Imputation of income data with generalized calibration procedure and GB2 law: illustration with SILC data

Eric Graf and Yves Tillé
Institut de Statistique, Université de Neuchâtel, Neuchâtel, SWITZERLAND
e-mail: eric.graf@unine.ch, yves.tille@unine.ch

Abstract

In sample surveys of households and persons, questions about income are often sensitive and thus subject to a higher non-response rate. Nevertheless, the household or personal incomes are among the important variables in surveys of this type. The distribution of such collected incomes is not normal, neither log-normal. Hypotheses of classical regression models to explain the income (or their log) are not fulfilled. Imputations using such models modify the original and true distribution of the data. This is not suitable and may conduct the user to wrong interpretations of results computed from data imputed in this way. The generalized beta distribution of the second kind (GB2) is a four parameters distribution. Empirical studies have shown that it adapts very well to income data. The advantage of a parametric income distribution is that there exist explicit formulae for the inequality measures like the Laeken indicators as functions of the parameters. We present a parametric method of imputation, based on the fit of a GB2 law on the income distribution by the use of suitably adjusted weights obtained by generalized calibration. These weights can compensate for non ignorable non-response that affects the variable of interest. We validate our imputation system on data from the Swiss Survey on Income and Living Conditions (SILC).

Keywords: GB2, generalized calibration, imputation, inequality measures, non-ignorable non-response, SILC

1 Introduction

In economics and social statistics, the study of income distribution is in the heart of inequality measures and more generally in the evaluations of social welfare. In official national sample surveys of households and individuals, non-response in income variables is often a difficult and important problem in the sense that the phenomenon affects the image of the reality that the survey reflects. Without proper treatment, the outcomes of political or social decision-making based on the results of the investigation may be wrong. We show that generalized calibration can be used to obtain weights allowing to compensate even for non ignorable non-response (see for instance Osier (2013)). Moreover these weights can be used to perform an imputation as well as for fitting a Generalized Beta distribution of the second kind (GB2) on the collected income data. The procedures developed are applied to the Swiss SILC data from year 2009 to assess their validity.

We aim to develop a method of imputation for income variables allowing direct analysis of the distribution of such data, but also the estimation of complex statistics such as indicators for poverty and social exclusion. We focus on five of them, formerly called *Laeken indicators* (Eurostat, 2005): the At risk of poverty threshold (ARPT), At risk of poverty rate (ARPR), Relative median at-risk-of poverty gap (RMPG), Quintile share ratio (QSR) and the Gini index (GINI).

It is of great importance that the imputation method is transparent and allows to produce reliable variance estimates for any statistic based on the imputed data including the complex statistics mentioned. Today the Swiss Federal Office of Statistics uses the IVEware package written in SAS® to conduct the necessary imputations for several income variables. IVEware can perform single or multiple

imputations of missing values using the Sequential Regression Imputation Method described in Raghunathan et al. (2001). Although IVEware is a very convenient tool, in our context it has showed several drawbacks: it produces a poor output documenting the models that were fitted (lack of transparency); if the data contain extreme values, often the case with income, these can destabilize the models; output information is insufficient for producing variance estimates for (complex) statistics; somehow a (multi-)normal distribution is used to impute; it does not allow to use the survey weights; the number of imputations to produce and then which measure take as final imputed value are not easy to decide.

Convincing results of empirical studies on income (see for instance Sepanski and Kong, 2008; Jenkins, 2007; Dastrup et al., 2007; Kleiber and Kotz, 2003) have shown that the GB2 fits well with such data and it is often more suitable than other four-parameter distribution. In addition the European project AMELI (2011) confirmed this fact for EU-SILC data. Encouraged by these works we decided to work out another parametric imputation method partially based on a GB2 fit.

The Swiss SILC data could by matched with a register, so the partially missing income data collected by the survey could be compared, on the unit level, with the official value of the register. The non-response mechanism observed in the survey has been applied to the register providing a unique occasion to test an imputation models and compare their outputs with the true values. We deal with about 30% of non-response rate.

2 Concepts

Through a household survey (example: SILC), we collect, among other measures, some income variable, say y. In our scenario y is affected by non ignorable non-response (worst case). That means that the missingness pattern depends on the (unknown) values of y itself. We place ourselves in an end user situation where we are dealing with a partially missing y-variable (item non-response) accompanied by a set of weights which may reflect the sampling design, different sorts of non-response adjustments and calibration on known population total. We intend to deliver an imputed variable y_{imp} without any missing values as well as doing inference to the target population in publishing estimated values with confidence intervals for several indicators of poverty and social exclusion based on this imputed variable.

To set the notations, let U represent the population, s the selected sample, d_k the sampling weight for unit k, $r \subseteq s$ the set of survey respondents, w the final survey weight defined on r, $r_y \subseteq r$ the set of respondents to the interest variable y. Let also $\mathbf{x}_k = (x_{k1}, ..., x_{kj}, ..., x_{kJ})$ be *auxiliary variables* used for calibrating. These are known on s, in addition population totals are available for $\mathbf{t}_x = \sum_U \mathbf{x}_k$. Moreover, let $\mathbf{z}_k = (z_{k1}, ..., z_{ki}, ..., z_{kI})$ be a set of *instrumental variables* known on r_y. We suppose that $J = I$. Note that some of the $\mathbf{z}_k$ can be identical to some of the $\mathbf{x}_k$. In short, one observes $(y_k, \mathbf{x}_k, \mathbf{z}_k)$ and disposes from $\mathbf{t}_x$.

2.1 Generalized calibration

Calibration and generalized calibration methods used are those developed and presented in reference articles like: Deville and Särndal (1992); Deville et al. (1993); Le Guennec and Sautory (2002); Sautory (2003); Deville (2002); Kott (2006). In short, one seeks for new weights, which are "near" the ones before calibration, specific to y, compensating for non-response and respecting some constraints. The objective is to estimate a population total $t_y = \sum_U y_k$, ideally by $\hat{t}_y = \sum_s d_k y_k$ (Horvitz-Thompson estimator of the total) if all the units of s are available, more

realistically by $\hat{t}_y^{cal} = \sum_{r_y} w_k^{cal} y_k$. The calibrated weights w^{cal} are near to the previous weights w according to some (pseudo-)distance G for every sample s_{ry}:

$$\underset{w_k^{cal}}{min} \sum_{k \in r_y} \frac{G_k(w_k^{cal}, w_k)}{q_k}$$

under the constraints $\mathbf{t}_x = \sum_{k \in r_y} w_k^{cal} \mathbf{x}_k$. One can give more or less importance to certain units in weighting each G_k by $1/q_k$. The calibrated weights are of the form:

calibration	**generalized calibration**
$w_k^{cal} = w_k F(q_k \mathbf{x}_k' \boldsymbol{\lambda})$	$w_k^{cal} = w_k F(\mathbf{z}_k' \boldsymbol{\lambda})$

where F depends on the form of the pseudo-distance G. For example, in the linear case, $G_k(w_k^{cal}, w_k) = \frac{(w_k^{cal} - w_k)^2}{2w_k}$ and

$F(q_k \mathbf{x}_k' \boldsymbol{\lambda}) = (1 + q_k \mathbf{x}_k' \boldsymbol{\lambda})$	$F(\mathbf{z}_k' \boldsymbol{\lambda}) = (1 + \mathbf{z}_k' \boldsymbol{\lambda})$

One solves for $\boldsymbol{\lambda}$ so that w_k^{cal} satisfies the constraints to obtain

$\hat{t}_{ylin} = \mathbf{t}_x' \hat{\mathbf{B}}_{r_y} + \sum_{k \in r_y} w_k e_k$ where $e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_{r_y}$ residuals of the regression of y on the J auxiliary variables x_k.	$\hat{t}_{ylinG} = \mathbf{t}_x' \hat{\mathbf{B}}_{r_y zx} + \sum_{k \in r_y} w_k e_k$ where $e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_{r_y zx}$ residuals of the instrumental regression of y on the J $\mathbf{x}_k$ defined on sample s, with the J instrumental variables $\mathbf{z}_k$
$\hat{\mathbf{B}}_{r_y} = \mathbf{T}_{r_y}^{-1} \sum_{k \in r_y} w_k q_k \mathbf{x}_k y_k$ is the J-parameters vector of the regression $\mathbf{T}_{r_y}^{-1} = \left(\sum_{k \in r_y} w_k \mathbf{x}_k q_k \mathbf{x}_k' \right)^{-1}$	$\hat{\mathbf{B}}_{r_y zx} = \mathbf{T}_{r_y zx}^{-1} \sum_{k \in r_y} w_k \mathbf{z}_k y_k$ is the J-parameters vector of the instrumental regression $\mathbf{T}_{r_y zx}^{-1} = \left(\sum_{k \in r_y} w_k \mathbf{x}_k \mathbf{z}_k' \right)^{-1}$

Note that in the case of generalized calibration, the vector of parameters depends also on the $\mathbf{z}_k$. A crucial point is that the $\mathbf{z}_k$ need only to be known on the subsample of r_y of respondent to y. Asymptotically all the choices of functions for the pseudo-distance are equivalent to the linear one presented here. Instrumental regression allows to offset the endogeneity of some of the auxiliary variables $\mathbf{x}_k$. If some $\mathbf{x}_k$ are endogenous, the linear model $y_k = \mathbf{x}_k' \hat{\mathbf{B}}_{r_y} + e_k$ will be misspecified, $\hat{\mathbf{B}}_{r_y}$ will be biased, the residuals e_k will be biased as well or show some extreme values which will be harmful for the desirable properties of the estimator $\hat{t}_y^{cal}$!

2.2 Estimating the variance

One needs some procedure to estimate the variance of statistics based on the imputed dataset. Note that the bias may also be a non negligible problem which we do not address in this work. Variance may originate from several sources (see for instance Eurostat, 2013, for a more complete description): sampling variance, non-response variance, imputation variance, over-coverage variance, response-variance. When one seeks to estimate the variance of complex statistics like Laeken indicators, additional problems arise from the fact that the income distribution is hard to model and that indicators for poverty and social exclusion are non linear functions of y. Many estimation methods (replication methods and linearization methods) can lead to adequate estimator of variance under the commonly used sampling design.

We have chosen for generalized linearization methods mainly introduced by Deville (1999) and Demnati and Rao (2004) as other authors did, see for example Osier

(2009); Goga et al. (2009); Verma and Betti (2011); Langel and Tillé (2013). In this fieldwork we revisited the results and brought some improvements in the estimation of the income density function (Graf and Tillé, 2013).

Linearization is not only used to linearize complex statistics, but also with respect to the calibrations applied to obtain the final survey weights. Indeed, Deville (1999) shows that estimating the variance of calibrated estimators (Deville and Särndal, 1992) is equivalent to estimate the design variance of the residuals from the regression of the variable of interest (for us income) on the auxiliary variables used to calibrate. In fact the residuals of that regression are the linearized variable with respect to the calibration. Massiani (2012) detailed the procedure to estimate the variance of a total taking all the procedure used to produce the final survey weight for the Swiss SILC taking into account for stratified random cluster sample, non-response corrections on several levels, weight sharing (see Lavalle, 2002), panels combinations and calibration on known totals (see Graf, 2008, for a detailed description of the Swiss SILC weighting scheme).

2.3 GB2 distribution

The Generalized beta distribution of the second kind (GB2) is a four parameters distribution: $GB2(a, b, p, q)$, it has been developed by McDonald (1984). Its density is given by

$$f_{GB2}(y; a, b, p, q) = \frac{a}{b \cdot B(p, q)} \frac{(y/b)^{ap-1}}{(1 + (y/b)^a)^{p+q}}$$

where $B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt$ is the beta function. Many other probability distributions can be seen as special cases of the GB2 by re-parametrizing or setting some of the parameters equal to suitably fixed values (Generalized Gamma, Dagum, Beta2, Singh-Maddala, Lognormal, Gamma, Fisk, Weibul, etc.).

The advantage of a parametric estimation for a distribution of income is that there are explicit formulae for the inequality measures as functions of the four parameters $\theta = (a, b, p, q)$ of the GB2 adjusted to the data. The detailed expressions can be found McDonald (1984); AMELI (2011); Graf and Nedyalkova (2011, 2013). The main inequality measures of concern in SILC are programmed and available in statistical software **R**, `package("GB2")`.

3 Imputation strategy

Our imputation system should fulfil the following conditions: transparency, the model(s) used must be known; robustness, extreme values must not destabilize the model. As far as possible it must be able to cope with non ignorable non-response and respect the natural distribution of the income; it should allow the use of survey weights; it should permit variance estimation. Our imputation strategy can be summarized as follow:

1. For the respondents, produce *wisely* adjusted weights through generalized raking compensating for the non-response that affects the variable of interest. The y variable is used, among others, as instrument: $z_{k1} = y_k$.
2. Use this set of weight to approach the best possible GB2, that is the one we would get by fitting on all units (if all would have responded). The poverty and inequality measures can already be estimated parametrically at this stage without doing any imputation.
3. Order the income variable by increasing (eventually weighted) ranks. This robustifies the system because we first impute ranks and will estimate the y-values later on (in step 7).

(4) Transform the ranks into normal quantiles. The quantiles are normally distributed by construction.
(5) Impute by predicting the missing quantiles with the help of a classical weighted multiple linear regression model based on auxiliary variables and the weights computed in step 1.
(6) Back transformation: obtain imputed ranks from the predicted normal quantiles.
(7) Knowing the ranks, estimate values for the variable of interest by either using the adjusted $GB2$ from step 2, or by some interpolation between the two nearest (with respect to ranks) known responding values.

4 Results and Conclusion

For the five Laeken indicators mentioned, first results show no significant difference between the empirical values estimated out of the register data (no non-response, no imputation), and, on the one hand side, parametric estimates based only on the GB2 fit (step 2) on the respondents, and on the other hand side, empirical estimates based on the partially imputed data. At the same time, empirical or parametric estimates based only on the respondent dataset with no attempt to adjust the survey weights and without imputation were all significantly different (biased).

Thus our imputation method reveals the importance and usefulness of a set of weights capable to compensate for non ignorable non-response. It adapts itself and respects much more the natural distribution of income variables than a procedure needing the hypothesis that income would be (log)normally distributed. Our method allows, through a GB2 fit, to deliver estimates of inequality measures without or before imputing. Its precision on the unit level depends on the explanation power of the auxiliary variables at disposal and from the weights obtained through generalized raking. For the poverty and inequality measures considered, empirical and GB2-parametric calculations provide very similar results.

The choice of the instruments in the generalized raking is crucial. Defining a systematic method providing valuable instruments among the variables at disposal is still work in progress as well as the estimation of variance due to imputations for several estimators (linear or non linear).

References

AMELI (2011). Advanced methodology for european laeken indicators. FP7. EU-project: http://www.uni-trier.de.

Dastrup, S. R., Hartshorn, R., and McDonald, J. B. (2007). The impact of taxes and transfer payments on the distribution of income: A parametric comparison. *Journal of Economic Inequality*, 5:353–369.

Demnati, A. and Rao, J. N. K. (2004). Estimateurs de variance par linarisation pour des donnes d'enqute. *Techniques d'enqute*, 30:17–27.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25:193–204.

Deville, J.-C. (2002). La correction de la non-rponse par calage gnralis. In *JMS INSEE*.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88:1013–1020.

Eurostat (2005). The continuity of indicators during the transition between ECHP and EU-SILC. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg.

Eurostat (2013). Handbook on precision requirements and variance estimation for ess household surveys. http://www.cros-portal.eu/content/handbook-precision-requirements-and-variance-estimation-ess-household-surveys.

Goga, C., Deville, J.-C., and Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika*, 96, 3:691–709.

Graf, E. (2008). Pondrations du silc pilote silc_i vague 2, silc_ii vague 1, silc_i et silc_ii combins. Technical report, Office Fdral de la Statistique, Neuchtel. Rapport de mthodes, ISBN 978-3-303-00406-7.

Graf, E. and Tillé, Y. (2013). Estimation de variance par linarisation pour des indices de pauvret et d'exclusion sociale. *Submitted paper*.

Graf, M. and Nedyalkova, D. (2011). *GB2: Generalized Beta Distribution of the Second Kind: properties, likelihood, estimation.* R package version 1.0.

Graf, M. and Nedyalkova, D. (2013). Modeling of income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. *Review of Income and Wealth*.

Jenkins, S. P. (2007). Inequality and the gb2 income distribution. *Discussion Paper IZA No. 2831*.

Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32:133–142.

Langel, M. and Tillé, Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society - Series A*, 176:521–540.

Lavalle, P. (2002). *Le sondage indirect ou la mthode gnralise du partage des poids*. Edition de l'Universit de Bruxelles.

Le Guennec, J. and Sautory, O. (2002). Calmar 2: Une nouvelle version de la macro calmar de redressement d'chantillon par calage. In *JMS Proceedings*.

Massiani, A. (2012). Estimation de la variance d'indicateurs transversaux pour l'enqute silc en suisse. *Techniques d'enqutes*. paratre.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, 52:647–663.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3:167–195.

Osier, G. (2013). Dealing with non-ignorable non-response using generalised calibration: Simulation study based on the luxemburgish household budget survey. *Economie et Statistiques: Working papers du STATEC*, 65.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27:85–95.

Sautory, O. (2003). Calmar 2: A new version of the calmar calibration adjustment program. In *Proceedings of Statistics Canada's Symposium 2003*.

Sepanski, J. H. and Kong, J. (2008). A family of generalized beta distributions for income. *Advances and Applications in Statistics*, 10:75–84.

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, 8:1549–1576.