}}_g) = [(1 - n/N)/n] \times [c\mu_3 A/V - A\beta)S_{3y}/S_y^2]. \qquad (A.25)$$

To obtain the unconditional variance of $\hat{\bar{Y}}_g$, we shall make use of the identities

$$E([\mathscr{E}(Z)]^2) = [E\mathscr{E}(Z)]^2 + V(\mathscr{E}(Z))$$
$$\text{Var}\,(Z) = E(\mathscr{V}(Z)) + V(\mathscr{E}(Z)) \qquad (A.26)$$

where the inside expectations are with respect to the model error distribution, given $y$'s, and the outside ones with respect to the random sampling of $y$'s. Letting, for the conditional bias,

$$L = \mathscr{E}(\hat{\bar{Y}}_g) - \bar{y}$$
$$= \beta\mathscr{E}(a)(\bar{Y} - \bar{y}) + \mathscr{C}ov\,(a, \bar{X} - \bar{x}). \qquad (A.27)$$

It follows from (A.3) – (A.5) that

$$E(\mathscr{V}(\hat{\bar{Y}}_g)) = \{[E\mathscr{E}(a)]^2 + \text{Var}\,(\mathscr{E}(a)) + E(\mathscr{V}(a))\}(1 - n/N)[\beta^2 S_y^2 + \sigma^2]/n + \beta^2\,\text{Cov}\,\{[\mathscr{E}(a)]^2, (\bar{Y} - \bar{y})^2\} + E[\mathscr{C}ov\,(a^2, (\bar{X} - \bar{x})^2)] - E(L^2) \qquad (A.28)$$

and

$$V(\mathscr{E}(\hat{\bar{Y}}_g)) = E([\mathscr{E}(\hat{\bar{Y}}_g) - \bar{Y}]^2) - [E(L)]^2 = (1 - n/N)[1 - 2\beta E\mathscr{E}(a)] \times S_y^2/n + E(L^2) - 2\beta\,\text{Cov}\,(\mathscr{E}(a), (\bar{Y} - \bar{y})^2) + 2c\mu_3\,\text{Cov}\,(a_0/v_0, (\bar{Y} - \bar{y}))/n - [E(L)]^2. \qquad (A.29)$$

Adding (A.28) and (A.29), the identity in (A.26) yields to $0(1/n^2)$

$$\text{MSE}\,(\hat{\bar{Y}}_g) = (1 - n/N)\{[E\mathscr{E}(a)]^2\sigma^2 + [1 - \beta E\mathscr{E}(a)]^2 S_y^2 + [\beta^2 S_y^2 + \sigma^2]E(\mathscr{V}(a))\}/n - 2\beta\,\text{Cov}\,(\mathscr{E}(a), (\bar{Y} - \bar{y})^2) + \beta^2\,\text{Cov}\,([\mathscr{E}(a)]^2, (\bar{Y} - \bar{y})^2) + E(\mathscr{C}ov\,(a^2, (\bar{X} - \bar{x})^2) - 2c\mu_3\,\text{Cov}\,(a_0/v_0, (\bar{Y} - \bar{y}))/n. \qquad (A.30)$$

The various terms in (A.30) are obtained using the expansion in (A.23) and the conditional results in (A.15)–(A.18). Their expressions given below maintain the mean square error in (A.30) to $0(1/n^2)$.

$$E[\mathscr{E}(a)] = A + (1 - n/N) \times [A^2(4A - 3/\beta)\sigma^2/(n-1)S_y^2 + c^2(A^3/\beta^2)(2\sigma^4 + K_4)/(n-1)S_y^4 - A\beta(1 - A\beta)(2S_y^4 + K_{4y})/(n-1)S_y^4] \quad (A.31)$$

$$\text{Var}(a) = V(\mathscr{E}(a)) + E(\mathscr{V}(a)) \quad (A.32)$$

where

$$V(\mathscr{E}(a)) = (1 - n/N)A^2(1 - A\beta)^2 \times (2S_y^4 + K_{4y})/(n-1)S_y^4 \quad (A.33)$$

$$E(\mathscr{V}(a)) = (1 - n/N) \times [A^2(1/\beta - 2A)\sigma^2/(n-1)S_y^2 + c(A^4/\beta^2)(2\sigma^2 + K_4)/(n-1)S_y^4 \quad (A.34)$$

$$\text{Cov}([\mathscr{E}(a)]^2, (\bar{Y} - \bar{y})) = 2(1 - n/N)A^2\beta^2(1 - A\beta)/n^2S_y^2 \times [K_{4y} + (1 - 3A\beta)S_{3y}^2/S_y^2] \quad (A.35)$$

$$E(\mathscr{C}ov\,(a^2, (\bar{X} - \bar{x})^2)) = (1 - n/N)[-2cA^3K_4/\beta S_y^2 + 3c^2(A^4/\beta^2)(2\mu_3^2/S_y^4)]/n^2 \quad (A.36)$$

$$\text{Cov}(\mathscr{E}(a), (\bar{Y} - \bar{y})^2) = (1 - n/N)A(1 - A\beta)/n^2S_y^2 + [K_{4y} - 2A\beta S_{3y}^2/S_y^2] \quad (A.37)$$

$$\text{Cov}(a_0/v_0, \bar{Y} - \bar{y}) = (1 - n/N) \times (1 - 2A\beta)AS_{3y}/nVS_y^2. \quad (A.38)$$

The unconditional variance of $\hat{\hat{Y}}_g$ is obtained by substitution from (A.31) – (A.38) and (A.25) into (A.30). The resulting expressions for the bias and MSE of $\hat{\hat{Y}}_g$ are as given in (3.1) and (3.2) in Section 3.

# Appendix B

## Conditional Expectations and Covariances for Differential Terms

We write the model in (1.1) as

$$\mathbf{x} = \alpha + \beta\mathbf{y} + \mathbf{e} \quad (B.1)$$

where **x**, **y**, and **e** are column vectors each of dimension $n$. Then one can write

$$b = \mathbf{h}'\mathbf{x} \quad (B.2)$$

with

$$\mathbf{h} = (\mathbf{y} - \mathbf{1}\bar{y})/(n-1)s_y^2. \quad (B.3)$$

Then for the differential $d$ and mean error $\bar{e}$, one has

$$d = b - \beta = \mathbf{h}'\mathbf{x} - \beta = \mathbf{h}'\mathbf{e} \quad (B.4)$$

$$d^2 = \mathbf{e}'\mathbf{h}\mathbf{h}'\mathbf{e} \quad (B.5)$$

$$\bar{e} = (1/n)\mathbf{1}'\mathbf{e} \quad (B.6)$$

$$\bar{e}^2 = (1/n^2)\mathbf{e}'\mathbf{1}\mathbf{1}'\mathbf{e}. \quad (B.7)$$

Making use of formulas in Tan and Cheng (1981), we obtain the conditional covariances, given $y$'s, as follows:

$$\mathscr{C}ov\,(d, \bar{e}) = \mathscr{C}ov\,(d, \bar{e}^2) = 0 \quad (B.8)$$

$$\mathscr{C}ov\,(d^2, \bar{e}^2) = 0(1/n^3) \quad (B.9)$$

$$\mathscr{C}ov\,(d^2, \bar{e}) = \mu_3/n^2s_y^2 \quad (B.10)$$

and formulas in Kendall and Stuart (1968, p. 286) lead to

$$\mathscr{C}ov\,(f, \bar{e}) = \mu_3/n \quad (B.11)$$

$$\mathscr{C}ov\,(f, \bar{e}^2) = K_4/n^2 \quad (B.12)$$

$$\mathscr{C}ov\,(f^2, \bar{e}) = 4\mu_3\sigma^2/n^2 + \text{terms involving } K_5 \quad (B.13)$$

where $K_5$ denotes the fifth cumulant for the error distribution. To evaluate $\mathscr{C}ov\,(f^2, \bar{e}^2)$, we use the following identity with respect to cumulants of order four for random variables, say $U_i$, with $E(U_i) = 0, i = 1, 2, 3, 4$:

$$K(U_1, U_2, U_3, U_4) = E(U_1 U_2 U_3 U_4) - \text{Cov}(U_1, U_2)\text{Cov}(U_3, U_4) - \text{Cov}(U_1, U_3)\text{Cov}(U_2, U_4) - \text{Cov}(U_1, U_4)\text{Cov}(U_2, U_3). \quad (B.14)$$

Then for the fourth order covariance,

$$\text{Cov}(U_1, U_2, U_3, U_4) = E(U_1 U_2 U_3 U_4) - \text{Cov}(U_1, U_2)\text{Cov}(U_3, U_4) \quad (B.15)$$

$$= K(U_1, U_2, U_3, U_4) + \text{Cov}(U_1, U_3) \times \text{Cov}(U_2, U_4) + \text{Cov}(U_1, U_4) \times \text{Cov}(U_2, U_3). \quad (B.16)$$

For $U_1 = U_2 = f$ and $U_3 = U_4 = \bar{e}$, we have

$$\mathscr{C}ov\,(f^2, \bar{e}^2) = 2\mu_3^2/n^2 + 0(1/n^3) \quad (B.17)$$

since $K(f, f^2, \bar{e}, \bar{e}^2)$ is $0(1/n^3)$. The above results are summarized in Table 4, where $Q$ denotes a differential variable from the set of variables, $d$, $f$, $d^2$, $f^2$ and $d \times f$, with $d = b - \beta$ and $f = s^2 - \sigma^2$.

Next, the covariances between $u = (s_y^2 - S_y^2)$, $(\bar{y} - \bar{Y})$, and $(\bar{y} - \bar{Y})^2$ with respect to the random sampling of $y$'s can be obtained by replacing $f$ by $u$, $\bar{e}$ by $(\bar{y} - \bar{Y})$, $\mu_3$ by $S_{3y}$, etc.

## 5. References

Berkson, J. (1969). Estimation of a Linear Function for a Calibration Line: Consideration of a Recent Proposal. Technometrics, 11, 649–660.

Cassel, C.-M., Särndal, C.-E. and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: John Wiley.

Chhikara, R.S. and Houston, A.G. (1984). Calibration or Inverse Regression: Which is Appropriate for Crop Surveys Using LANDSAT Data? Proceedings of the Second Annual NASA Symposium on Mathematical Pattern Recognition & Image Analysis, Johnson Space Center, Houston, Texas, 205–243.

Chhikara, R.S. and Deng, L.Y. (1988). Conditional Inference in Finite Population Sampling Under a Calibration Model. Communications in Statistics: Simulation and Computation, 17, 663–681.

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

Cochran, W.G. (1942). Sampling Theory When the Sampling Units Are of Unequal Sizes. Journal of the American Statistical Association, 37, 199–212.

Fuller, W.A. (1975). Regression Analysis for Sample Surveys. Sankhyā, ser. C, 37, 117–132.

Hartley, H.O. and Sielken Jr., R.L. (1975). A "Super-population Viewpoint" for Finite Population Sampling. Biometrics, 31, 411–422.

Holko, M.L. and Sigman, R.S. (1984). The Role of LANDSAT Data in Improving U.S. Crop Statistics. Paper presented at the Eighteenth International Symposium of Remote Sensing of Environment, Paris, France, October 1–5, 1984.

Hung, H.M. and Fuller, W.A. (1987). Regression Estimation of Crop Acreages with Transformed LANDSAT Data as Auxiliary Variables. Journal of Business and Economic Statistics, 5, 475–482.

Kendall, M.G. and Stuart, A. (1968). The Advanced Theory of Statistics, vol. 1. New York: Hafner.

Krutchkoff, R.G. (1967). Classical and Inverse Methods of Calibration. Technometrics, 9, 525–539.

Royall, R.M. (1970). On Finite Population Sampling Theory Under Certain Linear Regression Models. Biometrika, 57, 377–387.

Royall, R.M. and Herson, J.H. (1973). Robust Estimation in Finite Population I.

Journal of the American Statistical Association, 68, 880–889.

Sigman, R., Hanuschak, R., Craig, M., Cook, P., and Cardenas, M. (1978). The Use of Regression Estimation With LANDSAT and Probability Ground Sample Data. Proceedings of the American Statistical Association, Section of Survey Research Methods, 165–168.

Tan, W.Y. and Cheng, S.S. (1981). Cumulants of Bilinear Forms and Quadratic Forms and Their Applications. Communications in Statistics, ser. A, 10, 283–297.

Tin, M. (1965). Comparison of Some Ratio Estimators. Journal of the American Statistical Association, 60, 294–307.

U.S. Department of Agriculture (1975). Scope and Methods. Statistical Reporting Service, Misc. publication no. 1308, Washington, D.C., July.

Williams, E.J. (1969). A Note on Regression Methods in Calibration. Technometrics, 11, 189.

Received November 1987
Revised May 1990