A MONTE CARLO COMPARISON OF REGRESSION ESTIMATORS IN STRATIFIED RANDOM SAMPLING

Chien-Pai Han, University of Texas at Arlington
Department of Mathematics, Box 19408, Arlington, TX 76019

KEY WORDS: Separate regression estimator; combined regression estimator, preliminary test estimator; weighted estimator.

Abstract.

When auxiliary information is available it is advantegeous to use it to construct estimators. In stratified random sampling, one may construct either a separate regression estimator or a combined regression estimator for the population mean. The separate regression estimator is appropriate when the population regression coefficients are different from stratum to stratum; while the combined regression estimator is appropriate when the stratum regression coefficients are equal. In practice it may be uncertain whether the regression coefficients are equal. In such a case, we can use a preliminary test to test the equality of the population stratum regression coefficients. Then a preliminary test regression estimator can be constructed. Also we can use a weighted regression estimator which is a weighted average of the separate regression estimator and the combined regression estimator with the weights depending on the test statistic. A comparison of the various estimators is made by a Monte Carlo study.

1. INTRODUCTION.

When auxiliary information is available, it is advantegeous to use it to construct estimators. We consider regression estimators in stratified random sampling. Two commonly used estimators are the separate regression estimator and the combined regression estimator. Let y_{hi} be the values of the i*th* unit in the h*th* stratum, $i=1, 2, \cdots, N_h$ and $h=1, 2, \cdots, L$. Let $N=\Sigma N_h$, the total number of units in the population. Suppose an auxiliary variable is available and let x_{hi} be the corresponding value of the auxiliary variable for y_{hi}. We are interested in estimating the population mean

$$\bar{Y}=\sum_{h=1}^{L} W_h \bar{Y}_h \qquad (1)$$

where $W_h=N_h/N$ is the h*th* stratum weight, and $\bar{Y}_h=\sum_{i=1}^{N_h} y_{hi}/N_h$ is the h*th* stratum population mean. Suppose a simple random sample of size n_h is taken from the h*th* stratum. The separate regression estimator is obtained by first computing a separate regression estimator for each stratum, that is

$$\bar{y}_{lrh}=\bar{y}_h+b_h(\bar{X}_h-\bar{x}_h) \qquad (2)$$

where $\bar{y}_h$ and $\bar{x}_h$ are the sample means $\bar{X}_h=\sum_{i=1}^{N_h} x_{hi}/N_h$, and

$$b_h=\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)(x_{hi}-\bar{x}_h)/\sum_{i=1}^{n_h}(x_{hi}-\bar{x}_h)^2 \qquad (3)$$

Then the separate regression estimator is given as (see Cochran(1977)),

$$\bar{y}_{lrs}=\sum_{h=1}^{L} W_h \bar{y}_{lrh} \cdot \qquad (4)$$

The combined regression estimator is obtained by first computing the stratified sample estimators

$$\bar{y}_{st}=\sum_{h=1}^{L} W_h \bar{y}_h$$

$$\bar{x}_{st}=\sum_{h=1}^{L} W_h \bar{x}_h \cdot$$

The combined regression estimator is given as

$$\bar{y}_{lrc} = \bar{y}_{st} + b_c(\bar{X} - \bar{x}_{st}) \qquad (5)$$

where

$$b_c = \frac{\sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \sum_i (y_{hi}-\bar{y}_h)(x_{hi}-\bar{x}_h)}{\sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \sum_i (x_{hi}-\bar{x}_h)^2} \qquad (6)$$

$f_h = n_h/N_h$ is the sampling fraction.

It is known that the separate regression estimator is appropriate when the regression is linear and the true regression coefficients β_h vary from stratum to stratum, while the combined regression estimator is preferred if the true regression coefficients are the same in all strata. In practice it may happen that the investigator is uncertain whether the stratum population regression coefficients are equal. In such a case the investigator may resolve the uncertainty by using a preliminary test to test the null hypothesis that the population regression coefficients are equal. If the null hypothesis is rejected, i.e. the data indicates that the regression coefficients are different, we use the separate regression estimator. On the other hand, if the null hypothesis is not rejected, we use the combined regression estimator. This type of inference procedure is conditioning on the model specification. Hence it is termed as inference procedure based on conditional specification by Bancroft and Han (1977). Further references may be found in the bibliographies by Bancroft and Han (1977), Han, Rao and Ravichandran (1988). In Section 2 we discuss the preliminary test estimator and a weighted estimator. A comparison of the estimators is made in Section 3 by a Monte Carlo study.

2. PRELIMINARY TEST ESTIMATOR AND WEIGHTED ESTIMATOR.

Let us assume that y and x have a bivariate normal distribution. We test $H_0: \beta_1 = \beta_2 = \cdots = \beta_L$ by using the test statistic

$$F = \sum_{h=1}^{L} c_h(b_h - b_c)^2/[(L-1)S^2] \qquad (7)$$

where

$$c_h = \sum_{i=1}^{n_h}(x_{hi}-\bar{x}_h)^2$$

$$S^2 = \sum_h \sum_i [y_{hi}-\bar{y}_h - b_h(x_{hi}-\bar{x}_h)]^2/(N-2L).$$

The statistic F has an F distribution with $(L-1, N-2L)$ degrees of freedom under H_0.

We define two estimators of $\bar{Y}$ based on the statistic F. One estimator is the preliminary test estimator

$$\bar{y}_{PT} = \begin{cases} \bar{y}_{lrs} & \text{if } F > F_\alpha \\ \bar{y}_{lrc} & \text{if } F \le F_\alpha \end{cases} \qquad (8)$$

where F_α is the $100(1-\alpha)\%$ point of F(L-1, N-2L). The other estimator is the weighted estimator

$$\bar{y}_w = \frac{1}{1+F}\bar{y}_{lrc} + \frac{F}{1+F}\bar{y}_{lrs} \qquad (9)$$

The rationale of using $\bar{y}_w$ is that when H_0 is true, F will be small and the weight for $\bar{y}_{lrs}$ is small and the weight for $\bar{y}_{lrc}$ is large as $\bar{y}_{lrc}$ should be used in such a case. On the contrary, if H_0 is not true, one should use $\bar{y}_{lrs}$. Since F will be large and consequently the weight for $\bar{y}_{lrs}$ is large.

3. COMPARISON OF THE ESTIMATORS.

The derivations of the biases and mean square errors (MSE) for $\bar{y}_{PT}$ and $\bar{y}_w$ are very tedious. Hence we use a Monte Carlo study to compare the biases and MSE's. The simulation was done on the CRAY X-MP/24 at the University of Texas System Center for High Performance Computing. The subroutines in the International Mathematical and Statistical Library (IMSL) are used in the simulation. The subroutine RNMVN is used to generate the bivariate normal random variables for given population regression coefficients. Once a random sample is generated, the estimators $\bar{y}_{lrc}$, $\bar{y}_{lrs}$, $\bar{y}_{PT}$ and $\bar{y}_w$ are obtained. This

process is repeated 2000 times. The biases of the estimators are small since we take the sample from a bivariate normal distribution and the regression line is linear. Further the bias is a part of the mean square error (MSE). Hence we only compare the mean square errors of the estimators.

Let $\bar{y}_i$ denote any one of the estimators. The estimated mean square error is given as

$$\text{MSE}=\frac{1}{2000}\sum_i(\bar{y}_i-\bar{Y})^2$$

In the simulation, we let $\bar{Y}=1$. The relative efficency of $\bar{y}_{PT}$ to $\bar{y}_{lrc}$ is defined as

$$e_1=\frac{1/\text{MSE}(\bar{y}_{PT})}{1/\text{MSE}(\bar{y}_{lrc})}=\frac{\text{MSE}(\bar{y}_{lrc})}{\text{MSE}(\bar{y}_{PT})} \ .$$

Similarly we define

$$e_2=\frac{\text{MSE}(\bar{y}_{lrc})}{\text{MSE}(\bar{y}_w)}$$

$$e_3=\frac{\text{MSE}(\bar{y}_{lrc})}{\text{MSE}(\bar{y}_{lrs})} \ .$$

Table 1 gives the relative efficiencies of regression estimators for three strata. The stratum weights are set equal to .3, .3 and .4 respectively and the sample sizes are $(n_1, n_2, n_3)=(5, 5, 6)$, $(9, 9, 12)$ and $(30, 30, 40)$.

When the sample sizes are small, i.e. $(n_1, n_2, n_3)=(5, 5, 6)$, $\bar{y}_{lrc}$ has the smallest MSE when the β's are equal, the relative efficiencies are less than unity. The preliminary test estimator and the weighted estimator are about equally efficient. But $\bar{y}_w$ has the smallest MSE when the β's are very unequal and one coefficient has different sign, the relative efficiency e_2 is the largest.

When the sample sizes are moderate, i.e. $(n_1, n_2, n_3)=(9, 9, 12)$, $\bar{y}_{lrc}$ again has the smallest MSE when the regression coefficients are equal. When the regression coefficients are unequal, $\bar{y}_w$ has the highest relative efficiency, though $\bar{y}_{lrs}$ is not too far behind.

When the sample sizes are large, i.e. the case (30, 30, 40), and the regression coefficients are equal, the four estimator are essentially the same in efficiency. When the regression coefficients are unequal, the estimators $\bar{y}_{PT}$, $\bar{y}_w$ and $\bar{y}_{lrs}$ are better than $\bar{y}_{lrc}$. Further $\bar{y}_{PT}$, $\bar{y}_w$ and $\bar{y}_{lrs}$ all have similar relative efficiencies.

From the Monte Carlo study we conclude that the weighted estimator should be used when the sample sizes are moderately large since it has high relative efficiency. When the sample sizes are small, one may use the weighted estimator except when the regression coefficients are equal; in that case, the combined regression estimator should be used.

REFERENCES

Bancroft, T.A. and Han, Chien-Pai (1977). Inference based on conditional specification: A note and a bibliography. InternationalStatistical Review, 45, 117-127.

Cockran, W. G. (1977) Sampling Techniques, Wiley, New York.

Han, Chien-Pai, Rao, C.V. and Ravichandran, J. (1988) Inference based on conditional specification: A second bibliography. Communications in Statistics-Theory and Method 17, 1945-1964.

TABLE 1 Relative Efficiencies of Regression Estimators.

β_1	β_2	β_3	α	e_1	e_2	e_3
			$(n_1, n_2, n_3)=(5, 5, 6)$			
.5	.5	.5	.05	.95	.88	.76
			.25	.85	.88	.75
			.50	.81	.90	.79
−.5	−.5	−.5	.05	.96	.90	.78
			.25	.84	.88	.76
			.50	.80	.89	.77
.2	.5	.8	.05	.93	.95	.87
			.25	.88	.94	.85
			.50	.87	.95	.85
−.2	−.5	.8	.05	1.01	1.18	1.13
			.25	1.17	1.26	1.20
			.50	1.20	1.27	1.23
.4	.5	−.6	.05	.99	1.14	1.08
			.25	1.05	1.15	1.08
			.50	1.05	1.13	1.07
			$(n_1, n_2, n_3)=(9, 9, 12)$			
.5	.5	.5	.05	.98	.96	.91
			.25	.96	.98	.93
			.50	.93	.97	.92
−.5	−.5	−.5	.05	.99	.97	.92
			.25	.95	.97	.92
			.50	.92	.96	.90
.2	.5	.8	.05	.99	1.02	.99
			.25	.99	1.02	.98
			.50	.99	1.02	.99
−.2	−.5	.8	.05	1.35	1.41	1.41
			.25	1.29	1.31	1.30
			.50	1.34	1.35	1.34
.4	.5	−.6	.05	1.16	1.24	1.24
			.25	1.20	1.23	1.22
			.50	1.20	1.21	1.21
			$(n_1, n_2, n_3)=(30, 30, 40)$			
.5	.5	.5	.05	1.00	1.00	.99
			.25	.98	.99	.98
			.50	.99	1.00	.98
−.5	−.5	−.5	.05	.99	.98	.97
			.25	.99	.99	.98
			.50	.98	.99	.97
.2	.5	.8	.05	1.07	1.09	1.09
			.25	1.06	1.06	1.07
			.50	1.07	1.07	1.07
−.2	−.5	.8	.05	1.49	1.49	1.49
			.25	1.47	1.47	1.47
			.50	1.50	1.50	1.50
.4	.5	−.6	.05	1.32	1.32	1.32
			.25	1.36	1.36	1.36
			.50	1.28	1.28	1.28