Advances in Economics and Business Management (AEBM)
p-ISSN: 2394-1545; e-ISSN: 2394-1553; Volume 4, Issue 4; April-June, 2017, pp. 201-203

http://www.krishisanskriti.org/Publication.html

Regression Coefficient for Finite Population in case of Probability Proportional to Size without Replacement (PPSWOR)

Agarwal, Bhawna[1] and Sumer Singh[2]

[1]*Amity College of Commerce and Finance Amity University, Noida*
[2]*Vice-Chancellor, Sangam University, Bhilwara*

Abstract—*Estimation of population through samples is done in each and every field in market. For that, it is mandatory for a surveyor to understand that which kind of sampling methodology is to be considered and Why? Generally, when any researcher considers about random sampling, he considers SRSWR, SRSWOR, Startified sampling, Multi-stage sampling, Cluster sampling etc. But in few cases, it is required to consider Probability proportional to size sampling technique. So far, statisticians developed estimators and estimates for population mean, total, standard deviation, standard error etc. under Probability proportional to size sampling technique. Few researches have also been done for correlation coefficient but so far no proper attempt has been made to estimate regression coefficient. In this paper, the author has developed the estimator of regression coefficient and its bias.*

Keywords*: Regression Coefficient, Probability Proportional to Size without Replacement (PPSWOR), Finite Population*

1. NARRATION OF THE PROBLEM:

As discussed earlier also, no proper attempt has been made to study the regression coefficient for probability proportional to size without replacement (PPSWOR) sampling. So far, Gupta and Singh(1989) derived usual correlation coefficient in PPSWR sampling. In this paper, the authors are proposing the estimator and bias of regression coefficient where Y is a dependent variable and X is an independent variable. The expressions for bias, variance and estimator of the variance have been worked out for the regression coefficient in case of PPSWOR when the units in the sample are selected with unequal initial probabilities $\{P_i, \sum P_i = 1\}$ and the probability of drawing a specified unit of the population at a given draw changes with the draw.

Let the units in the given finite population be denoted by $U_1, U_2, \ldots, U_N$ and a sample of size n is taken with PPSWOR sampling and the measurements on variable X and Y are recorded. In what follows, we define the following:

$$t_i = \begin{cases} 1 \text{ if } U_i \text{ is included in the sample} \\ 0 \text{ otherwise} \end{cases}$$

Obviously, $E(t_i) = p_i$ (the prob. that U_i is selected) so that $E(t_i^2) = E(t_i^3) = E(t_i^4) = p_i$

$E(t_i t_j) = p_{ij}$ (the prob. that U_i and U_j both occur in the sample);

$E(t_i t_j t_k) = p_{ijk}$ (the prob. that U_i, U_j and U_k occur in the sample) and

$E(t_i t_j t_k t_l) = p_{ijkl}$ (the prob. that U_i, U_j, U_k and U_1 are included in the sample). (1)

We shall also observe the following notations throughout the paper:

$\sum_1$ for $\sum_{i=1}^{N}$; $\sum_2$ for $\sum_{i\neq j=1}^{N}$; $\sum_3$ for $\sum_{i\neq j\neq k=1}^{N}$; $\sum_4$ for $\sum_{i\neq j\neq k\neq l=1}^{N}$

2. REGRESSION COEFFICIENT

The regression coefficient of Y on X is given by

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{\sum XY - \frac{1}{n}(\sum X)(\sum Y)}{\sum X^2 - \frac{1}{n}(\sum X)^2}$$

………………. (3)

For PPSWOR, the estimator of above mentioned regression coefficient can be considered as

$$\hat{b}_{YX} = \frac{\hat{\theta}_1}{\hat{\theta}_2} \text{ ……….. (4)}$$

where $\hat{\theta}_1 = \sum_1 X_i Y_i t_i - \frac{1}{n}(\sum_1 X_i t_i)(\sum_1 Y_i t_i)$

$\hat{\theta}_2 = \sum_1 X_i^2 t_i - \frac{1}{n}(\sum_1 X_i t_i)(\sum_1 X_i t_i)$

Let $E(\hat{\theta}_1) = \theta_1$, this implies $\hat{\theta}_1 = \theta_1 + \varepsilon_1$ and $E(\hat{\theta}_2) = \theta_2$, which implies $\hat{\theta}_2 = \theta_2 + \varepsilon_2$.

Therefore, $E(\varepsilon_1)=0$, $E(\varepsilon_2)=0$, $V(\varepsilon_1)=E(\varepsilon_1^2)$, $V(\varepsilon_2)=E(\varepsilon_2^2)$, $Cov(\varepsilon_1,\varepsilon_2)=E(\varepsilon_1\,\varepsilon_2)$

Lemma 1:

The expected value of $\hat{\theta}_1$ is given by $E(\hat{\theta}_1)=\frac{1}{n}\left((n-1)\sum_1 X_iY_ip_i-\sum_2 X_iY_jp_{ij}\right)$

Proof: We have

$$E(\hat{\theta}_1)=E\left[\sum_1 X_iY_it_i-\frac{1}{n}(\sum_1 X_it_i)(\sum_1 Y_it_i)\right]$$

$$=E\left[\sum_1 X_iY_it_i-\frac{1}{n}\left(\sum_1 X_iY_i t_i^2+\sum_2 X_iY_j t_it_j\right)\right]$$

$$=E\left[\sum_1 X_iY_ip_i-\frac{1}{n}\left(\sum_1 X_iY_i p_i+\sum_2 X_iY_j p_{ij}\right)\right]$$

$$=\frac{1}{n}\left((n-1)\sum_1 X_iY_ip_i-\sum_2 X_iY_jp_{ij}\right)$$

Corollary 1:

$$E(\hat{\theta}_2)=\frac{1}{n}\left((n-1)\sum_2 X_i^2p_i-\sum_2 X_iX_jp_{ij}\right)$$

Lemma 2: The variance of ε_1 is given by

$$V(\varepsilon_1)=E(\varepsilon_1^2)=\frac{1}{n^2}\Bigg[(n-1)^2\sum_1 X_i^2Y_i^2p_i+(n^2-2n+2)\sum_2 X_iX_jY_iY_jp_{ij}+\sum_2 X_i^2Y_j^2p_{ij}-2(n-1)\sum_2 Y_i^2X_iX_jp_{ij}+\sum_3 Y_i^2X_jX_kp_{ijk}-2(n-1)\sum_2 X_i^2Y_iY_jp_{ij}+\sum_3 X_k^2Y_iY_jp_{ijk}-2(n-2)\sum_3 X_iX_kY_iY_jp_{ijk}+\sum_4 X_lX_kY_iY_jp_{ijkl}\Bigg]-\theta_1^2$$

Proof: We have $V(\varepsilon_1)=E(\varepsilon_1^2)$

$$=E\left(\sum_1 X_iY_it_i-\frac{1}{n}(\sum_1 X_it_i)(\sum_1 Y_it_i)\right)^2-\theta_1^2$$

$$=E\Big(\sum_1 X_i^2Y_i^2t_i^2+\sum_2 X_iX_jY_iY_jt_it_j+\frac{1}{n^2}\{\sum_1 X_i^2Y_i^2t_i^4+\sum_2 X_i^2Y_i^2t_i^2t_j^2+2\sum_2 X_iX_jY_i^2t_i^3t_j+\sum_3 X_jX_kY_i^2t_kt_jt_i^2+2\sum_2 X_i^2Y_iY_jt_i^3t_j+\sum_3 X_k^2Y_iY_jt_k^2t_it_j+4\sum_3 X_iX_kY_iY_jt_i^2t_jt_k+2\sum_2 X_iX_jY_iY_jt_i^2t_j^2+\sum_4 X_kX_lY_iY_jt_it_jt_kt_l\}-\frac{2}{N}\{\sum_1 X_i^2Y_i^2t_i^4+\sum_2 X_iX_jY_iY_jt_i^2t_j^2+\sum_2 X_i^2Y_iY_jt_i^3t_j+\sum_2 X_iX_jY_j^2t_j^3t_i+\sum_3 X_iX_jY_iY_kt_i^2t_jt_k\}\Big)-\theta_1^2$$

Now using the relations given in (1), $V(\varepsilon_1)$ is given by

$$V(\varepsilon_1)=\sum_1 X_i^2Y_i^2p_i+\sum_2 X_iX_jY_iY_jp_{ij}+\frac{1}{n^2}\{\sum_1 X_i^2Y_i^2p_i+\sum_2 X_i^2Y_j^2p_{ij}+\sum_3 Y_i^2X_jX_kp_{ijk}+2\sum_2 Y_i^2X_iX_jp_{ij}+2\sum_2 X_i^2Y_iY_jp_{ij}+\sum_3 X_k^2Y_iY_jp_{ijk}+4\sum_3 X_iX_kY_iY_jp_{ijk}+2\sum_2 X_iX_jY_iY_jp_{ij}+\sum_4 X_lX_kY_iY_jp_{ijkl}\}-\frac{2}{n}\{\sum_1 X_i^2Y_i^2p_i+\sum_2 X_iX_jY_iY_jp_{ij}+\sum_2 X_i^2Y_iY_jp_{ij}+\sum_2 Y_i^2X_iX_jp_{ij}+\sum_3 X_iX_jY_iY_kp_{ijk}\}-\theta_1^2$$

Corollary 2: In case of PPSWOR sampling, $V(\varepsilon_2)$ can be put as

$$V(\varepsilon_2)=E(\varepsilon_2^2)$$

$$=\frac{1}{n^2}\left[(n-1)^2\sum_1 X_i^4p_i+(n^2-2n+3)\sum_2 X_i^2X_j^2p_{ij}-4(n-1)\sum_2 X_i^3X_jp_{ij}-2(n-3)\sum_3 X_i^2X_jX_kp_{ijk}+\sum_4 X_iX_jX_kX_lp_{ijkl}\right]-\theta_2^2$$

Lemma 3:

The covariance between ε_1 and ε_2 can be easily obtained as

$$Cov(\varepsilon_1,\varepsilon_2)=\frac{1}{n^2}\Bigg[(n-1)^2\sum_1 X_i^3Y_ip_i+(n^2-2n+3)\sum_2 X_i^2X_jY_jp_{ij}-(n-1)\sum_2 X_i^3Y_jp_{ij}-(n-3)\sum_3 X_i^2X_j\,Y_k\,p_{ijk}-3(n-1)\sum_2 X_i^2X_jY_ip_{ij}-(n-3)\sum_3 X_iX_jX_kY_kp_{ijk}+\sum_4 X_iX_jX_kY_lp_{ijkl}\Bigg]-\theta_1\theta_2$$

On making use of the relations given in (1), $Cov(\varepsilon_1,\varepsilon_2)$ can be put as

$$Cov(\varepsilon_1,\varepsilon_2)=\sum_1 X_i^3Y_ip_i+\sum_2 X_i^2X_jY_jp_{ij}-\frac{1}{n}\Bigg\{\sum_1 X_i^3Y_ip_i+\sum_2 X_i^2X_jY_jp_{ij}+\sum_2 X_i^3Y_jp_{ij}+\sum_3 X_i^2X_jY_kp_{ijk}+\sum_2 X_i^2X_jY_ip_{ij}+\sum_1 X_i^3Y_ip_i+\sum_2 X_i^2X_jY_jp_{ij}+2\sum_2 X_i^2X_jY_ip_{ij}+\sum_3 X_iX_jX_kY_kp_{ijk}\Bigg\}$$

$$- \frac{1}{n^2}\left\{\sum_1 X_i^3 Y_j p_{ij} + \sum_1 X_i^3 Y_i p_i + 3\sum_2 X_i^2 X_j Y_j p_{ij} + 3\sum_2 X_i^2 X_j Y_i p_{ij} + 3\sum_3 X_i^2 X_j Y_k p_{ijk} + 3\sum_3 X_i X_j X_k Y_k p_{ijk} + \sum_4 X_i X_j X_i X_l p_{ijkl}\right\} - \theta_1\theta_2$$

3. BIAS OF B_{YX}:

Bias (b_{YX}) = $E(\hat{b}_{YX}) - \beta_{YX}$

$$= E\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right) - \beta_{YX}$$

$$= E\left(\frac{\theta_1 + \varepsilon_1}{\theta_2 + \varepsilon_2}\right) - \beta_{YX}$$

$$= \frac{\theta_1}{\theta_2} E\left(\frac{1+\frac{\varepsilon_1}{\theta_1}}{1+\frac{\varepsilon_2}{\theta_2}}\right) = \frac{\theta_1}{\theta_2} E\left[(1+\frac{\varepsilon_1}{\theta_1})(1+\frac{\varepsilon_2}{\theta_2})^{-1}\right] - \beta_{YX}$$

Now, it is assumed that the sample size is sufficiently large and expansion as a convergent series. $E(\hat{b}_{YX})$ will be

$$E(\hat{b}_{YX}) = \frac{\theta_1}{\theta_2}\left[1 + \frac{E(\varepsilon_1)}{\theta_1} - \frac{E(\varepsilon_2)}{\theta_2} - \frac{E(\varepsilon_1\varepsilon_2)}{\theta_1\theta_2} + \frac{E(\varepsilon_2^2)}{\theta_2^2}\right] - \beta_{YX}$$

By substituting expressions from Lemma and Corollary, we can get Bias (b_{YX}) = $E(\hat{b}_{YX}) - \beta_{YX}$.

4. CONCLUSION

The authors have proposed the estimator and bias of regression coefficient. These all are derived logically and recommended to the researcher for an empirical investigation of the derived estimators and bias.

REFERENCES

[1] Gupta, J.P. & Singh, R. (1989). Usual correlation coefficient in PPSWR sampling, Jour. Ind. Statist. Assoc., Vol.27, pp 13-16.

[2] Midzuno, H.(1950), An outline of the theory of sampling systems.Ann.Inst.Statist.Math., Vol. 1,pp 149-156.

[3] Sukhatme,P.V. & Sukhatme, B.V.(1970). Sampling theory of surveys with applications.Asia Publishing House, India.