Model based estimation of indicators of poverty and social exclusion

**Thomas Glaser, Statistics Austria, Vienna**

Statistics Austria started using mainly register data for the measurement of household income in EU-SILC from 2012 onwards which caused a break in time series. For EU-SILC 2011 household incomes from registers and from interviews are available. Thus household income can be calculated mainly using registers and a comparison between the two data sources is possible. At the time of the publication of the 2012 results no suitable way to link register and interview data was available yet for EU-SILC 2008-2010. However, the effect of the usage of register data on the household income could be modelled with 2011 data. The paper presents different versions of such models tested by Statistics Austria and compares their strengths and weaknesses. The model with the best fit was applied to the years 2008-2010. The estimates of the register household income facilitate a new time series of Europe 2020 headline indicators of poverty and social exclusion for 2008-2012. This is of importance, because the Europe 2020 strategy in Austria is aimed at reducing the number of persons at-risk-of-poverty or social exclusion by 235,000 within ten years, starting with 2008. To evaluate this strategy a continuing time series from 2008 onwards is necessary.

# Introduction

EU-SILC (European Community Statistics on Income and Living Conditions) is the major source for statistics on poverty, social inclusion and income of private households in Austria.

Since 2004 EU-SILC in Austria is carried out on a yearly basis by a sample survey that applies a rotational sampling design with four rotational subsamples. These subsamples are drawn from the central residence register (ZMR) by a probability sample. Participation in the survey is voluntary. The most important indicator based on EU-SILC is the number of “people at-risk-of-poverty of social exclusion” (AROPE) which is part of the indicators that are necessary to measure the Europe 2020 target to have “at least 20 million fewer people in or at-risk-of-poverty and social exclusion”.[[1]](#footnote-1) For Austria this translates to 235,000 people to be put out of poverty starting from 2008, which accounts for 23,500 people each year during the agreed ten year period of measurement. AROPE is defined as being affected by at least one of the following conditions: (1) “At-risk-of-poverty” (AROP) defined as persons in households with an equivalised household income[[2]](#footnote-2) below 60% of the median of referring distribution, (2) in a condition of severe material deprivation defined as a lack of resources defined by encountering at least four out of nine material deprivation indicators[[3]](#footnote-3) or (3) living in a household with very low work intensity[[4]](#footnote-4). Among these three indicators only AROP is a monetary indicator depending on income information. Data collection for EU-SILC in Austria is carried out by Computer Assisted Personal Interviewing (CAPI) and Computer Assisted Telephone Interviewing (CATI). However, in 2010 a national regulation was passed which provides the legal basis for using income data from registers for EU-SILC.[[5]](#footnote-5) EU‑SILC 2012 was the first year of the survey, were income was collected mainly from registers [2]. In 2011 only pension income was retrieved from registers. This change in methodology lead to a break in time-series which affected the results of the Europe 2020 indicator AROP and consequently also AROPE [2]. Since a continuing time series from 2008 onwards is necessary for evaluating the success of policies aimed at reducing the risk of poverty or social exclusion a revision of the time series of AROPE spanning the years 2008-2012 has become necessary. In order to distinguish AROP and AROPE based on the new methodology from the previous one, AROP(REG) and AROPE(REG) shall denote AROP and AROPE based on the new methodology using mainly register data for income measurement.

# Revision of the time series

The first possibility for providing a new time series of AROPE(REG) based on the new methodology would be to calculate household income based on registers also for SILC 2008-2011 and then recalculating AROP(REG)and consequently also AROPE(REG). At the time of the publication of the results from EU-SILC 2012 (December 2013) only register data for EU‑SILC 2011 were available and could be used for recalculating household income by matching register data to the sample of EU‑SILC 2011. Therefore, in 2011 two versions of the household income exist, one based mainly on data from interviews (HINC) and one based mainly on register income data (HINC(REG)). Indicators based on the recalculated equivalised income show similar results for AROP(REG) (14.5% in 2011 and 14.4% in 2012), meaning that by using the same new methodology for EU‑SILC 2011 and EU‑SILC 2012 the at-risk-of-poverty rate does not change significantly between the two years. For EU‑SILC 2008-2010 an ex post recalculation of income based on registers would be more difficult, but there exists another solution for gaining a time series of Europe 2020 indicators for 2008-2012 based on register income. Since in 2011 household income based on the interview and on registers is available, the effect of using mainly registers as a data source for income on the indicators AROP(REG) and AROPE(REG) could be modelled and a continuing time series from 2008-2012 based on the new methodology could be established by model based AROP(REG) and AROPE(REG) for 2008-2010.

# Choice of models

Several ways of estimating AROP(REG) and AROPE(REG) for the years 2008-2010 including the effect that register data have on these indicators were tested by Statistics Austria. Among these, three variants based on different approaches can be distinguished which shall be discussed in detail in this chapter.

*Variant 1) Direct estimation of indicators*

Since establishing a new continuous time series of AROPE(REG) including register income data is the main motivation for estimating Europe 2020 indicators for 2008-2010 the straightforward solution for this challenge is to directly estimate the likelihood of being “at-risk-of-poverty or social exclusion”. This can be easily carried out by a logistic regression using multiple predictors. The advantage of formulating a model where AROPE(REG) based on register data in 2011 is the estimand is that such a model can directly deliver an estimated value for these indicators. Furthermore, only the likelihood of a dichotomous variable and not an entirely new distribution of equivalised income has to be estimated. So the dependent variable in the model based on EU‑SILC 2011 refers exactly to the indicator that has to be estimated for the preceding years. Let ***x***denote the vector of *J* explanatory variables of the likelihood that AROPE(REG) =1 then the model which is used to predict the likelihood of a person to be at-risk-of-poverty or social exclusion can be defined as follows:

 (1)

If also the time series of AROP should be revised a separate model would be needed, defined similarly to the previous one:

 (2)

Depending on how well this different models are formulated and how good its fit turns out discrepancies and inconsistencies between AROP(REG) and AROPE(REG) can arise. For example, the rate of AROP(REG) cannot be smaller than AROPE(REG) by definition, but such an outcome is possible if the likelihoods of AROP(REG) and AROPE(REG) are estimated separately. A large variety of explanatory variables *xj* [*j*=1, … , J] which are related to AROPE(REG) or AROP(REG) were put into the models described in formulas (1) and (2). The selection of the explanatory variables for the model was carried out by a stepwise procedure in order to exclude insignificant explanatory variables.[[6]](#footnote-6) The reason for using an automatized procedure of variable selection is based on the fact that the important outcome of the model are the predicted indicators which should be estimated by a limited amount of the best predictors available. The fit of the model for AROPE(REG) resulted in a moderate value of Nagelkerke’s pseudo r-square amounting to 0.44 and a highly significant chi-square value (α<0.01) for the model. However, only AROPbased on survey data turned out to be a significant explanatory variable and remained in the model. For the separate estimation of AROP(REG) a better fit was obtained with a pseudo r-quare of 0.62 and also a highly significant chi-square value for the model. Instead of one explanatory variable 42 predictors remained in the model this time. Despite the not very good fit of the logistic regression models the estimated values resulted in an underestimation of AROPE(REG) and AROP(REG) for 2011 as can be seen in table 1. The estimate of the rate of AROPE(REG) and AROP(REG) is calculated as the mean value of all individual estimated probabilities from the logistic regression models weighted by the household weight.

Table 1: Comparison of model based estimates (variant 1)



*Variant 2) Estimation of household income*

Since AROPE contains only income information via AROP it would be sufficient to estimate the equivalised income based on register data and subsequently calculate AROPE(REG) and AROP(REG). Furthermore, only a new version of the household income is necessary, because the equivalised income then can be calculated using information available from interviews.

With a new household income, estimated by including the effect of register data, the at-risk-of-poverty threshold (60% of the median of equivalised income) and AROP(REG) as well as AROPE(REG) do not have to be estimated separately. Together with non-income household characteristics from the interview a new equivalised income based on the estimated household income can be calculated. Then also the at-risk-of-poverty threshold (60% of the median of equivalised income) as well as AROP(REG) and AROPE(REG) do not have to be estimated but can be calculated. The obvious advantage compared to variant 1 is that there is no danger of implausible differences between AROP(REG) and AROPE(REG) because they are both based on the same estimated household income distribution.

For EU-SILC 2011 a multiple linear regression model fitted by ordinary least squares (OLS) was used to predict HINC(REG),the household income based on registers. More specifically, the natural logarithm of HINC(REG) was used as the dependent variable because of the positively skewed distribution of household income. Let the vector HINC denote the household income based mainly on interview data and x*j* [*j*=1, … , J] the *J* additional explanatory variables, then the model which is used to predict the household income based mainly on registers can be defined as follows:

 (3)

The regression coefficients of the model in formula (3) were obtained by a stepwise procedure where insignificant predictors with a significance level smaller than 0.1 were not allowed to remain in the model. Since the linear regression model obtained by OLS is sensitive in regard to outliers, 124 households had to be omitted because of extreme values.[[7]](#footnote-7) Still, 6,063 households remained in the sample, so a sufficiently large number of cases were available for fitting the model. The final model delivered a very good fit with R2 amounting to 89%. This is mainly due to the strong predictive power of HINC since household income based mainly on interview data is highly correlated with household income based mainly on registers. The abovementioned strong predictive power of HINC is reflected in a comparatively high standardised regression coefficient of 0.774. Among the remaining predictors age groups show the strongest effect which is mainly due to the fact, that counting the number of persons belonging to a specific age group is an indirect measure of household size. The values predicted by the model described in formula (3) were used to calculate a distribution of equivalised income. This was carried out by exponentiation of the predicted logarithms and subsequently dividing by the consumption equivalents[[8]](#footnote-8) derived from the household size and age distribution of household members. Finally, the at-risk-of-poverty threshold was calculated and AROP(REG) as well as AROPE(REG) could be retrieved. The results presented in Table 2 indicate an underestimation of the indicators based on the model in formula (3).

Table 2: Comparison of model based estimates (Variant 2)



Since the regression model reduces variance by fitting the regression line the new estimated household income distribution will incorporate less variance than the original one of EU‑SILC 2011. Comparison of the original distribution of HINC(REG) (dependent variable in the model) with the distribution of estimates of shows that the latter one is indeed narrower and therefore fewer cases fall below the at-risk-of-poverty threshold leading to the underestimation of AROP(REG) as can be seen in Table 2. A solution to this issue is to add normally distributed stochastic error terms to the predicted values in order to reproduce the amount of variance found in the original register based household income. Such error terms are independent and identically distributed (i.i.d.) and therefore they only add noise and no additional information about the structure of the estimated household income. The distribution of the stochastic error terms was chosen as , where denotes the variance of the distribution of residuals excluding the highest and lowest percentiles. Thus the stochastic error terms added to the predicted values closely resemble the residuals excluding outliers. This methods delivers a distribution of estimated household income () that closely resembles HINC(REG). More importantly, also indicators based on now come very close to the indicators of EU-SILC 2011 as can be seen in Table 3.

Table 3: Comparison of model based estimates (Variant 2a)



In summary, the refined version of variant 2 including stochastic error terms – variant 2a – not only delivers estimates of AROP(REG) and AROPE(REG) that are based on the same income distribution, it also makes possible micro-data with estimated values of the household income based mainly on registers possible. However, caution is advised if these micro-data are used to analyse specific domains of the estimated indicators, because of i.i.d stochastic error terms added to estimated values of .

*Variant 3) Estimating the difference of household income versions*

Instead of estimating the entire household income distribution another approach would be to identify groups of households with a substantial difference of HINC and HINC(REG). The main goal of estimating a new household income is to model the effect that register data have on the already existing results on household income measured by interviews. If the difference between these two versions of household income is non-existent or negligible for specific households, only households with a notable difference of income versions are of concern. This approach would reduce the estimation effort to this latter category of households. In order to use such a model for predicting the register data effect for the years 2008-2010 two steps are necessary: Firstly, the likelihood of experiencing a substantial difference in income versions has to be classified and secondly, the amount of the difference between household income from registers or the interview has to be estimated. The advantage of variant 3 is that the register data effect can be explicitly modelled whereas it is only implicitly incorporated in variant 2 as an additional predictor. On the other hand, variant 3 requires two models that are dependent on each other because the outcome of the first model is a predictor in the second one. In order to assess the scope of the deviation of HINC from HINC(REG) the distribution of the difference HINC-HINC(REG) was analysed (excluding the lowest and highest percentile). About 10% of households had a zero difference between HINC and HINC(REG). However, also small differences should be left out of the estimation of the difference of household incomes. Based on the corresponding percentiles also differences of household incomes smaller than +/- 18.5 Euro where deemed as not relevant. In order to estimate if the difference of HINC and HINC(REG) is within these margins a stepwise discriminant analysis with a linear discriminant function was carried out. The results showed that 85% of the differences defined as relevant where also correctly classified as such by the model. For the households with
|HINC-HINC(REG)| ≥ 18.5 a stepwise linear regression model similar to the one in formula (3) was carried out resulting in a low model fit (R2=0.10). Applying the two modelling steps to the EU-SILC 2011 interview income data resulted in an overestimation of AROP(REG) and only a very slight overestimation of AROPE(REG) as is presented in Table 4.

Table 4: Comparison of model based estimates (Variant 3)



# The role of weighting

The usage of registers did not only change income measurement for EU-SILC, also the weighting scheme was expanded because of the availability of more marginal distribution that could be used in calibration of weights. EU-SILC facilitates case weights on household level that are also applied on personal level to each inhabitant. From EU‑SILC 2012 onwards weights are calibrated to the number of employees (aged 15 years or older) and retirees based on the wage tax register in addition to the already used marginal distribution.[[9]](#footnote-9)

However, all the models described in the previous section were fitted without these weights, because most of the relevant variables used in weighting, especially in calibration, are also predictors in the models [4]. Also linear regression models fitted by OLS are unbiased and are most efficient according to the Gauss-Markov theorem, so OLS should be used preferably in the case of linear regression [5]. Furthermore, a comparison of weighted and unweighted models showed that there were no notable differences in model fit and coefficients for all variants described above. Another obstacle would have arisen if weights were used in the models. Because of the lack of register information for the years 2008-2010 weights including new register information should have been estimated separately or the old version of the weights would have been applied to estimate the weighted distribution of household income based mainly on registers. Either way, estimation of indicators would have been overly complicated or not consistent with applied weights.

# Application of model based estimation for EU-SILC 2008-2010

The results presented in chapter 3 clearly speak in favour of variant 2a: estimation of household income in combination with stochastic error terms. The regression coefficients in formula (3) were applied to HINC and the other independent variables of the years 2008-2010. This way the modelled register income effect could be applied to the socio-economic structure of 2008-2010 reflected in the predictor variables. Stochastic error terms with independent distribution and taken from the 2011 model were added to the estimates of the years 2008-2010. Then the at-risk-of-poverty threshold as well as AROP(REG) and AROPE(REG) were calculated using the estimates and interview data from EU‑SILC 2008-2010. Figure 1 shows the time series AROP and AROPEcompared to AROP(REG) and AROPE(REG) for the years 2008-2012. The register based indicators include the model estimates for the years 2008-2010, thus providing an unbroken time series. The model based estimates show quite similar trends for both indicators compared to the time series based on the old methodology. As expected, both AROP(REG) and AROPE(REG) are higher than their non-register based counterparts.

**Figure 1: Comparison of indicators based on register an interview data**



# Conclusions and outlook

Estimating the Europe 2020 indicators at-risk-of poverty as well as at-risk-of-poverty and social exclusion based mainly on registers for the EU‑SILC surveys 2008-2010 which are currently not linked with in come register data proved to be a challenge. The effect of using mainly register information for income measurement had to be modelled and applied to the survey years were register data are not available. A variant that estimates the household register income in addition with a stochastic error term to reproduce the right amount of variation in the estimated household income distribution was chosen. This method allowed for using the existing information on household size and composition to calculate the at-risk-of-poverty rate and furthermore the rate of persons at-risk-of-poverty or social exclusion which is the main Europe 2020 indicator based on EU-SILC. The results provided a plausible timeline of these indicators based on the newly introduced register income methodology.

Still, household income based on registers would be the preferred choice also for EU‑SILC 2008‑2010. Adding register income data to these survey years is cumbersome, but Statistics Austria is currently working on the issue and it is planned to have results of Europe 2020 indicators based mainly on register income data ready by autumn/winter of 2014.

# References

[1] Glaser, Th./Till, M. (2010), Gewichtungsverfahren zur Hochrechnung von EU-SILC-Querschnittergebnissen. In: Statistische Nachrichten 7/2010. Statistics Austria. Vienna: 566-577.

[2] Heuberger, R./Glaser, T./Kafka, E. (2013), The use of register data in the Austrian SILC survey. In: Jäntti, M./Törmälehto, V.-M./Marlier, E., The use of registers in the context of EU-SILC: challenges and opportunities. Eurostat Statistical Working Papers. Luxembourg.

[3] Lamei, L./Glaser, T./Heuberger, R./Kafka, E./Oismüller, A./Skina-Tabue, M. (2013), Methodenbericht EU-SILC 2012, Vienna. Statistics Austria.

[4] Skinner, Ch./Mason, B. (2012), Weighting in the regression analysis of survey data with a cross-national application. Canadian Journal of Statistics, 40 (4): 697-711.

[5] Winship, Ch./ Radbill L. (1994), Sampling Weights and Regression Analysis. Sociological Methods & Research, 23 (2): 230-257.

1. Cf. <http://ec.europa.eu/europe2020/europe-2020-in-a-nutshell/targets/index_en.htm> (retrieved 2014-05-13) [↑](#footnote-ref-1)
2. The equivalised household income is defined as the available net household income divided by the number of consumption equivalents in the household which are calculated as a weighted sum of assumed resource requirements. A single adult is assigned a weight of 1, any additional adult receives a weight of 0.5 and every child below the age of 14 a weight of 0.3. [↑](#footnote-ref-2)
3. Cannot afford to pay rent or utility bills; keep home adequately warm; face unexpected expenses; eat meat, fish or a protein equivalent every second day; a week holiday away from home; a car; a washing machine; a colour TV; a telephone (including mobile phone). [↑](#footnote-ref-3)
4. In these households the work intensity of persons aged 18-59 is below 20% of the theoretically reachable work intensity. This indicator applies to household members younger than 60 years. [↑](#footnote-ref-4)
5. Cf. Einkommen- und Lebensbedingungen-Statistikverordnung – ELStV:

<http://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2010_II_277/BGBLA_2010_II_277.pdf> (retrieved 2014-05-13) [↑](#footnote-ref-5)
6. Explanatory variables with a p-value of the chi-square score statistic of p<0.1 were excluded from the model. [↑](#footnote-ref-6)
7. Using the natural logarithm of household incomes in formula (3) produced some very low values. Households with a yearly net household income lower than 2,210 Euro had to be excluded from the model. Additionally some households had to be excluded because they were identified as outliers during the analysis of residuals. [↑](#footnote-ref-7)
8. For a description see footnote 2. [↑](#footnote-ref-8)
9. Besides the number of employees (aged 15+) and retirees cross-sectional weights of EU-SILC in Austria are calibrated to marginal distribution on household level (province, household size, tenure status) and personal level (age and sex, citizenship, number of recipients of unemployment benefits for more than one month) [1]. [↑](#footnote-ref-9)