A New Method to Estimate the Low Income Proportion: Monte Carlo Simulations

Encarnación Álvarez, Rosa M. García-Fernández, Juan F. Muñoz

Open Science Index, Economics and Management Engineering Vol:8, No:10, 2014 publications.waset.org/9999583.pdf

Abstract—Estimation of a proportion has many applications in economics and social studies. A common application is the estimation of the low income proportion, which gives the proportion of people classified as poor into a population. In this paper, we present this poverty indicator and propose to use the logistic regression estimator for the problem of estimating the low income proportion. Various sampling designs are presented. Assuming a real data set obtained from the European Survey on Income and Living Conditions, Monte Carlo simulation studies are carried out to analyze the empirical performance of the logistic regression estimator under the various sampling designs considered in this paper. Results derived from Monte Carlo simulation studies indicate that the logistic regression estimator can be more accurate than the customary estimator under the various sampling designs considered in this paper. The stratified sampling design can also provide more accurate results.

Keywords—Poverty line, stratified sampling, Lahiri method, Midzuno method, Logistic regression estimator.

I. Introduction

MANY social and economic indicators are based upon binary variables. In addition, they may require the use of proportions to obtain such indicators. The aim of this paper is to estimate the low income proportion, which is defined as the proportion of individuals falling below the official poverty line. The low income proportion is an example of poverty indicator based upon binary variables.

As commented, the low income proportion is based upon the official poverty line, which can be defined as a threshold below which people are classified as poor. A relative poverty line is generally obtained by using indicators based on variables such as income or expenditures. In general, the relative poverty line is fixed as a percentage of the median of an economic variable. Some percentages used by many statistical agencies are the 50% and the 60%. For the problem of studying the severe poverty, a percentage of 30% is commonly considered.

In the literature, numerous references discuss about the low income proportion and related poverty indicators. For instance, some relevant references are [3], [6], [9], [11], etc.

Many countries and organizations carry out poverty studies. For instance, one of the most important objectives of the Millennium Development Goals is to eradicate the extreme poverty. Moreover, the Europe 2020 strategy establishes that at least 20 million people should lift out on poverty and social exclusion (see http://ec.europa.eu/europe2020). The main aim of this paper is to analyze the logistic regression estimator (see [1], [5] and [8]) for the problem of estimating the low income proportion and assuming different sampling designs. The logistic regression estimator is compared to the customary Horvitz-Thompson estimator ([7]). As far as the sampling designs is concerned, we consider the stratified sampling design (see [12], [13], [14], [15]) and the Lahiri, Midzuno and Poisson methods (see [15]).

The logistic regression estimator is an estimation method which assumes auxiliary information at the estimation stage. In particular, this estimator assumes quantitative auxiliary variables and a logistic model. Note that algorithms, such as Newton-Raphson, are generally used to estimate the parameter associated to this econometric model. This process consists on maximizing the likelihood function numerically ([8]).

Note that the auxiliary information can be given by auxiliary variables related to the variable of interest. Official surveys on income and living conditions may contain additional variables related to the variable used in the calculation of the low income proportion. Such additional variables could be used to improve the estimation of the low income proportion. For instance, [10] used various auxiliary variables with data derived from the European Survey on Income and Living Conditions and for the problem of estimating poverty indicators in small areas. Moreover, [1] analyze the performance of the logistic regression estimator of the low income proportion under simple random sampling without replacement. In this paper, we analyze the estimation of this parameter under additional sampling designs.

This paper is organized as follows. Estimators of the low income proportion are defined in Section II. In Section III, we introduce some sampling designs commonly used in survey sampling, and relevant references are also given. In Section IV, we analyze the performance of estimators defined in Section II and assuming the sampling designs described in Section III. The main conclusions are summarized in Section V.

II. The low income proportion

In this section, we first describe the main notation used in the context of survey sampling and define the low income proportion. Let $U = \{1, \ldots, N\}$ be a finite population with N individuals. The low income proportion P can be defined as $P = N_p/N$, where N_p is the number of poor in the population or the number of people at risk of poverty. Let y be a quantitative variable of welfare. For example, y can be income

E. Álvarez is with the Department of Quantitative Methods in Economics and Business, University of Granada, Granada, CP 18071, Spain (e-mail: encarniav@ugr.es).

R.M. García-Fernández is with the Department of Quantitative Methods in Economics and Business, University of Granada, Granada, CP 18071, Spain (e-mail: rosamgf@ugr.es).

J.F. Muñoz is with the Department of Quantitative Methods in Economics and Business, University of Granada, Granada, CP 18071, Spain (e-mail: jfmunoz@ugr.es).

or expenditure. Let L be a relative poverty line. L is commonly used to classify the population into poor and non-poor, i.e., an individual is considered as poor if its income or expenditure is less than the poverty line. Assuming this scenario, we can calculate the number of poor as $N_p = \sum_{i \in U} A_i$, where $A_i = \delta(y_i \leq L) = 1$ if the ith individual is classified as poor and $A_i = 0$ otherwise. $\delta(\cdot)$ is the indicator variable, which takes the value 1 if its argument is true and 0 otherwise.

We assume that the poverty line (L) is fixed by the corresponding authority at some official quantity. Note that statistical agencies usually define the poverty line as the 60% of the median income.

In the survey sampling context, the parameter P is estimated by using the information collected from a sample, which we denote as s, and which has size n. The sample s is selected from the population U according to a probability sampling design with sampling weights given by $d_i = 1/\pi_i$, where π_i is the first-order inclusion probability for the ith individual in the population.

In this paper, we analyze various probability sampling designs in such a way that the performance of estimators of the low income proportion can be compared under different scenarios. The probability sampling designs used in this paper are described in Section III. First, we describe the estimators of the low income proportion used in this paper.

The traditional estimator of P is the Horvitz-Thompson estimator (see [7]), which is given by

$$\widehat{p} = \frac{1}{N} \sum_{i \in s} d_i A_i. \tag{1}$$

The availability of auxiliary variables is quite common. For instance, official surveys on income and living conditions may contain additional variables, which could be used at the estimation stage to improve the estimation of the low income proportion. For this purpose, the auxiliary variables need to be observed at the population level. We assume the existence of a quantitative auxiliary variable related to y. This auxiliary variable will be denoted as x. The logistic regression estimator (see [5] and [8]) could be used to estimate the low income proportion in this situation, hence we propose to use this estimator for the problem of estimating the low income proportion. The main contribution of this paper consists on analyzing this estimator for the problem of estimating the low income proportion and assuming different sampling designs.

The logistic regression estimator is given by

$$\widehat{p}_L = \frac{1}{N} \left(\sum_{i \in U} \widehat{\mu}_i + \sum_{i \in s} d_i (A_i - \widehat{\mu}_i), \right) \tag{2}$$

where

$$\widehat{\mu}_i = \frac{\exp(x_i \widehat{\theta})}{1 + \exp(x_i \widehat{\theta})}, \quad i = 1, \ldots, N,$$

are the predicted values based upon the logistic regression model

$$Pr(A_i = 1) = \quad \mu_i = \frac{\exp(x_i \theta)}{1 + \exp(x_i \theta)}$$

$$Pr(A_i = 0) = \quad 1 - \mu_i = 1 - Pr(A_i = 1)$$

The maximum likelihood method can be used to estimate the parameter θ. The corresponding likelihood function is given by (see, also, [8], [13])

$$L(\theta) = \Pi_{i \in U_1} \mu_i \Pi_{i \in U_0} (1 - \mu_i),$$

where $U_1 = \{i : i \in U \text{and } A_i = 1\}$ and $U_0 = U - U_1$. The Newton-Raphson algorithm can be used to find the solution to the maximization problem.

III. Some sampling designs

In this section, we define various sampling designs used in the context of survey sampling.

First, we use simple random sampling without replacement, which can be considered as the simplest sampling design (see [13] and [15]). In this situation, the first inclusion probabilities are given by $\pi_i = n/N$, i.e., they are constant for the various individuals in the population. This also implies that the sampling weights are constant, and they are given by $d_i = N/n$. Furthermore, alternative expressions for the customary and logistic regression estimators can be given when the sample s is selected under simple random sampling without replacement. In particular, the customary Horvitz-Thompson estimator is given by $\widehat{p} = n_p/n$, where $n_p = \sum_{i \in s} A_i$ denotes the number of poor in the sample. As far as the logistic regression estimator is concerned, the expression under simple random sampling is given by

$$\widehat{p}_L = \widehat{p} + \overline{\mu} - \widehat{\overline{\mu}},$$

where $\overline{\mu} = N^{-1} \sum_{i \in U} \widehat{\mu}_i$ and $\widehat{\overline{\mu}} = n^{-1} \sum_{i \in s} \widehat{\mu}_i$.

We previously commented that the sampling weights are constants under simple random sampling without replacement, and such weights are generally unequal under alternative sampling designs. We now describe some sampling designs with unequal sampling designs. Stratified sampling is a very common sampling design used in practice ([15]). This method consists on dividing the population U into various strata, and then samples are selected from each stratum. Stratified sampling can give better results than simple random sampling without replacement, as discussed by [13]. Assuming stratified sampling, the concept of allocation plays an important role. Uniform and proportional allocations ([12]) are common types of allocation. In this paper, we also analyze numerically the logistic regression estimator (see Section IV) under different types of allocation.

In the literature, there exit other many sampling designs with unequal probabilities. In this paper, we use the Lahiri, Midzuno and Poisson methods. Such methods are described by [15], and analytical expressions for the inclusion probabilities can be also seen in this reference.

IV. Monte Carlo Simulations

In this section, the empirical performance of the various estimators of the low income proportion P is analyzed via Monte Carlo simulation studies. The various sampling design discussed in Section III are also considered. In this simulation study, we use a population based upon real data set obtained from an official survey on income and living conditions. In particular, we use microdata of the 2012 Spanish Survey on Income and Living Conditions (ES-SILC), which are supplied by the Spanish National Statistics Institute. For our simulation study, we considered the $N = 28210$ data collected from the survey as a population, from which samples are selected. The aim in this population is to estimate the low income proportion by using the definition given by Eurostat, i.e., the 60% of the median of the equivalised net income. We considered two auxiliary variables: the employee cash income (population named as ES-SILC-1) and the tax on income contributions (population named as ES-SILC-2).

We considered the customary estimator of the low income proportion defined by (1) and the logistic regression estimator defined by (2). As far as the sampling design is concerned, we considered simple random sampling without replacement (denoted as S), stratified sampling with uniform (T_U) and proportional (T_P) allocation, and the Lahiri (L), Midzuno (M) and Poisson (P) methods.

The empirical relative bias (RB) and the empirical relative root mean square error (RRMSE) are the measured used to compare the various estimators of P under the various sampling designs, where

$$RB = \frac{E[\widetilde{p}] - P}{P} \quad ; \quad RRMSE = \frac{\sqrt{MSE[\widetilde{p}]}}{P},$$

$\widetilde{p}$ is a given estimator of P, and $E[\cdot]$ and $MSE[\cdot]$ are, respectively, the empirical expectation and mean square error based on $R = 1000$ simulation runs, i.e.,

$$E[\widetilde{p}] = \frac{1}{R}\sum_{r=1}^{R}\widetilde{p}(r) \quad ; \quad MSE[\widetilde{p}] = \frac{1}{R}\sum_{r=1}^{R}(\widetilde{p}(r) - P)^2,$$

where $\widetilde{p}(r)$ denotes the value of the estimator $\widetilde{p}$ at the rth simulation run.

Note that the measures RB and RRMSE are very common for the problem of comparing the precision of estimators. For instance, such measures have been used by [2], [4], [12], [14], etc.

TABLE I
VALUES OF $RRMSE$ FOR THE VARIOUS ESTIMATORS OF P AND ASSUMING DIFFERENT SAMPLING DESIGNS (SD). $f = n/N$ IS THE SAMPLING FRACTION

		Sampling fraction f (%)							
		ES-SILC-1				ES-SILC-2			
SD	Estim.	1	5	10	20	1	5	10	20
S	$\widehat{p}$	11.6	5.0	3.6	2.5	11.6	5.0	3.7	2.4
	$\widehat{p}_L$	11.8	4.9	3.5	2.4	10.9	4.8	3.5	2.2
T_U	$\widehat{p}$	11.3	4.6	3.3	2.1	11.2	4.6	3.3	2.1
	$\widehat{p}_L$	11.3	4.5	3.2	2.0	10.9	4.5	3.3	2.0
T_P	$\widehat{p}$	11.2	4.4	3.1	2.0	11.0	4.5	3.1	2.0
	$\widehat{p}_L$	10.9	4.3	3.0	1.9	10.6	4.4	3.0	1.8
L	$\widehat{p}$	11.5	4.8	3.5	2.3	11.4	4.8	3.5	2.3
	$\widehat{p}_L$	11.6	4.7	3.4	2.2	11.0	4.7	3.4	2.2
M	$\widehat{p}$	11.6	5.0	3.5	2.5	11.5	4.9	3.6	2.3
	$\widehat{p}_L$	11.7	5.0	3.5	2.4	11.1	4.7	3.5	2.2
P	$\widehat{p}$	11.7	5.1	3.7	2.6	11.6	5.0	3.8	2.4
	$\widehat{p}_L$	11.9	5.0	3.6	2.5	11.1	4.9	3.6	2.3

Empirical biases of estimators discussed in this paper are negligible, since values of RB are all less than 1%. This implies that estimators have a good performance in terms of bias, and for this reason values of RB are omitted.

In Table I we can observe the values of RRMSE for the populations ES-SILC-1 and ES-SILC-2. First, we observe that the logistic regression estimator ($\widehat{p}_L$) is generally more efficient than the customary Horvitz-Thompson estimator ($\widehat{p}$). In addition, we observe that the best results are obtained when using stratified sampling (T_U and T_P). If we compare the uniform allocation (T_U) to proportional allocation (T_P), we observe that the most efficient estimators of the low income proportion are obtained when using proportional allocation. Results derived from the Lahiri and the Midzuno methods are slightly better than the results obtained under simple random sampling without replacement. Finally, we observe that the Poisson method is slightly less efficient than the results obtained under simple random sampling without replacement.

V. Conclusion

This paper discusses the estimation of the low income proportion. We propose to use the logistic regression estimator, which uses auxiliary information at the estimation stage and more accurate results can be obtained. The main purpose of this paper is to evaluate the performance of the logistic regression estimator of the low income proportion under different sampling designs. Results are also compared to the customary Horvitz-Thompson estimator.

The empirical performance of estimators of the low income proportion and assuming different sampling designs is a topic which has not been studied previously in the literature.

Monte Carlo simulation studies have been carried out to compare the performance of the various estimators under different scenarios. Simulation studies are based upon a real data set extracted from the ES-SILC, and they are used to evaluate the performance of various estimators under this real situation.

First, we observed that the various estimators have a good performance in terms of relative biases. We also observed that the logistic regression estimator performs better than the customary estimator. Stratified sampling with proportional allocation is the sampling design that provide the best results in terms of RRMSE.

In summary, results derided from Monte Carlo simulation studies indicate that the logistic regression estimator can be an alternative estimation method for the problem of estimating the low income proportion. Better results can be also obtained if complex sampling designs are used.

Acknowledgment

This work is supported by grant P11-SEJ-7090 (University of Granada) of the Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía).

References

[1] E. Álvarez, R.M. García-Fernández, J.F. Muñoz and F.J. Blanco-Encomienda, "On estimating the headcount index by using the logistic regression estimator". *International Journal of Mathematical, Computational, Physical and Quantum Engineering*, 8(8),pp. 1039–1041, 2014.

[2] J. Chen and R.R. Sitter, ”A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys”. *Statistica Sinica*, 9, pp. 385-406, 1999.

[3] E. Crettaz and C. Suter, ”The impact of Adaptive Preferences on Subjective Indicators: An Analysis of Poverty Indicators”. *Social Indicators Research*, 114, pp. 139-152, 2013.

[4] J.C. Deville and C.E. Särndal, ”Calibration estimators in survey sampling”. *Journal of the American Statistical Association*, 87, pp. 376-382, 1992.

[5] P. Duchesne, ”Estimation of a proportion with survey data”. *Journal of Statistics Education*, 11, pp. 1-24, 2003.

[6] F. Giambona and E. Vassallo, ”Composite Indicator of Social Inclusion for European Countries”. *Social Indicators Research*, 116, pp. 269-293, 2014.

[7] D.G. Horvitz and D.J. Thompson, ”A generalization of sampling without replacement from a finite universe”. *Journal of the American Statistical Association*, 47, pp. 663-685, 1952.

[8] R.Lehtonen and A. Veijanen, ”On multinomial logistic generalized regression estimators”, *Survey Methodology*, 24, pp. 51-55, 1998.

[9] M. Medeiros, ”The Rich and the Poor: the Construction of an Affluence Line from the Poverty line”. *Social Indicators Research*, 78, pp. 1-18, 2006.

[10] I. Molina and J.N.K. Rao, ”Small area estimation of poverty indicators”, *The Canadian Journal of Statistics*, 38, pp. 369-385, 2010.

[11] J. Navicke, O. Rastrigina and H. Sutherland, ”Nowcasting Indicators of Poverty Risk in the European Union: A Microsimulation Approach”. *Social Indicators Research*, doi: 10.1007/s11205-013-04918. 2013

[12] J.N.K. Rao, J.G. Kovar and H.J. Mantel, ”On estimating distribution function and quantiles from survey data using auxiliary information”. *Biometrika*, 77, pp. 365-375, 1990

[13] C.E. Särndal, B. Swensson and J. Wretman, *Model Assisted Survey sampling*, Springer Verlag, 1992.

[14] P.L.D. Silva and C.J. Skinner, ”Estimating distribution function with auxiliary information using poststratification”. *Journal of Official Statistics*, 11, pp. 277-294, 1995.

[15] S. Singh, *Advanced sampling theory with application: how Michael selected Amy*, Kluver Academic Publisher, 2003.

Encarnación Álvarez is a lecturer in the Department of Quantitative Method in Economics and Business at the University of Granada in Granada, Spain. Her research is about the estimation of proportions and applications in poverty.

Rosa M. García-Fernández is associate professor in the Department of Quantitative Method in Economics and Business at the University of Granada in Granada, Spain. Her research is about the analysis and study of the poverty and inequality.

Juan F. Muñoz is associate professor in the Department of Quantitative Method in Economics and Business at the University of Granada in Granada, Spain. His research is about quantitative methods used for the estimation of parameters.