PEDITOR Estimation Formulae Compiled Across The Years

Mike Craig

February 18, 1998

1. ESTIMATE.DOC - print of 1989 Xywrite file kept in-house by Richard Sigman as documetation for estimation (use MEC-CART.PRN driver to print)

2. ESTIMAT1.XYW - print of 1994 Xywrite file created by Mike Bellow to document large domain alternative estimators (use 3PDP-P2.PRN driver to print).

3. Copy of "Sampling Theory When the Sampling Units are of Unequal Sizes", by W.G. Cochran, JASA, Vol. 37, pp. 199-212, 1942. This paper is the source of the 1/(n-3) factor seen in the PEDITOR regression estimator variance formula per Richard Sigman.

4. One page (p.425) excerpt from "Frequency-Based Contextual Classification" by Gong & Howarth, PERS, Vol 58, No.4, April 1992; source of KAPPA measure added to CORREC module in 1994.

5. Memo from Raj Chhikara, University of Houston-Clear Lake, dated June 19, 1986, containing a report on varinace estimation named "Regression Estimates of Population Mean in Stratified Sampling: Case of Combined Regression".

6. Paper on robustness of PEDITOR regression estimator from Raj Chhikara and Jim J. McKeon, University of Houston-Clear Lake, dated May 8, 1986, titled "Regression Estimators in Crop Surveys: Case of Non-Normal and NonHomogeneous Errors".

ESTIMATE.DOC - 1989

Estimation Procedures Used in PEDITOR

Following is a condensed listing of the formulas used in the various programs in PEDITOR.

Analysis district estimator for $\hat{y}$ in a single stratum.

$\hat{y}_{ha} = N_{ha}\left[\bar{y}_{ha} + b_{ha}(\bar{X}_{ha} - \bar{x}_{ha})\right]$ where:

N_{ha} = The number of frame units in the stratum in the analysis district

$\bar{y}_{ha}$ = The mean of reported acres in the stratum in the analysis district

b_{ha} = The slope from the regression model of the stratum in the analysis district

$\bar{X}_{ha}$ = The population pixel mean in the stratum in the analysis district

$\bar{x}_{ha}$ = The sample pixel mean in the stratum in the analysis district

Analysis district estimator for $\hat{y}$ when combining strata with more than 2 segments per strata.

Strata should only be combined when the strata to be combined have the same characteristics ie. the same land use stratification and the same expected segment size. Strata can be combined in Iowa and with care in Missouri.

$\hat{y}_{ca} = \Sigma_{ha}\left[N_{ha}(\bar{y}_{ha} + b_{ca}(\bar{X}_{ca} - \bar{x}_{ha}))\right]$ where the differences from above are:

b_{ca} = The slope from the regression model using all segments in the combined strata in the analysis district

$\bar{X}_{ca}$ = The population pixel mean in the combined strata in the analysis district

The following changes must be made in the calculations for strata that are to be combined but have fewer than 2 segments.

$\left.\begin{matrix}\bar{y}_{ha}\\ \bar{x}_{ha}\end{matrix}\right]$ Must be the weighted mean of the strata that do have more than 2 segments. ie. weighted by frame units.

b_{ca} Should be the slope from the regression model for the combined strata. Strata with less than 2 segments should be excluded from the model for developing the slope.

Single Stratum Variance and R²

$$Var_h = N_h^2 * (1 - n_h/N_h) * \left[\Sigma(y_i - \bar{y})^2/(n - 2)\right] * (1 - R^2) * \left[1 + 1/(n_h - 3)\right]$$

$R^2 = (S2XY)^2/(S2Y*S2X)$ where:

$$S2XY = (\Sigma xy - n\overline{xy})/(n - 1)$$

$$S2Y = (\Sigma y^2 - n(\bar{y})^2)/(n - 1)$$

$$S2X = (\Sigma x^2 - n(\bar{x})^2)/(n - 1)$$

Combined Strata Variance and R²

When all combined strata have more than 2 segments.

$Var_{ca} = \Sigma_h\left[a_h * s_h^2 * (1 + 2/(n - k - 2))\right]$ where:

$$a_h = (N_h^2/n_h) * (1 - (n_h/N_h))$$

$$s_h^2 = \Sigma_i\left[(y_{hi} - \bar{y}_h) - b_c(x_{hi} - \bar{x}_h)\right]^2/(n_h - 1)$$

$$b_c = (\Sigma a_h * S2XY)/(\Sigma a_h * S2X)$$

$R^2 = \left[Var(DE) - Var(REG)\right]/Var(DE)$ where:

$$Var(REG)_c = \Sigma_h Var_{ch}$$

$Var(DE)_c = \left[\Sigma_{\#segs \geq 2} Var_h\right] * \left[1 + (\Sigma N_{h'}/\Sigma N_{h''})\right]^2$ where:

$N_{h'}$ = strata with # segs < 2
$N_{h''}$ = strata with # segs ≥ 2

Changes needed when some of the strata to be combined have less than 2 segments.

1 - Calculate the variance for all strata with # of segs > 2 including the adjustment factor $1 + 2/(n - k - 2)$
2 - Σ variances for strata with 2 or more segments
3 - The varaince for strata with less than 2 segments is calculated as follows:

$Var_{h'} = \left[2N_{h'}/\Sigma N_{h''} + (N_{h'}/\Sigma n_{h''})^2\right] * \Sigma Var_{h''}$ where:

h' = Strata with less than 2 segments
h'' = Strata with 2 or more segments
$N_{h'}$ = Frame units for strata h'
$N_{h''}$ = Frame units for strata h''

Regression County Estimate Procedure in PEDITOR

To determine the county level ŷ for a county using the state level regression estimate, the number of frame units in a county are multiplied times the adjusted county mean.

$$\hat{Y}_{h\ cty} = N_{h\ cty} * \bar{Y}_{h\ adj\ cty}$$

$\bar{Y}_{h\ adj\ cty} = \bar{Y}_1$ or $\bar{Y}_0$ or $\bar{Y}_\delta$ where:

$$\bar{Y}_1 = \bar{Y}_{cty} + b_{hAD}(\bar{X}_{cty} - \bar{x}_{cty})$$

$\bar{Y}_0 = b_{1hAD} * \bar{X}_{cty} - b_{0hAD}$ where:

b_{0hAD} = analysis district intercept by stratum
b_{1hAD} = analysis district slope by stratum

$$\bar{Y}_\delta = (1 - \delta)\bar{Y}_0 + \delta\bar{Y}_1$$

There are 5 rules use to determine the proper δ value to use:

1) never use $\delta = 1$
2) always use $\delta = 0$ when no county in the analysis district has more than 2 segments
3) use $\delta = \Gamma = 0$ when $\sigma^2_{between} = 0$
4) if $\sigma^2_{within} = 1.0$ then $\Gamma = 1$ and use $\delta = 0$
5) otherwise use $\delta = \Gamma$

where:

$\Gamma = \sigma^2_{between} / (\sigma^2_{between} + \sigma^2_{within}/n)$ where:

$\sigma^2_{between}$ = The variance between county means within an analysis district by stratum
σ^2_{within} = The variance of reported data within a county by stratum

If $\sigma^2_{between} = 0$ then $\bar{Y}_0$ is used

$\sigma^2_{within} = 0$ then $\bar{Y}_1$ is used

otherwise Y_δ is used.

The variance for the county level estimate is

$$Var_{h\ cty} = \sigma^2_{between} + \sigma^2_{within}$$

[1] From: MIKE BELLOW 3/16/94 4:00PM (7226 bytes: 1 ln, 1 fl)
To: MIKE CRAIG
Subject: New Formulas added to PEDITOR Estimation Programs
------------------------ Message Contents ------------------------

******** CAUTION use 3PDP-P2.PRN driver in XYWRITE **********

Text item 1: Text_1

The attached Xywrite document contains formulas for several large domain crop area estimators that have recently been programmed in PEDITOR. The separate regression estimator is the one that RSS has always used. The other estimators can be computed using special research programs known as ESTST and ESTLT, residing on Martin Ozga's disk but not in the main PEDITOR directory. The corresponding modules ESTS and ESTL still exist in the main directory and are unchanged (i.e., only the separate regression estimator is available). The alternative estimators will be studied and compared with the current one using the past three years' Delta data. The results will be presented at the ASA Conference in Toronto.

File item 2: LDEST.TXT 3/7/94 3:33P

FIVE CROP AREA ESTIMATORS

1. Raw Pixel Count Estimator

$$\hat{Y}^{(RPC)} = \lambda X$$

$$= \sum_{h=1}^{H} \lambda X_h$$

where:

λ = conversion factor (area units per pixel)

H = number of strata

X = number of pixels classified to crop in analysis district

X_h = number of pixels classified to crop in stratum h

2. Separate Ratio Estimator

$$\hat{Y}^{(SR)} = \sum_{h=1}^{H} [\bar{y}_{h.}/\bar{x}_{h.}]X_h$$

$$= \sum_{h=1}^{H} \hat{R}_h X_h$$

where:

$\bar{y}_{h.}$ = sample mean reported crop acreage in stratum h

$\bar{x}_{h.}$ = sample mean pixels classified to crop in stratum h

Variance Estimator -

$$v[\hat{Y}^{(SR)}] = \sum_{h=1}^{H} [N_h(N_h-n_h)/n_h][s_{yh}{}^2+\hat{R}_h{}^2 s_{xh}{}^2-2\hat{R}_h s_{xyh}]$$

where:

N_h = number of population units in stratum h

n_h = number of sample segments in stratum h

$$s_{xh}{}^2 = [1/(n_h-1)]\sum_{i=1}^{n_h} (x_{hi}-\bar{x}_{h.})^2$$

$$s_{yh}{}^2 = [1/(n_h-1)]\sum_{i=1}^{n_h} (y_{hi}-\bar{y}_{h.})^2$$

$$s_{xyh} = [1/(n_h-1)]\sum_{i=1}^{n_h} (x_{hi}-\bar{x}_{h.})(y_{hi}-\bar{y}_{h.})$$

y_{hi} = reported crop acreage in stratum h, sample segment i

x_{hi} = number of pixels classified to crop in stratum h, sample segment i

3. Combined Ratio Estimator

$$\hat{Y}^{(CR)} = [\sum_{h=1}^{H} N_h\bar{y}_{h.} / \sum_{h=1}^{H} N_h\bar{x}_{h.}]X$$

$$= \hat{R}X$$

Variance Estimator -

$$v[\hat{Y}^{(CR)}] = \sum_{h=1}^{H} [N_h(N_h-n_h)/n_h][s_{yh}{}^2+\hat{R}^2 s_{xh}{}^2-2\hat{R}s_{xyh}]$$

4. Separate Regression Estimator

$$\hat{Y}^{(SRG)} = \sum_{h=1}^{H} N_h[\bar{y}_{h.} + \hat{b}_h(\bar{X}_h-\bar{x}_{h.})]$$

where:

$$\hat{b}_h = s_{xyh}/s_{xh}{}^2$$

$\bar{X}_h$ = mean pixels per population unit classified to crop in stratum h

Variance Estimator -

$$v(\hat{Y}^{(SRG)}) = \sum_{h=1}^{H} [N_h(N_h-n_h)(n_h-1)/n_h(n_h-2)][s_{yh}{}^2-\hat{b}_h s_{xyh}]$$

5. Combined Regression Estimator

$$\hat{Y}^{(CRG)} = N[\bar{y}_{st} + \hat{b}(\bar{X} - \bar{x}_{st})]$$

where:

N = number of population units in analysis district

$$\bar{y}_{st} = \sum_{h=1}^{H} (N_h/N)\bar{y}_h$$

$$\bar{x}_{st} = \sum_{h=1}^{H} (N_h/N)\bar{x}_h$$

$$\hat{b} = [\sum_{h=1}^{H} A_h s_{xyh}]/[\sum_{h=1}^{H} A_h s_{xh}{}^2]$$

$$A_h = N_h(N_h-n_h)/n_h$$

$$\bar{X} = X/N$$

Variance Estimator -

$$v(\hat{Y}^{(CRG)}) = \sum_{h=1}^{H}[N_h(N_h-n_h)/n_h][s_{yh}{}^2+\hat{b}^2 s_{xh}{}^2-2\hat{b}s_{xyh}]$$

U.S. DEPARTMENT OF AGRICULTURE

REFERENCE SLIP

DATE 3/30/87

TO Mike Craig

□ ACTION

□ APPROVAL

☒ AS REQUESTED

□ FOR COMMENT

☒ FOR INFORMATION

□ INITIALS

□ NOTE AND FILE

□ NOTE AND RETURN

□ PER PHONE CALL

□ RECOMMENDATION

□ REPLY FOR SIGNATURE OF

□ RETURNED

□ SEE ME

□ YOUR SIGNATURE

REMARKS Mike,

Attached is the Cochran paper that is the source of the $\frac{1}{n-3}$ in the regression-estimator variance formula (See p. 201.)

FROM Richard Sigman

U.S.G.P.O. 1985-526-216

28

SAMPLING THEORY WHEN THE SAMPLING-UNITS ARE OF UNEQUAL SIZES*

By W. G. Cochran
Iowa State College

In sampling, the sampling-units are usually chosen so as to be similar in size and structure. With some types of population, however, it is convenient or necessary to use sampling-units that differ in size. Thus the farm is often the sampling-unit for collecting agricultural data, though farms in the same county may vary in land acreage from a few acres to over 1,000 acres. Similarly, when obtaining information about sales or prices, the sampling-unit may be a dealer or store, these ranging from small to large concerns.

In such cases the question arises: Should differences between the sizes of the sampling-units be ignored or taken into account in selecting the sample and in making estimates from the results of the sample? This paper contains a preliminary discussion of the problem, though further research is needed, many of the results given below being only large-sample approximations. It is convenient to consider first the problem of estimation, since it appears that the best method of distributing the sample depends on the process of estimation that is to be used.

THE PROBLEM OF ESTIMATION

To state the problem of estimation in mathematical terms, we assume that sampling units are drawn at random without regard to their sizes, and consider how to estimate the population total of some quantity y which can be measured on each sampling-unit. Associated with each sampling-unit is also a quantity x, which is called its *area* rather than its *size*, to avoid possible confusion between the terms "size of sample" and "size of sampling-unit." Some knowledge is assumed to be available about the values of x in the sample, and possibly also in the population.[1] In order to apply results from the statistical theory of estimation, it is also assumed that the number of sampling-units in the population may be considered infinite. Formulae applicable to the

* A paper presented at the 103rd Annual Meeting of the American Statistical Association in joint session with the Institute of Mathematical Statistics, New York, December 30, 1941.

Journal paper No. J989 of the Iowa Agricultural Experiment Station, Ames, Iowa. Project No. 611.

[1] For some populations an alternative method of specification may be more appropriate. For instance, each sampling may consist of an integral number of sub-units, as in the case of human populations where the sampling-unit is a household and the sub-unit is a single person. The specification may be made in terms of the value of y per sub-unit and the number n of sub-units per sampling-unit (cf. Hanson and Hurwitz, 1942). Since this approach would not apply in the examples given at the beginning of this paper, it will not be considered here.

Reproduced with permission from the *Journal of the American Statistical Association*, Vol. 37, pages 199–212. Copyright © 1942.

practical situation of sampling from finite populations can be obtained by adding suitable correction terms.

Stated in this way, the problem of estimation is a familiar one in mathematical statistics. If the joint frequency distribution of x and y in the population is known, the theory of estimation provides a routine technique leading to an efficient estimate of the population total of y and using to best advantage any available information about x. There are, however, difficulties in utilizing this method of approach. The joint frequency distribution is often known at best only vaguely from the available data, and may not appear to follow any of the few types of bivariate frequency distribution that have been studied. Further, there are strong administrative arguments for keeping the computations involved in making the estimate as simple as possible; these requirements may impose a bar on the use of estimates which, while highly efficient statistically, are rather difficult to compute.

Both difficulties can be met to some extent by restricting the estimates to those derived from the regression of y on x. For the calculation of regression equations, it is not necessary to describe completely the joint frequency distribution of x and y; we need only know how the mean value and the variance of y change as x changes. These can be examined from a graph or two-way table of the pairs of values of x and y constructed from any available data. If the form of the regression line and the relative weights assigned to different values of y are correct, the regression estimate is a *best unbiased linear estimate* as defined by David and Neyman (1938), though it is not a maximum likelihood estimate unless in addition the values of y are normally distributed within arrays in which x is fixed. The computations required for the simpler types of regression line are well-known and not unduly laborious.

ESTIMATES DERIVED FROM LINEAR REGRESSION

In the following sections it will be assumed that the quantity to be estimated is the population total of y; any formulae can easily be altered so as to refer to the estimation of the population mean per sampling-unit.

The simplest case occurs when the mean value of y is linearly related to the area of the sampling-unit, with constant variance; i.e. y is of the form $\alpha+\beta x+e$, where e has mean value zero and constant variance in arrays in which x is fixed. In this case, the linear regression estimate Y_l for the population total of y is

$$Y_l = N\{\bar{y}_s + b(\bar{x}_p - \bar{x}_s)\} \tag{1}$$

where N is the number of sampling-units in the population, b is the sample regression coefficient $S(y-\bar{y}_s)\,(x-\bar{x}_s)/S(x-\bar{x}_s)^2$, and the suffixes p and s refer to the population and sample respectively. It will be noted that this estimate requires a knowledge both of the total number N of sampling-units and of the mean value of x in the population.

In samples in which the x's remain fixed, the sampling variance of Y_l is

$$V(Y_l) = N^2\sigma_y^2(1 - \rho^2)\left\{\frac{1}{n} + \frac{(\bar{x}_p - \bar{x}_s)^2}{S(x - \bar{x}_s)^2}\right\} \tag{2}$$

n being the number of sampling-units in the sample and ρ the correlation coefficient between y and x. The distribution of Y_l tends to normality as n increases, being exactly normal for any size of sample if y is normally distributed for fixed x. A sample estimate of this variance is obtained by substituting for $\sigma_y^2\,(1-\rho^2)$ the mean square s_d^2 of deviations from the sample regression line.

For comparison with other estimates we may require the average variance of the regression estimate under random sampling. From (2), this clearly depends on the form of the frequency distribution of the areas. Since the areas are essentially positive, their distribution will not in general be normal, except perhaps as an approximation. The mean value of (2) may be expanded in a series of inverse powers of n, the sample size. Retaining the two leading terms, we obtain

$$\overline{V}(Y_l) = \frac{N^2\sigma_y^2(1 - \rho^2)}{n}\left\{1 + \frac{1}{n} + \frac{3 + 2\gamma_1^2}{n^2}\right\} \tag{3}$$

where γ_1 is Fisher's (1941) measure of relative skewness ($\gamma_1^2 = \kappa_3^2/\kappa_2^3$). If the areas were normally distributed, γ_1 would of course be zero, and the exact value for the term in curled brackets would be $(n-2)/(n-3)$, $= 1 + \frac{1}{n-3}$ which agrees with the value given above to this order of approximation. With large samples the factor is close to unity.

In many problems the true regression line must pass through the origin, as for example when y represents corn acreage and x farm acreage. Even in such cases, it may be advisable to use the preceding type of regression, if it appears on examination that a straight-line regression not passing through the origin will provide a satisfactory fit, whereas it would be necessary to use a curvilinear regression in order to include the origin. If a straight line through the origin can be used, y being of the form $(\beta\bar{x}+e)$, with constant residual variance, the regression estimate Y_o ($_o$ for origin) of the population total is

$$Y_o = N \frac{S(xy)}{S(x^2)} \bar{x}_p = \{\Sigma(x)\} \frac{S(xy)}{S(x^2)} \tag{4}$$

where $\Sigma(x)$ is the population total of the areas. The variance of Y_o is

$$V(Y_o) = \{\Sigma(x)\}^2 \sigma_y^2 (1 - \rho^2)/S(x^2). \tag{5}$$

The number of sampling-units in the population does not enter into either of these formulae, which require only the population total of the areas.

The expression for the average value of this variance, under repeated random sampling, is rather complicated. If the distribution of the areas is not far from normal, the leading terms give

$$\bar{V}(Y_o) = \frac{N^2 \sigma_y^2 (1 - \rho^2)}{n(1 + c_x)} \left\{ 1 + \frac{2c_x(2 + c_x)}{n(1 + c_x)^2} \right\} \tag{6}$$

$c_x = \sigma_x^2 / \bar{x}_p^2$ being the square of the coefficient of variation of x.

From formulae (3) and (6), we may compare the sampling errors of Y_l and Y_o with that of the estimate Y_s (s for sampling-unit) which is obtained by multiplying the sample mean per sampling-unit by the total number of sampling-units, and is commonly used where sampling-units are equal in size. Since the variance of Y_s is $N^2 \sigma_y^2 / n$, the ratios of the three pairs of variances in large samples are as follows:

$$\frac{\bar{V}(Y_l)}{V(Y_s)} = (1 - \rho^2); \frac{\bar{V}(Y_o)}{V(Y_s)} = \frac{(1 - \rho^2)}{(1 + c_x)}; \frac{\bar{V}(Y_o)}{\bar{V}(Y_l)} = \frac{1}{(1 + c_x)} \cdot \tag{7}$$

The additional factors involving $1/n$ and $1/n^2$ have been omitted from these expressions; they should be included in practical applications unless they are negligible. In large samples, both regression estimates are more accurate than the sample-mean estimate, the gain in accuracy being considerable if ρ is high. As would be expected, Y_o is more accurate than Y_l when the true regression line is straight and passes through the origin, the increase in accuracy depending on the coefficient of variation of the areas.

These results must be interpreted with care. They indicate that in large samples Y_l can never be less accurate, on the average, than Y_s. This statement was proved under the assumption that the true regression is linear (whether it passes through the origin or not); in the following section it will be shown to hold substantially even if the true regression is not linear. The conclusions about Y_o have a much more restricted validity, holding only if the true regression is linear and however, a loss of efficiency which remains fixed in large samples. While no exact small-sample theory has been reached, it appears that both the estimate itself and the estimated variance are biased in small samples.

WEIGHTED REGRESSIONS

Thus far we have considered the case in which only the mean value of y changes as x changes. The variance of y may also change, particularly so if there is considerable variation in the areas of the sampling-units. The theory of regression has been extended to meet this case, provided that the ratios of the variances of different values of y are known exactly, a condition which rarely if ever holds in problems of this type. If the true residual variance of y_i is σ_i^2, and $w_i = 1/\sigma_i^2$ the best unbiased linear estimate Y_{wl} is

$$Y_{wl} = N\{\bar{y}_w + b_w(\bar{x}_p - \bar{x}_w)\} \tag{12}$$

where $\bar{y}_w = S(w_i y_i)/S(w_i)$, $\bar{x}_w = S(w_i x_i)/S(w_i)$ are weighted sample means and $b_w = Sw_i(x_i - \bar{x}_w)(y_i - \bar{y}_w)/Sw_i(x_i - \bar{x}_w)^2$ is the weighted sample regression coefficient. For a fixed set of x's the sampling variance of Y_{wl} is

$$V(Y_{wl}) = N^2\left\{\frac{1}{S(w_i)} + \frac{(\bar{x}_p - \bar{x}_w)^2}{Sw_i(x_i - \bar{x}_w)^2}\right\}. \tag{13}$$

It will be noticed in (12) that Y_{wl} remains unchanged if instead of the correct weights w_i we use numbers $w_i' = \lambda w_i$ which are proportional to the weights; i.e. only the *relative* weights assigned to different values of y need be known in order to calculate Y_{wl}. Formula (13) for the sampling variance cannot be used however unless the actual values of the weights are known. If only relative weights w_i' are known, an unbiased sample estimate of (13) is given by

$$S^2(Y_{wl}) = N^2\frac{Sw_i'(y_i - Y_i)^2}{(n-2)}\left\{\frac{1}{S(w_i')} + \frac{(\bar{x}_p - \bar{x}_w)^2}{Sw_i'(x - \bar{x}_w)^2}\right\} \tag{14}$$

where $Sw_i'(y_i - Y_i)^2/(n-2)$ is the weighted mean square of deviations from the sample regression, using w_i' as weights.

In practiee, before these formulae can be used, it will be necessary to estimate the residual variances, and hence the weights, from the results of the sample and any other comparable data. Baker (1941) has recently discussed this problem for the case in which the x's fall into a number of distinct groups, all x's having the same value within each group. More generally, the x's will show a continuous range of variation. Space does not permit a detailed investigation of the best procedure for estimating the weights in this case. If the weights are presumed to change *continuously* as x changes, the first step seems clearly to subdivide the range of variation of x into a number of groups. The residual variance of y within each group can then be estimated by fitting an unweighted linear regression of y on x separately for each group. From these results, the relation between the residual variance and the area can be studied, and a smooth curve drawn to give the variance as a function of x. The weight to be assigned to any value of y is then obtained by noting the area of the sampling-unit, reading the curve, and taking the inverse of the variance.

The greater the number of groups, the more points are available for appraising the relation between variance and area. A further advantage of having many groups is that if the range of x is small within the groups, the within group correlation between y and x may be negligible, so that the total within-group mean square of y may be used as equivalent to the residual mean square, thus obviating the necessity of fitting a regression within each group. However, as the grouping is made finer, the number of observations within each group decreases, leading to less accurate estimates of the within-group variances. The optimum number of groups is not clear without further examination, though at a guess it seems advisable to have at least 20 observations in each group.

The estimated weights are, of course, subject to sampling errors. These errors have two consequences. The estimate Y_{wt} is not as accurate as it could have been made if the true weights had been known. This loss of accuracy is unavoidable, the somewhat laborious process described above for estimating the weights being an attempt to reduce the loss to a minimum. Secondly, and somewhat more seriously, both formulae (13) and (14) give biased estimates of the sampling variance of *Y_{wt} even in large samples,* i.e. even ignoring the correction terms of order $1/n$ which have appeared in previous formulae. If w_i' are the estimated and w_i the true weights, the correct sampling variance of Y_{wt} in large samples appears to be

$$N^2S\left(\frac{w_i'^2}{w_i}\right) \Bigg/ (Sw_i')^2. \qquad (15)$$

Substituting w_i' for w_i, formula (13) gives for large samples

$$N^2/S(w_i') \qquad (16)$$

while (14) gives, on the average

ESTIMATION BY USING POPULATION WEIGHTS

If a complete tabulation of the areas of all sampling-units in the population is available, the areas can be sub-divided into groups or strata, an estimate of the total of y being made for each stratum. While this procedure could be carried out with all the estimates previously discussed, this investigation will be confined to the simplest estimate Y_s. If $n_1, \ldots n_k$, $N_1, \ldots N_k$ are the numbers of sampling-units in the sample and population respectively for the k groups, the estimate Y_{gs} of the population total over all strata is

$$Y_{gs} = (N_1\bar{y}_1 + \cdots + N_k\bar{y}_k) \tag{23}$$

the sampling variance being

$$V(Y_{gs}) = \left(\frac{N_1{}^2\sigma_1{}^2}{n_1} + \frac{N_2{}^2\sigma_2{}^2}{n_2} + \cdots + \frac{N_k{}^2\sigma_k{}^2}{n_k}\right) \tag{24}$$

where $\sigma_1{}^2 \ldots \sigma_k{}^2$ are the within-strata variances of y.

As shown by Neyman (1934), this variance is smallest, for a fixed total size of sample, when the sample is distributed amongst the groups so that n_i is proportional to $N_i\sigma_i$. To retain comparability with previous estimates, however, we will assume that the sample is chosen at random.

In large samples, the average value of (24) works out approximately as

$$\bar{V}(Y_{gs}) = \frac{N^2}{n}\left\{\frac{(N_1\sigma_1{}^2 + \cdots + N_k\sigma_k{}^2)}{N}\left(1 - \frac{1}{n}\right) + \frac{k}{n}\,\frac{(\sigma_1{}^2 + \cdots + \sigma_k{}^2)}{k}\right\} \tag{25}$$

where n and N are as before the total numbers of sampling-units in the sample and population respectively. The expression inside the curled brackets contains both a weighted and an unweighted mean of the within-strata variances of y.

If all within-strata variances are the same, this reduces to

$$\bar{V}(Y_{gs}) = \frac{N^2\sigma^2}{n}\left(1 + \frac{k-1}{n}\right). \tag{26}$$

By increasing the number of x-groups with a given sample, the within-group variances are presumably decreased, since that portion of the variances of y which is due to variation in x is decreased by cutting down the range of x within each group. However, the factor involving k is of course increased as k increases, so that a point is reached beyond which a further increase in the number of groups will result in less accuracy. From (26) it follows that in the case of equal variances the factor involving k is relatively unimportant provided that $(k-1)/n$ is less than say .05, which holds if the average number of observations per group exceeds 20.

On comparing (26) with (3), Y_{gs} is found to be somewhat less accurate than the linear regression estimate Y_l if the true population regression is linear with equal variances. This follows because the within-stratum variance σ^2 cannot be less than $\sigma_y^2(1-\rho^2)$, while the additional factor in $1/n$ is also larger for Y_{gs} than for Y_l. This conclusion was to be expected, since under the conditions mentioned Y_l is a best unbiased linear estimate. If however the relation between y and x is markedly curvilinear or discontinuous, Y_{gs} may be superior to Y_l, since the variation in y arising from *any* type of relation with x can be reduced by a suitable choice of strata, whereas Y_l eliminates only the effects of the linear component of the relationship. Moreover, Y_{gs} is an unbiased estimate for any type of relation between y and x and any size of sample. Similarly, an unbiased estimate of the variance of Y_{gs} is always obtained by substituting the sample within-strata mean squares in (24).

Similar comparisons can be made between Y_{gs} and the weighted linear regression estimates by means of the formulae given for the sampling errors. Goldberg (1942) has discussed briefly the properties of Y_{ga}, the corresponding weighted estimate derived from the sample mean per unit area within each group.

FURTHER NOTES

Some apology is needed for presenting in the previous sections a number of large-sample approximations without guidance as to the limits within which these apply. Unfortunately these limits depend on the form of the joint frequency distribution of x and y, and could not be specified more definitely without a classification of the types of frequency distribution. Moreover, in extensive surveys, where problems of organization are difficult, biases may arise through the method of selecting the sample, incompleteness in the returns, and errors in reporting or recording the data. Such biases, while affecting the accuracy of the estimates, may not be measured by the formula for the sampling error, so that a rough approximation to the sampling error is often sufficient for practical purposes.

If the correct form of regression is used, population estimates derived from regressions remain unbiased in non-random sampling, provided that all sampling-units *with the same area* have an equal chance of selection. Thus the large sampling-units might be allotted a greater chance of inclusion in the sample, this procedure giving a more accurate estimate whenever the variance of y increases as x increases. On the other hand, if the method of selection discriminates in favor of certain sampling-units amongst those of the same area, bias may arise.

The formulae in this paper will of course apply to any variable x which is correlated with y. For example, in agricultural sampling, where the sampling-unit is sometimes a fixed area of land, x may be taken as the number of farms in the area, the total farm land or the total crop land, according to which gives the highest correlation with y.

In developing correction terms to be applied where an appreciable fraction of the population is sampled, the initial difficulty is that of defining a regression in a finite population. Writing $y=\alpha+\beta x+e$, we may suppose that e has no linear correlation with x in the finite population, but if we attempt to postulate that e is uncorrelated with any power of x, the number of conditions to be satisfied is greater than the number of values of e available, so that e and x cannot be independently distributed in the sense in which this term is applied with infinite populations. An alternative approach is to regard the finite population as a random sample from an infinite population in which e and x are independent. From a preliminary investigation and from Goldberg's (1942) work, it appears that the first approximation consists in multiplying formulae (2), (3) and (22) for the sampling-variances, and formulae (11) and (21) for the biases by $(N-n)/N$, this being the same correction as in the case of the sample-mean estimate Y_s. In formula (24), each term is multiplied by the corresponding factor $(N_i-n_i)/N_i$. For Y_o, the estimate derived from a straightline regression through the origin, and Y_{wl}, the weighted linear regression estimate, further investigation is needed. The difficulty arises because these two estimates do not equal the true population total when the sample consists of the whole finite population; i.e. they are *inconsistent* in the sense of Fisher (1941) whereas Y_s, Y_l, Y_a, and Y_{gl} are always *consistent.*

For sampling surveys in which the areas x of the sampling-units are unequal, the properties of various estimates of the population total of some observed quantity y are discussed, these estimates being mostly derived from the regression of y on x. In order of ease of calculation, the estimates are as follows; Y_s, derived from the sample mean per sampling-unit; Y_a, derived from the sample mean per unit area; Y_{ps}, a weighted form of Y_s, using population weights; Y_o and Y_l, based on unweighted linear regressions; and Y_{wo} and Y_{wl}, using weighted regressions. The conditions under which each estimate is most efficient are described, with various comparisons of their relative efficiencies.

REFERENCES

Baker (1941) this JOURNAL, 36, 500.

David and Neyman (1938) *Statistical Research Memoirs*, II, 105.

Fisher (1929) *Proceedings of the London Mathematical Society*, 30, 199.

Fisher (1941) *Statistical Methods for Research Workers*, Edinburgh, Oliver and Boyd, 8th ed.

Goldberg (1942) (Not yet published).

Hanson and Hurwitz (1942) this JOURNAL, 37, 89.

Neyman (1934) *Journal of the Royal Statistical Society*, 97, 588.

proposed ADDITION of KAPPA measure to % Correct (CORREC) 4/20/

changed, the smaller the within-class variations and the more separable are the two classes for a given vector. On the other hand, if the within-class variations are constant, the higher the inter-class deviation and again the more separable the two classes are. The separability for all vectors is obtained by summing up all the divisions between inter- and within-class deviations.

In order to examine the separability power of a given pixel window size, an average is taken from separabilities for all possible class pairs. This is calculated from

$$\mathrm{Sep}_l = \frac{1}{c\,(c-1)} \sum_{s=1}^{c-1} \sum_{t=s+1}^{c} \mathrm{sep}_l\,(s,\,t)$$

where Sep_l is the average separability for all the classes with a given pixel window size. By comparing the average separabilities from different window sizes, a pixel window size is selected which has the greatest average separability.

THE CLASSIFIER

The classifier used in this study is the minimum-distance classifier with the city-block metric (Gonzalez and Wintz, 1987). A city-block distance between two vectors is calculated by first obtaining a difference between every two corresponding vector elements, and then summing all the absolutes of these differences. There are two reasons for selecting the city-block distance. The first is that this distance is the simplest one in terms of computation, and therefore it could be used to handle occurrence frequencies extracted from an image with more gray-level vectors. Second, because we are comparing frequencies to make the classification decision, the use of Euclidian distance or other metrics is meaningless. In fact, some preliminary tests have been made in this study to compare the performances of the city-block metric and the Euclidian metric. Overall accuracies were on average 5 percent higher in favor of the city-block metric.

For given mean histograms of all c land-use classes, $h_u = (f_u(1), f_u(2), \ldots, f_u(N_v))$, $u = 1, 2, \ldots, c$, the city-block distance between a new histogram $h_l(i,j)$ and h_u is calculated from the following:

$$d_u = \sum_{v=0}^{N_v-1} \left| f_u\,(v) - f\,(i,\,j,\,v) \right|$$

The classifier compares all the c distances and assigns pixel (i,j) to the class which has a minimum distance to $h_l(i,j)$.

PERFORMANCE ASSESSMENT

To evaluate the performance of the classification methods, two criteria are often used: final classification accuracy and the time consumed during the classification process. There are two types of time included in the "time consumed during the classification process:" the hours of human labor and the CPU times used by computers. While it is difficult to estimate accurately and compare the time consumed by human labor because of the different skill levels of different image analysts, it is relatively easy to determine and compare the CPU times required by computers. In this study, the CPU time was used as an index for the "time consumed during the classification process."

The most commonly used accuracy-assessment method is test-sample checking. It requires three steps: determination of sample size and sampling strategy, sample identification (ground confirmation) to generate reference data, and comparison of the reference data with classification results to derive classification accuracies. The first two steps are described in the experimental design section. The third step is discussed below.

For a classified image (or a map), a confusion matrix (also called an error matrix or a contingency table) can be made by comparing the classification results with reference data. In this matrix, the reference data are represented by the columns of the matrix while the classified data are represented by the rows, or *vice versa*. The major diagonal of the confusion matrix indicates the agreement between these two data sets. The confusion matrix allows various accuracy indices to be derived.

In this research, the Kappa coefficient K (Cohen, 1960) and its estimated variance $\hat{V}$ (Fleiss *et al.*, 1969) were calculated for each confusion matrix to evaluate the overall agreement between the classification results and the reference data. The Kappa coefficient was used as an overall accuracy index for each classification. It has been recommended as a suitable accuracy measure in thematic classification for representing the whole confusion matrix (Fung and LeDrew, 1988; Rosenfield and Fitzpatrick-Lins, 1986; Congalton and Mead, 1983). It takes all the elements in the confusion matrix into consideration, rather than just the diagonal elements, which occurs with the calculation of overall classification accuracy. The variance was used when significance tests were made.

For an m by m confusion matrix, let p_{ij} be the proportion of subjects placed in the i, j^{th} cell; let p_{i+} and p_{+j} be the proportions of subjects placed in the i^{th} row and j^{th} column respectively. Then, with

$$p_o = \sum_{i=1}^{m} p_{ii} \text{ and } p_c = \sum_{i=1}^{m} p_{i+}\,p_{+i},$$

the Kappa coefficient $\hat{K}$ is defined by

$$\hat{K} = \frac{p_o - p_c}{1 - p_c}$$

where p_o and p_c indicate the proportion of units which agree, and the proportion of units for expected chance agreement, respectively. With the above definition, Fleiss *et al.* (1969) showed that the most appropriate method to estimate the variance of $\hat{K}$ is

$$\hat{V} = \frac{1}{N(1 - p_c)^4} \left\{ \sum_{i=1}^{m} p_{ii} \left[(1 - p_c) - (p_{i+} + p_{+i})\,(1 - p_o)\right]^2 + (1 - p_o)^2 \sum_{i=1}^{m}\sum_{j=1}^{m} p_{ij}\,(p_{i+} + p_{+j})^2 - (p_o p_c - 2\,p_c + p_o)^2 \right\}$$

To determine the difference between two $\hat{K}$ s, the significance test proposed by Cohen (1960) for comparing two classification results was adopted. With this method, the difference between two Kappa coefficients resulting from two classifications is first obtained. The square root of the sum of the variances Var between the two classifications is then calculated. A z-value can be determined by dividing the difference by the square root. A z-value greater than 2.58 indicates a significant improvement at the 0.99 probability confidence level.

In order to examine classification accuracies on a class-by-class basis, the conditional Kappa coefficient (Bishop *et al.*, 1975) was derived by comparing the classification results and the reference data. The conditional Kappa coefficient K_i is a class accuracy index which is derived from the agreement between the i^{th} row and i^{th} column in the confusion matrix for a particular land-use class i. The formula used to calculate the conditional Kappa coefficient is

$$K_i = \frac{p_{ii} - p_{i+}\,p_{+i}}{p_{i+} - p_{i+}\,p_{+i}}$$

Notations in this formula are the same as above.

EIGEN-BASED GRAY-LEVEL VECTOR REDUCTION

As explained in the above section, in order to make better use of the frequency-based classification technique, the number

PE+RS Vol. LVIII #4 April, 1992 pg. 425

University of Houston

Clear Lake

PROGRAMS IN SCIENCES

June 19, 1986

Mr. Richard Sigman
Remote Sensing Branch
National Agricultural Statistical Service
U.S. Department of Agriculture
Room 4833, South Building
14th and Independence Ave.
Washington, D.C. 20250

Dear Richard:

Attached is a report on variance estimation for the regression estimator in stratified sampling. We conducted certain simulations to evaluate the Cochran estimator and the other two estimators that are discussed in this report. We have recommended one of these two variance estimators for you to consider. This is very similar to the usual small sample variance estimator and hence, easy to implement. Let me know if you have any questions.

With respect to AI applications, you may be interested in the new book, Artificial Intelligence & Statistics, edited by William A. Gale and published by Addison-Wesley.

Sincerely yours,

Raj Chhikara

Raj S. Chhikara
Associate Professor of
Mathematical Sciences

REGRESSION ESTIMATES OF POPULATION MEAN IN STRATIFIED SAMPLING: (CASE OF COMBINED REGRESSION)

1. Variance Formulas:

For the h^{th} stratum, let N_h and n_h be the ~~stratum~~ Population and sample size, respectively, for stratum h and

$$E(y_h | x_h) = \alpha_h + \beta x_h$$

$$Var(y_h | x_h) = \sigma_h^2 \qquad , h = 1,2,\ldots,K.$$

For the samples, first compute the stratified sample means

$$\bar{y}_{st} = \Sigma W_h \bar{y}_h \qquad , \bar{x}_{st} = \Sigma W_h \bar{x}_h$$

where

$$W_h = N_h / \Sigma N_h .$$

Following Cochran (1977), consider the combined regression estimate of population mean $\bar{Y}$ given by

$$\hat{\hat{Y}} = \boxed{\bar{y}_{st}} + b_c \; (\bar{X}_? - \boxed{\bar{x}_{st}})$$

$$x \cdot y - \bar{x} \cdot y \cdot r$$

where

$$b_c = \Sigma \, a_h s_{xy} / \Sigma \, a_h s_x^2$$

with

$$\boxed{a_h} = (1 - f_h) W_h^2 / n_h \qquad a_h' = (1 - f_h) N_h^2 / n_h$$

and s_{xy} and s_x^2 are sample covariance and variance~~,~~ respectively. It then follows that the conditional variance of $\hat{\hat{Y}}$, given x's , is

$$Var(\hat{\hat{Y}}) = \Sigma \, a_h \sigma_h^2 + (\bar{X} - \bar{x}_{st})^2 \; Var(b_c) \tag{1}$$

where

$$Var(b_c) = [\, \Sigma a_h^2 s_{xh}^2 \; \sigma_h^2 / (n_h - 1)] / [\Sigma a_h s_{xh}^2]^2 . \tag{2}$$

Cochran ignores the second term in (1), assuming var (b_c) is negligible.

The unconditional variance of $\hat{\hat{Y}}$, obtained by taking expectation of the expression in (1), is

$$V(\hat{\hat{Y}}) = \Sigma \, a_h \, \sigma_h^2 + (\Sigma \, a_h S_{xh}^2) E[Var(b_c)] . \tag{3}$$

Let

$$L = \Sigma a_h^2 s_{xh}^2 \sigma_h^2/(n_h - 1)$$
$$M = \Sigma a_h s_{xh}^2$$

so that

$$\mathrm{Var}(b_c) = L/M^2.$$

By considering a Taylor series expansion of L/M^2 as a function of s_{xh}^2, it can be shown that up to the order of n^{-1}, the unconditional variance of b_c is

$$E[\mathrm{Var}(b_c)] = [L_0/M_0^2][1 + Q_1] \qquad (4)$$

where

$$L_0 = \Sigma a_h^2 S_{xh}^2 \sigma_h^2/(n_h - 1)$$
$$M_0 = \Sigma a_h S_{xh}^2$$

which are obtained by replacing s_x^2 by S_x^2 in L and M, and

$$Q_1 = [(2/M_0^2)\,\Sigma a_h^2 S_{xh}^4/(n_h-3)][3 - 2a_h^2 \sigma_h^2 M_0/(n_h-1)a_h L_0].$$

Although a substitution from (4) into (3) provides the unconditional variance to $O(n^{-2})$, it is an highly involved expression. Considering a further approximation of Q_1 under the assumption that the parameters are equal for all strata, one obtains

$$V(\hat{\hat{Y}}) = (\,\Sigma a_h \sigma_h^2)[1 + 1/(n-K-2)] \qquad (5)$$

where $n = \Sigma n_h$ and K is the number of strata.

In the special case of K=1, the variance in (5) reduces to the usual formula,

$$(1-f)(\sigma^2/n)[1 + 1/(n-3)]$$

as given in Cochran (1977).

2. Variance Estimates

An estimate of any of the variances given above in (3) and

(5) is obtained by replacing the unknown parameters by their estimates. Denote the estimates of variances in (3) and (5) by v_1 and v_2 obtained by replacing σ_h^2 by

$$s_h^2 = \Sigma[(y_{hi} - \bar{y}_h) - b_c(x_{hi} - \bar{x}_h)]^2/(n_h-1)$$

and S_x^2 by s_x^2. In addition, if only the leading term in these estimates is retained, one obtains the variance estimate given by

$$v_0(\hat{\bar{Y}}) = \Sigma\, a_h s_h^2. \tag{6}$$

This estimator is the same as in Cochran, p.203.

The stratum error variances σ_h^2, h = 1,2,..,K, are estimated with n_h-1 degrees of freedom. This overlooks the degree of freedom required to estimate β and thereby causes underestimation of $V(\hat{\bar{Y}})$. This underestimation can be compensated for by multiplying v_0, v_1 and v_2 by (n-K)/(n-K-1). (Note that n-K = $\Sigma(n_h-1)$.). The resulting variance estimates will be denoted by v_0', v_1' and v_2'. In particular, the variance estimator corresponding to (5) simplifies to

Recommended.

$$v_2'(\hat{\bar{Y}}) = (\Sigma\, a_h' s_h^2)[1 + 2/(n-K-2)]. \tag{7}$$

Σn_h # strata

3. Simulation Results

A population consisting of six strata was considered. Each stratum mean was taken to be 100. The other stratum parameters (size and error variance) and the sample sizes were considered equal as well as unequal. The following was the input data:

Case 1: (All input same)

N_h = 1,000 , n_h = 10, σ_h^2 = 100, h = 1,2,3,4,5,6.

Case 2: (Only stratum sizes different)

h	1	2	3	4	5	6
N_h	1500	1500	500	500	200	200

$n_h = 10$, $\sigma_h^2 = 100$

Case 3: (Stratum sizes and sample sizes different)

h	1	2	3	4	5	6
N_h	1500	1500	500	500	200	200
n_h	15	15	10	10	5	5

Case 4: (Unequal stratum parameters and sample sizes)

h	1	2	3	4	5	6
N_h	1500	1500	500	500	200	200
n_h	15	15	10	10	5	5
σh^2	150	120	100	100	80	80

Case 5: Input as in case 4 and different correlation coefficients for strata

For the correlation coeficient, four different cases were considered corresponding to $\rho^2 = .25, .50, .70$ and $.90$.

The attached table lists the simulation results obtained for the ratio of estimated variance to observed variance. Given are these ratios for the three variance estimators v_0', v_1', and v_2' as well as v_0. The simulation results are based on 1000 replications.

These results show that v_0 (and also, v_0') provides an underestimate of the variance in this example as much as 10 percent. The performance of the other two estimators is about the same. However, v_2' is much simpler to compute and hence, recommended over the other one.

Table: Ratio of the estimated to observed variance of the combined regression estimator

ρ^2	Case	Observed Variance	Variance Ratio			
			v_1'	v_2'	v_0'	v_0
.25	1	1.279	.962	.961	.943	.926
	2	1.906	1.032	1.004	.986	.967
	3	1.502	.938	.932	.914	.897
	4	1.896	.962	.953	.935	.918
	5	1.934	1.043	1.029	1.019	.991
.50	1	.901	.970	.970	.952	.934
	2	1.383	.954	.928	.910	.894
	3	.981	.984	.978	.959	.941
	4	1.268	.986	.977	.958	.940
	5	1.650	.998	.984	.966	.948
.70	1	.447	1.036	1.036	1.017	.997
	2	.794	1.000	.973	.955	.937
	3	.633	.942	.935	.918	.901
	4	.685	1.064	1.054	1.034	1.015
	5	1.114	1.000	.986	.968	.950
.90	1	.153	1.078	1.078	1.057	1.038
	2	.278	.968	.942	.924	.907
	3	.198	.942	.936	.918	.902
	4	.236	1.011	1.001	.982	.964
	5	.487	.960	.948	.930	.913

REGRESSION ESTIMATORS IN CROP SURVEYS:
CASE OF NON-NORMAL AND NONHOMOGENEOUS ERRORS

RAJ S. CHHIKARA

AND

JIM J. MCKEON

UNIVERSITY OF HOUSTON-CLEAR LAKE

HOUSTON, TEXAS

MAY 8, 1986

OBJECTIVES

1. INVESTIGATE THE ROBUSTNESS OF REGRESSION ESTIMATOR WHEN THE ERRORS HAVE

   - NON-NORMAL DISTRIBUTION

   - NONHOMOGENEOUS VARIANCES

2. INVESTIGATE THE VARIANCE ESTIMATION

   - VARIANCE FORMULA FOR LARGE SAMPLES

   - V_N (SMALL SAMPLE VARIANCE) AS A SUITABLE VARIANCE ESTIMATOR

VARIANCE ESTIMATORS FOR $\hat{X}$

- LARGE SAMPLE APPROXIMATION:

$$V_L = (1 - f)\, s_e^2/n$$

WHEN

$$s_e^2 = \Sigma(x_i - \hat{x}_i)^2/(n - 2).$$

- COCHRAN FORMULA:

$$V_C = (1 - f)(s_e^2/n)[1 + 1/(n-3) + 2g_1^2/n^2]$$

- NORMAL APPROXIMATION:

$$V_N = (1 - f)(s_e^2/n)[1 + 1/(n-3)]$$

- VARIANCE UNDER THE PRESENT MODEL:

$$V_O = (1 - f)(s_e^2/n)[1 + 1/(n-3) + \rho^2(1 - \rho^2)(K_X + K_e)/(n-1)]$$

K = kurtosis

10/31/85

TABLE : RATIO OF ESTIMATED TO ACTUAL VARIANCE OF $\hat{\hat{X}}_R$
(CASE OF NORMAL ERRORS)

		VARIANCE ESTIMATOR			
ρ^2	n	V_A	V_N	V_O	V_C
.25	4	.588	1.175	1.245	1.244
	10	.881	1.007	1.048	1.016
	25	1.015	1.061	1.073	1.062
.70	4	.504	1.007	1.074	1.073
	10	.832	0.951	1.012	0.962
	25	.878	.918	.943	.920
.90	4	.434	.868	.913	.930
	10	.766	.875	.907	.898
	25	.915	.957	.972	.960

TABLE : RATIO OF ESTIMATED TO ACTUAL VARIANCE OF $\hat{\hat{X}}_R$ (CASE OF NONNORMAL ERRORS $- \gamma_1 = 1.1$, $\gamma_2 = 1.1$)

ρ^2	n	VARIANCE ESTIMATOR			
		V_A	V_N	V_O	V_C
.25	4	.400	.800	.849	.851
	10	.799	.913	.998	.928
	25	.896	.937	.989	.940
.70	4	.504	1.009	1.076	1.069
	10	.799	.913	.998	.928
	25	.896	.937	.989	.940
.90	4	.345	.690	.724	.735
	10	.779	.890	.912	.905
	25	.858	.897	.896	.900

TABLE : RATIO OF ESTIMATED TO ACTUAL VARIANCE OF $\hat{\bar{X}}_R$ (CASE OF NONHOMOGENOUS BUT NORMAL ERRORS*)

		VARIANCE ESTIMATOR			
ρ^2	n	V_A	V_N	V_O	V_C
.25	4	.501	1.002	1.063	1.070
	10	.761	.870	.905	.887
	25	.936	.978	.987	.982
.70	4	.335	.670	.710	.720
	10	.776	.887	.938	.908
	25	.909	.950	.975	.955
.90	4	.387	.773	.814	.820
	10	.813	.929	.960	.950
	25	.943	.985	1.000	.990

* $\mathrm{Var}\ (e|x) \propto x$

CONCLUSIONS

ESTIMATORS ARE ROBUST WITH RESPECT TO DEPARTURE FROM NORMALITY OF THE ERROR DISTRIBUTION.

NONHOMOGENEITY OF ERROR VARIANCES AFFECTS SIGNIFICANTLY THE RELATIVE EFFICIENCY OF THE CLASSICAL ESTIMATORS.

THE EFFICIENCY OF THE REGRESSION ESTIMATOR IS REDUCED SLIGHTLY (AT MOST 15%) DUE TO NONHOMOGENEITY OF ERROR VARIANCES.

NO EFFECT ON BIAS (WHICH IS NEGLIGIBLE AND OR INSIGNIFICANT) DUE TO NONNORMALITY AND NONHOMOGENEITY.

VARIANCE ESTIMATOR V_N IS FAIRLY ROBUST WITH RESPECT TO

- LACK OF NORMALITY
- NONHOMOGENEITY OF ERROR VARIANCES

VARIANCE ESTIMATOR V_O IS OVERALL THE BEST THOUGH MARGINALLY

IN THE PRESENCE OF NONNORMALITY AND/OR NONHOMOGENEITY.

CONCLUSIONS

- BIAS IS INSIGNIFICANT AND/OR NEGLIGIBLE FOR ALL ESTIMATORS.
- REGRESSION ESTIMATOR ($\hat{\bar{X}}$) IS THE MOST EFFICIENT.
- RELATIVE EFFICIENCY OF $\hat{\bar{X}}$ DEPENDS UPON ρ AND SAMPLE SIZE n:
  - RE < 1 FOR $\rho^2 = .25$
  - $1 <$ RE < 2 FOR $\rho^2 = .70$
  - $2 <$ RE < 5.5 FOR $\rho^2 = .90$

  $n = 4$

  - $1 \leq$ RE < 1.5 FOR $\rho^2 = .25$
  - $2.5 \leq$ RE < 4 FOR $\rho^2 = .70$
  - $7.5 <$ RE < 11 FOR $\rho^2 = .90$

  $n = 10, 25$

- LARGE SAMPLE VARIANCE PROVIDES SUBSTANTIAL UNDERESTIMATION FOR $n = 4$
- OTHER VARIANCE ESTIMATORS PERFORM EQUALLY WELL:
  - SKEWNESS AND KURTOSIS HAVE NEGLIGLIBLE EFFECTS
  - VARIANCE BASED ON NORMAL APPROXIMATION (V_N) AND THE ALTERNATIVE VARIANCE ESTIMATOR (V_O) BASED ON THE APPROPRIATE MODEL HAVE SIMILAR PERFORMANCE

10/31/85