REVIEW OF REGRESSION ESTIMATORS

By

W. Edwards Deming and Morris H. Hansen

Reprinted from

DIE STATISTIK IN DER WIRTSCHAFTSFORSCHUNG

Festgabe für Professor Rolf Wagenführ

zum 60. Geburtstag

DUNCKER & HUMBLOT BERLIN

Review of Regression Estimators

By W. Edwards Deming and Morris H. Hansen

Reason to Use Regression Estimators

There are a number of ways to use supplementary information concerning the frame, for calculation of an estimate of some characteristic of the frame. For example, the estimate $X = N\bar{x}$ of the total population of a frame of N sampling units makes use of knowledge of N. The ratio-estimate $X' = Bf$ also makes use of supplementary information concerning the frame. B is here the total y-population in the frame, already known, and f is the ratio $\bar{x} : \bar{y}$ derived from the sample. Stratified sampling is another way to make use of supplementary information concerning the frame.

Still other ways are suggested by a study of regression estimators. We shall introduce here a general form of estimator that includes $X = N\bar{x}$ and $X' = Bf$ as special cases, but which leads also to other forms of estimators, sometimes highly useful. It is important to note at the outset that any of the estimators that take advantage of supplementary information may have considerable advantage over the simple estimate $\bar{x}$ if the correlation ρ between x_i and y_i is high, but that this condition is not always sufficient.

We assume simple random sampling with replacement, and write the regression estimator in the form

$$\bar{x}_i = \bar{x} + m_i(b - \bar{y}) \tag{1}$$

wherein $b = E\bar{y}$ is the y-population per sampling unit, known independently from some source such as the Census. Later, we shall note the circumstance in which b is estimated from a larger sample, or from an independent sample, with a variance to reckon with. We consider here 4 cases ($i = 1, 2, 3, 4$), from amongst a number of possible alternatives, taken largely from Hansen, Hurwitz, and Madow (Wiley, 1953).

Preliminary Remarks

The estimate $\bar{x}_1 = \bar{x}$ is unbiased. We include the estimator $\bar{x}_1$ on our list here for comparison, because it yields the variance $\sigma_{\bar{x}}^2$ which appears in the other variances to follow.

The estimator $\bar{x}_2$ is also unbiased for any choice of the constant m_2. Estimators $\bar{x}_3$ and $\bar{x}_4$ are subject to mathematical biases that are usually too trivial to cause concern, but which may under certain circumstances be troublesome when the number n of sampling units is only 1, 2, 3, or some other small number. For a discussion of these biases see Cochran *Sampling Techniques* (Wiley, 1950, 2d ed., 1963) Chapter 7.

Table 1

Some possible regression estimators

Case i	m_i	Equation	Remarks
1	0	$\bar{x}_1 = \bar{x}$	This choice of m_i makes no use of supplementary information.
2	m_2	$\bar{x}_2 = \bar{x} + m_2 (b - \bar{y})$	Here m_2 is any constant not derived from the sample under consideration. This estimator is sometimes called the difference estimator.
3	$m_3 = \dfrac{\hat{\sigma}_{xy}}{\hat{\sigma}_y^2} = \hat{\varrho}\dfrac{\hat{\sigma}_x}{\hat{\sigma}_y}$	$\bar{x}_3 = \bar{x} + m_3 (b - \bar{y})$	This is the so-called least-squares regression estimator.
4	$m_4 = \bar{x}/\bar{y} = f$	$\bar{x}_4 = f\, b$	This is the ratio-estimator.

Variances (assuming N large relative to n)

(2) Estimator 1 $\quad \text{Var}\ \bar{x}_1 = \sigma_{\bar{x}}^2$; likewise $\text{Var}\ \bar{y}_1 = \sigma_{\bar{y}}^2$.

Estimator 2 $\quad \text{Var}\ \bar{x}_2 = \sigma_{\bar{x}}^2 (1 - \varrho^2) + \sigma_{\bar{y}}^2 (m_2 - \beta)^2$

(3) $\qquad = \sigma_{\bar{x}}^2 (1 - \varrho^2 + \varrho^2 e^2)$

wherein $\beta = \varrho\, \sigma_{\bar{x}} : \sigma_{\bar{y}}$ and $e = (m_2 - \beta)/\beta$

(Note that

$$\varrho\, e = (\sigma_{\bar{y}} / \sigma_{\bar{x}})(m_2 - \beta)$$

even if $\varrho = 0$.)

(4) Estimator 3 $\quad \text{Var } \bar{x}_3 = \sigma_{\bar{x}}^2 (1 - \varrho^2) + R$

(5) Estimator 4 $\quad \text{Var } \bar{x}_4 = \sigma_{\bar{x}}^2 (1 + C_{\bar{y}}^2 / C_{\bar{x}}^2 - 2\varrho\, C_{\bar{y}} / C_{\bar{x}}) + R'$

(6) $\qquad \doteq 2\, \sigma_{\bar{x}}^2 (1 - \varrho) \quad [n \text{ large; and } C_{\bar{x}} \doteq C_{\bar{y}}]$

R and R' are remainders in the Taylor's series, involving $1/n^2$ and higher powers, and will be negligible if n is large.

It follows that

(7) $$\frac{\text{Var } \bar{x}_4}{\text{Var } \bar{x}_2} = \frac{1 + C_{\bar{y}}^2 / C_{\bar{x}}^2 - 2\varrho\, C_{\bar{y}} / C_{\bar{x}} + R'}{1 - \varrho^2 + (\sigma_{\bar{y}}^2 / \sigma_{\bar{x}}^2)(m_2 - \beta)^2}$$

(8) $$\doteq \frac{2}{1 + \varrho} \quad [n \text{ large; } m_2 \doteq \beta \text{ and } C_x \doteq C_y]$$

(9) $$\frac{\text{Var } \bar{x}_2}{\text{Var } \bar{x}_3} = \frac{\sigma_{\bar{x}}^2 (1 - \varrho^2) + \sigma_{\bar{y}}^2 (m_2 - \beta)^2}{\sigma_{\bar{x}}^2 (1 - \varrho^2) + R}$$

(10) $$\doteq 1 \quad [n \text{ large}, m_2 - \beta \text{ small}]$$

(11) $$\frac{\text{Var } \bar{x}_4}{\text{Var } \bar{x}_3} = \frac{1 + C_{\bar{y}}^2 / C_{\bar{x}}^2 - 2\varrho\, C_{\bar{y}} / C_{\bar{x}} + R'}{1 - \varrho^2 + R}$$

(12) $$\doteq \frac{2}{1 + \varrho} \quad [n \text{ large, and } C_{\bar{x}} \doteq C_{\bar{y}}].$$

Choice of Estimator

Comparison of variances is important, of course, but the burden of computation of estimates is also a factor in the choice of estimator. Estimator $\bar{x}_3$ involves much more arithmetic than the others. This may not be important to an electronic computer, or if the size of sample is small and if there are only a few estimators to prepare. Estimators $\bar{x}_2$ and $\bar{x}_4$, on the other hand, are extremely simple to apply, requiring but little more effort than $\bar{x}_1 = \bar{x}$.

The variances of $\bar{x}_2$ and of $\bar{x}_3$ will be about equal if an acceptable value for m_2 is known from prior studies. It will usually suffice if m_2 does not differ from β by more than 30 % or 40 % (i. e., if $e < .3$ or $.4$). In a continuing series of surveys, we may often adopt m_2 as the least squares estimator of the regression coefficient from prior studies, or as some

simple approximation thereto. The choice of estimator will then depend on a variety of local competitive arguments that will involve the subject-matter and facilities for computation, not usefully discussed in a general treatment.

If the correlation ϱ between x_i and y_i be moderate or high positive, and if the line of regression of x on y misses the origin by a wide margin, then the estimator $\bar{x}_3$ will show substantial advantages over $\bar{x}_4$ and $\bar{x}_1$. If the correlation be moderate or high negative, then whether or not the regression of x on y goes through the origin, the estimator $\bar{x}_3$ will show substantial advantages over x_4 and $\bar{x}_1$. If the y-variate shows relatively wide spread (i. e., if $C_{\bar{y}}$ is much greater than $C_{\bar{x}}$), the ratio-estimator $\bar{x}_4$ may actually be far less precise than the simple estimator $\bar{x}_1 = \bar{x}$, even if ϱ be high positive, especially if the line of regression misses the origin by a wide margin. On the other hand, if $\varrho > 0$ and if the line of regression passes through the origin $(\varrho C_{\bar{x}} = C_{\bar{y}})$ or nearly so, $\bar{x}_3$ and $\bar{x}_4$ will have about the same variance, but $\bar{x}_4$ may be much the easier one to compute.

Estimator $\bar{x}_2$ is practicable if we have at hand from prior knowledge (as from prior surveys of a related type) a rough approximation to the regression coefficient $\beta = \varrho \sigma_{\bar{x}} / \sigma_{\bar{y}}$: otherwise the estimator $\bar{x}_2$ may lead to high variance.

Estimate of b Subject to Sampling Error

It often happens that the y-population per sampling unit is not known with the reliability of a census, but comes instead from a sample. This circumstance introduces additional terms into the variances. Let n be the size of the present sample, n' the size of the other sample, which gives an estimate of b with variance $n \sigma_{\bar{y}}^2 / n'$.

We distinguish between two cases: I. the present sample of size n is a subsample of a sample of size n'; II. the two samples are independent. Approximate variances are in the table on the next page.

Table 3 shows numerical comparisons for Estimators 1, 3 and 4, under the assumption that $C_{\bar{x}} = C_{\bar{y}}$. It will be observed that when $C_{\bar{x}} \doteq C_{\bar{y}}$ and when ϱ is high and positive, the variances of Estimators 3 and 4 are almost equal, and that both of them yield considerable gain in precision over Estimator 1. The gain is especially striking when n' is large compared with n. The gap between Estimators 1 and 4 closes at $\varrho = .5$, and Estimator 4 actually loses precision for $\varrho < .5$. On the other hand, Estimator 3 continues to show gains over Estimator 1, even when ϱ is very low, at the expense of extra calculation. (Estimator 2 is not usefully compared in a general table, because of the wide latitude of choices open for m_2, which will depend heavily on local information.)

Table 2

Rel-variances of estimators

Estimator	Case I: sample of size n drawn as a subsample of n'	Case II: samples of size n and n' are independent
$\bar{x}_1$	$C_{\bar{x}}^2$ [No supplementary information used]	Same as in Case I
$\bar{x}_2$	$C_{\bar{x}}^2\,[1-\varrho^2(1-e^2)(1-n/n')]$	$C_{\bar{x}}^2\,[1-\varrho^2(1-e^2)+\varrho^2(1+e)^2 n/n']$
$\bar{x}_3$	$C_{\bar{x}}^2\,[1-\varrho^2(1-n/n')]$	Same as in Case I
$\bar{x}_4$	$C_{\bar{x}}^2-(2\varrho\,C_{\bar{x}}C_{\bar{y}}-C_{\bar{y}}^2)(1-n/n')$	$C_{\bar{x}}^2-(2\varrho\,C_{\bar{x}}C_{\bar{y}}-C_{\bar{y}}^2)(1-n/n')+(2\,C_{\bar{y}}\,n/n')(C_{\bar{y}}-\varrho\,C_{\bar{x}})$

Table 3

Ratio of Var $\bar{x}_3$ and Var $\bar{x}_4$ to Var $\bar{x}_1$

Case I: the sample of size n is a subsample of n'

Assume $C_{\bar{x}}=C_{\bar{y}}$

ϱ	Estimator	$n'/n=5$	$n'/n=10$	$n'/n=\infty$
.95	3	.278	.188	.0975
	4	.280	.190	.100
.9	3	.352	.271	.190
	4	.360	.280	.200
.8	3	.488	.424	.360
	4	.520	.460	.400
.7	3	.608	.559	.510
	4	.680	.640	.600
.5	3	.800	.775	.750
	4	1.000	1.000	1.000
.3	3	.928	.919	.910
	4	1.320	1.360	1.400
.1	3	.992	.991	.990
	4	1.640	1.720	1.800

Example

In a study of the cost of hauling goods by motor truck, one aim was to estimate the average number of actual miles per shipment. It was a fairly simple matter to look up in a table, for any shipment the so-called revenue-miles between origin and destination. The actual miles that a shipment moved over, however, was difficult, and required specialized study, as the actual origin and destination might be from 1 to 50 miles away from the point from which revenue is computed. Moreover, detours, regular and irregular, figure in the actual miles.

Pilot studies showed that the correlation ϱ between revenue-miles and actual miles on a shipment was almost never below .8, and often much higher.

Let $f = \bar{x}/\bar{y}$ be the ratio of actual miles to revenue-miles in a small subsample of size n, and let b be the estimate of the average revenue-miles per shipment in the main sample of size n'.

If actual miles be placed only on a random subsample of size $n = n'/10$, and if $\bar{x}_4 = f\,b = (\bar{x}/\bar{y})\,b$ be calculated, then it turns out that

$$\text{Var } \bar{x}_4 = .424 \text{ Var } \bar{x}_1$$

It follows that the standard error of $\bar{x}_4$ would be only $(10 \times .424)$ or 2.1 times the standard error that would come from placing actual miles on every shipment of the sample of size n', and using $\bar{x}_1$.

Var $\bar{x}_3$ would be only about 10 % lower than Var $\bar{x}_4$, and in this example would not be worth the additional labor of computation.

In this instance, the larger sample of size n' was in use for other purposes. Consequently, the only problem was to establish a subsampling fraction n/n' that would give sufficient precision for the purpose. If the entire procedure were being designed for the purpose of estimating average actual miles, and the independent information were not obtainable from other sources, one would take into account comparative unit costs of filling in the information on x and y, and adjust both n and n' to optimum sizes. Costs would enter the formulas in somewhat the same way as they do in stratified sampling.

Zusammenfassung

Übersicht über Regressions-Schätzfunktionen

Die Verwendung von Regressions-Schätzfunktionen, bei denen man sich ergänzender Informationen bedient, kann erhebliche Vorteile gegenüber einfachen Schätzverfahren haben. Die Bedingungen hierzu werden von den Verfassern systematisch untersucht und für folgende 4 Methoden verglichen:

1. Einfache Schätzfunktionen ohne Verwendung von Nebeninformationen
2. Differenzen-Schätzfunktionen mit einem konstanten Koeffizienten
3. Regressions-Schätzfunktionen, die auf der Methode der kleinsten Quadrate basieren
4. Verhältnis-Schätzfunktionen

Sie kommen zu dem Ergebnis, daß die Methoden 2 und 4 die Einfachheit der Berechnung für sich haben und auch meistens günstiger sind als die Methode 1, daß aber in allen Fällen mittlerer oder hoher positiver Korrelation und ins Gewicht fallender Regressionskonstante sowie fast immer bei negativer Korrelation die Regressions-Schätzung nach 3 vorzuziehen ist.

Die Schätzungen werden oft dadurch erschwert, daß der Erwartungswert der zur Nebeninformation herangezogenen Daten (b) selbst einer Stichprobenvariabilität unterworfen ist. Dann sind die beiden Fälle zu unterscheiden, daß die beiden Stichproben voneinander abhängig (die eine Stichprobe stellt eine Teilmenge der anderen dar) oder unabhängig sind. Für die Methode 1 und 3 macht diesbezüglich der Stichprobenfehler keinen Unterschied, wohl aber für 2 und 4. Im Fall abhängiger Stichproben erweist sich die Überlegenheit von 3 über 1 bei mittleren und schwachen Korrelationsverhältnissen, während 4 dann vergleichsweise schlechtere Ergebnisse zeitigt.