BOOTSTRAP CONFIDENCE INTERVAL FOR MODIFIED LINEAR REGRESSION ESTIMATION OF THE POPULATION MEAN

Sri Haryatmi Kartiko
Math Dept, Gadjahmada University, Yogyakarta, Indonesia

Abstract

Ratio estimator is used to estimate the population mean, but it has biased property. Linear regression estimator is proposed to solve this bias problem. Several modified version of this estimator that were proposed, some has smaller variance.

Performance of the estimator can always be improve using using values of the population parameter of the auxiliary variable under study which are positively correlated with the study variable.A class of modified linear regression estimators for the population mean of the form

$$\hat{\bar{Y}}_{SK} = \alpha\frac{S_y}{C_y} + (1-\alpha)(\bar{y} - \frac{b_{yx}}{\rho}(\bar{x} - \bar{X}))$$

where $b_{yx} = \frac{s_{yx}}{s_x^2}$, is proposed by Subramani and Kumarapandiyan(2012). This form of estimator gives smaller value of variance and therefore perform better. Bootstrap confidence interval as well as its coverage probability for population mean using this estimator is constructed in this paper.

Key words : sample random sampling, auxiliary variable, linear regression estimator, mean squared error, bootstrap confidence interval

I. Introduction

The linear regression estimate is designed to increase precision by the use of an auxiliary variate x_i that is correlated with y_i. When the relation between y_i and x_i is examined, it may be found that although the relation is approximately linear, the line does not go through the origin. This suggests an estimate based on the linear regression of y_i on x_i rather than on the ratio of the two variables.

Suppose that y_i and x_i are obtained for every unit in the sample and that population mean $\bar{X}$ of the x_i is known. The linear regression estimate of $\bar{Y}$, the population mean of the y_i is

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) \tag{1.1}$$

where the subscript lr denotes *linear regression* and b is an estimate of the change in y when x is increased by unity. The rationale of this estimate is that if $\bar{x}$ is below average we should expect $\bar{y}$ also to be below average by an amount $b(\bar{X} - \bar{x})$ because of the regression of y_i on x_i. For an estimate of the population total Y, we take $\hat{Y}_{lr} = N\bar{Y}_l r$.

The regression estimate

$$\bar{y} + b(\bar{X} - \bar{x}) \tag{1.2}$$

adjusts the sample mean of the actual measurements by the regression of the actual measurements on the rapid estimates. The rapid estimates need not be free from bias. If $x_i - y_i = D$, so that the rapid estimate is perfect except for a constant bias D, then with $b = 1$ the regression estimate becomes

$$\bar{y} + (\bar{X} - \bar{x}) = \bar{X} + (\bar{y} - \bar{x}) \tag{1.3}$$

If no linear regression model is assumed, our knowledge of the properties of the regression estimate is of the same scope as our knowledge for the ratio estimate. The regression estimate is consistent, in the trivial sense that when the sample comprises the whole population, $\bar{x} = \bar{X}$, and the regression estimate reduces to $\bar{Y}$. The estimate is in general biased, but the ratio of the bias to the standard error becomes small when the sample is large. We possess a large-sample formula for the variance of the estimate, but more information is needed about the distribution of the estimate in small samples and about the value of n required for the practical use of large-sample results.

By a suitable choice of b, the regression estimate includes as particular cases both the mean per unit and the ratio estimate. Obviously if b is taken as zero, $\bar{y}_{lr}$ reduces to $\bar{y}$. If $b = \bar{y}/\bar{x}$,

$$\bar{y}_{lr} = \bar{y} + \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \frac{\bar{y}}{\bar{x}}\bar{X} = \hat{\bar{Y}}_R \tag{1.4}$$

Estimation of the population mean using linear regression estimator, its variance and its variance estimator for preassign value of b as well as its estimated value from the sample are written clearly in Cochran(1997). Kadilar and Cingi (2006) make some improvement in estimating the population mean by using the corelation coefficient. Some improvement of product method of

estimation in sample surveys is discussed by Singh (2003), while an improvement of estimator of population mean using power transformation is proposed by Singh et all (2004). Improvement of ratio type estimator using jacknife methods of estimation is done by Banerjie and Tiwari (2011), while specifically a class of modified linear regression estimators for estimation of finite population mean is done by Subramani and Kumarapandyan (2012).

II. Linear Regression Estimate With Preassign b

In most applications, b is estimated from the results of the sample and sometimes it is reasonable to choose the value of b in advance. In repeated surveys, the sample values of b remain fairly constant; or, if x is the value of y at a recent census, general knowledge of the population may suggest that b is not far from unity, so that $b = 1$ is chosen. Since the sampling theory of regression estimates when b is preassigned is both simple and informative, this case is considered first.

Theorem 2.1. *In simple random sampling, in which b_0 is preassigned constant, the linear regression estimate*

$$\bar{y}_{lr} = \bar{y} + b_0(\bar{X} - \bar{x}) \tag{2.5}$$

is unbiased, with variance

$$V(\bar{y}_{lr}) = \frac{1-f}{n} \frac{\sum_{i=1}^{N}[(y_i - \bar{Y}) - b_0(x_i - \bar{X})]^2}{N-1} \tag{2.6}$$

$$= \frac{1-f}{n} (S_y^2 - 2b_0 S_{yx} + b_0^2 S_x^2) \tag{2.7}$$

Note that no assumption is required about the relation between y and x in the finite population.

Corollary 2.2. *An unbiased sample estimate of $V(\bar{y}_{lr})$ is*

$$v(\bar{y}_{lr}) = \frac{1-f}{n} \frac{\sum_{i=1}^{n}[(y_i - \bar{y}) - b_0(x_i - \bar{x})]^2}{n-1} \tag{2.8}$$

$$= \frac{1-f}{n} (s_y^2 - 2b_0 s_{yx} + b_0^2 s_x^2) \tag{2.9}$$

Theorem 2.3. *The value of b_0 that minimizes $V(\bar{y}_{lr})$ is*

$$b_0 = B = \frac{S_{yx}}{S_x^2} \tag{2.10}$$

$$= \frac{\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})}{\sum_{i=1}^{N}(x_i - \bar{X})^2} \tag{2.11}$$

which may be called the linear regression coefficient of y on x in the finite population. Note that B does not depend on the properties of any sample that is drawn, and therefore could theoretically be preassigned. The resulting minimum variance is

$$V_{min}(\bar{y}_{lr}) = \frac{1-f}{n}\ S_y^2(1-\rho^2) \tag{2.12}$$

where ρ is the population correlation coefficient between y and x.

The same analysis may be used to show how far b_0 can depart from B without incurring a substantial loss of precision.

$$V(\bar{y}_{lr}) = \frac{1-f}{n}\left(S_y^2(1-\rho^2) + (b_0 - B)^2 S_x^2\right) \tag{2.13}$$

$$= V_{min}(\bar{y}_{lr})\left(1 + \frac{(b_0 - B)^2 S_x^2}{S_y^2(1-\rho^2)}\right) \tag{2.14}$$

Since $BS_x = \rho S_y$, this may be written

$$V(\bar{y}_{lr}) = V_{min}(\bar{y}_{lr}) + [1 + (\frac{b_0}{B} - 1)^2\left(\frac{\rho^2}{(1-\rho)^2}\right)] \tag{2.15}$$

Thus, if the proportional increase in variance is to be less than α, we must have

$$|\frac{b_0}{B} - 1| < \sqrt{\alpha(1-\rho^2)/\rho^2} \tag{2.16}$$

To ensure a small proportional increase in variance b_0/B must be close to 1 if ρ is very high but can depart substantially from 1 if ρ is only moderate.

III. Linear Regression Estimate With b From The Sample

Theorem 2.3 suggests that if b must be computed from the sample an effective estimate is likely to be the familiar least squares estimate of B, that is,

$$b = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.17}$$

The theory of linear regression plays a prominent part in statistical methodology. The standard results of this theory are not entirely suitable for sample surveys because they require the assumptions that the population regression of y on x is linear, that the residual variance

of y about the regression line is constant, and that the population is infinite. If the first two assumptions are violently wrong, a linear regression estimate will probably not be used. However, in surveys in which the regression of y on x is thought to be approximately linear, it is helpful to be able to use $\bar{y}_{lr}$ without having to assume exact linearity or constant residual variance.

Consequently we present an approach that makes no assumption of any specific relation between y_i and x_i. As in the analogous theory for the ratio estimate, only large-sample results are obtained.

With b as in (3.17), the linear regression estimator of $\bar{Y}$ in simple random samples is

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) = \bar{y} - b(\bar{x} - \bar{X}) \tag{3.18}$$

The estimator $\bar{y}_{lr}$ like $\bar{y}_R$ have a bias of order $1/n$. In finding the sampling error of $\bar{y}_{lr}$, replace the sample b by the population regression coefficient B. In Theorem 3.1 the error committed in this approximation will be shown to be of order $1/\sqrt{n}$ relative to the terms retained. We first examine the relation between b and B.

Introduce the variate e_i defined by the relation

$$e_i = y_i - \bar{Y} - B(x_i - \bar{X}) \tag{3.19}$$

Two properties of the e_i are that $\sum^N e_i = 0$ and

$$\sum^N e_i(x_i - \bar{X}) = \sum^N (y_i - \bar{Y})(x_i - \bar{X}) - B\sum^N (x_i - \bar{X})^2 = 0 \tag{3.20}$$

by definition of B. Now

$$b = \sum^N y_i(x_i - \bar{x}) / \sum^N (x_i - \bar{x})^2 = \sum^N [\bar{Y} + B(x_i - \bar{X}) + e_i] / \sum^N (x_i - \bar{x})^2 = B + \sum^N e_i(x_i - \bar{x}) / \sum^N (x_i - \bar{x})^2 \tag{3.21}$$

A result needed in Theorem 3.1 is that $(b - B)$ is of order $1/\sqrt{n}$. It is known that $\sum^N e_i(x_i - \bar{x})/(n-1)$ is an unbiased estimate of $\sum^N e_i(x_i - \bar{X})/(N-1)$. Thus, $\sum^N e_i(x_i - \bar{x})/(n-1)$ is distributed about a zero mean in repeated samples. Since the standard error of a sample covariance is known to be of order $1/\sqrt{n}$, $\sum^N e_i(x_i - \bar{x})/(n-1)$ is of order $1/\sqrt{n}$. But $\sum^N (x_i -$

$\bar{x})^2/(n-1) = s_x^2$ is of order unity. Hence $(b-B)$ is the ratio of these two quantities, is of order $1/\sqrt{n}$.

Theorem 3.1. *If b is the last squares estimate of B and*

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) \tag{3.22}$$

then in simple random samples of size n, with n large

$$V(\bar{y}_{lr}) = \frac{1-f}{n}\ S_y^2(1-\rho^2) \tag{3.23}$$

where $\rho = S_{yx}/S_yS_x$ is the population correlation between y and x.

IV. Modified Linear Regression Estimate

The performance of the linear regression estimator can be improved using the value of the known values of the population parameter of the auxiliary variable, which is positively correlated with the study variable. In this section we will present several modified linear regression estimators. Subramani et. al. (2012) proposed class of modified linear regression estimators for population mean $\bar{Y}$ is

$$\hat{\bar{Y}}_{SK} = a\frac{S_y}{C_y} + (1-a)\left(\bar{y} - \frac{b_{yx}}{\rho}(\bar{x} - \bar{X})\right) \tag{4.24}$$

where $b_{yx} = \frac{S_{yx}}{S_x^2}$, S_{xy}, S_x, S_y are population covariance and standard deviation a is a chosen scalar. For the sake of deriving its variance the estimator is written in the different form as given below

$$\hat{\bar{Y}}_{SK} = \bar{y} - S_y\left(\frac{a}{C_y}e_0 + \frac{1-a}{C_x}e_1\right) \tag{4.25}$$

where $e_0 = \frac{\bar{y}-\bar{Y}}{\bar{Y}}$, $e_1 = \frac{\bar{x}-\bar{X}}{\bar{X}}$, C_y and C_x are the coefficient of variation. Further we write $\bar{y} = \bar{Y}(1+e_0)$ and $\bar{x} = \bar{X}(1+e_1)$. It can be easily seen that $E(e_0) = E(e_1) = 0$, $E(e_0^2) = \frac{1-f}{n}C_y^2$, $E(e_1^2) = \frac{1-f}{n}C_x^2$ $E(s_0e_1) = \frac{1-f}{n}\rho C_yC_x$ Taking expectation on both sides of equation 4.25, the expected value of the proposed estimator is

$$E() \tag{4.26}$$

V. Simulation Study

Simulation study is conducted to study the performance of the percentile bootstrap confidence interval for the population mean, its coverage probability as well as the average length. Sample of size 20, 30, 40 and 50 are taken from Bivariate $(X, Y)'$ data with known mean $(10, 100)'$, known covariance matrix, and high score correlation($\rho = 0.8$). From each sample, 100 bootstrap re sample is taken, while replication is done 200 times for every case. The coverage probability is calculated and shown in Table 1

Table 1. Coverage Probability

	Sample Size n			
1-α	20	30	40	50
85%	.856	.859	.844	.853
90%	.919	.910	.913	.899
95%	.967	.942	.962	.953
99%	.977	.984	.986	.988

The length of each bootstrap confidence interval is calculated and their averaged are reported in Table 2

Table 2. Average Length

	Sample Size n			
1-α	20	30	40	50
85%	4.342	5.659	4.554	5.231
90%	4.342	5.659	4.554	5.231
95%	5.067	4.942	4.062	5.043
99%	5.136	4.541	4.086	5.222

From Table 1 we can see that the coverage probability is closer to the confidence level. The average length is about the same for every confidence level as shown in Table 2.

8

References

[1] Banerjie, J. and Tiwari, N. (2011). Improve ratio type estimator using jacknife methods of estimation. Journal of Reliability and Statistical Studies, 4(1), p. 53-63.

[2] Cochran, W. G. (1977). Sampling Techniques. Third Edition, Wiley Eastern Limited

[3] Kadilar, C. and Cingi, H. (2006) An improvement in estimating the population mean by using the corelation coefficient. Hacettepe Journal of Mathematics and Statistics, 38(2), p. 217-215

[4] Singh, G.N. (2003) On the improvement of product method of estimation in sample surveys, Journal of the Indian Society of Agricultural statistics, 56(3), p. 267-265

[5] Singh, H.P., Taylor, R. and Kakran, M.S. (2004), An improve estimator of population mean using power transformation. Journal of the Indian Society of Agricultural Statistics, 58(2), p. 223-230

[6] Subramani, J. and Kumarapandyan, G. (2012). A Class of Modified Linear Regression Estimators for Estimation of Finite Population Mean.