Regression and ratio estimators on two occasion

G. Diana, C. Tommasi, M. Zuin

2002.14

Dipartimento di Scienze Statistiche
Università degli Studi
Via C. Battisti 241-243
35121 Padova

Ottobre 2002

REGRESSION AND RATIO ESTIMATORS ON TWO OCCASION

Giancarlo Diana   Chiara Tommasi   Marino Zuin

REGRESSION AND RATIO ESTIMATORS ON TWO OCCASION

1 Introduction

In two occasion sampling scheme the regression estimator is usually used for estimating a population mean. On the other hand, Kumar and Dayal (1998) consider the ratio estimator too. Specifically, they show that, for a fixed sample size, the regression estimator is more efficient but the ratio estimator is "cheaper". In this work ratio and regression estimator are compared through their efficiencies, but for fixed costs instead of for fixed sample size. The details are given in Section 2.

Let $U=\{1,\ldots,i,\ldots,N\}$ be a finite population where the interest variable Y is observed on two subsequent occasions. Let Y_{1i} and Y_{2i} be the values of Y on the first and second occasion respectively. A sampling scheme on two occasions is based on two steps:

I step: a simple random sample without replacement (SRSWR) of n units is selected from N units, on the first occasion;

II step: out of these n units, m are retained for the second occasion (matched portion) and a new SRSWR of $u=n-m$ units is selected from $N-n$ units (unmatched portion). In this way the sample size is constant on both occasions. Let y_{1i} and y_{2i} be the sampling realisations on the first and second occasion respectively. The following quantities can be considered

$$\bar{Y}_h=\frac{1}{N}\sum_{i=1}^{N}Y_{hi}\ ;\ \ S_{Y_h}^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{hi}-\bar{Y}_h\right)^2;\ \ CV\left(Y_h\right)=\frac{S_{Y_h}}{\bar{Y}_h}\quad \text{with}\quad h=1,2$$

$$S_{Y_1Y_2}=\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i}-\bar{Y}_1\right)\left(Y_{2i}-\bar{Y}_2\right);\ \ \rho=\frac{S_{Y_1Y_2}}{S_{Y_1}S_{Y_2}};$$

$$\bar{y}_h=\frac{1}{n}\sum_{i=1}^{n}y_{hi}\ ;\ \ \bar{y}_{hm}=\frac{1}{m}\sum_{i=1}^{m}y_{hi}\ ;\ \ s_{y_h}^2=\frac{1}{m-1}\sum_{i=1}^{m}\left(y_{hi}-\bar{y}_{hm}\right)^2\quad \text{with}\quad h=1,2$$

$$s_{y_1y_2} = \frac{1}{m-1}\sum_{i=1}^{m}(y_{1i} - \bar{y}_{1m})(y_{2i} - \bar{y}_{2m});\ \bar{y}_{2u} = \frac{1}{u}\sum_{1=1}^{u} y_{2i} \text{ and } V(\bar{y}_{2u}) = \frac{S_{Y_2}^2}{u}.$$

In the above expression of $V(\bar{y}_{2u})$ we assume that the finite population correction can be ignored. From now on we will use this approximation.

In order to estimate the population mean of Y on the second occasion, $\bar{Y}_2$, the following estimators can be used

$$\bar{y}'_{\mathrm{Re}} = \frac{u\bar{y}_{2u} + k\bar{y}_{\mathrm{Re}}}{u+k} \quad \text{with} \quad V(\bar{y}'_{\mathrm{Re}}) = \frac{S_{Y_2}^2(n - u\rho^2)}{(n^2 - u^2\rho^2)}$$

$$\bar{y}'_{Ra} = \frac{u\bar{y}_{2u} + k'\bar{y}_{Ra}}{u+k'} \quad \text{with} \quad V(\bar{y}'_{Ra}) = \frac{S_{Y_2}^2(n - ut)}{(n^2 - u^2t)}$$

where

$$t = 2\rho\frac{CV(Y_1)}{CV(Y_2)} - \left[\frac{CV(Y_1)}{CV(Y_2)}\right]^2;\ \frac{1}{k} = \frac{1-\rho^2}{m} + \frac{\rho^2}{n};\ \frac{1}{k'} = \frac{1-t}{m} + \frac{t}{n}$$

and $\bar{y}_{\mathrm{Re}}$ and $\bar{y}_{Ra}$ are the regression and ratio estimators in double sampling when the auxiliary variable is Y measured on the first occasion. More specifically,

$$\bar{y}_{\mathrm{Re}} = \bar{y}_{2m} + b(\bar{y}_1 - \bar{y}_{1m}) \quad \text{with} \quad b = \frac{s_{y_1y_2}}{s_{y_1}^2}$$

and

$$\bar{y}_{Ra} = \bar{y}_1\frac{\bar{y}_{2m}}{\bar{y}_{1m}}.$$

Kumar and Dayal (1998) compare $\bar{y}'_{\mathrm{Re}}$ and $\bar{y}'_{Ra}$ for fixed n assuming $CV(Y_1) = CV(Y_2)$ (for a justification of this hypothesis see Zuin, 2002). Specifically, they minimize $V(\bar{y}'_{\mathrm{Re}})$ and $V(\bar{y}'_{Ra})$ with respect to m and u, getting

$$V_{\min}(\bar{y}'_{\mathrm{Re}}) = \frac{S_{Y_2}^2}{2n}\left(1 + \sqrt{1-\rho^2}\right) \quad \text{and} \quad V_{\min}(\bar{y}'_{Ra}) = \frac{S_{Y_2}^2}{2n}\left(1 + \sqrt{1-t}\right).$$

Since $V_{\min}(\bar{y}'_{\mathrm{Re}}) < V_{\min}(\bar{y}'_{Ra})$ regression estimator is more efficient than ratio estimator. On the other hand, $\bar{y}'_{Ra}$ is better than $\bar{y}'_{\mathrm{Re}}$ from an economic point of view, since $m_{Ra} < m_{\mathrm{Re}}$ and it is usually more expensive to get information about the unmatched portion than the matched one.

2 Efficiency comparison for fixed costs

In practical setting, it is very common to have cost limitations. Thus, in this section we compare ratio and regression estimators through their efficiencies under a cost constraint instead of for a fixed sample size.

- **Theorem**: ratio estimator is never more efficient than regression estimator, under the linear cost function,

$$C = C_0 + C_m m + C_u u\,, \tag{1}$$

where C_0 is a fixed cost while C_m and C_u ($C_m < C_u$) denote the unit costs for the retained and replaced units, respectively.

Proof

Let us denote the sampling cost as $C^* = C - C_0$. Of course we are assuming that C^* is fixed to some suitable value. Since $m = n - u$, from (1) we have that

$$u = a - bn \quad \text{where} \quad a = \frac{C^*}{C_u - C_m} \quad \text{and} \quad b = \frac{C_m}{C_u - C_m}. \tag{2}$$

Assuming $CV(Y_1) = CV(Y_2)$, as Kumar and Dayal (1998), $V(\bar{y}'_{\mathrm{Re}})$ and $V(\bar{y}'_{Ra})$ can be written as

$$V(\bar{y}'_{\mathrm{Re}}) = \frac{S^2_{Y_2}\left[n - (a-bn)\rho^2\right]}{n^2 - (a-bn)^2\rho^2} \quad \text{and} \quad V(\bar{y}'_{Ra}) = \frac{S^2_{Y_2}\left[n - (a-bn)(2\rho - 1)\right]}{n^2 - (a-bn)^2(2\rho-1)}. \tag{3}$$

In what follows we prove that $V(\bar{y}'_{\mathrm{Re}}) - V(\bar{y}'_{Ra}) < 0$ for any value of $n > 0$ and $0 \le u \le n$, thus

$$\left.\begin{array}{l} u \ge 0 \Rightarrow a - bn \ge 0 \Rightarrow n \le \dfrac{a}{b} \\ u \le n \Rightarrow a - bn \le n \Rightarrow n \ge \dfrac{a}{b+1} \end{array}\right\} \Rightarrow \frac{a}{b+1} \le n \le \frac{a}{b}. \tag{4}$$

From (3) we have

$$V(\bar{y}'_{\mathrm{Re}}) - V(\bar{y}'_{Ra}) = \frac{n\,S^2_{Y_2}(\rho-1)^2\left[b(b+1)n^2 - a(2b+1)n + a^2\right]}{\left[(b^2\rho^2 - 1)n^2 - 2ab\rho^2 n + a^2\rho^2\right]\left\{\left[(2\rho-1)b^2 - 1\right]n^2 - 2ab(2\rho-1)n + (2\rho-1)a^2\right\}} \tag{5}$$

The goal is to study the sign of the difference $V(\bar{y}'_{\mathrm{Re}}) - V(\bar{y}'_{Ra})$. It is straightforward to prove that the numerator of expression (5) is negative for any values of n in

$[a(b+1)^{-1}, ab^{-1}]$. On the other hand for studying the sign of the denominator we have to consider two cases, $0<\rho\le 0{,}5$ and $0{,}5<\rho<1$. For computational reasons, let us factorise the first term of the denominator as $(n-n_1)(n-n_2)$ where $n_1 = a\rho(b\rho+1)^{-1}$ and $n_2 = a\rho(b\rho-1)^{-1}$.

a) $0<\rho\le 0{,}5$. We have three possibilities:

1) $\rho^2 < b^{-2} \quad \Rightarrow \quad n_2 < n_1 < \dfrac{a}{b+1}$;

2) $\rho^2 > b^{-2} \quad \Rightarrow \quad n_1 < \dfrac{a}{b+1} < \dfrac{a}{b} < n_2$;

3) $\rho^2 = b^{-2} \quad \Rightarrow$ the first factor becomes $-2ab^{-1}n + a^2b^{-2}$.

In all the cases the first factor is negative in $[a(b+1)^{-1}, ab^{-1}]$. On the other hand, the second factor of the denominator is always negative, thus the denominator is positive when $0<\rho\le 0{,}5$.

b) $0{,}5<\rho<1$. The second term of the denominator in (5) can be factorised as $(n-n_3)(n-n_4)$ where $n_3 = a\sqrt{2\rho-1}\left(b\sqrt{2\rho-1}+1\right)^{-1}$ and $n_4 = a\sqrt{2\rho-1}\left(b\sqrt{2\rho-1}-1\right)^{-1}$ and we have three possibilities:

1) $2\rho-1<\rho^2<b^{-2} \quad \Rightarrow \quad n_2 < n_1 < \dfrac{a}{b+1}$ and $n_4 < n_3 < \dfrac{a}{b+1}$;

2) $2\rho-1<b^{-2}<\rho^2 \quad \Rightarrow \quad n_1 < \dfrac{a}{b+1} < \dfrac{a}{b} < n_2$ and $n_4 < n_3 < \dfrac{a}{b+1}$;

3) $b^{-2}<2\rho-1<\rho^2 \quad \Rightarrow \quad n_1 < \dfrac{a}{b+1} < \dfrac{a}{b} < n_2$ and $n_3 < \dfrac{a}{b+1} < \dfrac{a}{b} < n_4$.

In all the cases as well as when $\rho^2 = b^{-2}$ or $2\rho-1=b^{-2}$, both the first and the second factor are negative for any n in $[a(b+1)^{-1}, ab^{-1}]$, thus the denominator is positive when $0{,}5<\rho<1$.

Therefore for any n in $[a(b+1)^{-1}, ab^{-1}]$, $V(\bar{y}'_{Ra}) \ge V(\bar{y}'_{\mathrm{Re}})$.

3 Numerical simulation

Even if, for fixed costs, regression estimators are more efficient than ratio estimators, it may be interesting to have an idea about the efficiency loss when

$[\ldots]$. On the other hand, to evaluate the sign of the denominator we have to consider two cases: $a \leq b$ and $a > b$. For computational reasons, let us factorise the first term of the denominator as $(x - x_1)(x - x_2)$ where x_1 [illegible]

a) $a \leq b$: we have three possibilities:

1) [illegible]

2) [illegible]

3) [illegible] the same results [illegible]

[illegible] and [illegible] both negative in [illegible] so the denominator is positive when [illegible]

b) $a > b$: the second term of the denominator in (9) can be factorised as $(x - x_3)(x - x_4)$ where [illegible] and [illegible]

and we have three possibilities:

1) [illegible] and [illegible]

2) [illegible] and [illegible]

3) [illegible] and [illegible]

In all the cases as well as when [illegible], both the first and the second factor are negative for any x in [illegible], thus the denominator is positive when $0.5 < \rho < 1$.

Therefore for any x in [illegible]

3 Numerical simulation

Even if, for fixed costs, regression estimators are more efficient than ratio estimators, it may be interesting to have an idea of the efficiency loss when using the ratio estimator. For this reason we have performed a simulation study. For different values of C^*, C_u/C_m and $\rho\in(0.5,1)$ we have computed the following quantities: n_{re}^{ott} and n_{ra}^{ott}, that is the optimal values of n minimizing $V(\bar{y}'_{\mathrm{Re}})$ and $V(\bar{y}'_{Ra})$, respectively; the corresponding optimal values for the variances, i.e. $V^{ott}(\bar{y}'_{\mathrm{Re}})$ and $V^{ott}(\bar{y}'_{Ra})$ (with $S^2_{Y_2}=1000$) and finally the efficiency loss rate.

From the simulation results, which are summarized in Table 3.1, it seems that

i) the precision loss rate for using $\bar{y}'_{Ra}$ instead of $\bar{y}'_{\mathrm{Re}}$ essentially does not change whatever the fixed value for $C*$;

ii) the precision loss rate is always less than 2,7%.

Remark 1: in order to ensure a sampling cost not greater than the fixed value, we have rounded the optimal value of n to the closest integer and the optimal value for u to its integer part. That must cause an efficiency gain for $\bar{y}'_{Ra}$. For instance in Table 3.1 we have six cases.

Remark 2: In this simulation study we have considered only $\rho\in(0.5,1)$ since regression and ratio estimators are not useful when ρ is "low".

4 Conclusions and practical considerations

With or without cost constraints, regression estimators are more efficient than ratio estimators. The precision loss rate for using ratio estimators is however quite small, thus the simplicity can justify its use.

From a practical point of view, for using $\bar{y}'_{\mathrm{Re}}$ and/or $\bar{y}'_{Ra}$ we need for n_{re}^{ott} and n_{ra}^{ott}, which are given by

$$n_{re}^{ott}=\frac{a\rho\left[\rho\sqrt{b^2\rho^2-1}+\sqrt{1-\rho^2}\right]\sqrt{b^2\rho^2-1}}{b^3\rho^4+b^2\rho^2-b\rho^2-1}\quad\text{and}\quad n_{ra}^{ott}=\frac{a\left[k\left(b^2k-1\right)+\sqrt{k(1-k)\left(b^2k-1\right)}\right]}{b^3k^2+b^2k-bk-1},\ \text{where } k=2\rho-1.$$

It is easy to prove that whenever n_{re}^{ott} and n_{ra}^{ott} exist they are greater then the lower bound $a(a+b)^{-1}$, whatever the value for ρ. Furthermore for any $\rho\in(0.5,1)$, if $2\rho-1>b^{-1}$, i.e. $\rho>0.5\left(1+b^{-1}\right)$, then both n_{re}^{ott} and n_{ra}^{ott} exist and are less than the upper bound ab^{-1}, otherwise if $b^{-2}<2\rho-1\le b^{-1}<\rho^2$, then both n_{re}^{ott} and n_{ra}^{ott} exist

TABLE 3.1: Efficiency comparison for fixed costs between $\bar{y}'_{\mathrm{Re}}$ and $\bar{y}'_{Ra}$.

C^*	C_u/C_m	ρ	n_{Re}^{ott}	n_{Ra}^{ott}	$V(\bar{y}'_{\mathrm{Re}})$	$V(\bar{y}'_{Ra})$	$\frac{V(\bar{y}'_{Ra})-V(\bar{y}'_{\mathrm{Re}})}{V(\bar{y}'_{\mathrm{Re}})}\%$
100	1,2	0,6	95	100	9,773	10,000	2,326
		0,7	93	94	9,424	9,674	2,659
		0,8	91	91	8,905	9,057	1,714
		0,9	89	89	8,125	8,180	0,681
	1,4	0,7	94	100	9,930	10,000	0,701
		0,8	88	90	9,598	9,709	1,152
		0,9	83	83	8,970	9,018	0,533
	1,6	0,8	96	100	10,025	10,000	- 0,250
		0,9	82	83	9,625	9,678	0,551
	1,8	0,9	98	100	10,039	10,000	- 0,386
200	1,2	0,6	189	200	4,886	5,000	2,340
		0,7	185	188	4,711	4,837	2,678
		0,8	182	183	4,452	4,528	1,691
		0,9	178	178	4,063	4,090	0,681
	1,4	0,7	188	200	4,965	5,000	0,701
		0,8	176	179	4,799	4,859	1,245
		0,9	166	167	4,476	4,505	0,637
	1,6	0,8	191	200	4,992	5,000	0,157
		0,9	165	166	4,817	4,839	0,452
	1,8	0,9	195	200	5,004	5,000	- 0,085
400	1,2	0,6	378	400	2,443	2,500	2,340
		0,7	370	376	2,356	2,419	2,678
		0,8	364	365	2,226	2,264	1,690
		0,9	356	357	2,031	2,045	0,672
	1,4	0,7	377	400	2,484	2,500	0,651
		0,8	353	358	2,401	2,427	1,094
		0,9	332	333	2,238	2,251	0,582
	1,6	0,8	382	400	2,496	2,500	0,157
		0,9	329	332	2,407	2,416	0,352
	1,8	0,9	390	400	2,502	2,500	- 0,085
800	1,2	0,6	757	800	1,221	1,250	2,341
		0,7	741	751	1,178	1,209	2,679
		0,8	728	731	1,113	1,132	1,687
		0,9	712	713	1,016	1,022	0,671
	1,4	0,7	753	800	1,242	1,250	0,677
		0,8	705	716	1,200	1,214	1,120
		0,9	664	667	1,119	1,125	0,557
	1,6	0,8	764	800	1,248	1,250	0,157
		0,9	659	664	1,203	1,208	0,402
	1,8	0,9	781	800	1,251	1,250	- 0,060
1600	1,2	0,6	1.513	1.600	0,611	0,625	2,341
		0,7	1.482	1.502	0,589	0,605	2,679
		0,8	1.455	1.462	0,557	0,566	1,688
		0,9	1.425	1.427	0,508	0,511	0,671
	1,4	0,7	1.507	1.600	0,621	0,625	0,689
		0,8	1.411	1.433	0,600	0,607	1,145
		0,9	1.328	1.333	0,560	0,563	0,545
	1,6	0,8	1.528	1.600	0,624	0,625	0,157
		0,9	1.317	1.328	0,602	0,604	0,339
	1,8	0,9	1.561	1.600	0,625	0,625	- 0,022

but only n_{re}^{ott} is less than ab^{-1}, while $n_{ra}^{ott} \geq ab^{-1}$. For this reason we have considered only values of ρ in $[0.5(1+b^{-1}),1)$.

For practical purposes we suggest to round n_{re}^{ott} and n_{ra}^{ott} to their closest integers. Of course, the unmatched portion is given by (2) replacing n with n_{re}^{ott} or n_{ra}^{ott}. Specifically, we suggest to use the integer part of such obtained values in order to ensure a sampling cost not greater than the fixed value.

5 Bibliography

Kumar R. and Dayal S. (1998), *Regression and ratio estimators in sampling on two occasion*, "Journal of the Indian Statistical Association", 36, pp. 85-89.

Zuin M. (2002), *Il campionamento in due occasioni*, Graduation thesis, Department of Statistics, University of the study of Padova.

SUMMARY

Recent studies in sampling on two occasion have highlighted the preference for regression estimator in terms of precision and for ratio estimators in terms of economic convenience. These conclusions are valid until no cost constraint is considered. In this paper an efficiency comparison for fixed costs is performed: it is shown that ratio estimator is always worse than regression estimator even if the efficiency loss is often small.

RIASSUNTO

Recenti studi nell'ambito del campionamento in due occasioni hanno evidenziato la preferibilità dello stimatore per regressione in termini di precisione e dello stimatore per rapporto in termini di convenienza economica. Si trattava, però, di un'analisi di tipo marginale, valida quando la numerosità campionaria è fissata. In questo lavoro si propone un confronto di efficienza tra i due stimatori a parità di costi. Da tale confronto emerge che lo stimatore per rapporto non è mai preferibile anche se la perdita di efficienza è spesso trascurabile.