International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:8, No:10, 2014

The Effect of Outliers on the Economic and Social Survey on Income and Living Conditions

Encarnación Álvarez, Rosa M. García-Fernández, Francisco J. Blanco-Encomienda, Juan F. Muñoz

Abstract—The European Union Survey on Income and Living Conditions (EU-SILC) is a popular survey which provides information on income, poverty, social exclusion and living conditions of households and individuals in the European Union. The EU-SILC contains variables which may contain outliers. The presence of outliers can have an impact on the measures and indicators used by the EU-SILC. In this paper, we used data sets from various countries to analyze the presence of outliers. In addition, we obtain some indicators after removing these outliers, and a comparison between both situations can be observed. Finally, some conclusions are obtained.

Keywords—Headcount index, poverty line, risk of poverty, skewness coefficient.

I. Introduction

POVERTY studies are very common among countries and statistical agencies, and this issue is due to the fact that the fight against poverty and social exclusion is in a priority place in the political agendas. For instance, the problem of eradicating the extreme poverty is the first of the Millennium Development Goals, whereas the Europe 2020 strategy establishes that at least 20 million people should lift out on poverty and social exclusion.

Poverty studies are generally based on surveys. For instance, the European Union Survey on Income and Living Conditions (EU-SILC), coordinated by Eurostat, provides information on income, poverty, social exclusion and living conditions of households and individuals in the European Union. Specifically, the survey provides two types of data:

1) cross-sectional data pertaining to a certain time period with variables on income and other living conditions, and
2) longitudinal data pertaining to individual-level changes over time, observed periodically - usually over four years.

According to the EU-SILC, the proportion of people falling below the poverty line is around 17% in the Euro area. This proportion is commonly named as the headcount index, risk of poverty or proportion of poor. An individual is considered as poor if his/her income is less than the official poverty line. Poverty line and risk of poverty are common concepts used by poverty studies. Relevant references that define and describe such concepts related to poverty are [2], [3], [8], [12], [13], [16], [19], [22] and [23].

It is quite common that surveys used by poverty studies contain various variables, and some of them can have a strong relationship with respect to the variable of interest. For example, this situation can be observed by the EU-SILC. The variable of interest in this survey is the equivalised net income, since this variable is used for the problem of estimating the poverty line and the poverty risk. The EU-SILC also contains information related to the income sources on which taxes are paid. In this paper, we can observe a strong relationship between such variables for various countries from the European Union. Note that the relationship between two variables is measured in terms of the linear correlation coefficient.

The linear correlation coefficient can play an important role in poverty studies, since populations with large values of the linear correlation coefficient can provide more accurate estimation methods for the various poverty indicators. In particular, the additional variables with a strong relationship with respect to the variable of interest can be used at the estimation stage to improve the estimation of poverty indicators. This technique of using estimation methods based on auxiliary variables is quite common in the context of survey sampling, and the main reason is probably due to the fact that desirable results are generally obtained.

Many estimation methods based on additional variables exist. For the problem of estimating a population mean or a population total, the most known estimation methods are the ratio type estimator and the regression type estimator (see [26]). Recently, the calibration method [9] and the pseudo empirical likelihood method [6] were also proposed for the problem of estimating a population mean.

For the problem of estimating the distribution function and quantiles, relevant estimation methods based on auxiliary variables are presented by [5], [25] and [27]. [10] conducted an extensive review of estimators of the distribution function and quantiles based on auxiliary variables.

For the problem of estimating a proportion, the logistic regression estimator (see, for example, [11] and [21]) is also quite common in the context of survey sampling.

Finally, [23] proposed estimators of some poverty measures based on auxiliary variables. The poverty measures discussed by [23] are based on the family of Foster-Greer-Thorbecke (FGT) poverty measures, which was proposed by [14], and

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).

F.J. Blanco-Encomienda is with the Department of Quantitative Methods in Economics and Business, University of Granada, Granada, CP 18071, Spain (e-mail: jble@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).

subsequently have been widely used in many references related to poverty studies (see, for example, [32], [20], [18], [15], etc). Note that the poverty risk discussed in this paper is also included into the family of FGT poverty measures.

Among the variables related to poverty studies we can find some of them, such as income or expenditure, that have highly skewed distributions, and the presence of outliers is quite common in these situations. For instance, [7], [17] and [30] deal with distributions in the presence of outliers, and they trimmed 1% of the upper and lower tails of the income distributions in order to reduce the impact of outliers on various poverty measures. From this issue we can conclude that outliers can have an impact on various indicators and measures used in poverty studies. For example, if we are interested in using a measure of location, the median is less sensitive to outliers than the mean. The most known indicator used to measure the skewness of a given population is the popular skewness coefficient. This indicator can also be affected by the presence of outliers in the population. In this paper, we also analyze the impact on the skewness coefficient when outliers are removed.

The various studies discussed in this paper are based on real data sets extracted from the 2011 European Union Survey on Income and Living Conditions. In particular, we considered data selected from various countries. This paper is organized as follows. First, we use data from various countries of the Eurozone to analyze the presence of outliers. Then, some statistical and poverty measures are obtained after removing outliers, in such a way that the impact of outliers on the various measures can be observed. Specifically, we analyze the impact of outliers on the linear correlation coefficient, the risk of poverty and the skewness coefficient. Finally, some conclusions are also given.

II. Results derived from the EU-SILC

Assuming the 2011 EU-SILC, we considered real data from various countries in the European Union. In particular, we considered data collected by the EU-SILC in the following countries: Belgium, Bulgaria, Italy, Lithuania, Poland, Slovenia, Spain and United Kingdom (UK). For each country, we used information related to two variables: the equivalised net income and the tax on income. The equivalised net income is the variable used to obtain the risk of poverty, whereas the tax on income can be used as auxiliary variable by estimation methods based on this type of variables. In this regard, the regression and the logistic regression estimators are examples of methods based on auxiliary variables (see [1], [2], [21], [26] and [28]).

In this section, we analyze the presence of outliers in the variable equivalised net income and for the various countries previously commented. Outliers appeared in the data sets from Belgium, Bulgaria, Italy, Poland and UK, whereas the variable equivalised net income does not contain outliers for the data sets from Slovenia, Lithuania and Spain.

In Fig. 1 and Fig. 2 we can see a scatter plot for the data set obtained from UK and Belgium. We observe the presence of outliers in these populations, which is shown here to illustrate

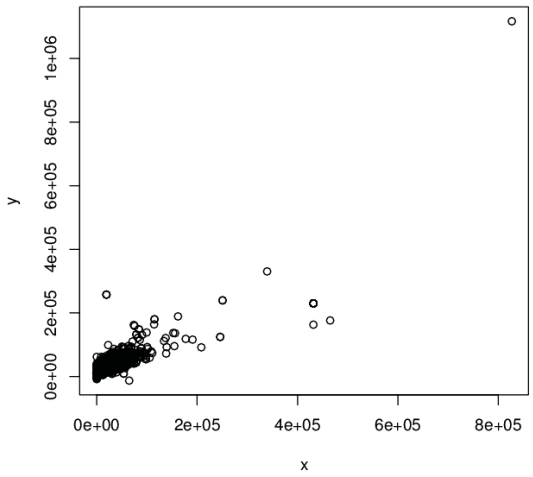

Fig. 1. Scatter plot for the data set from UK.

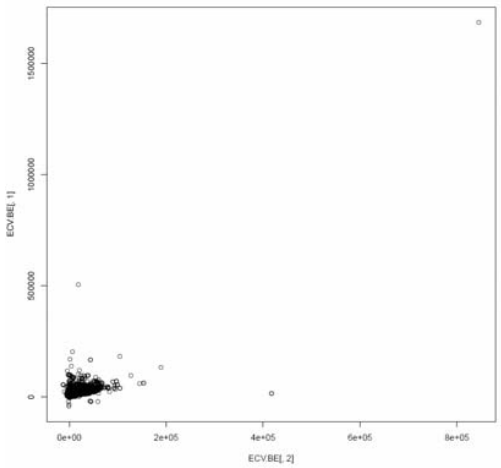

Fig. 2. Scatter plot for the data set from Belgium.

the situation for given countries. Outliers can have an impact on various measures. For this reason, we remove outliers in the various populations and compare results.

Table I gives statistical and poverty measures for the various countries from the EU-SILC analyzed in this paper. The sample size is denoted as n. ρ indicates the linear correlation coefficient between the variables, i.e., the equivalised net income and the tax on income. The risk of poverty is given by P. References related to proportions and estimation of proportions are [2], [4], [24], [29] and [31]. Finally, g is the skewness coefficient. From Table I and for the various countries, we observe a large relationship between the equivalised net income and the tax on income. The largest value ($\rho = 0.88$) is observed for the data set from Italy. This implies that the use of methods based on auxiliary variables can give desirable results, since such methods assume a strong relationship between the variables. However, outliers can have an impact on the linear correlation coefficient, and

the performance of methods based on auxiliary variables can be affected by this issue.

TABLE I
VALUES OF n, ρ, P AND g FOR VARIOUS COUNTRIES FROM THE EU-SILC.

Country	n	ρ	P	g
Belgium	9305	0.65	14.6	51.6
Bulgaria	12469	0.70	22.1	18.3
Italy	27553	0.88	18.9	23.1
Lithuania	8421	0.66	18.7	5.6
Poland	22113	0.82	17.7	11.0
Slovenia	17612	0.83	12.2	2.6
Spain	28210	0.65	17.7	1.9
UK	10586	0.83	17.2	20.5

As we previously commented, the risk of poverty is 17% in the Euro area. If we analyze this poverty measure at a national level, we observe (from Table I) that Bulgaria, Italy, Lithuania, Poland, Spain and UK have a risk of poverty larger than 17% observed in the Euro area. Only Belgium and Slovenia have a risk of poverty smaller than the average observed in this area. The risk of poverty is based on the median, which is a measure that is not affected by outliers. For this reason, we expect that the various countries have similar risks of poverty after removing outliers.

Finally, we computed the skewness coefficient g. Results can also be observed in Table I. The countries with large values of g are Belgium, Bulgaria, Italy, Poland and UK. We can see that countries with large values of g are the same than countries with outliers. For this reason, we expect that the outliers observed in the various countries can have an impact on the skewness coefficients. This issue is analyzed in the next section.

III. RESULTS AFTER REMOVING OUTLIERS

Assuming data sets from various countries of the EU-SILC, the main aim of this paper is to analyze the effect of outliers on various common statistical and poverty measures. In Section II we observed that Belgium, Bulgaria, Italy, Poland and UK present outliers in the variable equivalised net income. A scatter plot for the data set from UK and from Belgium were given in Fig. 1 and Fig. 2. On the other hand, in Fig. 3 and Fig. 4 we can observe a scatter plot for the data set from these countries after removing outliers.

Results derived after removing outliers can be seen in Table II. Brackets in Table II are the absolute differences between results from Tables I and II.

The number of outliers for each population is given by the absolute differences of values of n. We observe that Poland contains 6 outliers, whereas Belgium and UK only have 1 outlier each population.

There is not a big difference between values of ρ, except for the case of Belgium. The linear correlation coefficient in Belgium is 0.65, whereas this coefficient after removing outliers is 0.48. This implies that the difference is 0.17 in absolute terms and 35.4% in relative terms. In summary, outliers do not have an important impact on the linear correlation coefficient, except for the case of Belgium.

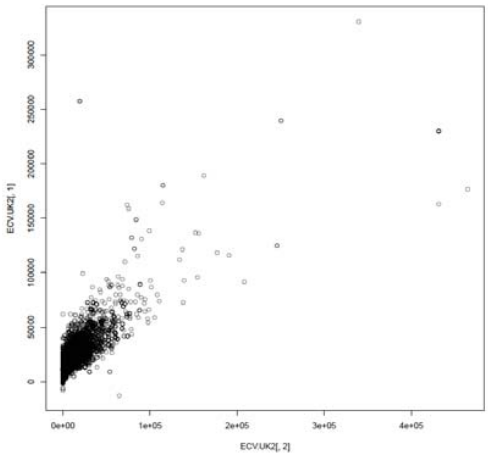

Fig. 3. Scatter plot for the data set from UK after removing outliers.

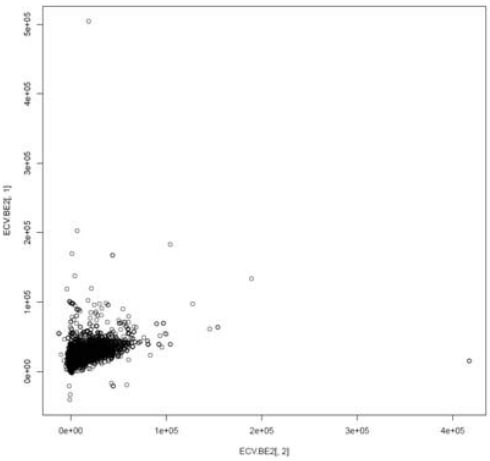

Fig. 4. Scatter plot for the data set from Belgium after removing outliers.

As we expected, outliers do not have an impact on the risk of poverty. This is due to the fact that this poverty measure is based on the median, and the median is not affected by outliers. For the various countries, the risk of poverty keeps similar after removing outliers.

Finally, we observe that outliers have a relevant impact on the skewness coefficient. For example, the skewness coefficient is 51.6 in Belgium, whereas this measure is 8.2 after removing outliers. In the case of Italy, there is also a great difference (17.8). Similar conclusions can be derived for the remainder countries.

In this table brackets indicate the absolute differences in comparison to data from Table I.

IV. CONCLUSION

Assuming real data derived from the European Union Survey on Income and Living Conditions, this paper analyzes the presence of outliers in various countries of the Euro area. In particular, we analyzed two variables: the equivalised

TABLE II
VALUES OF n, ρ, P AND g FOR VARIOUS COUNTRIES FROM THE EU-SILC AFTER REMOVING OUTLIERS.

Country	n	ρ	P	g
Belgium	9304	0.48	14.6	8.2
	(1)	(0.17)	(0.0)	(43.4)
Bulgaria	12465	0.66	22.1	2.7
	(4)	(0.04)	(0.0)	(15.6)
Italy	27550	0.81	18.9	5.3
	(3)	(0.07)	(0.0)	(17.8)
Poland	22107	0.77	17.7	4.5
	(6)	(0.05)	(0.0)	(6.5)
UK	10585	0.82	17.2	6.3
	(1)	(0.01)	(0.0)	(14.2)

net income and the tax on income. On the one hand, the equivalised net income is used for the calculation of the poverty measure commonly named as the headcount index, the risk of poverty or the proportion of poor. On the other hand, the tax on income can be used by estimation methods based on auxiliary variables, such as the regression and the logistic regression methods.

We observed that 5 from the 8 countries analyzed in this paper contain outliers in the variable equivalised net income, which is the variable of interest used by many poverty studies. The presence of outliers can have an impact on the various measures, hence we analyze the impact of outliers on various measures.

As far as the linear correlation coefficient is concerned, outliers have an impact on the data sets obtained from Belgium. These results indicate that the estimation methods based upon auxiliary variables can be affected by this issue, since they assume a strong relationship between the variable of interest and the auxiliary variable.

As we expected, outliers do not have an impact on the headcount index, whereas outliers have an important effect on the skewness coefficient.

ACKNOWLEDGMENT

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

REFERENCES

[1] T.N. Achia, A. Wangombe and N. Khadioli, "A logistic regression model to identify key determinants of poverty using demographic and health survey data". *European Journal of Social Sciences*, *13*(1), pp. 38–45, 2010.

[2] 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.

[3] A.B. Atkinson, "On the measurement of poverty". *Econometrica*, *55*(4), pp. 749-764, 1987.

[4] C.R. Blyth and H.A. Still, "Binomial confidence intervals". *Journal of the American Statistical Association*, 78, pp. 108–116, 1983.

[5] R.L. Chambers and R. Dunstan, "Estimating distribution functions from survey data". *Biometrika*, 73, pp. 597–604. 1986.

[6] 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.

[7] F.A. Cowell and M.P. Victoria-Feser, "Welfare ranking in the presence of contaminated data". *Econometrica*, 70, pp. 1221–1233, 2002.

[8] 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.

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

[10] A.H. Dorfman, "Inference on distribution functions and quantiles, in *Handbook of Statistics 29B Sample suveys: Inference and Analysis*, D. Pffefermann and C.R. Rao, Eds. Amsterdam: North-Holland, 2009, pp. 371-395.

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

[12] EUROSTAT, *"Laeken" indicators-detailed calculation methodology*, Directorate E: Social Statistics, Unit E-2: Living Conditions, DOC.E2/IPSE/2003. http://www.cso.ie/en/media/csoie/eusilc/documents/Laeken%20Indicators%20-%20calculation%20algorithm.pdf, 2003.

[13] J.E. Foster, "Absolute versus relative poverty". *The American Economic Review*, 88, pp. 335-341, 1998.

[14] J.E. Foster, J. Greer and E. Thorbecke, "A class of decomposable poverty measures". *Econometrica*, 52, pp. 761-766, 1984.

[15] J.R. Frick, M.M. Grabka and O. Groh-Samberg, "Dealing with incomplete household panel data in inequality research". *Sociological Methods and Reserch*, 41, pp. 89-123, 2012.

[16] F. Giambona and E. Vassallo, "Composite indicator of social inclusion for European countries". *Social Indicators Research*, 116, pp. 269-293, 2014.

[17] H. Gravelle and M. Sutton, "Income relative income, and seft-reporter health in Britain 1979-2000". *Center for Health Economics Research Paper*, 10, 2006.

[18] J. Haughton and S.R. Khandker, *Handbook on poverty and inequality*. Washington, DC: The World Bank, 2009.

[19] D. Jolliffe, "Measuring absolute and relative poverty. The sensitivity of estimated household consumption to survey design". *Journal of Economics and Social Measurement*, 27, pp. 1-23, 2001.

[20] S.R. Khandker, *Introduction to Poverty Analysis*. Washington, DC: World Bank Institute, 2005.

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

[22] 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.

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

[24] R.G. Newcombe, "Two-sided confidence intervals for the single proportion: comparison of seven methods". *Statistic in Medicine*, 17, pp. 857–872, 1998.

[25] 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.

[26] C.E. Särndal, B. Swensson and J. Wretman, *Model assisted survey sampling*. New York: Springer Verlag, 1992.

[27] 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.

[28] A. Tarozzi and A. Deaton, "Using census and survey data to estimate poverty and inequality for small areas". *Review of Economics and Statistics*, *91*(4), pp. 773–792, 2009.

[29] S.E. Vollset, "Confidence interval for a binomial proportion". *Statistic in Medicine*, 12, pp. 809–824, 1993.

[30] S. Weich, G. Lewis and S.P. Jenkins, "Income inequality and self-rated health in Britain". *Journal of Epidemiology and Community Health*, 56, pp. 436–441, 2002.

[31] E.B. Wilson, "Probable inference, the law of succession, and statistical inference". *Journal of the American Statisitical Association*, 22, pp. 209–212, 1927.

[32] B. Zheng, "Statistical inference for poverty measures with relative poverty lines". *Journal of Econometrics*, 101, pp. 337-356, 2001.

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.

Francisco J. Blanco-Encomienda 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 in economics and business.

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 parameter.