LINEAR REGRESSION ESTIMATORS IN SAMPLE SURVEYS UNDER CALIBRATION

James J. McKeon and Raj S. Chhikara,
University of Houston-Clear Lake

Thomas Boullion, University of Southwestern Louisiana

ABSTRACT

The problem of regression estimation in sample surveys is addressed in the context of calibration and three regression estimators of the finite population mean, called classical, modified classical and inverse regression are studied. It is shown that the linear (inverse) regression estimator is more efficient than the classical estimators. The bias and variance of the inverse regression estimator are analytically derived up to the order of 1/n, where n is the sample size. Four variance estimators for the inverse regression estimator are compared using simulations.

1. INTRODUCTION

Suppose a population consists of N units and for some characteristic of interest, let x_i represent the exact value for the $\underline{ith}$ unit and let y_i be an estimate of this value measured by some convenient but fallible device. For a random sample of n units, the pairs (x_i, y_i), i = 1,2,...n, are observed.

On the basis of the error structure, the model

$$E(y|x) = \alpha + \beta x,$$

or equivalently,

$$y = \alpha + \beta x + \varepsilon \qquad (1.1)$$

is the most plausible choice of models. The model error sum of squares is minimized by the classical estimator of the mean $\bar{X}$ of the x population given by

$$\hat{\bar{X}}_c = \bar{x} + (1/b)(\bar{Y} - \bar{y}) \qquad (1.2)$$

where $\bar{Y}$ is the population mean of the y population, $\bar{x}$ and $\bar{y}$ are the sample means and

$$b = s_{yx}/s_x^{\,2},$$

$$s_{yx} = \sum_1^n (Y_i - \bar{y})(x_i - \bar{x})/(n - 1)$$

$$s_x^{\,2} = \sum_1^n (x_i - \bar{x})^2/(n - 1)$$

It has been pointed out by Williams (1969) that the classical estimator has infinite variance. This serious defect can be removed by using a modified classical estimator proposed by Naszodi (1978). Letting

$$s^2 = \sum_1^n (y_i - \bar{y} - b(x_i - \bar{x}))^2/(n - 2)$$

and

$$q = s^2/b^2(n-1)s_x^{\,2},$$

it can be shown that

$$E[1/b(1 + q)] = 1/\beta + 0(1/n^2)$$

Thus, a modified estimator of $\bar{X}$ is obtained by

$$\hat{\bar{X}}_u = \bar{x} - (\bar{Y} - \bar{y})/b(1 + q). \qquad (1.3)$$

This estimator has finite variance and is unbiased to 0(1/n).

On the other hand, the expected estimation error sum of squares is minimized by the regression estimator of the mean,

$$\hat{\bar{X}} = \bar{x} + \hat{t}\,(\bar{Y} - \bar{y}), \qquad (1.4)$$

where $\hat{t} = s_{yx}/s_y^{\,2}$. This minimizing property does not assume that the regression of x on y is linear even though $\hat{t}$ is an estimate of the regression coefficient in the linear inverse model,

$$x = \gamma + \delta y + \xi , \qquad (1.5)$$

If (1.1) represents the true error structure and if the error distribution and the marginal distribution of x are both normal, then

$$E(x|y) = \gamma + \delta y,$$

but for a general x-distribution, the regression of x on y is non-linear, and var (x|y) may be a function of y.

Cochran (1973) has discussed the bias and variance of $\hat{\bar{X}}$ with and without assuming the linear model in (1.5), but he does not consider the model in (1.1) or discuss the classical estimator.

In the context of calibration theory, the two estimators corresponding to (1.2) and (1.4) have been extensively studied. Krutchkoff (1967) advocated the use of the linear inverse regression estimator whereas Berkson (1969) favored the classical estimator. The inverse estimator is biased toward the mean by an amount proportional to the distance from the mean. Since we are considering estimation of the mean only, all three estimators are unbiased when errors are normal.

Assuming normal errors, Shukla (1972) obtained expressions for the expected value and variance of the classical and the inverse estimators. Lwinn and Maritz (1982) extended these results to the case of non-normal error.

2. SIMULATIONS

Simulations were carried out to evaluate the bias and variance of the three estimators. In these simulations the x values were generated using a family of Beta-like random variables over the interval (0,1). The y values were obtained using the model in (1.1) with $\beta = 1$ and errors generated according to a normal distribution. The y values were truncated at 0 and 1, which produced some non-normality in the error distribution.

Populations were constructed with $N = 25, 100$ and 500 with means $\mu = .10, .33, .50, .67, .90$ for the x-distributions. Samples of size $n = 4$, 10 and 25 were drawn. The variance of the error was chosen to make

$\rho^2 = .25, .70$ and $.90$, where ρ is the correlation coefficient between x and y.

The estimates of $\bar{X}$ and other parameters were calculated and these values averaged over 500 replications. The ratio $f = n/N$ was required to be less than 1/3.

A comparison of relative efficiencies is given in Table 1. The relative efficiency is defined to be the ratio of the variance of sample mean to that of an estimator. The results clearly show

that the classical estimator $\hat{\hat{X}}_c$ is less efficient than the linear regression

estimator $\hat{\hat{X}}$; in fact it is highly inefficient when $n = 4$ since its variance is substantially larger than that of the sample mean. The unbiased

classical estimator $\hat{\hat{X}}_u$ compares well with the linear regression estimator

$\hat{\hat{X}}$ except when ρ is small.
Even though in Section 3 it is showed that, in theory, the inverse estimator is biased when the error distribution is skewed, the expected value of this bias was negligibly small and the inverse estimator showed less bias overall than the classical estimators.

Since the inverse regression

estimator of the mean, $\hat{\hat{X}}$ is clearly the best among the three considered, the remainder of this paper will be concerned only with this estimator. The

mean and variance of $\hat{\hat{X}}$ to $O(1/n)$ will be obtained in Section 3 for fixed x values and under sampling from a finite population. The variance estimator is compared with simulation results.

3. MEAN AND VARIANCE OF $\hat{\hat{X}}$.

It is easily seen that

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

$$= b^2 s_x^2 + (n-2)s^2/(n-1)$$

and

$$s_{xy} = b s_x^2. \qquad (3.1)$$

Thus, the regression coefficient in (1.3) can be rewritten as

$$\hat{t} = b/(b^2+h) \qquad (3.2)$$

where

$$h = (n-2)s^2/(n-1)s_x^2$$

For the given x_i's,

$$E(h) = (n-2)\sigma^2/(n-1)s_x^2,$$

$$Var(h) = [2+(\beta_2-3)]\sigma^4/(n-1)s_x^4 \qquad (3.3)$$

up to $O(1/n)$; where $\beta_2 = \mu_4/\sigma^4$ is the measure of kurtosis.

Expanding $\hat{t}$ in a Taylor series in b and h around β and E(h), and denoting small deviation $e = h - E(h)$, and $d = (b - \beta)$ as before, one gets

$$\hat{t} = t + t(1/\beta - 2t)d - (t^2/\beta)e + t^2 (4t-3)/\beta)d^2 + (t^3/\beta^2) e^2 - (t^2/\beta^2) (1 - 4t)de + \dots \qquad (3.4)$$

$$t = s_x^2/(\beta^2 s_x^2 + \sigma^2) \qquad (3.5)$$

Next, the mean and variance of t up to $O(1/n)$ are

$$E(\hat{t}) = t + t^2(4t - 3/\beta) \text{ var } (b) + (t^3/\beta^2) \text{ var } (h) \qquad (3.6)$$

and

$$Var(\hat{t}) = (t/\beta - 2t^2)^2 \text{ var } (b) + (t^4/\beta^2) \text{ var } (h) \qquad (3.7)$$

where $Var(b) = \sigma^2/(n-1)s_x^2$ and $Var(h)$ is given in (3.3).

To determine the bias and variance of

$\hat{\hat{X}}$, we first observe that

$$\hat{\hat{X}} - \bar{x} = (1-f)\hat{t}(\bar{Y}_1 - \bar{y}) \qquad (3.8)$$

where

$$\bar{Y}_1 = \sum_{1}^{N-n} y_i/(N-n)$$

$$= (N\bar{Y} - n\bar{y})/(N-n)$$

so that

$$E(\hat{\bar{X}} - \bar{x}) = (1 - f)[E(\hat{t})E(\bar{Y}_1 - \bar{y}) + \mathrm{Cov}(\hat{t}, \bar{Y}_1 - \bar{y})] \quad (3.9)$$

and

$$\mathrm{Var}(\hat{\bar{X}} - \bar{x}) = (1 - f)^2[E(\hat{t}^2)E(\bar{Y}_1 - \bar{y})^2 - (E(\hat{t}(\bar{Y}_1 - \bar{y})))^2 + \mathrm{Cov}\ (\hat{t}^2, (\bar{Y}_1 - \bar{y})^2)]. \quad (3.10)$$

The first covariance term in (3.9) can be evaluated using the theorem of Tan and Cheng (1981). Determination of the covariance given in the last term in (3.10) is quite involved. As shown in Chhikara and McKeon (1985), it follows that given sampled x_i's,

$$E((\hat{\bar{X}} - \bar{x})) = \beta(\bar{X} - \bar{x})E(t) + (1 - f)t^2\mu_3/\beta n s_x^2 \quad (3.11)$$

and

$$\mathrm{Var}(\hat{\bar{X}} - \bar{x}) = (1 - f)(\sigma^2/n)[(E(\hat{t}))^2 + \mathrm{Var}(\hat{t})] + \beta^2(\bar{X} - \bar{x})^2\mathrm{Var}\ (\hat{t}) + 2(1 - f)(\bar{X} - \bar{x})\ \mu_3 t^2(2t - E(\hat{t}))/ns_x^2 \quad (3.12)$$

with t, $E(\hat{t})$ and Var $(\hat{t})$ given in (3.5), (3.6) and (3.7) respectively.

By averaging with respect to the sampling distribution of the x_i's, the bias up to O(1/n) is

$$\mathrm{Bias}(\hat{\bar{X}}) = (1 - f)\ \mu_3 T/n(\sigma^2 + \beta^2 S_x^2) \quad (3.13)$$

where T is obtained by replacing s_x^2 by S_x^2 in t. Here S_x^2 is the finite population variance for the x values. Next, the variance can be obtained by using the standard result

$$\mathrm{Var}(\hat{\bar{X}}) = E_x(\mathrm{Var}(\hat{\bar{X}}|x_i\text{'s})) + \mathrm{Var}_x(E(\hat{\bar{X}}|x_i\text{'s})). \quad (3.14)$$

Below we provide the final expression for the variance; the full details of its derivation are given in Chhikara and McKeon (1985).

$$\mathrm{Var}(\hat{\bar{X}}) = [(1 - f)(1 - \rho^2)S_x^2/n]. [1 + 1/(n - 1) + \rho^2(1 - \rho^2)(K_x + K_e)/(n - 1)] \quad (3.15)$$

where K_x is the kurtosis (standardized fourth cumulant) of the x-distribution and K_e is the kurtosis of the error distribution.

4. VARIANCE ESTIMATORS

A variance estimator based on the expression in (3.15) which assumes the linearity of y on x in (1.1) is compared with three estimators of the variance of $\hat{\bar{X}}$ given by, Cochran (1973) who assumes that the inverse regression of x on y is linear.

First, the large sample estimator is given by

$$V_A = (1 - f)s_r^2/n \quad (4.1)$$

where s_r^2 is the residual mean square error,

$$s_r^2 = \sum_{i-1}^{n} (x_i - \hat{x}_i)^2/(n - 2) \quad (4.2)$$

with

$$\hat{x}_i = \bar{x} + \hat{t}(y_i - \bar{y}). \quad (4.3)$$

Another estimator is obtained using Equation (7.36) of Cochran (1973) which is of order $1/n^2$ and takes the skewness of the distribution of y into consideration. It is given by

$$V_C = (1 - f)(s_r^2/n)[1 + 1/(n - 3) + 2g_1^2/n^2] \quad (4.4)$$

where g_1 is the estimated relative skewness of the distribution of y. If the distribution of y is nearly normal, then this estimator is approximately

$$V_N = (1 - f)(s_r^2/n)[1 + 1/(n - 3)] \quad (4.5)$$

Let V_I be the variance estimator obtained by replacing ρ^2, K_x and Ke by their sample estimates. The term 1/(n-1) in (3.15) is replaced by 1/(n-3) to improve the small sample properties. This difference is of $O(1/n^2)$ and can be justified by noting that var($\hat{t}$) contains the factor $1/s_x^2$ and $E_x(1/s_x^2) = [(n - 1)/(n - 3)]\ S_x^2$.

We then have

$$V_I = [(1 - f)/n]s_r^2[1 + 1/(n - 3) + \hat{\rho}^2(1 - \hat{\rho}^2)(\hat{K}_x + \hat{K}_e)/(n - 1)] \quad (4.6)$$

The ratio of each of the estimates V_A, V_N. V_C and V_I to the actual variance from the simulations are given in Table 2 for N = 500 and n = 10. There is very little difference between V_N, V_C amd V_I indicating that the effects of the kurtosis of the x-distribution and the error distributions are slight. When estimating the mean it does not appear to matter significantly whether the linear regression of y on x or of x on y is assumed.

Table 1. Relative efficiencies of the estimators when N = 100 and n = 10

	Estimator		
μ	$\hat{\hat{X}}$	$\hat{\hat{X}}_u$	$\hat{\hat{X}}_c$
	$\rho^2 = .25$		
.10	1.00	.44	.01
.33	1.08	.49	.00
.50	1.05	.47	.00
	$\rho^2 = .70$		
.10	2.66	1.87	1.40
.33	2.83	1.92	.48
.50	2.82	1.86	.00
	$\rho^2 = .90$		
.10	9.31	7.46	6.69
.33	8.88	8.42	8.19
.50	9.66	8.82	3.38

Table 2. Ratios of the estimated variance to actual variance of $\hat{\hat{X}}$ when N = 100 and n = 10

	Variance Estimator			
μ	V_A	V_N	V_C	V_I
		$\rho^2 = .25$		
.10	.82	.93	.95	.94
.33	.89	1.02	1.03	1.02
.50	.81	.92	.93	.92
		$\rho^2 = .70$		
.10	.80	.92	.93	.92
.33	.90	1.03	1.04	1.03
.50	.94	1.07	1.08	1.07
		$\rho^2 = .90$		
.10	.92	1.05	1.06	1.05
.33	.79	.90	.91	.90
.50	.95	1.08	1.09	1.08

5. REFERENCES

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

Chhikara, R.S. and McKeon, J.J., "Regression Estimation in Sample Surveys". Technical Report, University of Houston-Clear Lake, Houston, Texas

Cochran, W.G. (1973), Sampling Techniques, 3rd Edition, John Wiley & Sons, New York.

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

Lwin, T. and Maritz, J.S. (1982), "An Analysis of the Linear-Calibration Controversy from the Perspective of Compound Estimation," Technometrics, 24, 235-242.

Naszodi, L.J. (1978), "Elimination of the Bias in the Course of Calibration," Technometrics, 20, 201-205.

Shukla, G.K. (1972), "On the Problem of Calibration," Technometrics. 14, 547-553.

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

Williams, E.J. (1969). A note on regression methods in calibration. Technometrics, 11, 189.