Variance estimation of change of poverty based upon the Turkish EU-SILC survey

Yves G. Berger[1], Melike Oguz Alper[2]

[1] University of Southampton, UK, E-mail: Y.G.Berger@soton.ac.uk
[2] University of Southampton, UK, E-mail: M.OguzAlper@soton.ac.uk

Abstract

Interpreting differences between point estimates at different waves may be misleading, if we do not take the sampling variation into account. It is therefore necessary to estimate the standard error of these differences in order to judge whether or not the observed differences are statistically significant. A major problem is to take into account of temporal correlations between estimators. Correlations play an important role in estimating the variance of a change between cross-sectional estimates. The standard correlation can be biased, because of the rotation of the design used for the European Union Statistics on Income and Living Conditions (EU-SILC) surveys. Furthermore, poverty rates depend on poverty thresholds which are estimated. We propose to use a multivariate linear regression approach to estimate the correlations. We also show how this approach can be adjusted to account for the estimation of poverty thresholds. The proposed estimator is not a model-based estimator, as this estimator is valid even if the model does not fit the data. We implemented the proposed approach to the Turkish EU-SILC survey data.

Keywords: Linearisation, multivariate regression, stratification, unequal inclusion probabilities.

Acknowledgements: This work was supported by consulting work for the Net-SILC2 project (Atkinson and Marlier 2010).

1. Introduction

In order to monitor the process towards agreed policy goals, particularly in the context of the Europe 2020 strategy, the evolution of social indicators plays an important role. However, interpreting differences between indicators at different waves may be misleading. It is therefore necessary to estimate the standard error of these differences in order to judge whether or not the observed differences are statistically significant.

The poverty rate is an important policy indicator, especially within the context of the Europe 2020 strategy. This rate is defined as “the proportion of people with an equivalised total net income below 60% of the national median income” (Eurostat 2003 p.2). This indicator is calculated from the European Union Statistics on Income and Living Conditions (EU-SILC) surveys (Eurostat 2012) which collect yearly information on income, poverty, social exclusion and living conditions from approximately 300,000 households across Europe. The poverty rate is a complex statistics unlike population totals or means; since, it is based on a poverty threshold computed from the median of the income distribution. Hence, there exist two sources of variability: one is due to the estimated threshold and the other one comes from the estimated proportion given the estimated threshold (Berger and Skinner 2003; Verma and Betti 2011). Berger Osier and Goedemé (2012) proposed an estimator for the variance of change. However, this estimator ignores the sampling variability due to the poverty threshold. In this case, the poverty rate is treated as a ratio. In Section 4, we show how this approach can be adjusted to take into account of the sampling variability of the poverty threshold. In Section 5, we compare the proposed approach with the variance estimates produced using the simpler approach proposed by Berger Osier and Goedemé (2012) (see also Berger and Priam 2010, 2013).

2. Rotating sampling designs

As the EU-SILC surveys use rotating designs to select samples at different waves, the samples of two consecutive waves are different. However, there are units which are selected at both waves. We consider that the sample design is such that the common sample has a fixed number of units. With panel surveys, it is common practice to select new units in order to replace old units that have been in the survey for a specified number of waves (e.g. Gambino and Silva 2009; Kalton 2009). The units sampled on both waves usually represent a large fraction of the first wave sample. This fraction is called the fraction of the common sample. For example, for the EU-SILC surveys, this fraction is 75%. For the Canadian labour force survey and the British labour force survey, this fraction is 80%. For the Finish labour force survey, this fraction is 60%.

3. Estimation of change in poverty

Suppose, we wish to estimate the absolute change $\Delta = \theta_2 - \theta_1$ between two population poverty rates θ_1 and θ_2, from wave 1 and wave 2 respectively. Suppose that Δ is estimated by $\hat{\Delta} = \hat{\theta}_2 - \hat{\theta}_1$; where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the cross-sectional estimators of poverty rates. The design-based variance of the change $\hat{\Delta}$ is given by

$$\operatorname{var}(\hat{\Delta}) = \operatorname{var}(\hat{\theta}_1) + \operatorname{var}(\hat{\theta}_2) - 2\operatorname{corr}(\hat{\theta}_1, \hat{\theta}_2)\sqrt{\operatorname{var}(\hat{\theta}_1)\operatorname{var}(\hat{\theta}_2)}\,.$$

Standard design-based estimators can be used to estimate the cross-sectional variances $\operatorname{var}(\hat{\theta}_1)$ and $\operatorname{var}(\hat{\theta}_2)$. The correlation $\operatorname{corr}(\hat{\theta}_1, \hat{\theta}_2)$ is the most difficult part to estimate because $\hat{\theta}_1$ and $\hat{\theta}_2$ can be estimated from different samples.

Berger and Priam (2010, 2013) proposed a multivariate approach to estimate the variance of the change between functions of totals. This approach can be used to estimate the variance of change between poverty rates when they are treated as ratios; that is, when we ignore the sampling variability of the poverty threshold. If we consider that the poverty threshold is fixed, $\hat{\theta}_1$ and $\hat{\theta}_2$ are ratios; that is $\hat{\theta}_1 = \hat{\tau}_1 / \hat{\tau}_2$ and $\hat{\theta}_2 = \hat{\tau}_3 / \hat{\tau}_4$. Therefore the change is also a smooth function of totals; that is, $\hat{\Delta} = g(\hat{\boldsymbol{\tau}})$; where $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4)'$ is a vector of four totals. Berger and Priam (2010, 2013) showed that using a Taylor linearisation approach, the design-based variance of $\hat{\Delta}$ can be estimated by

$$\hat{\mathrm{var}}(\hat{\Delta}) = \mathbf{grad}(\hat{\boldsymbol{\tau}})' \, \hat{\mathbf{var}}(\hat{\boldsymbol{\tau}}) \, \mathbf{grad}(\hat{\boldsymbol{\tau}}), \qquad (1)$$

where $\mathbf{grad}(\hat{\boldsymbol{\tau}})$ is the gradient of $g(\hat{\boldsymbol{\tau}})$, and $\hat{\mathbf{var}}(\hat{\boldsymbol{\tau}})$ is the covariance matrix which is computed using a multivariate regression (general linear) model (see Berger and Priam 2010, 2013). The covariates of this model are the stratification variables and suitable interactions which account for the rotation of the sampling design. Note that the approach proposed by Berger and Priam (2010, 2013) also account for multi-stage sampling, using an ultimate cluster approach. Correlations in $\hat{\mathbf{var}}(\hat{\boldsymbol{\tau}})$ are estimated by taking into account of the whole sample; not only the common part. This gives an approximately unbiased estimator for the variance of change (Berger and Priam 2010, 2013).

In a series of simulations based on the Swedish Labour Force Survey, Andersson et al. (2011a) (see also Andersson et al. 2011b) showed that for estimation of change within strata domains, the estimator proposed by Berger (2004) is more accurate than standard estimators of variance of change (e.g. Tam 1984; Qualité and Tillé 2008). Therefore, based on Andersson et al. (2011b) simulation studies, the estimator proposed by Berger (2004) is recommended when we are interested in change within strata domains. The estimator (1) has the same property, as it reduces to the Berger (2004) estimator when the sampling fractions are negligible (see Berger and Priam 2013).

4. Allowing for the variability of the poverty threshold

Note that in (1), the variability of the poverty threshold is not taken into account because we treat $\hat{\theta}_1$ and $\hat{\theta}_2$ as ratios. Treating the poverty threshold as fixed might lead over-estimation of the cross-sectional variances (Preston 1995; Berger and Skinner 2003; Verma and Betti 2011). Verma and Betti (2011) compared the ratio variance estimator (i.e. when the poverty threshold is treated as fixed) with linearisation and Jackknife repeated replication. They found that the ratio variance estimator over-estimated the standard errors for all the poverty measures and several complex statistics. However, these findings are related to cross-sectional estimators and do not necessarily hold for variance of change.

Taking into account the whole variability means that the sampling variation of the poverty threshold is also considered. However, the poverty rate is more complex than a ratio and cannot be expressed as a function of totals. Hence, the Taylor method (described in Section 3) cannot be used if we want to consider the whole variability. We propose to use the linearisation approach proposed by Deville (1999). The implementation of this approach for the poverty rate and inequality measures can be found in the literature (e.g. Verma and Betti 2005; Osier 2009; Münnich and Zins 2011; Verma and Betti 2011).

Osier (2009) proposed the following linearised variable for the poverty rate.

$$L_{t;i} = \frac{1}{\hat{N}_t}(\delta\{y_{t;i} \leq 0.6\hat{Y}_{t;0.5}\} - \hat{\theta}_t) - \frac{0.6}{\hat{N}_t}\frac{\hat{f}(0.6\hat{Y}_{t;0.5})}{\hat{f}(\hat{Y}_{t;0.5})}(\delta\{y_{t;i} \leq \hat{Y}_{t;0.5}\} - 0.5) \quad (t=1,2) \quad (2)$$

where $\hat{f}(\cdot)$ is the kernel estimator of the income density function (Preston 1995). The function $\delta\{A\}=1$, when A is true, and $\delta\{A\}=0$ otherwise. The quantity $\hat{N}_t$ is the estimator of the population size at wave t $(t=1,2)$ and $\hat{Y}_{t;0.5}$ is the estimator of the median of the income distribution.

The proposed estimator for the variance of change is given by

$$\mathrm{var}(\hat{\Delta}) = \mathrm{var}(\hat{\theta}_1^L) + \mathrm{var}(\hat{\theta}_2^L) - 2\,\mathrm{corr}(\hat{\theta}_1^L, \hat{\theta}_2^L)\sqrt{\mathrm{var}(\hat{\theta}_1^L)\,\mathrm{var}(\hat{\theta}_2^L)}\,,$$

where

$$\hat{\theta}_t^L = \sum_{i \in s_t} w_{t;i}\, L_{t;i}\,. \quad (3)$$

Berger and Priam (2010, 2013) proposed an estimator for the correlation between two totals. This estimator is also based upon the residual variance of a multivariate regression model. We propose to use the approach proposed by Berger and Priam (2010, 2013) by treating (3) as estimators of totals. The resulting variance estimator is different from (1), because in (1) a different multivariate regression model with more variables is used.

5. Numerical results based on the Turkish EU-SILC survey

For the purpose of analysis, the 2007 and 2008 cross-sectional Turkish EU-SILC data sets were used. The personal cross-sectional survey weights (RB050 in R-file) were used. The effect of calibration was not taken into account because we did not have any information about the auxiliary variables. The effect of imputation was also ignored.

Table 1 gives the estimates for several domains when the poverty threshold is treated as fixed. We observe a significant change for the domain “tenants” at the 95% confidence level. Therefore, the absolute change (i.e. 6.7%) is statistically significant. Table 2 gives the estimates obtained using the linearisation approach described in Section 4. We also observe a highly significant change for the domain “tenant”. We do not observe major differences in the p-values between Table 1 and 2, except for the domain “owner” for which we observe a smaller p-value when the sampling variation of the poverty threshold is taken into account. This is due to the fact that the variance of change is larger in Table 2.

Table 1. Estimates when the poverty threshold is treated as fixed (see (1))

Domain	Pov '07 (%)	Var '07	Pov '08 (%)	Var '08	Change (in % point)	Var Change	Corr	p-value
Overall	23.4	0.616	24.1	0.644	0.7	0.447	0.65	0.297
Male	23.0	0.650	23.7	0.665	0.7	0.494	0.62	0.328
Female	23.8	0.639	24.6	0.678	0.7	0.465	0.65	0.299
Owner	24.9	0.739	23.8	0.872	-1.1	0.593	0.63	0.140
Tenant	18.5	1.395	25.3	1.511	6.7	1.522	0.48	0.000
0_14	33.5	1.164	34.5	1.258	1.1	0.882	0.64	0.263
15_24	24.2	1.162	25.3	1.181	1.1	1.118	0.52	0.296
25_49	19.8	0.527	20.7	0.548	0.9	0.405	0.62	0.178
50_64	14.4	0.568	15.0	0.719	0.6	0.569	0.56	0.404
65+	17.7	1.077	16.2	0.929	-1.5	0.988	0.51	0.120

Source: 2007 and 2008 cross-sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

Table 2. Estimates when the sampling variation of the poverty threshold taken into account (see Section 4)

Domain	Pov '07 (%)	Var '07	Pov '08 (%)	Var '08	Change (in % point)	Var Change	Corr	p-value
Overall	23.4	0.281	24.1	0.275	0.7	0.338	0.39	0.230
Male	23.0	0.382	23.7	0.386	0.7	0.375	0.51	0.262
Female	23.8	0.375	24.6	0.403	0.7	0.354	0.55	0.234
Owner	24.9	0.362	23.8	0.420	-1.1	0.450	0.43	0.090
Tenant	18.5	1.123	25.3	1.242	6.7	1.357	0.43	0.000
0_14	33.5	0.919	34.5	0.986	1.1	0.762	0.60	0.228
15_24	24.2	0.984	25.3	1.023	1.1	1.013	0.50	0.273
25_49	19.8	0.332	20.7	0.351	0.9	0.325	0.52	0.133
50_64	14.4	0.482	15.0	0.615	0.6	0.516	0.53	0.380
65+	17.7	0.990	16.2	0.856	-1.5	0.938	0.49	0.111

Source: 2007 and 2008 cross-sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

We observe smaller estimates of the correlations when the variability of the poverty threshold is taken into account. Indeed, the correlations in Table 2 are less than the correlations in Table 1 throughout. Moreover, there are noticeable decreases in the correlations for the overall population and for the domain “owners”. This reduction may be explained by the fact that some part of the correlations has been captured by the underlying variables in (2). We can attempt to explain this situation by viewing (2) as residuals. For example, Andersson *et al.* (2011a, 2011b) showed that the correlation estimated with a generalised regression estimator, which is based on the residuals, is lower than the correlation between the actual variables of interest. In other words, underlying variables created some kind of confounding effect on the correlation. This result depends on the data used. Hence, how the variability of the poverty threshold affects the correlation should be studied more deeply through simulation studies.

By comparing Table 1 with Table 2, we also found that all variances were estimated more conservatively when the threshold is treated as fixed. Preston (1995), Berger and Skinner (2003) and Verma and Betti (2011) demonstrated that cross-sectional variances are more conservative when the poverty threshold is treated as fixed. However, for variance of change, we cannot anticipate an increase in the variance when the poverty threshold is treated as fixed. Let assume that the cross-sectional variances are equal $\hat{\mathrm{var}}(\hat{\theta}_1) = \hat{\mathrm{var}}(\hat{\theta}_2)$. Then, the variance estimator of change is given by $\hat{\mathrm{var}}(\hat{\Delta}) = 2\,\hat{\mathrm{var}}(\hat{\theta}_1)(1 - \mathrm{corr}(\hat{\theta}_1, \hat{\theta}_2))$. Hence, variance of change is affected in the same direction by the variance term and in the opposite direction by the correlation term. Thus, when both the variance and the correlation terms increase or decrease concurrently, the direction of the effect on the variance of change cannot be predicted. We may not necessarily have more conservative estimates of variance of change when the poverty threshold is treated as fixed. However, with the data we used, we found that the variances of changes were more conservative (see Table 1).

Table 3. Estimates when the sampling variation of the poverty threshold taken into account (see Section 4). The smoothing parameter is based on the inter-quartile range of the income distribution.

Domain	Pov '07 (%)	Var '07	Pov '08 (%)	Var '08	Change (in % point)	Var Change	Corr	p-value
Overall	23.4	0.292	24.1	0.290	0.7	0.372	0.36	0.252
Male	23.0	0.361	23.7	0.350	0.7	0.368	0.48	0.257
Female	23.8	0.350	24.6	0.354	0.7	0.346	0.51	0.228
Owner	24.9	0.347	23.8	0.385	-1.1	0.457	0.38	0.092
Tenant	18.5	1.088	25.3	1.171	6.7	1.325	0.41	0.000
0_14	33.5	0.815	34.5	0.828	1.1	0.708	0.57	0.211
15_24	24.2	0.973	25.3	0.988	1.1	1.003	0.49	0.270
25_49	19.8	0.320	20.7	0.324	0.9	0.319	0.50	0.129
50_64	14.4	0.505	15.0	0.630	0.6	0.525	0.54	0.384
65+	17.7	0.989	16.2	0.876	-1.5	0.940	0.50	0.111

Source: 2007 and 2008 cross-sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

As shown by Verma and Betti (2005), probability density functions are quite sensible to the chosen bandwidth parameter in (2). The larger value of the bandwidth parameter is, the smoother density functions will be. We also investigate the situation when the smoothing parameter is based on the inter-quartile range of the income distribution (Berger and Skinner 2003). The results are given in Table 3. By comparing Table 1 with Table 3, we also observed smaller cross-sectional variances, variances of changes and correlations when the bandwidth parameter based on the inter-quartile range. When we compare Table 2 with Table 3, variance estimates do not differ so much between two linearisation methods based on different smoothing parameters. However, the estimates slightly vary from each other for the age group: 0-14. However, differences between variance estimates of change calculated from two linearisation methods are negligible although correlations seem to differ a little bit more over some domains.

References

Andersson, C., Andersson, K. and Lundquist, P. (2011a). Estimation of Change in a Rotation Panel Design. IN: Proceeding of the 58th World Statistics Congress. Dublin: International Statistical Institute.

Andersson, C., Andersson, K. and Lundquist, P. (2011b) Variansskattningar avseende förändringsskattningar i panelundersökningar (variance estimation of change in panel surveys). Methodology reports from Statistics Sweden (Statistiska centralbyrån).

Atkinson, A. and Marlier, E. (2010). Income and Living Conditions in Europe. Luxembourg: Office for Official Publications. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-31-10-555/EN/KS-31-10-555-EN.PDF.

Berger, Y. G. (2004). Variance estimation for measures of change in probability sampling. Canadian Journal of Statistics, 32, 451–467.

Berger, Y. G., Osier, G. and Goedemé, T. (2012). Standard error estimation and related sampling issues. Proceeding of the 2012 International conference on comparative EU statistics on income and living conditions, Vienna, Austria.

Berger, Y. G. and Priam, R. (2010). Estimation of Correlations Between Cross-Sectional Estimates from Repeated Surveys – an Application to the Variance of Change. IN: Proceedings of the Statistics Canada Symposium, 2010.

Berger, Y. G. and Priam, R. (2013). A simple variance estimator of change for rotating repeated surveys: an application to the EU-SILC household surveys. Pre-print. http://eprints.soton.ac.uk/347142

Berger, Y. G. and Skinner, C. J. (2003). Variance Estimation of a Low-Income Proportion. Journal of the Royal Statistical Society: Series C (Applied Statistics), 52, 457-468.

Deville, J. C. (1999). Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques. Survey Methodology, 25, 193–203.

Eurostat (2003). ‘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

Eurostat (2012). European Union Statistics on Income and Living Conditions (EU-SILC). http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc (Accessed 26 October 2012).

Gambino, J. G. and Silva, P. L. N. (2009). Sampling and Estimation in Household Surveys. Handbook of Statistics: Design, Method and Applications, (ed). D. Pfeffermann and C.R. Rao, Elsevier, 29A, 407–439.

Kalton, G. (2009). Design for Surveys Over Time. Handbook of Statistics: Design, Method and Applications, (ed). D. Pfeffermann and C.R. Rao, Elsevier, 29A, 89–108.

Münnich, R. and Zins, S. (2011). Variance Estimation for Indicators of Poverty and Social Exclusion. Work-Package of the European Project on Advanced Methodology for European Laeken Indicators (AMELI). http://www.uni-trier.de/index.php?id=24676

Osier, G. (2009). Variance Estimation for Complex Indicators of Poverty and Inequality Using Linearization Techniques. Survey Research Method, 3, 167–195.

Preston, I. (1995). Sampling Distributions of Relative Poverty Statistics. Applied Statistics, 44, 91-99.

Qualité, L. and Tillé, Y. (2008). Variance Estimation of Changes in Repeated Surveys and Its Application to the Swiss Survey of Value Added. Survey Methodology, 34, 173-181.

Tam, S. M. (1984). On covariances from overlapping samples. American Statistician, 38, 288–289.

Verma, V. and Betti, G. (2005). Sampling Errors and Design Effects for Poverty Measures and Other Complex Statistics. Working Paper 53. Siena: Dipartimento di Metodi Quantitativi, Università degli Studi.

Verma, V. and Betti, G. (2011). Taylor Linearization Sampling Errors and Design Effects for Poverty Measures and Other Complex Statistics. Journal of Applied Statistics, 38, 1549-1576.