The Philippine Statistician
Volume XVI, Nos. 3-4, 1967

RATIO AND REGRESSION-TYPE ESTIMATORS IN DOMAIN ESTIMATION

By

JOSE S. GUTIERREZ[1]

1. **Introduction.** For many estimation problems in sample surveys, it is sufficient to devise from the sample unbiased estimators of the population parameters. Biased estimators, however, are acceptable provided they are consistent and in comparison with the available unbiased estimators may be proved to be more precise. As a working rule, Cochran[1] stated that the effect of bias on the accuracy of an estimate is negigible if the bias is less than 0.1 of the standard deviation of the estimate.

A class of relatively simple estimators can be devised as linear functions of the sample elements and involves no stronger assumption than finiteness of the first and second-order moments of the components of the sample[6]. The tendency, however, in sample survey theory of estimation is toward the utilization of estimating techniques which are independent of the form of the distributions under study. This is due to the following reasons: only vague knowledge of the actual distribution is usually available and the sample sizes are quite often adequate for statement based on limiting distributions[5].

In finite sampling theory, information collected on a concomitant variate is often used to create more precise estimators of population parameters. A general class of estimators designed to utilize this supplementary information includes ratio and regression estimators, although the validity of a

[1] Assistant Professor of Statistics and Economics, Statistical Center, University of the Philippines.

regression model base on maximum likelihood principles and on the assumption of linear relations will be often in doubt, the use of regression estimation may still result in gain of precision. Nevertheless, the theory is still inappropriate and requires considerable development before it can be applied to finite population. Besides, there are also the problems of constant residual variance about the regression line and the assumption of infinite population[4]. An approach which does not demand that the regression in the population be linear has been discussed by Cochran[1] but the results hold only in large samples.

In analytic studies, the question arises as to which estimators should be used in estimating the domain means when sample surveys provide data for concomitant variables. Ratio estimators are easily adopted to the estimation of domain means. These estimators, however, are likely to be effective only if the scales of both variables can be so chosen that the population line will intersect near the origin[3]. This paper will present two ratio and regression-type estimators formulated on the basis of the logic that the probability of the number falling in each sub-group is a linear function of the independent variable.

2. **Development of General Regression Procedures for the Estimation of the Domain Mean.** Suppose we draw a sample of n elements from a population of N elements by simple random sampling without replacement. However, we don't know the number N_j of elements in each domain. The best estimate we can use for this domain number is n_j/n. In other words, we can only identify an element to which domain it belongs after the element was drawn. Let us consider the following variables:

$$w_{ji} = \begin{cases} y_i, \text{ if the } i^{th} \text{ unit is in the } j^{th} \text{ domain} \\ 0, \text{ otherwise} \end{cases}$$

$$z_{ji} = \begin{cases} x_i, \text{ if the } i^{th} \text{ unit is in the } j^{th} \text{ domain} \\ 0, \text{ otherwise} \end{cases}$$

$$c_{ji} = \begin{cases} 1, \text{ if the } i^{th} \text{ unit is in the } j^{th} \text{ domain} \\ 0, \text{ otherwise} \end{cases}$$

This means if the i^{th} sample element is in the j^{th} domain, the w_{ji} takes a value y_i; the z_{ij}, x_i and the count variable c_{ij}, 1. Consider the simplest form of a ratio-type estimator, as follows:

$$\hat{\bar{y}}_{j\cdot R} = \frac{\bar{w}_{j\cdot}}{c_{j\cdot}}, \quad j = 1, 2, \ldots R$$

where $\hat{\bar{y}}_{j\cdot R}$ is an estimate of the mean of the dependent variable Y_i for j^{th} domain,

$$\bar{w}_{j\cdot} = \sum_{i=1}^{n} w_{ji}/n \text{ and } c_{j\cdot} = \sum_{i=1}^{n} c_{ji}/n. \quad \text{Unike in}$$

the case of a simple random sample estimation of the mean $\bar{y} = \Sigma\, x_i/n$ where n is given, in a domain estimation the denominator is a random variable (i. e. $\bar{c}_{j\cdot}$).

The approximate variance of this estimator following a general procedure[2] is given by

$$\text{Var } (\hat{\bar{y}}_{j\cdot R}) = \frac{E^2(\bar{w}_{j\cdot})}{E^2(\bar{c}_{j\cdot})} \left[\frac{\text{Var } \bar{w}_{j\cdot}}{E^2(\bar{w}_{j\cdot})} + \frac{\text{Var } \bar{c}_{j\cdot}}{E^2(\bar{c}_{j\cdot})} - \frac{2\,\text{Cov}\,(\bar{w}_{j\cdot}, \bar{c}_{j\cdot})}{E(\bar{w}_{j\cdot})\,E(\bar{c}_{j\cdot})} \right]$$

The bias of this estimator is given by

$$\text{Bias in } \bar{y}_{j\cdot R} = -\frac{\text{Cov}\,(\bar{y}_{j\cdot R}, \bar{c}_{j\cdot})}{E(\bar{c}_{j\cdot})}$$

The proportional bias is

$$\frac{(\text{Bias in } \bar{y}_{j\cdot R})}{[\text{Var } \bar{y}_{j\cdot R}]^{1/2}} \leq \frac{\frac{1\text{-f}}{n}\,\frac{N_j}{N\text{-1}}\,(1 - \frac{N_j^{1/2}}{N})}{E(\bar{c}_{j\cdot})}$$

(when $f = \frac{n}{N}$) which takes a decreasing value with an increasing n.

Suppose we consider the ratio of two linear functions as follows:

$$\bar{y}_{j\cdot L} = \frac{\bar{w}_{j\cdot} - b_{wj}\,(\bar{x} - \bar{X})}{\bar{c}_{j\cdot} - b_{cj}\,(\bar{x} - \bar{X})} = \frac{\bar{w}_{j\cdot L}}{\bar{c}_{j\cdot L}}\,.$$

where

$$b_{wj} = \sum_{i=1}^{n} (w_{ji} - \bar{w}_j)\,(x_i - \bar{x}) / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$b_{cj} = \sum_{i=1}^{n} (c_{ji} - \bar{c}_{j\cdot})\,(x_i - \bar{x}) / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

that is b_{wj} and b_{cj} are estimators of the population parameters B_{wj} and B_{cj}, respectively.

Following the same procedure as before, the approximate variance formula is given by

$$\text{Var } y_{j\cdot L} = \frac{E^2(\bar{w}_{j\cdot L})}{E^2(\bar{c}_{j\cdot L})} \left[\frac{E^2(\bar{w}_{j\cdot L})}{E^2\,(\bar{w}_{j\cdot L})} + \right.$$

$$\left. \frac{\text{Var } \bar{c}_{j\cdot L}}{E^2(\bar{c}_{j\cdot L})} - \frac{2\,\text{Cov}(\bar{w}_{j\cdot L}, \bar{c}_{j\cdot L})}{E(\bar{w}_{j\cdot L})\,E(\bar{c}_{j\cdot L})} \right]$$

This bias of this estimator $\bar{y}_{j \cdot L}$ is

$$\text{Bias in } \bar{y}_{j \cdot L} = \frac{\text{Cov } (\bar{y}_{j \cdot L}, \bar{c}_{j \cdot L})}{E(\bar{c}_{j \cdot L})}$$

and with a proportional bias of

$$\frac{|\text{ Bias in } \bar{y}_{j \cdot L}|}{(\text{Var } \bar{y}_{j \cdot L})^{1/2}} \leqq \frac{[\frac{1-f}{n} \text{ Var } (\bar{c}_{j \cdot}) - B^2_{cj} \text{ Var } (\bar{x})]^2}{E(\bar{c}_{j \cdot L})}$$

For illustrative applications of the estimators developed above the following were considered

Domains

Tenant operated farms

Livestock farms

100 — 219 hectare farms

Independent variable

Total farm area, x_1

Total sales, x_4

Hog sales, x_2

Crop sales, x_3

Dependent variable

Net farm income, Y

The results of the application are given in the following table:

Domain/ Estimator	Independent Variable	$\bar{y}_{j.}$	Var $\bar{y}_{j.}$	Cov ($\bar{y}_{j.}$ $\bar{c}_{j.}$)
Tenant				
$\bar{y}_{j.R}$	—	2,155.92	298,398	— 17.46
$\bar{y}_{j.L}$	x_1	2,139.99	211,448	— 17.76
	x_2	2,416.54	391,828	— 25.62
	x_3	2,205.74	332,225	— 15.88
	x_4	2,197.17	311,539	— 41.24
Livestock				
$\bar{y}_{j.R}$	—	2,722.96	295,094	— 7.90
$\bar{y}_{j.L}$	x_1	2,743.42	295,862	— 6.58
	x_2	3,073.26	337,219	— 4.92
	x_3	2,644.82	354,722	— 2.45
	x_4	2,740.32	279,372	— 4.63
100 - 219				
$\bar{y}_{j.R}$	—	2,036.36	215,217	— 1.65
$\bar{y}_{j.L}$	x_1	2,043.86	116,560	— 6.31
	x_2	2,244.14	158,883	— 22.21
	x_3	2,087.70	148,217	— 19.21
	x_4	2,066.93	152,971	— 21.29

The above procedures are limited to the availability of the population mean of the independent variables. However, an extension of the study to the use of suitable estimators of the population mean indicate the feasibility of the use of the aforementioned estimators when the population means are unknown. This aspect of the study will be illustrated using the following tenant domain, total farm area and using the following cases:

domain means, domain totals and domain numbers are known and using the estimator $\bar{Y}_{j \cdot L}$.

The estimation of the population means is done as follows:

Case Situation	Estimate of Population Mean
Domain mean known	Domain mean
Domain total known	$(J/N)x_j$ (where J is the total number of domains)
Domain number known	$JN_j - / n$ (where $\bar{x}$ is the sample value)

(This case seldom exists)

The results of the application of the above procedures are as follows:

Case Situation	$\bar{y}_{j.L}$	$\text{Var}\,\bar{y}_{j.L}$	$\text{Cov}(\bar{y}_{j.L}\, c_j)$
Domain mean known	1,948.02	133,181	— 10.13
Domain total known	1,535.71	92,192	— 0.77
Domain Number known	2,174.27	288,320	— 17.61

3. **Summary and Conclusion.** The development of ratio and regression-type estimators appropriate for the estimation of domain means was presented. Two models were presented, a ratio-type estimator and a ratio-regression type estimator (other types of the ratio regression type estimators are presented i nthe unpublished Ph.D. thesis of the writer). The use of the population means of the independent variables and the use of estimates of these means were also examined.

4. **Acknowledgement.** This article is part of the dissertation submitted by the author in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Iowa State University.

REFERENCES

[1]. COCHRAN, William G. *Sampling Techniques.* 2nd Edition. New York: John Wiley and Sons, 1963.

[2]. GUTIERREZ, Jose S. *Regression Analysis of Cross-Section Survey Data for Planning and Evaluation of Economic Development Programs.* Unpublished Ph.D. Dissertation. Ames: Iowa Library, Iowa State University of Science and Technology, 1966.

[3]. HARTLEY, H. O. *Advanced Survey Designs.* Mimeographed Notes. Ames: Iowa Statistical Laboratory, Iowa State University of Science and Technology, 1958.

[4]. HARTLEY, H. O. *Analytic Studies of Survey Data.* Reprint Series No. 63. Ames: Iowa Statistical Laboratory, Iowa State University of Science and Technology, 1959.

[5]. HOWITZ, Daniel. *Ratio Method of Estimation in Sample Surveys.* Unpublished Ph.D. Dissertation. Ames: Iowa Library, Iowa State University of Science and Technology, 1953.

[6]. WILKS, S. S. *Mathematical Statistics.* New York: John Wiley and Sons, 1962.