Comparison of Regression Estimator and Ratio Estimator: A Simulation Study

Dixi M. Paglinawan
School of Statistics
University of the Philippines Diliman

We compared ratio and regression estimators empirically based on bias and coefficient of variation. Simulation studies accounting for sampling rate, population size, heterogeneity of the auxiliary variable x, deviation from linearity and model misspecification were conducted. The study shows that ratio estimator is better than regression estimators when regression line is close to the origin. Ratio and regression estimators still work even if there is a weak linear relationship between x and y, provided that there is minimal, if not absent, model misspecification. When the relationship between the target variable and the auxiliary variable is very weak, bootstrap estimates yield lower bias. Regression estimator is generally more efficient than ratio estimator.

Keywords: auxiliary variable, ratio estimator, regression estimator, bootstrap estimator

1. Introduction

Ratio and regression estimators produce more precise estimates than the ordinary mean by invoking an auxiliary variable x in the estimation process. These estimators provide more efficient estimates among those that utilize auxiliary variables.

However, these estimators rely heavily on certain assumptions, e.g., linear dependence of the target variable y on the auxiliary variable x. According to Freedman and Navidi (1986), it is common for applied workers not to do empirical testing to verify important assumptions. Statistical inference may not be reliable if the assumptions do not hold true.

It is therefore important to evaluate the performance of ratio and regression estimators, especially when the assumption of linearity is violated. Researchers are still using these estimators even when they can only postulate the relationship

between the target and the auxiliary variables. This study aims to compare the two estimators, in terms of bias and coefficient of variation, in estimating the population mean. The comparison will provide insights on scenarios where one estimator outperforms the other. The bootstrap estimator is also considered as an alternative in cases where ratio and regression estimators fail.

2. Auxiliary Information

Statisticians studied the use of auxiliary information to increase precision in the estimation of population descriptive parameters. The auxiliary variable must be strongly correlated with the variable of interest. This information must be inexpensive to obtain or else, they defy the purpose of using an auxiliary variable. Time and energy spent to obtain the information may well be spent in obtaining large samples (Gregoire and Salas, 2009).

There are already numerous surveys conducted that utilize the effect of having an auxiliary information to estimate its desired population parameter. Even if bias is introduced, estimators that take advantage of the correlation between the auxiliary information and the variable of interest provide estimates of the population mean that have smaller variances. Therefore, the higher the correlation between x and y, the more optimal the estimate is (Lohr, 1999).

3. Estimation of Population Mean

Let $U = \{Y_1, Y_2, ..., Y_N\}$ be a finite population. The population mean, defined as

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} \tag{1}$$

with variance

$$S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \tag{2}$$

can be estimated by one or more suitable estimators for a given sampling design.

3.1. Ordinary mean

In a simple random sampling without replacement, the population mean is estimated by the sample mean defined as

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3}$$

The variance of $\bar{y}$ is

$$V(\bar{y}) = \frac{s_y^2}{n}(1 - \frac{n}{N}) \tag{4}$$

where

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{5}$$

The factor $(1-\frac{n}{N})$ is called the finite population correction. The factor disappears when the sample size is very small compared to the population size (Lohr, 1999).

3.2. Ratio estimator

Ratio estimator uses auxiliary information about the population to estimate the unknown parameter of interest. The linear relationship between the auxiliary variable and the variable of interest increases the precision of the estimate.

The ratio estimate of the population mean $\overline{Y}$ in simple random sampling is defined as

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}}\overline{X} \tag{6}$$

where $\bar{y}$ is the sample mean of Y, and $\bar{x}$ and $\overline{X}$ are the sample and the population means of X, respectively. Its variance is defined as

$$Var(\bar{y}_r) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{y}^2\left(C_x^2 + C_y^2 - 2\rho C_x C_y\right) \tag{7}$$

where

C_x is the coefficient of variation of x_i
C_y is the coefficient of variation of y_i
ρ is the sample correlation value between x and y
$\bar{y}$ is the sample mean

3.3. Regression estimator

Other than the ratio estimator, the linear regression estimator is one of the most commonly used estimators of $\overline{Y}$ given an auxiliary variable. The linear regression estimator is defined by

$$\bar{y}_{rg} = \bar{y} + b(\overline{X} - \bar{x}) \tag{8}$$

where

$$b = \frac{\sum(y_j - \bar{y})(x_j - \bar{x})}{\sum(x_j - \bar{x})^2} \tag{9}$$

is the sample regression coefficient of y on x. The variance is best approximated by

$$Var(\bar{y}_{rg}) = \frac{1-f}{n} s_y^2 [1-\rho^2] \tag{10}$$

where

$$f = \frac{n}{N} \tag{11}$$

3.4. Bootstrap

The bootstrap method could be used to make inferences about a population parameter where some of the parametric assumptions such as normality are not justified (Mooney and Duval, 1993).

Bootstrapping relies on the sampling distribution of $\hat{\theta}$ based on the samples to estimate the population parameter θ. Bootstrap creates an empirical estimate of the entire sampling distribution of a statistic by "resampling" the data with replacement an infinite number of times. The bootstrap method assumes that the sample gathered represent the population.

For case resampling, a Monte Carlo sampling technique is implemented to create an empirical estimate of the sampling distribution of $\hat{\theta}$. This method takes a random sample of size n with replacement from the population repeatedly. From each of the drawn samples, the statistic $\hat{\theta}^*$ is computed to estimate the relative frequency distribution of the sampling distribution of $\hat{\theta}$. The empirical distribution function (EDF) may not approximate the population distribution function (PDF) due to small sample size, bias sampling design, or by chance. In this case, the bootstrapped estimate is inaccurate. Unless prior information is available and parametric inference will be used, only then can the fit between EDF and PDF be improved (Mooney and Duval, 1993).

4. Simulation Study

We investigate the performance of both estimators with simulated data sets generated from the model

$$y_i = a + bx_i + h\varepsilon_i \; i = 1, 2, \ldots, N \tag{12}$$

where

- a is the intercept of the regression line
- b is the slope of the regression line, known to describe the linear relationship between the response and auxiliary variables
- x_i is the value of the auxiliary variable
- ε_i is used to introduce misspecification error
- h is used to incorporate the minimal or dominating effect of the error term
- N is the number of population

The following cases are considered:

Case 1: Comparison of the estimators considering the heterogeneity of the auxiliary variable used in the estimation process.

Case 2: Comparison of the estimators when model is correctly or severely misspecified.

Case 3: Comparison of the estimators when the linear relationship between the target variable and the auxiliary variable is weak. Two sub-cases are considered:

Case 3.1: When the relationship between the target variable and the auxiliary variable is weak but model is correctly specified.

Case 3.2: When the relationship between the target variable and the auxiliary variable is severely misspecified.

Under these scenarios, we compare the performance of both estimators based on their ability to estimate the true parameter $\overline{Y}$.

To estimate $\overline{Y}$, Monte Carlo samples are drawn at specified sampling rate. Let *M* be the number of Monte Carlo iterations. The Monte Carlo estimate of the mean is defined as:

$$\phi\hat{M}\frac{1}{M}\sum_{i=1}^{M}\hat{\phi}_i \tag{13}$$

where

$\hat{\phi}_i$ is the mean of the estimator in the ith iteration.

To evaluate the performance of the estimators, the empirical bias and coefficient of variation CV are considered. Estimators with lower empirical bias and CV are preferred. The empirical bias and CV are computed as follows:

$$bias_{est} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\phi}_i - \overline{Y}) \tag{14}$$

$$CV_{est} = \frac{1}{M}\sum_{i=1}^{M}\frac{std_i}{\hat{\phi}_i} \tag{15}$$

where

$\overline{Y}$ is the population mean

$\hat{\phi}_i$ is the mean of the estimator in the i^{th} iteration

std_i is the standard deviation of the estimator in the i^{th} iteration

From here on, we will drop the term "empirical" and use only bias and CV to refer to the empirical bias and CV of the estimators, respectively.

5. Results and Discussion

We first investigate the performance of both estimators in terms of the heterogeneity of the auxiliary variable used in the estimation process (Case 1). As we can see in Table 5.1, the estimates obtained by the ratio estimator is closer to the true value whether x have similar or different characteristics as long as the *x*'s are normally distributed and the sample size is of moderate number. A difference ranging from 0.03 to 0.18 can be observed from the bias of the estimates. When *x* is skewed, there is no significant difference between the bias of ratio and regression estimator except when the auxiliary variable is heterogeneous and the sample size used in the estimation is small. Examples of data that exhibit skewed populations are financial data such as income and assets.

Table 5.1 Bias of the Estimators per Heterogeneity of the X's for Varying Sample Size

Distribution of X	Heterogeneity of X	Small-size n = 1%		Medium-size n = 3%		Large-size n = 5%	
		Ratio	Regression	Ratio	Regression	Ratio	Regression
Symmetric	Homogeneous	0.38	0.45	0.04	0.07	0.09	0.07
	Heterogeneous	0.27	0.45	0.05	0.12	0.09	0.07
Skewed	Homogeneous	0.22	0.22	0.32	0.29	0.07	0.02
	Heterogeneous	0.13	0.24	0.31	0.32	0.04	0.01

In addition, the CV of ratio estimator is preferred when the auxiliary variable are heterogeneous and the sample size is small. Result shows that the CV of regression estimator is 0.44 while CV of ratio estimator is only 0.05 in a normally distributed heterogeneous x. When set to have a skewed distribution, CV of both ratio and regression estimators are comparable in small-size population.

Table 5.2 CV of the Estimators per Heterogeneity of the X's for Varying Sample Size

Distribution of X	Heterogeneity of X	Small-size n = 1%		Medium-size n = 3%		Large-size n = 5%	
		Ratio	Regression	Ratio	Regression	Ratio	Regression
Symmetric	Homogeneous	0.22	0.20	0.09	0.13	0.13	0.02
	Heterogeneous	0.05	0.44	0.02	0.01	0.04	0.01
Skewed	Homogeneous	1.25	0.07	0.07	0.02	0.01	0.02
	Heterogeneous	0.03	0.04	0.14	0.01	0.04	0.05

It is noticeable that regression estimator demonstrates great performance in both bias and CV when the sample size is large compared to ratio estimator. As can be seen from the previous tables, regression estimator generally outperforms ratio estimator when sample size is large.

Table 5.3 Bias of the Estimators per Model Misspecification for Varying *n* when Auxiliary Variable comes from a Symmetric or Skewed Distribution

Misspecification of the Model	Small-size n = 1%				Large-size n = 5%			
	Symmetric		Skewed		Symmetric		Skewed	
	Ratio	Regression	Ratio	Regression	Ratio	Regression	Ratio	Regression
Correctly Specified	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.00
Severely Misspecified	1.10	1.48	0.57	0.76	0.30	0.22	0.16	0.02

Next, we compare the performance of both estimators for a correctly specified or severely misspecified model (Case 2). This is the case where the current auxiliary data is sufficient to estimate $\overline{Y}$ or if there are still other variables that should have been in the model. It is apparent that the bias of both estimators blows up when the model is severely misspecified. This behavior is expected since both estimators rely on the auxiliary data to obtain the estimate of $\overline{Y}$. We also note that if the model is correctly specified, both estimators show preferable bias. Results also show that ratio estimator outperforms regression estimator in a small size setting while regression estimator has better bias in a large size setting especially when the model is severely misspecified. In terms of precision, regression estimator has better CV compared to ratio estimator.

Table 5.4 CV of the Estimators per Model Misspecification for Varying *n* when Auxiliary Variable comes from a Symmetric or Skewed Distribution

Misspecification of the Model	Small-size n = 1%				Large-size n = 5%			
	Symmetric		Skewed		Symmetric		Skewed	
	Ratio	Regression	Ratio	Regression	Ratio	Regression	Ratio	Regression
Correctly Specified	0.03	0.03	0.08	0.04	0.01	0.01	0.06	0.01
Severely Misspecified	0.16	0.11	0.05	0.01	0.33	0.05	0.00	0.03

Finally, we investigate the performance when the linear relationship between the target variable and the auxiliary variable is weak. Results show that indeed regression estimator outperforms ratio estimator when the regression line is away

from the origin (Table 5.5 and Table 5.6). In Case 3.1, ratio or regression estimators still work even if the linear relationship is weak, provided that there is minimal model misspecification. Table 5.5 shows that there is only a slight increase of bias for both estimators compared to the ordinary mean. Both estimators are also more precise than the ordinary mean.

Table 5.5 Comparison of the Estimators for a Correctly Specified Model where *X* and *Y* are Weakly Correlated

Distance of the Regression Line From the Origin	Ordinary Mean		Bootstrap		Ratio		Regression	
	Bias	CV	Bias	CV	Bias	CV	Bias	CV
Close	0.00	1.17	0.00	1.29	0.00	0.07	0.00	0.02
Away	0.00	0.09	0.00	0.09	0.01	0.02	0.00	0.02

Note: Average sample correlation based on 36 populations is $\rho = 0.02$.

In Case 3.2, however, the bias of ratio and regression estimators are greater than the ordinary mean. This is the case then where the use of ratio or regression estimator no longer gives us an efficient estimate. If there is a weak linear relationship between the auxiliary data and a different set of x's are expected to be in the model, ordinary mean would outperform both ratio and regression estimators.

It is quite noting, however, that aside from the ordinary mean, there is already a substantial decrease with the amount of bias when using the bootstrap estimate (see Table 5.6). In fact, the bias of the bootstrap estimate outperforms all other estimators. More so, the bootstrap estimator has lower CV than all other estimators when the regression line is close to the origin.

Table 5.6 Comparison of the Estimators for a Severely Misspecified Model where *X* and *Y* are Weakly Correlated

Distance of the Regression Line From the Origin	Ordinary Mean		Bootstrap		Ratio		Regression	
	Bias	CV	Bias	CV	Bias	CV	Bias	CV
Close	0.11	2.94	0.09	0.07	0.18	0.43	0.14	0.09
Away	0.11	1.25	0.09	0.93	0.19	0.08	0.14	0.03

Note: Average sample correlation based on 36 populations is $\rho = 0.01$.

Therefore, it is apparent that bootstrap estimator can provide efficient estimate compared to all other estimators discuss when there is a weak linear relationship between the auxiliary and the target variable and the regression line is close to the origin.

In general, however, regression estimator is still the most efficient estimator compared to ordinary mean, bootstrap and ratio estimators.

6. Conclusions and Recommendations

This study shows that ratio and regression estimator can still provide comparable bias to the ordinary mean when there is a weak linear relationship between x and y, given that x captures the whole of the model. Field experts should be consulted to determine which prior information should be used in estimating the population. As an alternative to ratio and regression estimator, bootstrap estimator provides better bias when there is a weak linear relationship between x and y, or that the model is greatly misspecified. In general, however, regression estimator provides more precise estimates.

It is recommended that more empirical comparisons be done for situations when the relationship of the target and auxiliary variable is not linear. Furthermore, the effect of stratification and unequal probability of selection on the performance of ratio, regression, and bootstrap estimators are also worth investigating.

ACKNOWLEDGMENT

We would like to thank Prof. Angela Nalica and Prof Kevin Carl Santos of the UP School of Statistics for their insightful ideas, and the Statistical Research and Training Center for the financial support.

REFERENCES

COCHRAN, W. G., 1977, *Sampling Techniques* (3rd ed.), New York: John Wiley.

FREEDMAN D.A. and NAVIDI W.C., 1986, R*egression Models for Adjusting the 1980 Census*, Vol. 1, No. 1, p. 311.

GREGOIRE, T. G. and SALAS, C., 2009, Ratio Estimation with Measurement Error in the Auxiliary Variate, *Biometrics*, Vol. 65, No. 2, pp. 590-598.

LOHR, S.L., 1999, *Sampling Design and Analysis*, Duxbury Press.

MOONEY, C.Z. and DUVAL R.D., 1993, *Bootstrapping A Nonparametric Approach to Statistical Inference*, Sage Publications, Inc.