# A note on double sampling estimate and post estimation with BLS outlier adjustment

Seok Dong Kim*,
Kyonggi University, Republic of Korea, poya0799@kgu.ac.kr
Sang Eun Lee,
Kyonggi University, Republic of Korea, sanglee62@kgu.ac.kr

**Abstract**

In sample design, a certain auxiliary information related to the variable of interest should be used for attaining more efficient sampling. However, some cases, the required or related information are not available in advance from the population. In that case, double sampling technique could be one of the reasonable approaches. In double sampling, getting a useful variable for second stage sample design is the most important thing but it is not quite easy and even it costs. Therefore we suggest the post estimate method applying the BLS outlier weight adjustment after the survey which is done by sample design with lack of related information from population. In this study, simulation study is performed with survey of truck transportation data, the results of double sample and suggested post estimation method are compared with RSE.

keywords : double sample, weight adjustment, post estimation, RSE

## 1. Introduction

Most of official statistics are obtained from sample survey. In sample survey, sample selection arises when the observed sample is not a random draw from the population of interest. Failure to take this selection into account can potentially lead to inconsistent and biased estimates of the parameters of interest. The first step of the sample survey is the sample design. And most important thing is the representative of sample and efficiency of estimation. So the first procedure of sample design is finding variables which are strongly related with estimating variable of interest to attain more efficient sampling. If required or related information are not available in advance from the entire population, double sampling is recommended.
For double sampling, it needs to collect certain items for a large, preliminary sample as first sample to improve on the estimations. Even from the large preliminary sample, a difficulty often encountered in first sample is still hard to collect the required information and also it costs. If required or related information from first sampling is not still enough for the representative of sampling then stratification for second sampling may not be suitable and the result estimation of that sample survey is quite often over/under estimate. So, the information from first sample should be reasonable and suitable and if not, the weight adjustment is still needed. Therefore in this study we suggest the post estimate method with BLS outlier weight adjustment after the survey which is done by simple sample design with lack of related information from population. And the simulation study is done using the survey of truck transportation data. The results of double sample and suggested post estimation method are compared with RSE. In section 2, double sampling is described, in section 3, suggested method is explained, section4 contains the simulation study and finally summary is in section5.

## 2. Estimation of Doubling Sampling

### 2.1 Double Sampling

Double sampling is a sampling method which makes use of auxiliary data where the auxiliary information is obtained through sampling. More precisely, we first take a sample of units strictly to obtain auxiliary information for sample design, and then take a second sample from first stage sample with proper sample design. It will often be the case that this second sample is a subsample of the preliminary/first sample used to acquire auxiliary information.

### 2.2 Estimation

The estimator of population total, $\tau_{dst}$ is as following

$$\hat{\tau}_{dst} = N \sum_{h=1}^{L} w_h \, \overline{y}_{dst}$$

where $\overline{y}_{dst} = \sum_{h=1}^{n_h} w_h \, \overline{y}_h$ , $\overline{y}_h = \frac{1}{n_h} \sum_{h=1}^{n_h} y_{hi}$

$w_h = \frac{n_{1h}}{n_1}$ , $n_1$ ; Total sample size of first phase sample

$n_{1h}$ ; Sample size of $h$ stratum if first phase sample

and variance estimation of $\tau_{dst}$ is

$$\hat{V}(\hat{\tau}_{dst}) = N(N - n_1)\frac{s^2}{n_1} + N^2 \sum_{h=1}^{L} \frac{n_{1h} - n_h}{n_h} \frac{w_h {s_h}^2}{n_1}$$

where $n_h$ ; sample size of $h$ stratum in second phase sample

$$\mathrm{s}^2 = \sum_{h=1}^{L} \frac{n_h - 1}{n - 1} {s_h}^2 + \sum_{h=1}^{L} \frac{n_h}{n - 1} (\overline{y}_h - \overline{y}_{dst})^2$$

$$\mathrm{s}_h{}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \overline{y}_h)^2$$

## 3. Post Estimation Method with BLS outlier Weight Adjustment

Sampling weights for the main survey are readjusted to account for non-response rate or population adjustment call it as raking. In this study, because of the not being used for required or related information at first sample design stage, stratification may not be suitable. For example, in each strata, variation of estimating variable may be very high. That is each stratum is not quite homogeneous at all for a estimating variable/parameter of interest. It is possible because related or required information are not considered from the first stage of sampling. Thus the results of estimates with sample design weight may cause the over/under estimates which are getting the unbiased estimates.

In this study we suggest that BLS outlier weight adjustment is applied to each strata for reducing the variation of the data.

3.1 Estimation

3.1.1 Estimation for usual stratified sampling

In general, estimator of population total, $\hat{\tau}_{st}$, called as Horvitz-Thompson estimator (HT) is as following

$$\hat{\tau}_{st} = N\overline{y}_{st}$$

where $\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \overline{y}_h$, $\overline{y}_h = \frac{1}{n_h}\sum_{h=1}^{n_h} y_{hi}$ ; sample mean of $h$ stratum

$L$ = number of strata
$n_h$ = sample size of $h^{th}$ strata
$y_{hi}$= i$^{th}$ observation of $h^{th}$ strata
$w_h$= design weight of $i^{th}$ strata

and variance estimation of $\hat{\tau}_{st}$ is

$$\hat{V}(\widehat{\tau_{st}}) = \sum_{h=1}^{L} {N_h}^2 \frac{N_h - n_h}{N_h}\frac{{s_h}^2}{n_h}$$

where ${s_h}^2 = \sum_{i=1}^{n_h} \frac{(y_{hi}-\overline{y}_h)^2}{n_h-1}$, $\overline{y}_h = \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h}$

$N$ = population size
$N_h$= population size of $h^{th}$ strata
$n_h$= sample size of $h^{th}$ strata

3.1.2 Post Estimation with BLS outlier

Now post estimation is adjusting weight by using outlier idea to data.
First, in each stratum, the observation which is the over the 95 percentile value is treated as outlier and adjusts the weight as

$$Outlier\ factor\ weight = w_{hi}^{OAF} = \frac{1}{w_{hi}}$$

where $h^{th}$ strata, $i^{th}$ observation treated as outlier
$w_{hi}$ is the design weight of $h^{th}$ strata, $i^{th}$ observation

This means if there exists the observation which is over the 95 percentile in that stratum then we consider that observation as belonging in take–all stratum which give to the weight becomes equal to 1.

And next step we have to reweight the rest of observations except the observations which become at take-all strata call it as non-outlier factor is as following

$$Non-outlier\ factor\ weight = w_{hi}^{NOAF} = w_{hi} + \frac{\sum_{h\in C}^{a}(w_{hi}-1)}{n_h - \alpha_h}$$

where $w_{hi}$ is the design weight on non-outlier sample
$n_h$ is the number of sample on h$^{th}$ strata
$\alpha_h$ is the number of outlier in h$^{th}$ strata
$C$ is the group treated outlier

Therefore final post estimation is as following

$$\hat{\tau}_{pst} = N\overline{y}_{pst} = N \times \frac{\sum_{h=1}^{L}\left(\sum_{i\in C}^{n_h} w_{hi}^{OAF} w_{hi} y_h + \sum_{i\in C'}^{n_{h_h}} w_{hi}^{NOAF} y_{hi}\right)}{\sum_{h}^{L}\sum_{i=1}^{n_h}\left(w_{hi}^{OAF} w_{hi} + hi\right)}$$
$$= \left(\sum_{h=1}^{L}\sum_{i\in C}^{n_h} y_{hi} + \sum_{h=l}^{L}\sum_{i\in C'}^{n_{h_h}} w_{hi}^{NOAF} y_{ij}\right)$$

where $C$ is the group of treated as outlier, $C'$ is the group treated non-outlier and variance estimation of $\hat{\tau}_{pst}$ is

$$\hat{V}\left(\hat{\tau}_{pst}\right) = \sum_{h=1}^{L-1} {N_{hc'}}^2 \frac{(N_{hc'} - n_{hc'})}{N_{hc'}} \frac{s_{hc'}^2}{n_{hc'}}$$

where $s_{hc'}^2 = \sum_{i\in C'} \frac{(y_{hi} - \bar{y}_{hc'})^2}{n_{hc'}-1}$ ; sample variance of non-outlier of $h$ stratum
$N_{hc'} = N_h - \alpha_h$ ; total number of outlier in stratum $h$
$n_{hc'} = n_h - \alpha_h$ ; sample size of non-outlier in stratum $h$

## 4. Simulation Study and Summary

Simulation study is done with transportation O/D survey data. For double sampling, in $1^{st}$ phase sample, sample sizes are allocated proportionally by region, industrial classification and the number of employee and selected by SRS. And for $2^{nd}$ phase, get the range of amounts of transportation and use them as stratification variable. Finally samples are selected by stratified SRS with region, industrial classification, the number of employees and amounts of transportation stratification variables.
And for suggested method, using the $2^{nd}$ phase sample size, sample is designed with proportional allocation by region and industrial classification and the number of employees and from each stratum, samples are selected by SRS. The results are done by 1000 replications and are shown table 4.1.

Table4.1 RSE of HT, BLS adjustment and Double sampling

| First sample | Second sample | HT | BLS | Double |
|---|---|---|---|---|
| 600 | 300 | 0.2694 | 0.1608 | 0.1899 |
| | 200 | 0.2933 | 0.1854 | 0.1931 |
| | 100 | 0.3281 | 0.1850 | 0.1905 |
| 500 | 300 | 0.2695 | 0.1720 | 0.1921 |
| | 200 | 0.3012 | 0.1820 | 0.1900 |
| | 100 | 0.3238 | 0.1801 | 0.1882 |
| 400 | 300 | 0.2710 | 0.1692 | 0.1955 |
| | 200 | 0.3002 | 0.1805 | 0.1890 |
| | 100 | 0.3318 | 0.1854 | 0.1893 |

※ note: HT – Estimation using design weight
BLS – Suggested post estimation
Double – Double sampling estimation

From table 4.1 obviously we can see the case of HT which is using design weight for estimation is worst among three of them. And between suggested post estimation and double sampling, the suggested post estimation is little bit more efficient than double sampling estimation by RSE measurement. For case of double we adopted the range of the amounts on each company for $2^{nd}$ phase sampling stratification variable. Generally in double sampling, if auxiliary variables obtained at first phase sampling stage are efficiently applied to the second phase sampling, double sampling estimation can be more efficient than others. Finally this study shows before adopting the double sampling we have to make sure we can surveyed and get the variable which highly correlated with variable of interest for $2^{nd}$ phase sampling, otherwise it will be better off only using a simple sample design and do the post adjustment for estimation.

**References**

1.Cochran,W.G., 1977. Sampling Techniques, 3rd edn. John Wiley, NY.

2.Robothama, H., Young, Z.I., Saavedra-Nievas, J.C., 2008. Jackknife method for estimating the variance of the age composition using two-phase sampling with an application to commercial catches of swordfish, Fisheries Research 93, 135-139.