A COMPARISON OF DOUBLE SAMPLING REGRESSION ESTIMATORS

Dennis L. Clason

G. Morris Southward

A COMPARISON OF DOUBLE SAMPLING REGRESSION ESTIMATORS.

Dennis L. Clason
G. Morris Southward

New Mexico State University
Department of Experimental Statistics
Las Cruces, NM 88003

ABSTRACT

We investigate three alternative models for estimating the mean of a population using double sampling survey techniques. One estimator was found in the range science literature (Cook and Stubbendieck, 1986), another is the estimator presented by Cochran (1977). The third estimator uses method-of-moments estimators with measurement error regression models. Simulation studies suggest that the measurement error model does not work well when the slope is appreciably different from unity. Delta method variance estimators of the measurement error model may give negative variance estimates under these circumstances. The other estimators have better small sample performance (both are approximately unbiased, and have similar variances), but the two estimators have very different estimated variances under some circumstances.

1. INTRODUCTION

Double sampling can be viewed as a calibration problem in a sampling context. The essence of the idea is to substitute an inexpensive measurement for an expensive measurement. Expense may be measured in monetary units, in time, or in terms of resource consumption. It is usually assumed that the expensive measurement is either more precise or more accurate.

Several methods of using the inexpensive variable are available. Cochran (1977) presents stratified, ratio and regression estimators exploiting double sampling. Cook and Stubbendieck (1986) present a regression method in a widely used research methods handbook. It is notable that these two works do not agree on the variance of the estimator, although they agree on the choice of estimator.

Following Cochran's notation, we refer to the cheap measurement as X_i and the expensive measurement as Y_i. The double sampling method takes a sample of size n, observing both X and Y. A further sample of size n' > n is then taken, observing only X. The mean of the (larger) sample x' is adjusted using the regression of Y on X

$$\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x}). \tag{1}$$

Cochran gives the variance of this estimator as

$$V(\bar{Y}_{lr}) = \frac{(1-\rho^2)\sigma^2_Y}{n} + \frac{\rho^2\sigma^2_Y}{n'} - \frac{\sigma^2_Y}{N}. \tag{2}$$

Cook and Stubbendieck give the variance as:

$$V(\bar{Y}_{lr}) = s_{Y \cdot X}{}^2 \left[\frac{1}{n} + \frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] + \beta^2 \frac{s_X{}^2}{n'} \left(1 - \frac{n'}{N}\right). \quad (3)$$

Neither reference gives any rationale for preferring the regression of the accurate measurement on the inaccurate measurement. Reflection on the joint bivariate nature of the response under Cochran's model reveals the selection to be appropriate in that case. Also, Brown (1979) notes that this estimator is optimal in terms of integrated mean square error (IMSE), if the first two moments of the calibration sample match those of the population to be calibrated. This will clearly be the case when the calibration sample is a simple random sample of the population.

Cook and Stubbendieck recommend that cases be chosen "... as to provide a good (i.e. small variance) estimate of the population regression coefficient β" (pg 247). In some sense, then, either the X_i or the Y_i are fixed, and a bivariate model is no longer appropriate. This may explain the differing variance formulae, although we are not able to reproduce the formula from any reasonable assumptions. In fact, the formula appears to be the result of taking the variance of the adjusted mean assuming x' is fixed and adding the variance assuming that β is fixed.

Now, if Y is a fixed variable, it is reasonable to consider a direct estimator from the regression of X on Y, that is:

$$\bar{Y}'_{adj} = \frac{\bar{x}' - \hat{\alpha}}{\hat{\beta}}. \quad (4)$$

If the error distribution is normal, Y'_{adj} has no moments -- its sampling distribution is related to the Cauchy. This is unrealistic in a finite population case. An obvious way to estimate the variance of Y'_{adj} is to use a Taylor series expansion (aka delta method) and substitute moment estimators for parameters. The Taylor series variance for this case is

$$V\left(\frac{\bar{X} - \hat{\alpha}}{\hat{\beta}}\right) \approx \left(\frac{\mu_X - \alpha}{\beta}\right)^2 \left(\frac{\sigma^2_{\bar{X}'} + \sigma^2_{\hat{\alpha}}}{(\mu_X - \alpha)^2} + \frac{\sigma^2_{\hat{\beta}}}{\beta^2} - 2\frac{cov(\hat{\alpha},\hat{\beta})}{\beta(\mu_X - \alpha)}\right). \quad (5)$$

Fuller (1986) notes that if the regressor is measured with error, β is biased (shrunken toward 0). The attenuation of β can be corrected by using

$$\tilde{\beta} = \frac{S_{XY}}{S^2_Y - \sigma^2_U}, \quad \tilde{\alpha} = \bar{x} - \tilde{\beta}\bar{y}. \quad (6)$$

Here S_{XY} and S^2_Y are the covariance of X and Y and the variance of Y, respectively. σ^2_U is the variance of the measurement error in Y. Fuller's monograph provides a calibration estimator. That estimator is not used in this study. We are investigating the properties of the measurement error adjusted estimators in a naive solution of this calibration problem.

2. SIMULATION STUDY

Wylie (1991) performed ground truth surveys supporting a satellite remote sensing project in the central pastoral zone of Niger from 1986 to 1988. The surveys used various sampling designs and estimation strategies: of particular interest her are the 1987 and 1988 surveys. These years used randomly placed clusters of quadrats to assess measurement error and within site variation. Analysis of Wylie's data shows the biomass distribution to be heavily skewed.

To examine the merits of these estimators and their variance estimates we performed a simulation study, using Gauss v1.49B (Edlefson and Jones, 1987) with parameters based on Wylie's (1991) field work in the Sahel region of Niger. This work suggested that the slope is usually near 1, and certainly no smaller than about 1/3 nor greater than about 3. Each plot was generated using a gamma distribution for the plot biomass and normal errors were saved for use in generating clipped weights and ocularly estimated weights and measurement errors. Parameters used in generating the data were α=1, β=(0.3, 1, 3) and $\kappa = \frac{\sigma^2_U}{\sigma^2_{Y \bullet X}}$ = (0, 0.3, 1, 3). The parameter values were selected to approximately conform to Wylie's data with respect to the mean and variance. One thousand iterations were performed n=(16, 64) and n'=128. The results are summarized below.

3. RESULTS

Table 1 presents a summary of the simulation results. When there is no measurement error, the Cochran estimator is unbiased, as one expects from theory. The direct estimator is slightly biased for small β. The measurement error corrected estimator is disastrously biased when the slope is well away from unity. For most practical purposes there is little to choose between the Cochran and direct estimators in terms of variances.

Measurement error appears to have no appreciable effects on either the Cochran or the direct estimator themselves. Both remain approximately unbiased. There is a tendency for increasing measurement error variance (relative to the conditional variance $\sigma^2_{Y.X}$ to result in slightly larger standard errors for these estimators. Measurement error does not appear to result in appreciable bias for either of these estimators. With added measurement error, the measurement error estimator loses its disastrous bias, although it remains biased.

The slope of the regression relationship between the variables has a tremendous influence on the standard error of the estimators. Figure 1 shows the standard errors of the direct and inverse estimators for various measurement error ratios and slopes. Increasing slopes result in increased standard errors of the adjusted means. The relationship between slope and standard error of the adjusted mean (SEAM) is nonlinear. The relationship between slope and SEAM is reversed for the measurement error estimator -- large slopes result in small standard errors. Similar relationships between slopes and standard errors have been observed in simulation studies of ratio estimators (Royall and Cumberland, 1981).

The Taylor series variance estimator for the direct adjustment exhibits unacceptable behavior for large slopes. It is not bounded below by zero, and when the covariance between α and β is high, the variance estimator is negative. For lower slopes (0.3, 1), the estimator, although positive, is too small by an order of magnitude for small samples. The properties of the Taylor series variance estimator are better for the larger sample size. Measurement error seriously degrades its performance.

The Cook and Stubbendieck and Cochran variances (henceforth V_{CS} and V_C, respectively) both give reasonable approximations to the true variance *on the average* (i.e. in expectation). They are not functionally related, nor is the stochastic relationship a simple linear regression. The simulation results clearly show a tendency for V_{CS} and V_C to fill a region bounded by a right isoceles triangle in the first quadrant. So, when V_{CS} is large V_C is likely to be small, and vice versa. Increasing measurement error makes this trend rather fuzzy, but does not otherwise seem to affect the variance estimators.

4. CONCLUSIONS

There is little to choose between the direct adjustment and Cochran's adjustment in terms of small sample properties. Cochran's estimator is on sounder ground theoretically. This is true from the viewpoint of a Best Linear Unbiased (BLU) estimator of the conditional mean of Y given X, and from the viewpoint of the regression of X (the more variable quantity) on a fixed Y. The conditions are essentially filled to use Brown's (1979) result on the IMSE optimality of the inverse regression.

Simply trying to adjust out the attenuation of β using moment estimators is not fruitful. The resulting estimator is badly biased whenever the slope is removed from 1. The failure of the Taylor series approximation to the variance of the direct estimator does not bode well for a similar attempt on the measurement error adjusted estimator. In any event, measurement error has little influence on the properties of Cochran's estimator and the direct estimator.

At this time, the conditions leading to large and small values of V_{CS} and V_C have not been investigated, although this is high on our list of priorities. It is clearly important to know when these estimators are likely to overestimate and underestimate the true variance.

The relationship between slope and estimator sampling variances has implications for training purposes. It is very important that the field crews be trained and cross checked to insure that the slope of the regression be near 1. Results giving slopes well in excess of 1 have potentially deleterious effects on the sampling properties of the adjusted means.

Finally, if we were making recommendations for a ground truth survey in a semi-arid area we would:

1) Recommend the use of Cochran's sampling procedure. The procedure performs well in simulation and is relatively unaffected by measurement errors. It also has the best theoretical support;

2) Recommend the use of Cochran's variance estimator, pending characterization of the conditions leading to disagreement between V_{CS} and V_C.

5. LITERATURE CITED

Brown, G.H. **1979** An optimization criterion for linear inverse estimation. Technometrics **21**:575-578.

Cochran, W.G. **1977** *Sampling Techniques, 3rd edition*. Wiley, New York.

Cook, C.W. and Stubbendieck, J. **1986** *Range Research: Basic problems and Techniques*. Society for Range Management, Denver, CO.

Edlefson, L.E. and Jones, S.D. **1986** *GAUSS Programming Language Manual*. Aptech Systems, Inc Kent, WA 98064.

Fuller, W.A. **1987** *Measurement Error Models*. Wiley, New York.

Mood, A.M.M., Graybill, F.A., Boes, D.C. **1974** *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

Royall, R.M. and Cumberland, W.G. **1981** An empirical study of the ratio estimator and estimators of its variance. Journal of the American Statistical Association **76**:66-77.

Wylie, B. **1991** *Herbaceous biomass assessment in the central pastoral zone of the Sahelian country of Niger, 1968-1988*. Unpublished New Mexico State University Doctoral Dissertation, University Microfilms, Ann Arbor MI.

Table 1. Estimated means and variances of the Cochran, direct, and measurement error corrected regression estimators.

n=16, n′=128

		Slope 0.3		Slope 1		Slope 3	
Kappa		Mean	Variance	Mean	Variance	Mean	Variance
0	Cochran	-0.0054	0.2361	0.0116	1.7133	-0.0829	14.1820
	Direct	0.0819	0.3616	0.0391	1.7640	-0.0741	14.2276
	M. Error	23.1996	15.6648	-1.9985	1.5193	-27.9232	0.1719
0.3	Cochran	-0.0065	0.2134	-0.0507	1.6624	0.1429	13.0255
	Direct	0.0807	0.3233	-0.0181	1.6890	0.0430	13.1267
	M. Error	23.2499	15.7784	-2.0261	1.5258	-27.9117	0.1683
1	Cochran	0.2822	0.2114	-0.0775	1.6514	-0.0204	14.1949
	Direct	0.1108	0.3043	-0.0327	1.7194	0.0548	14.3165
	M. Error	23.5864	16.2897	-2.0374	1.4916	-27.9055	0.1820
3	Cochran	0.0136	0.2426	-0.0622	1.8473	-0.2194	15.3204
	Direct	0.1122	0.3964	0.0280	1.9529	-0.0170	15.8447
	M. Error	23.5165	16.6271	-2.0027	1.6187	-27.4498	0.2023

n=64, n′=128

		Slope 0.3		Slope 1		Slope 3	
Kappa		Mean	Variance	Mean	Variance	Mean	Variance
0	Cochran	-0.0036	0.1528	0.0195	1.7062	0.0576	13.3672
	Direct	0.0097	0.1728	0.0231	1.7279	0.0588	13.3869
	M. Error	23.7681	14.0424	-1.9486	1.5486	-27.9122	0.1615
0.3	Cochran	0.0157	0.1535	0.0369	1.4834	-0.0512	12.9217
	Direct	0.0276	0.1760	0.0417	1.5065	-0.0471	12.9840
	M. Error	24.0114	14.7210	-1.9298	1.3848	-27.9263	0.1570
1	Cochran	0.0121	0.1664	0.0254	1.6483	-0.0333	14.2019
	Direct	0.0260	0.1918	0.0319	1.6889	-0.0226	14.3885
	M. Error	23.8801	15.9760	-1.9485	1.5684	-27.9154	0.1720
3	Cochran	-0.0028	0.5314	-0.0130	1.5797	-0.0107	14.4789
	Direct	0.0118	0.1784	0.0035	1.6526	0.0256	15.0449
	M. Error	23.8559	14.1316	-1.9617	1.5001	-27.9188	0.1856

Figure 1. Estimated standard errors of the Cochran (inverse) and direct estimators. The measurement error corrected estimator is omitted to clarify details about the two useful estimators.

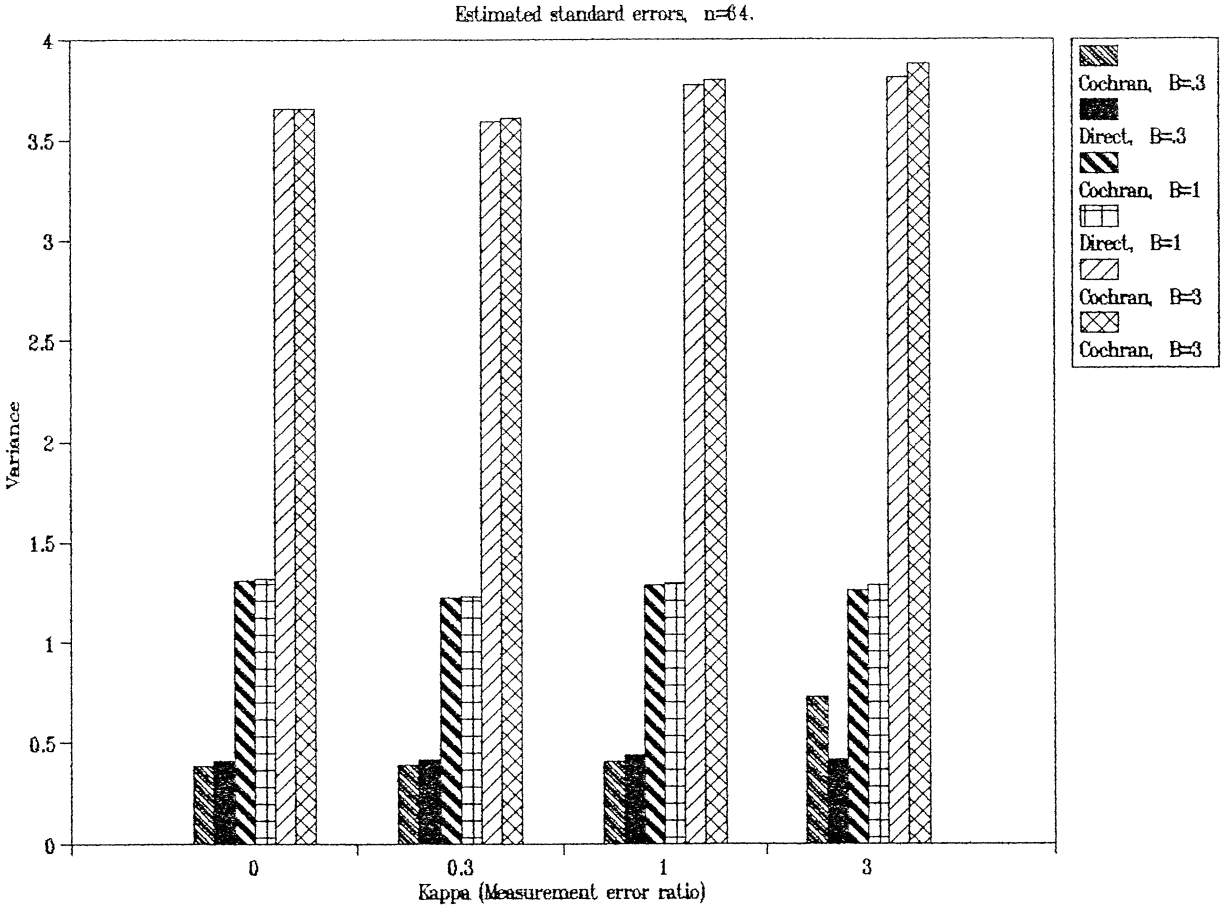

Figure 2. Estimated biases of the Cochran (inverse) and direct estimators.