Imputation of income variables in a survey context and estimation of variance for indicators of poverty and social exclusion

Eric Graf

Institute of Statistics
University of Neuchâtel, Switzerland
www.unine.ch/statistics

Neuchâtel, 25th of November 2014

Outline

Motivation & Introduction
- Goal
- Nonresponse in SILC
- SDAP at the SFSO
- IVEware
- Swiss SILC09 : match with a register file

Imputation of Income Data with Generalized Calibration and GB2 Distribution : Illustration with SILC Data

Variance Estimation Using Linearization for Laeken Indicators

Goal

Through a household survey, we collect some income variable y.
We intend to do inference to the population in publishing estimated values with confidence intervals for several indicators of poverty and social exclusion.

Press release

```
in this country, the Gini index
is of 29% ± 1.5%.
```

- Diff. levels of nonresponse & sources of variance.

Need of procedure to impute the missings and estimate the variance.

SILC : Statistics on Income and Livings Conditions

- European project yielding comparable indicators on poverty, social welfare and exclusion and generally living conditions.
- In CH : about 7,500 responding households and 12,500 responding individuals, rotative panel on 4 years.
- Among the missing values, those for income variables are the most problematic :
  - On individual level : >100 income components are surveyed,
  - on HH-level : 30 income components.
  - Without proper treatments (re-weighting and/or imputation), inequality measures are biaised.
  - Informative knowledge about the income distribution is a great advantage.

SILC : Statistics on Income and Livings Conditions

- European project yielding comparable indicators on poverty, social welfare and exclusion and generally living conditions.
- In CH : about 7,500 responding households and 12,500 responding individuals, rotative panel on 4 years.
- Among the missing values, those for income variables are the most problematic :
  - On individual level : >100 income components are surveyed,
  - on HH-level : 30 income components.
  - Without proper treatments (re-weighting and/or imputation), inequality measures are biaised.
  - Informative knowledge about the income distribution is a great advantage.

SILC : Statistics on Income and Livings Conditions

- European project yielding comparable indicators on poverty, social welfare and exclusion and generally living conditions.
- In CH : about 7,500 responding households and 12,500 responding individuals, rotative panel on 4 years.
- Among the missing values, those for income variables are the most problematic :
  - On individual level : >100 income components are surveyed,
  - on HH-level : 30 income components.
  - Without proper treatments (re-weighting and/or imputation), inequality measures are biaised.
  - Informative knowledge about the income distribution is a great advantage.

SILC : Statistics on Income and Livings Conditions

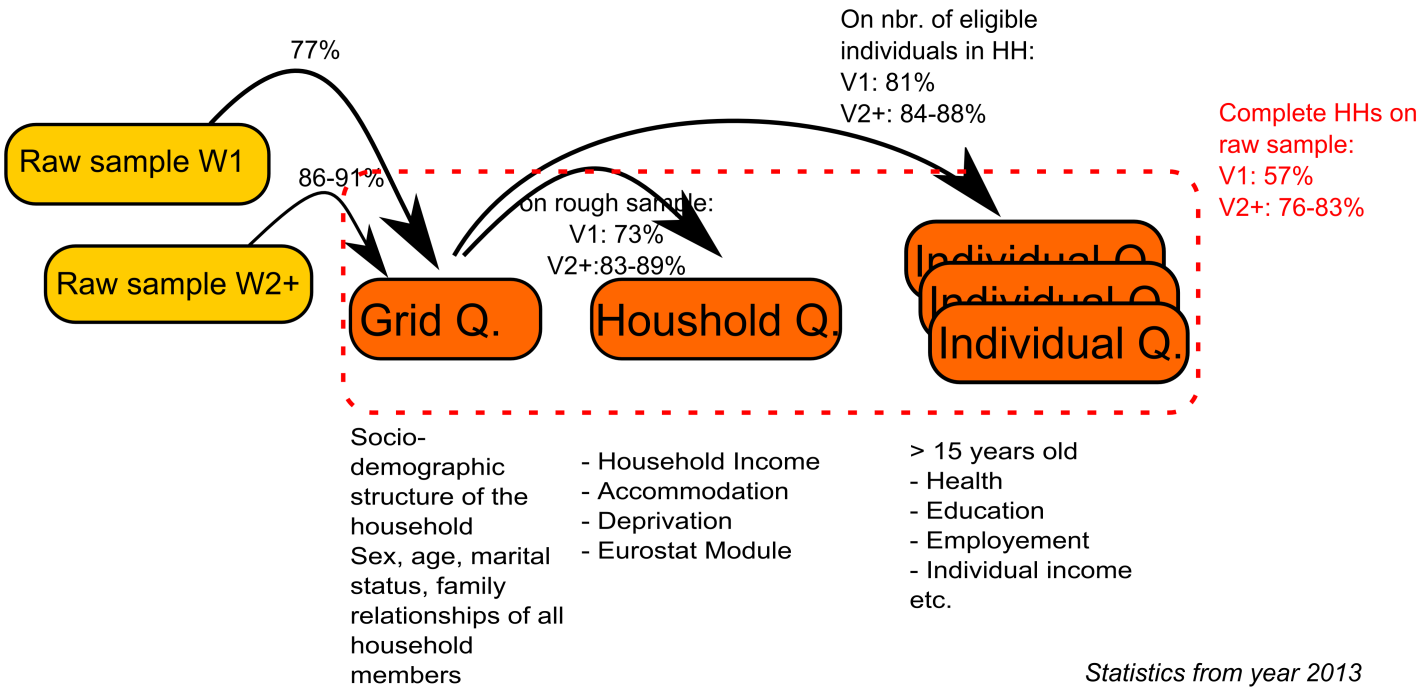
- European project yielding comparable indicators on poverty, social welfare and exclusion and generally living conditions.
- In CH : about 7,500 responding households and 12,500 responding individuals, rotative panel on 4 years.
- Among the missing values, those for income variables are the most problematic :
  - On individual level : >100 income components are surveyed,
  - on HH-level : 30 income components.
  - Without proper treatments (re-weighting and/or imputation), inequality measures are biaised.
  - Informative knowledge about the income distribution is a great advantage.

SILC : Statistics on Income and Livings Conditions

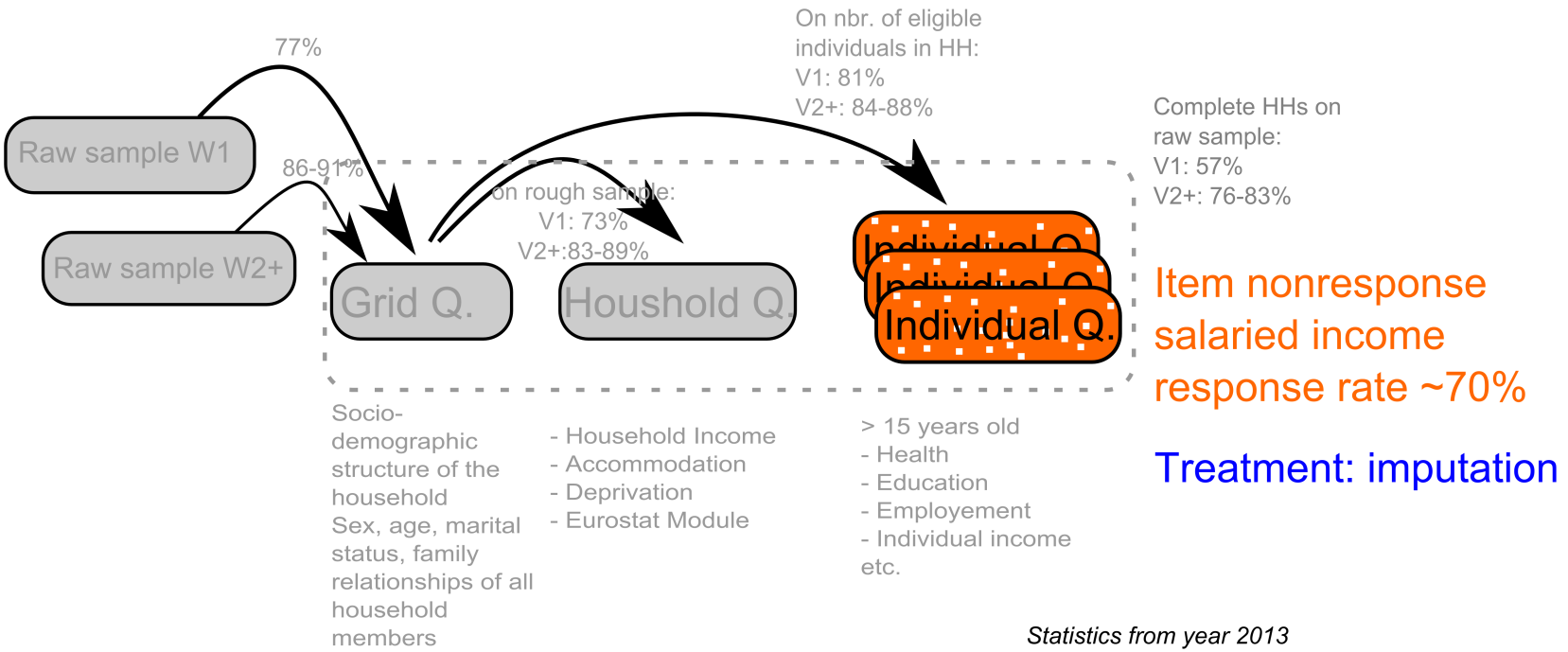
- European project yielding comparable indicators on poverty, social welfare and exclusion and generally living conditions.
- In CH : about 7,500 responding households and 12,500 responding individuals, rotative panel on 4 years.
- Among the missing values, those for income variables are the most problematic :
  - On individual level : >100 income components are surveyed,
  - on HH-level : 30 income components.
  - Without proper treatments (re-weighting and/or imputation), inequality measures are biaised.
  - Informative knowledge about the income distribution is a great advantage.

Response rates in SILC

Questionnaire NR appears at many different stages in the survey process, treated by reweighting.



Response rates in SILC. Example : salaried income



Statistics from year 2013

Statistical Data Preparation at the SFSO (regarding nonresponse to monetary variables)

- Quality control & editing of survey data (coherence control rules, validity rules, comparison with other sources at micro-/macro-level, matching with registers)
- Treatment for extreme-values and/or outliers
- Total/questionnaire NR : **re-weighting** (model NR via segmentation trees, calibrations),
- Item NR : multiple **imputations** (50 values) via regression model with IVEware software (Raghunathan et al., 2001) among imputation classes
- Imputation rate in SILC : 1.4 - 48.8% depending on the income component.

Statistical Data Preparation at the SFSO (regarding nonresponse to monetary variables)

- Quality control & editing of survey data (coherence control rules, validity rules, comparison with other sources at micro-/macro-level, matching with registers)
- Treatment for extreme-values and/or outliers
- Total/questionnaire NR : **re-weighting** (model NR via segmentation trees, calibrations),
- Item NR : multiple **imputations** (50 values) via regression model with IVEware software (Raghunathan et al., 2001) among imputation classes
- Imputation rate in SILC : 1.4 - 48.8% depending on the income component.

Statistical Data Preparation at the SFSO (regarding nonresponse to monetary variables)

- Quality control & editing of survey data (coherence control rules, validity rules, comparison with other sources at micro-/macro-level, matching with registers)
- Treatment for extreme-values and/or outliers
- Total/questionnaire NR : **re-weighting** (model NR via segmentation trees, calibrations),
- Item NR : multiple **imputations** (50 values) via regression model with IVEware software (Raghunathan et al., 2001) among imputation classes
- Imputation rate in SILC : 1.4 - 48.8% depending on the income component.

Statistical Data Preparation at the SFSO (regarding nonresponse to monetary variables)

- Quality control & editing of survey data (coherence control rules, validity rules, comparison with other sources at micro-/macro-level, matching with registers)
- Treatment for extreme-values and/or outliers
- Total/questionnaire NR : **re-weighting** (model NR via segmentation trees, calibrations),
- Item NR : multiple **imputations** (50 values) via regression model with IVEware software (Raghunathan et al., 2001) among imputation classes
- Imputation rate in SILC : 1.4 - 48.8% depending on the income component.

Statistical Data Preparation at the SFSO (regarding nonresponse to monetary variables)

- Quality control & editing of survey data (coherence control rules, validity rules, comparison with other sources at micro-/macro-level, matching with registers)
- Treatment for extreme-values and/or outliers
- Total/questionnaire NR : **re-weighting** (model NR via segmentation trees, calibrations),
- Item NR : multiple **imputations** (50 values) via regression model with IVEware software (Raghunathan et al., 2001) among imputation classes
- Imputation rate in SILC : 1.4 - 48.8% depending on the income component.

Problems met with IVEware

- The income distribution is not normal, neither log-normal.
- Cannot cope with NMAR nonresponse (Little and Rubin, 2002).
- Lack of transparency, insufficient output about adjusted models.
- Extreme values can destabilize the adjusted models : lack of robustness (Chambers, 1986 ; Hulliger, 1999).
- Cannot take the survey weights into account.
- Choice of number of multiple imputations to be conducted not easy.

Problems met with IVEware

- The income distribution is not normal, neither log-normal.
- Cannot cope with NMAR nonresponse (Little and Rubin, 2002).
- Lack of transparency, insufficient output about adjusted models.
- Extreme values can destabilize the adjusted models : lack of robustness (Chambers, 1986 ; Hulliger, 1999).
- Cannot take the survey weights into account.
- Choice of number of multiple imputations to be conducted not easy.

Problems met with IVEware

- The income distribution is not normal, neither log-normal.
- Cannot cope with NMAR nonresponse (Little and Rubin, 2002).
- Lack of transparency, insufficient output about adjusted models.
- Extreme values can destabilize the adjusted models : lack of robustness (Chambers, 1986 ; Hulliger, 1999).
- Cannot take the survey weights into account.
- Choice of number of multiple imputations to be conducted not easy.

Problems met with IVEware

- The income distribution is not normal, neither log-normal.
- Cannot cope with NMAR nonresponse (Little and Rubin, 2002).
- Lack of transparency, insufficient output about adjusted models.
- Extreme values can destabilize the adjusted models : lack of robustness (Chambers, 1986 ; Hulliger, 1999).
- Cannot take the survey weights into account.
- Choice of number of multiple imputations to be conducted not easy.

Problems met with IVEware

- The income distribution is not normal, neither log-normal.
- Cannot cope with NMAR nonresponse (Little and Rubin, 2002).
- Lack of transparency, insufficient output about adjusted models.
- Extreme values can destabilize the adjusted models : lack of robustness (Chambers, 1986 ; Hulliger, 1999).
- Cannot take the survey weights into account.
- Choice of number of multiple imputations to be conducted not easy.

Problems met with IVEware

- The income distribution is not normal, neither log-normal.
- Cannot cope with NMAR nonresponse (Little and Rubin, 2002).
- Lack of transparency, insufficient output about adjusted models.
- Extreme values can destabilize the adjusted models : lack of robustness (Chambers, 1986 ; Hulliger, 1999).
- Cannot take the survey weights into account.
- Choice of number of multiple imputations to be conducted not easy.

SILC09 : match with a register file

CATI survey

salary
15,000
.
2,506
45,000
110,000
.
84,100

CCO register

salary
15,900
67,140
3,050
40,152
115,000
33,000
84,100

- The SILC CATI-data could be matched with the register of the Central Compensation Office (CCO).

- We applied the true NR mechanism observed for CATI-data -salaried income- to the register-data. $\Rightarrow$ 30% missing values. We impute and compare to the true known values.
- That NR mechanism is non ignorable (NMAR).
- $corr(y_{CATI}, y_{CCO}) = 84\%$.

SILC09 : match with a register file

CATI survey

salary
15,000
.
2,506
45,000
110,000
.
84,100

CCO register

salary
15,900
67,140
3,050
40,152
115,000
33,000
84,100

- The SILC CATI-data could be matched with the register of the Central Compensation Office (CCO).

- We applied the true NR mechanism observed for CATI-data -salaried income- to the register-data. $\Rightarrow$ 30% missing values. We impute and compare to the true known values.
- That NR mechanism is non ignorable (NMAR).
- $corr(y_{CATI}, y_{CCO}) = 84\%$.

SILC09 : match with a register file

CATI survey

salary
15,000
.
2,506
45,000
110,000
.
84,100

CCO register

salary
15,900
67,140
3,050
40,152
115,000
33,000
84,100

- The SILC CATI-data could be matched with the register of the Central Compensation Office (CCO).

- We applied the true NR mechanism observed for CATI-data -salaried income- to the register-data. $\Rightarrow$ 30% missing values. We impute and compare to the true known values.
- That NR mechanism is non ignorable (NMAR).
- $corr(y_{CATI}, y_{CCO}) = 84\%$.

Outline

Motivation & Introduction

Imputation of Income Data with Generalized Calibration and GB2 Distribution : Illustration with SILC Data

- Imputation strategy
- GB2 distribution and inequality (Laeken) indicators
- Generalized calibration
- How to choose & combine auxiliary/instrumental variables
- Results
- Conclusions - Imputation

Variance Estimation Using Linearization for Laeken Indicators

Income distribution, GB2 & imputation strategy

Two main ideas :

1. use the good fit and properties of the generalized beta distribution of the second kind *GB2* fitted on the income data in a regression imputation strategy,
2. use generalized calibration to compensate for NMAR nonresponse.

Imputation procedure I

1. For the respondents to y, produce adjusted weights, w_k^{cal}, through generalized calibration. These weights can compensate even for NMAR NR. A GB2-adjustment based on these weights is used to evaluate their relevance.
2. Order the income variable by increasing weighted ranks $R_m^w \in [0, n_r]$.
3. Compute normal scores by using the van der Waerden's method (Conover, 1999), based on ranks :

$$Q^w = \Phi^{-1}\left(\frac{R^w}{n_r + 1}\right)$$

Imputation procedure I

1. For the respondents to y, produce adjusted weights, w_k^{cal}, through generalized calibration. These weights can compensate even for NMAR NR. A GB2-adjustment based on these weights is used to evaluate their relevance.
2. Order the income variable by increasing weighted ranks $R_m^w \in [0, n_r]$.
3. Compute normal scores by using the van der Waerden's method (Conover, 1999), based on ranks :

$$Q^w = \Phi^{-1}\left(\frac{R^w}{n_r + 1}\right)$$

Imputation procedure I

1. For the respondents to y, produce adjusted weights, w_k^{cal}, through generalized calibration. These weights can compensate even for NMAR NR. A GB2-adjustment based on these weights is used to evaluate their relevance.
2. Order the income variable by increasing weighted ranks $R_m^w \in [0, n_r]$.
3. Compute normal scores by using the van der Waerden's method (Conover, 1999), based on ranks :

$$Q^w = \Phi^{-1}\left(\frac{R^w}{n_r + 1}\right)$$

Imputation procedure II

4. Imputation : adjust a classical weighted multiple linear regression model on the normally shaped Q^{wcal} and predict $\hat{Q}_o^{wcal}$ for the nonrespondents.
5. Back transformation : obtain imputed ranks from the predicted normal scores (quantiles) : $\hat{R}_o = \Phi(\hat{Q}_o^{wcal}) \in]0, 1[$.
6. The imputed values $\hat{R}_o$ of each nonrespondents intercalate between the respondent's ranks, by renumbering all these ranks from 0 to $n = n_r + n_o$, one defines the imputed ranks. The predicted income is then estimated for example by simple interpolation.

Imputation procedure II

4. Imputation : adjust a classical weighted multiple linear regression model on the normally shaped Q^{wcal} and predict $\hat{Q}_o^{wcal}$ for the nonrespondents.
5. Back transformation : obtain imputed ranks from the predicted normal scores (quantiles) : $\hat{R}_o = \Phi(\hat{Q}_o^{wcal}) \in]0, 1[$.
6. The imputed values $\hat{R}_o$ of each nonrespondents intercalate between the respondent's ranks, by renumbering all these ranks from 0 to $n = n_r + n_o$, one defines the imputed ranks. The predicted income is then estimated for example by simple interpolation.

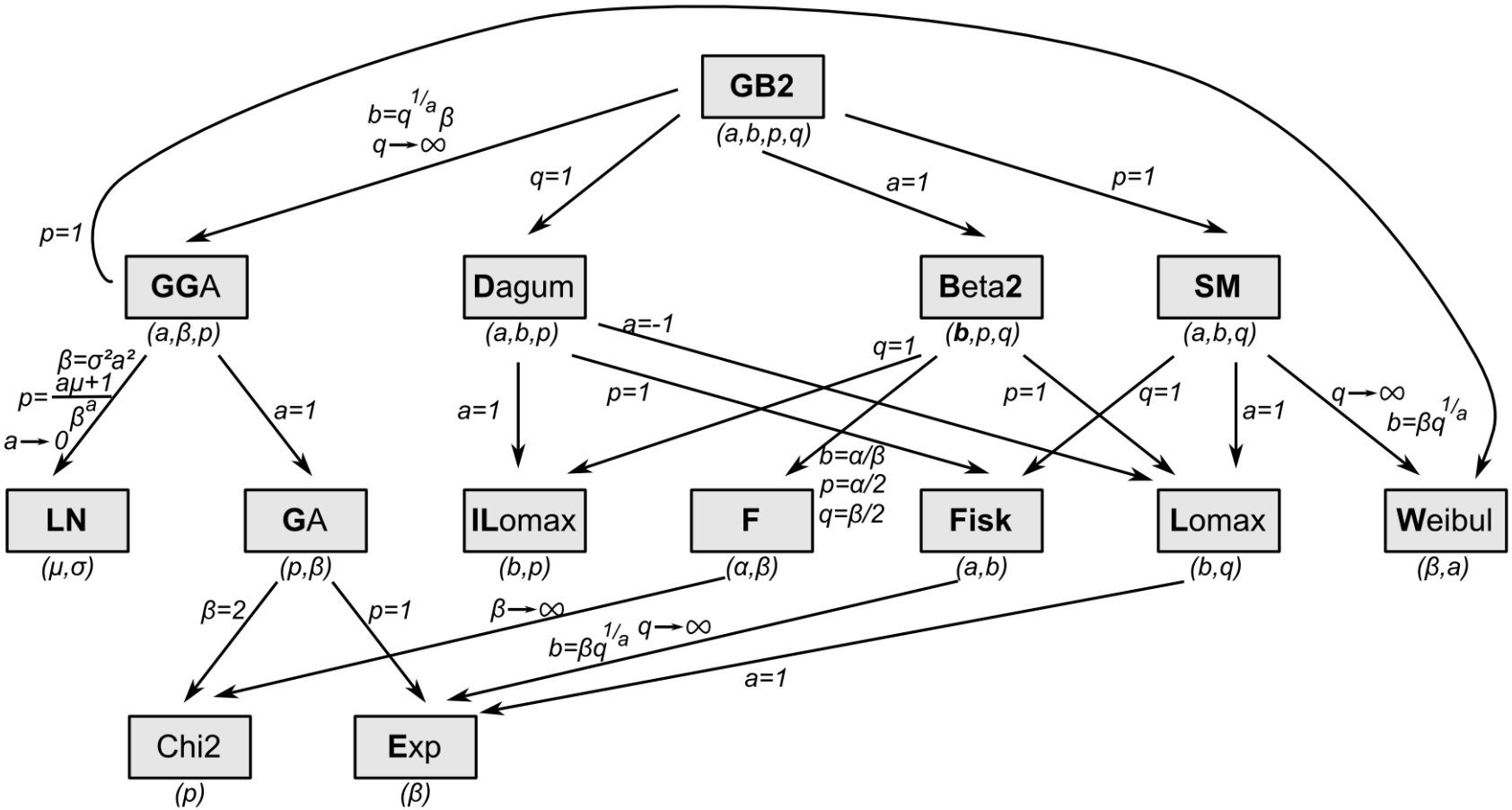
Imputation procedure II

4. Imputation : adjust a classical weighted multiple linear regression model on the normally shaped Q^{wcal} and predict $\hat{Q}_o^{wcal}$ for the nonrespondents.
5. Back transformation : obtain imputed ranks from the predicted normal scores (quantiles) : $\hat{R}_o = \Phi(\hat{Q}_o^{wcal}) \in]0, 1[$.
6. The imputed values $\hat{R}_o$ of each nonrespondents intercalate between the respondent's ranks, by renumbering all these ranks from 0 to $n = n_r + n_o$, one defines the imputed ranks. The predicted income is then estimated for example by simple interpolation.

Generalized beta distribution of the second kind (GB2)

- four parameters distribution : $GB2(a, b, p, q)$ (McDonald, 1984).
- Empirical studies on income - see for ex. Kleiber and Kotz (2003) ; Jenkins (2007) ; Dastrup et al. (2007) ; Sepanski and Kong (2008) ; Jones et al. (2011) ; McDonald et al. (2013) - show that the GB2 fits well with such data and it is often more suitable than other four-parameter distribution.
- Results from AMELI 2011 confirms for EU-SILC.
- Explicit formulae for the inequality measures, McDonald (1984) ; Graf (2009) ; AMELI 2011.

Many other probability distributions can be seen as special cases of the GB2



Generalized calibration I

- Calibration methods have been introduced by the reference articles of Deville and Särndal (1992) and Deville et al. (1993).
- Generalized calibration has been presented and applied by Le Guennec and Sautory (2002) ; Sautory (2003) ; Deville (2002).
- Kott (2006) ; Osier (2013) ; Lesage and Haziza (2013b,a) ; Park and Kim (2014) studied generalized calibration and its ability to correct nonresponse.
- The objective is to estimate a population total $t_y = \sum_U y_k$, ideally by $\hat{t}_y = \sum_s w_k y_k$ if all the units of s are available,
- more realistically based on the respondents' subset : $\hat{t}_y^{cal} = \sum_{k \in r} w_k^{cal} y_k$, where w_k^{cal}, $k \in r$, are the calibrated weights satisfying to the constraints $\sum_{k \in r} w_k^{cal} \mathbf{x}'_k = \hat{\mathbf{t}}_x$.

Generalized calibration I

- Calibration methods have been introduced by the reference articles of Deville and Särndal (1992) and Deville et al. (1993).
- Generalized calibration has been presented and applied by Le Guennec and Sautory (2002) ; Sautory (2003) ; Deville (2002).
- Kott (2006) ; Osier (2013) ; Lesage and Haziza (2013b,a) ; Park and Kim (2014) studied generalized calibration and its ability to correct nonresponse.
- The objective is to estimate a population total $t_y = \sum_U y_k$, ideally by $\hat{t}_y = \sum_s w_k y_k$ if all the units of s are available,
- more realistically based on the respondents' subset : $\hat{t}_y^{cal} = \sum_{k \in r} w_k^{cal} y_k$, where w_k^{cal}, $k \in r$, are the calibrated weights satisfying to the constraints $\sum_{k \in r} w_k^{cal} \mathbf{x}'_{k.} = \hat{\mathbf{t}}_x$.

Generalized calibration II

- the calibrated weights are of the form $w_k^{cal} = w_k F(\mathbf{z}_k.\boldsymbol{\lambda})$
  F : calibration function from $\mathbf{R}^J \rightarrow \mathbf{R}$ (can even be observation-dependent), usual choices are available : linear, raking ratio, logit, truncated (Deville et al., 1993).
- Simplest case : $F(u) = 1 + u$ (linear), leads to

$$\hat{t}_{ylinG} = \hat{\mathbf{t}}_x' \hat{\mathbf{B}}_{rzx} + \sum_{k \in r} w_k e_k^g,$$

where $e_k^g = y_k - \mathbf{x}_k.\hat{\mathbf{B}}_{rzx}$ are the residuals of the instrumental regression of y on the J auxiliary variables $\mathbf{x}_k.$ on sample r, with the J instrumental variables $\mathbf{z}_k.$ (Deville, 2000, 2002).

Generalized calibration II

- the calibrated weights are of the form $w_k^{cal} = w_k F(\mathbf{z}_k.\boldsymbol{\lambda})$
  F : calibration function from $\mathbf{R}^J \to \mathbf{R}$ (can even be observation-dependent), usual choices are available : linear, raking ratio, logit, truncated (Deville et al., 1993).
- Simplest case : $F(u) = 1 + u$ (linear), leads to

$$\hat{t}_{ylinG} = \hat{\mathbf{t}}_x' \hat{\mathbf{B}}_{rzx} + \sum_{k \in r} w_k e_k^g,$$

where $e_k^g = y_k - \mathbf{x}_k.\hat{\mathbf{B}}_{rzx}$ are the residuals of the instrumental regression of y on the J auxiliary variables $\mathbf{x}_k.$ on sample r, with the J instrumental variables $\mathbf{z}_k.$ (Deville, 2000, 2002).

Choice of auxiliary/instrumental variables combination

- z_k. only needs to be known on the sub-sample r of respondents to y.
- Asymptotically, in the case of full response, all the choices of calibration functions are equivalent to the linear.
- But : in the presence of nonresponse, the choice of calibration function reflects implicit assumptions on the nonresponse model (Lesage and Haziza, 2013a,b).
- Although we could verify this statement, praxis has shown that a good choice of the combination of auxiliary/instrumental variables is much more influential !

Choice of auxiliary/instrumental variables combination

- z_k. only needs to be known on the sub-sample r of respondents to y.
- Asymptotically, in the case of full response, all the choices of calibration functions are equivalent to the linear.
- But : in the presence of nonresponse, the choice of calibration function reflects implicit assumptions on the nonresponse model (Lesage and Haziza, 2013a,b).
- Although we could verify this statement, praxis has shown that a good choice of the combination of auxiliary/instrumental variables is much more influential !

Choice of auxiliary/instrumental variables combination

- $\mathbf{z}_k$. only needs to be known on the sub-sample r of respondents to y.
- Asymptotically, in the case of full response, all the choices of calibration functions are equivalent to the linear.
- But : in the presence of nonresponse, the choice of calibration function reflects implicit assumptions on the nonresponse model (Lesage and Haziza, 2013a,b).
- Although we could verify this statement, praxis has shown that a good choice of the combination of auxiliary/instrumental variables is much more influential !

Choice of auxiliary/instrumental variables combination

- z_k. only needs to be known on the sub-sample r of respondents to y.
- Asymptotically, in the case of full response, all the choices of calibration functions are equivalent to the linear.
- But : in the presence of nonresponse, the choice of calibration function reflects implicit assumptions on the nonresponse model (Lesage and Haziza, 2013a,b).
- Although we could verify this statement, praxis has shown that a good choice of the combination of auxiliary/instrumental variables is much more influential !

Why generalized calibration ?

- calibration $\leftrightarrow$ WLS regression
  generalized calibration $\leftrightarrow$ WIV regression,
- If the NR mechanism is imperfectly modelled, or if NMAR NR, or errors in some aux. var. : calibration estimator is biased,
- generalized calibration can be better if one finds a good combination of auxiliary and instrumental variables.
- ↯ no procedure available to identify good instruments a priori, only a posteriori checks are possible.

WLS regression ($\leftrightarrow$ calibration) :

$$
\begin{array}{rl}
U-level: & y_k = \mathbf{x}_k.\hat{\mathbf{B}}_U + e_k^U,\ k \in U, \\
p(s)\downarrow & \text{sampling} \\
s \subset U: & y_k = \mathbf{x}_k.\hat{\mathbf{B}}_s + e_k^s,\ k \in s, \\
\Psi(r|s)\downarrow & \text{NR mechanism} \\
r \subset s: & y_k = \mathbf{x}_k.\hat{\mathbf{B}}_r + e_k^r,\ k \in r.
\end{array}
$$

$$\hat{\mathbf{B}}_r = (\mathbf{X}_r'\hat{\mathbf{D}}_r\mathbf{X}_r)^{-1}\mathbf{X}_r'\hat{\mathbf{D}}_r\mathbf{y}$$
$$\hat{\mathbf{D}}_r = \mathrm{diag}(1/\pi_k \cdot 1/\widehat{\psi}_k)$$

Problem : ψ_k unperfectly modelled / NMAR NR / errors in X_r
Consequence : $\hat{\mathbf{B}}_r$ biased estimator of $\hat{\mathbf{B}}_U$.

WIV regression ($\leftrightarrow$ generalized calibration) :

$$r \subset s: y_k = \mathbf{x}_k.\hat{\mathbf{B}}_{rzx} + e_k^g,\ k \in r, \qquad \hat{\mathbf{B}}_{rzx} = (\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r)^{-1}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{y}$$

WLS regression ($\leftrightarrow$ calibration) :

$$
\begin{array}{rl}
U-level: & y_k = \mathbf{x}_k.\hat{\mathbf{B}}_U + e_k^U,\ k \in U, \\
p(s)\downarrow & \text{sampling} \\
s \subset U: & y_k = \mathbf{x}_k.\hat{\mathbf{B}}_s + e_k^s,\ k \in s, \\
\Psi(r|s)\downarrow & \text{NR mechanism} \\
r \subset s: & y_k = \mathbf{x}_k.\hat{\mathbf{B}}_r + e_k^r,\ k \in r.
\end{array}
$$

$$\hat{\mathbf{B}}_r = (\mathbf{X}_r'\hat{\mathbf{D}}_r\mathbf{X}_r)^{-1}\mathbf{X}_r'\hat{\mathbf{D}}_r\mathbf{y}$$
$$\hat{\mathbf{D}}_r = \mathrm{diag}(1/\pi_k \cdot 1/\widehat{\psi}_k)$$

Problem : ψ_k unperfectly modelled / NMAR NR / errors in X_r
Consequence : $\hat{\mathbf{B}}_r$ biased estimator of $\hat{\mathbf{B}}_U$.

WIV regression ($\leftrightarrow$ generalized calibration) :

$$r \subset s: y_k = \mathbf{x}_k.\hat{\mathbf{B}}_{rzx} + e_k^g,\ k \in r, \qquad \hat{\mathbf{B}}_{rzx} = (\mathbf{z}_r'\hat{\mathbf{D}}_r\mathbf{x}_r)^{-1}\mathbf{z}_r'\hat{\mathbf{D}}_r\mathbf{y}$$

WLS regression ($\leftrightarrow$ calibration) :

$$
\begin{array}{rl}
U-level: & y_k = \mathbf{x}_k.\hat{\mathbf{B}}_U + e_k^U,\ k \in U, \\
p(s)\downarrow & \text{sampling} \\
s \subset U: & y_k = \mathbf{x}_k.\hat{\mathbf{B}}_s + e_k^s,\ k \in s, \\
\Psi(r|s)\downarrow & \text{NR mechanism} \\
r \subset s: & y_k = \mathbf{x}_k.\hat{\mathbf{B}}_r + e_k^r,\ k \in r.
\end{array}
\qquad
\begin{array}{l}
\hat{\mathbf{B}}_r = (\mathbf{X}_r'\hat{\mathbf{D}}_r\mathbf{X}_r)^{-1}\mathbf{X}_r'\hat{\mathbf{D}}_r\mathbf{y} \\
\hat{\mathbf{D}}_r = \mathrm{diag}(1/\pi_k \cdot 1/\widehat{\psi}_k)
\end{array}
$$

Problem : ψ_k unperfectly modelled / NMAR NR / errors in X_r
Consequence : $\hat{\mathbf{B}}_r$ biased estimator of $\hat{\mathbf{B}}_U$.

WIV regression ($\leftrightarrow$ generalized calibration) :

$$r \subset s : y_k = \mathbf{x}_k.\hat{\mathbf{B}}_{rzx} + e_k^g,\ k \in r, \qquad \hat{\mathbf{B}}_{rzx} = (\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r)^{-1}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{y}$$

Conditions on instruments

However, if we dispose of some other variables grouped in matrix $\mathbf{Z}_r$, such that, on r :

(i) $\frac{1}{n_r}\sum_{k\in r}\frac{1}{\pi_k}\frac{1}{\widehat{\psi}_k}z_{kj}e_k^U = O(n_r^{-\frac{1}{2}})$ for all $j = 1, ..., J$: the instrument $\mathbf{z}_{.j}$ is valid or *exogenous*.

(ii) $\left(\frac{1}{n_r}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r\right)^{-1}$ exists and is finite.

Then $\hat{\mathbf{B}}_U$ can still be estimated via the so-called *weighted instrumental variable regression estimator* $\hat{\mathbf{B}}_{rzx}$ and e_k^g can be taken as an estimate of e_k^U, $k \in r$.

$\not\downarrow e_k^U$ unknown, conditions cannot be checked a priori...

Conditions on instruments

However, if we dispose of some other variables grouped in matrix $\mathbf{Z}_r$, such that, on r :

(i) $\frac{1}{n_r}\sum_{k\in r}\frac{1}{\pi_k}\frac{1}{\widehat{\psi}_k}z_{kj}e_k^U = O(n_r^{-\frac{1}{2}})$ for all $j = 1, ..., J$: the instrument $\mathbf{z}_{.j}$ is valid or *exogenous*.

(ii) $\left(\frac{1}{n_r}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r\right)^{-1}$ exists and is finite.

Then $\hat{\mathbf{B}}_U$ can still be estimated via the so-called *weighted instrumental variable regression estimator* $\hat{\mathbf{B}}_{rzx}$ and e_k^g can be taken as an estimate of e_k^U, $k \in r$.

$\not\downarrow e_k^U$ unknown, conditions cannot be checked a priori...

Conditions on instruments

However, if we dispose of some other variables grouped in matrix $\mathbf{Z}_r$, such that, on r :

(i) $\frac{1}{n_r}\sum_{k\in r}\frac{1}{\pi_k}\frac{1}{\widehat{\psi}_k}z_{kj}e_k^U = O(n_r^{-\frac{1}{2}})$ for all $j = 1, ..., J$: the instrument $\mathbf{z}_{.j}$ is valid or *exogenous*.

(ii) $\left(\frac{1}{n_r}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r\right)^{-1}$ exists and is finite.

Then $\hat{\mathbf{B}}_U$ can still be estimated via the so-called *weighted instrumental variable regression estimator* $\hat{\mathbf{B}}_{rzx}$ and e_k^g can be taken as an estimate of e_k^U, $k \in r$.

$\lightning$ e_k^U unknown, conditions cannot be checked a priori...

Conditions on instruments

However, if we dispose of some other variables grouped in matrix $\mathbf{Z}_r$, such that, on r :

(i) $\frac{1}{n_r}\sum_{k\in r}\frac{1}{\pi_k}\frac{1}{\widehat{\psi}_k}z_{kj}e_k^U = O(n_r^{-\frac{1}{2}})$ for all $j = 1, ..., J$: the instrument $\mathbf{z}_{.j}$ is valid or *exogenous*.

(ii) $\left(\frac{1}{n_r}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r\right)^{-1}$ exists and is finite.

Then $\hat{\mathbf{B}}_U$ can still be estimated via the so-called *weighted instrumental variable regression estimator* $\hat{\mathbf{B}}_{rzx}$ and e_k^g can be taken as an estimate of e_k^U, $k \in r$.

$\lightning$ e_k^U unknown, conditions cannot be checked a priori...

Novel idea : check via GB2-fit

Best possible GB2-adjustment = the one obtained on the full dataset (without NR)

Iterative process until satisfactory GB2-fit obtained, given two lists :

- auxiliary variables : explain y & total known on s.
- instruments : explain the NR & observed at least on r.

1. With some combination of aux. and instr. variables produce weights compensating for NR through generalized calibration.
2. Relying on these weights, adjust a GB2 on the income data.
3. Graphically evaluate how near the obtained GB2-density is from the best possible GB2-adjustment.

Novel idea : check via GB2-fit

Best possible GB2-adjustment = the one obtained on the full dataset (without NR)

Iterative process until satisfactory GB2-fit obtained, given two lists :

- auxiliary variables : explain y & total known on s.
- instruments : explain the NR & observed at least on r.

1. With some combination of aux. and instr. variables produce weights compensating for NR through generalized calibration.
2. Relying on these weights, adjust a GB2 on the income data.
3. Graphically evaluate how near the obtained GB2-density is from the best possible GB2-adjustment.

Novel idea : check via GB2-fit

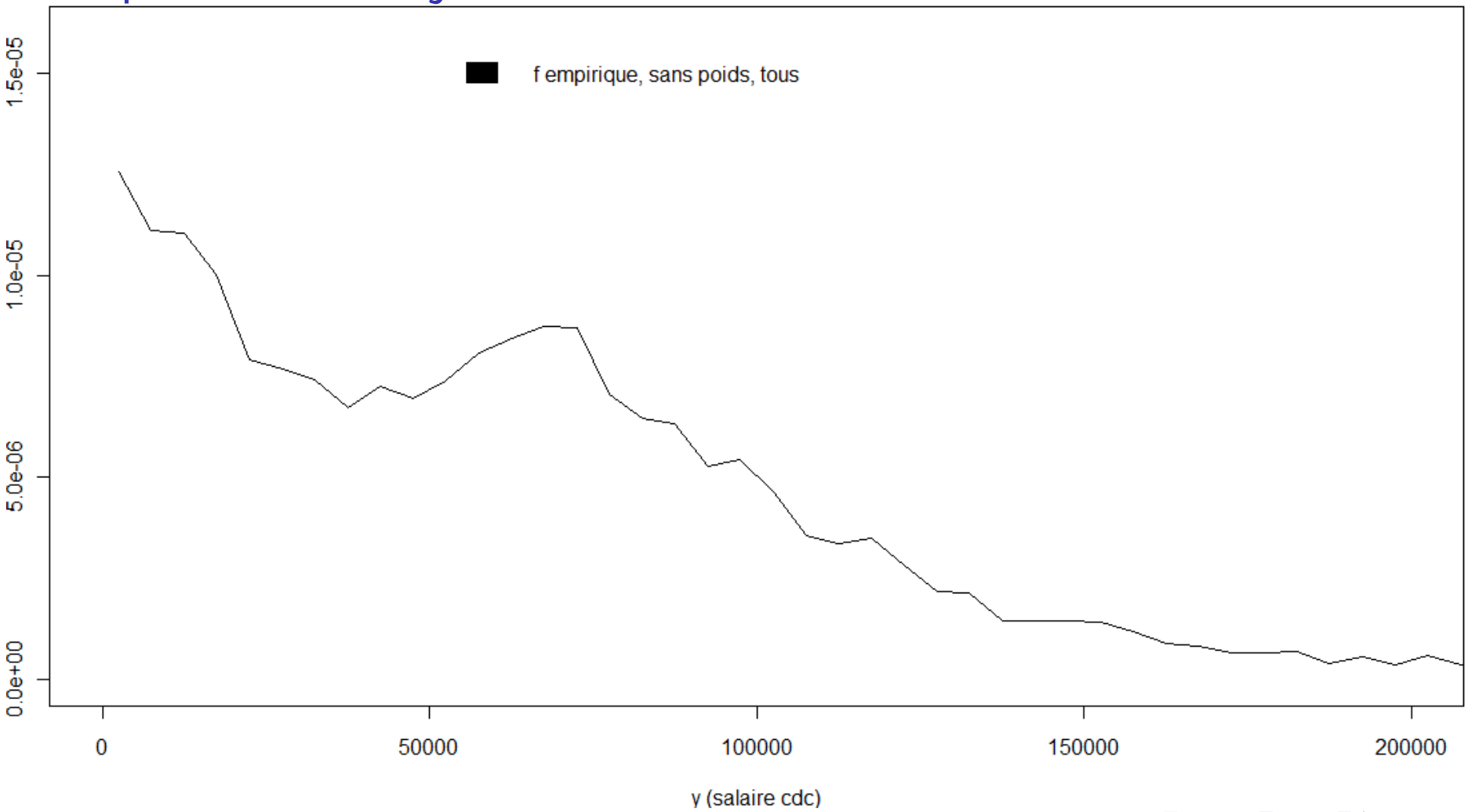
Best possible GB2-adjustment = the one obtained on the full dataset (without NR)

Iterative process until satisfactory GB2-fit obtained, given two lists :

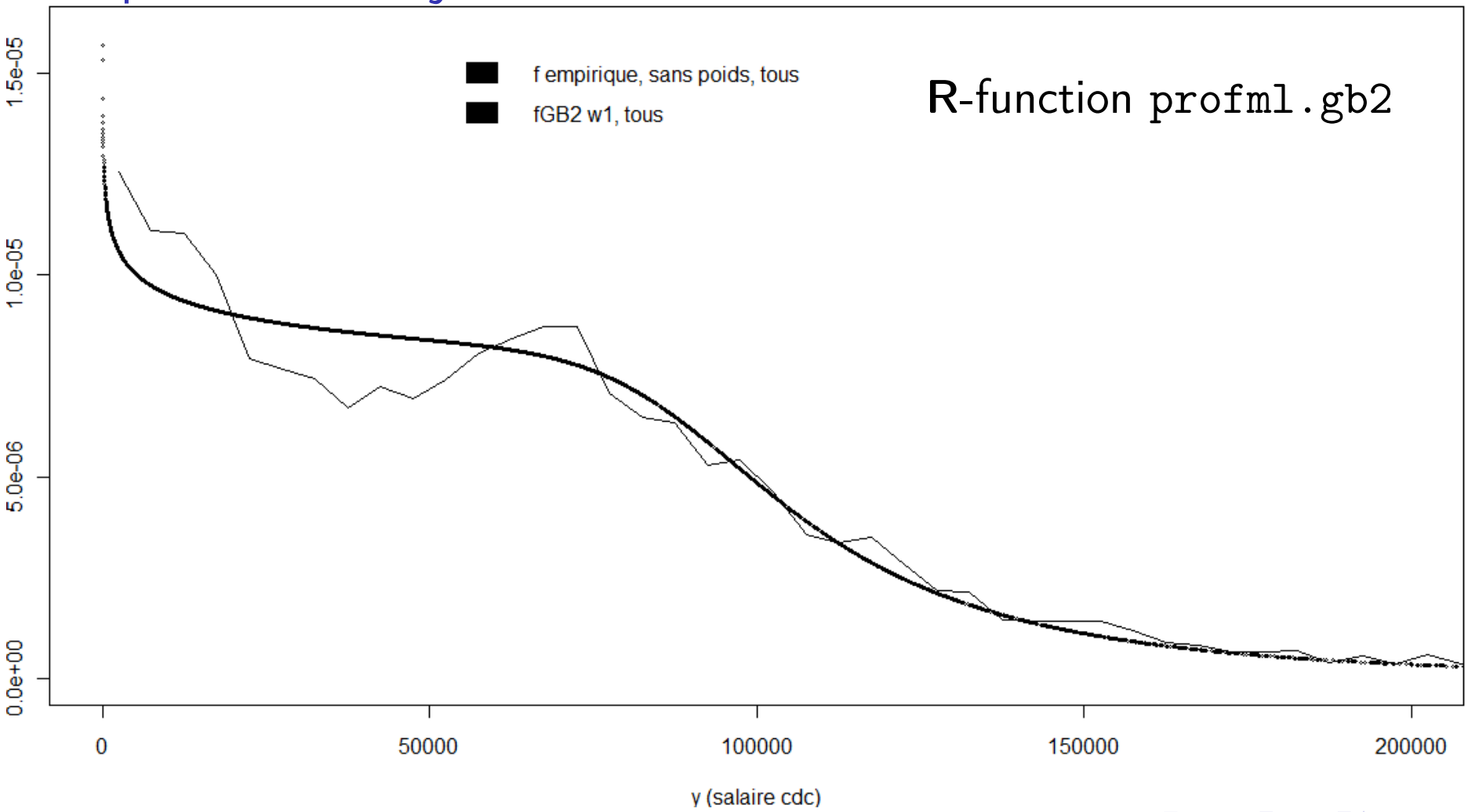
- auxiliary variables : explain y & total known on s.
- instruments : explain the NR & observed at least on r.

1. With some combination of aux. and instr. variables produce weights compensating for NR through generalized calibration.
2. Relying on these weights, adjust a GB2 on the income data.
3. Graphically evaluate how near the obtained GB2-density is from the best possible GB2-adjustment.

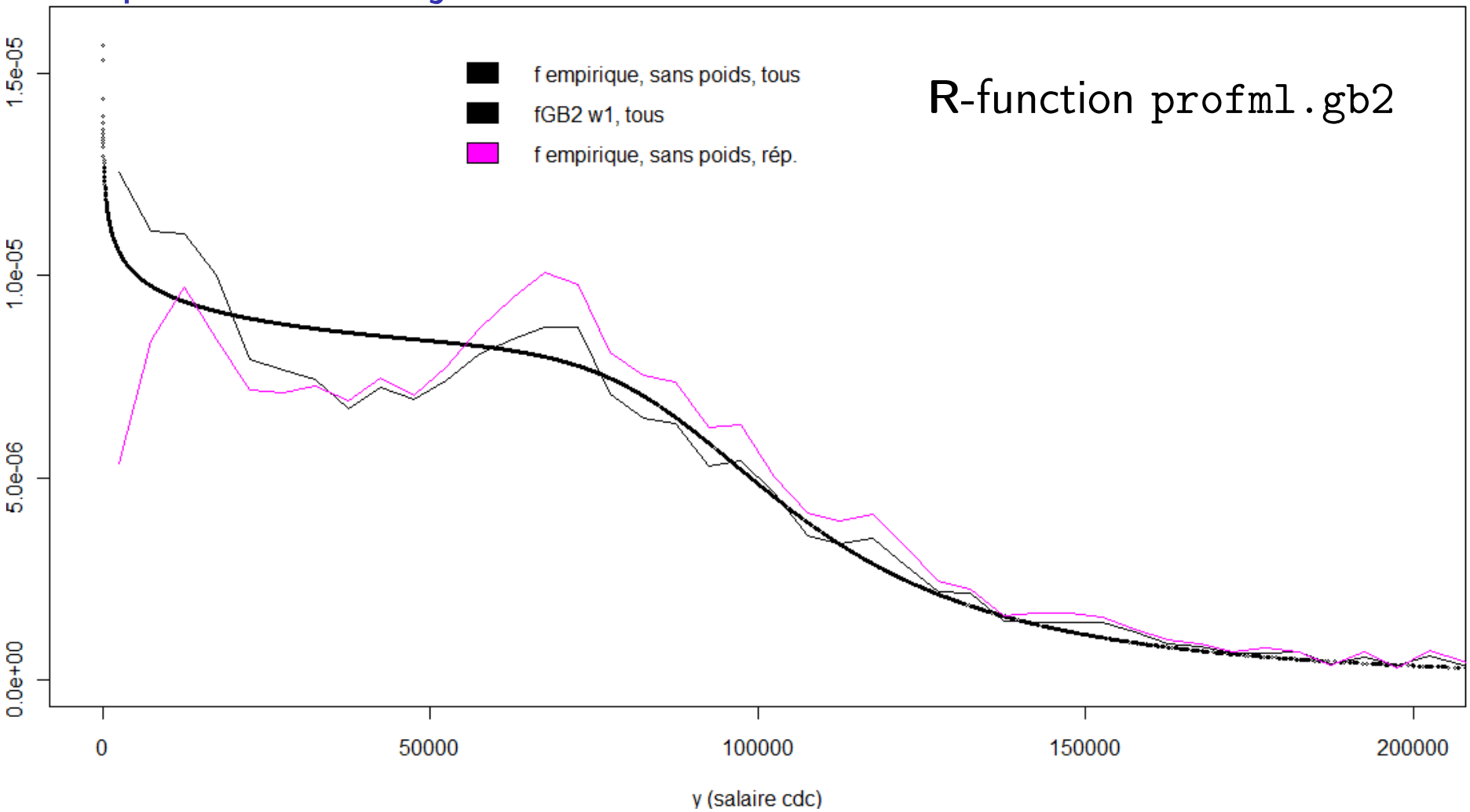
Example : GB2 adjustments on SILC09, salaried income



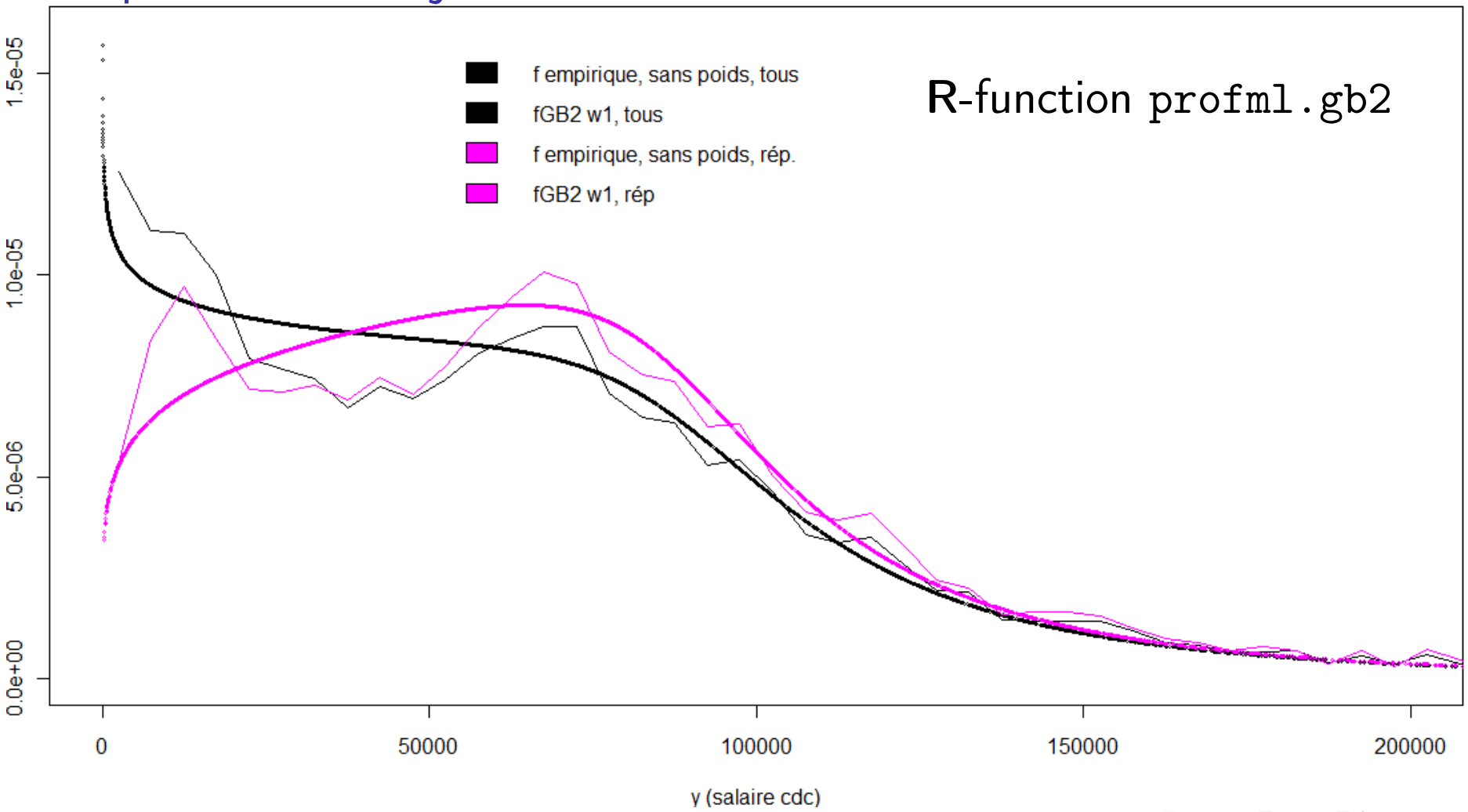
Example : GB2 adjustments on SILC09, salaried income



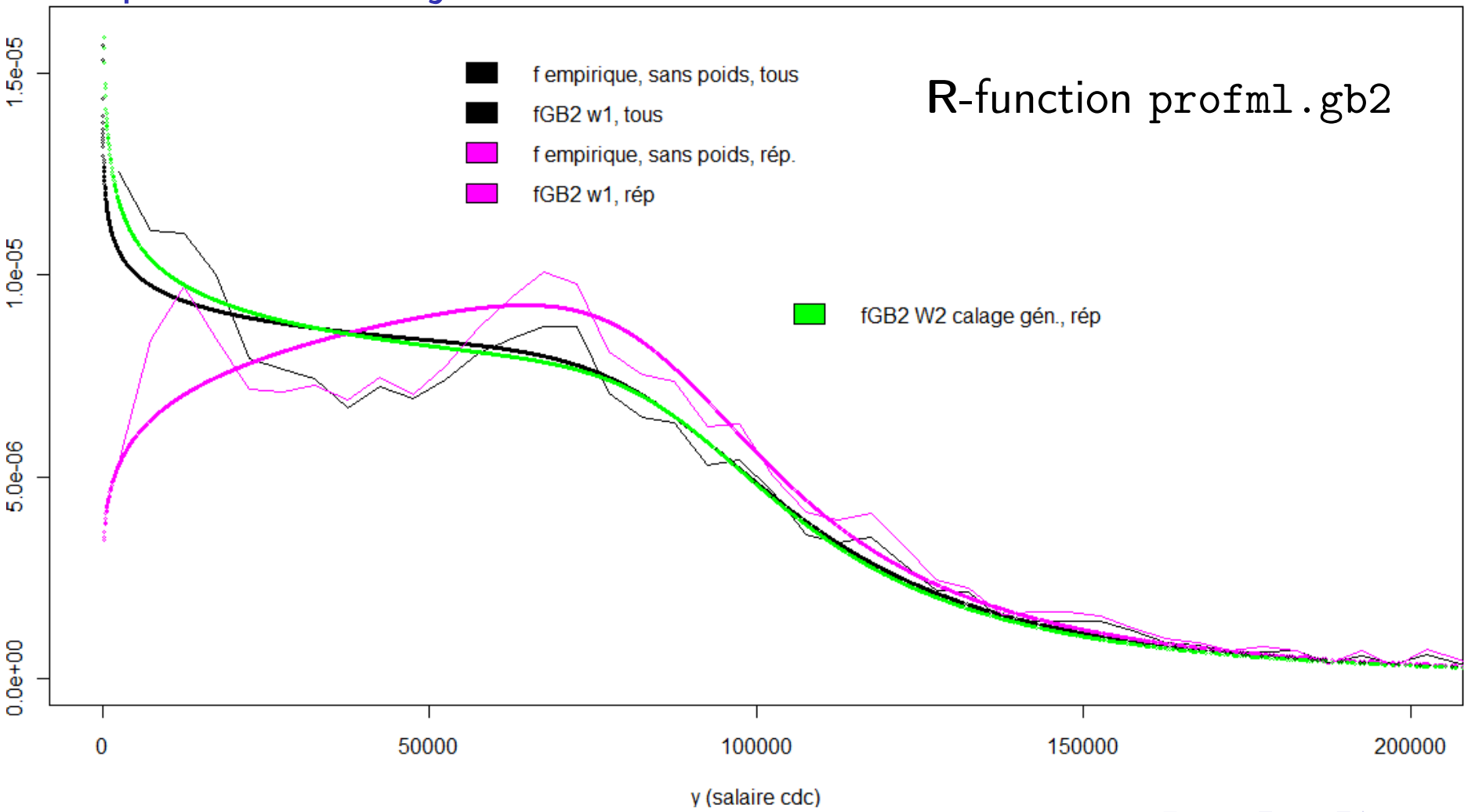
Example : GB2 adjustments on SILC09, salaried income



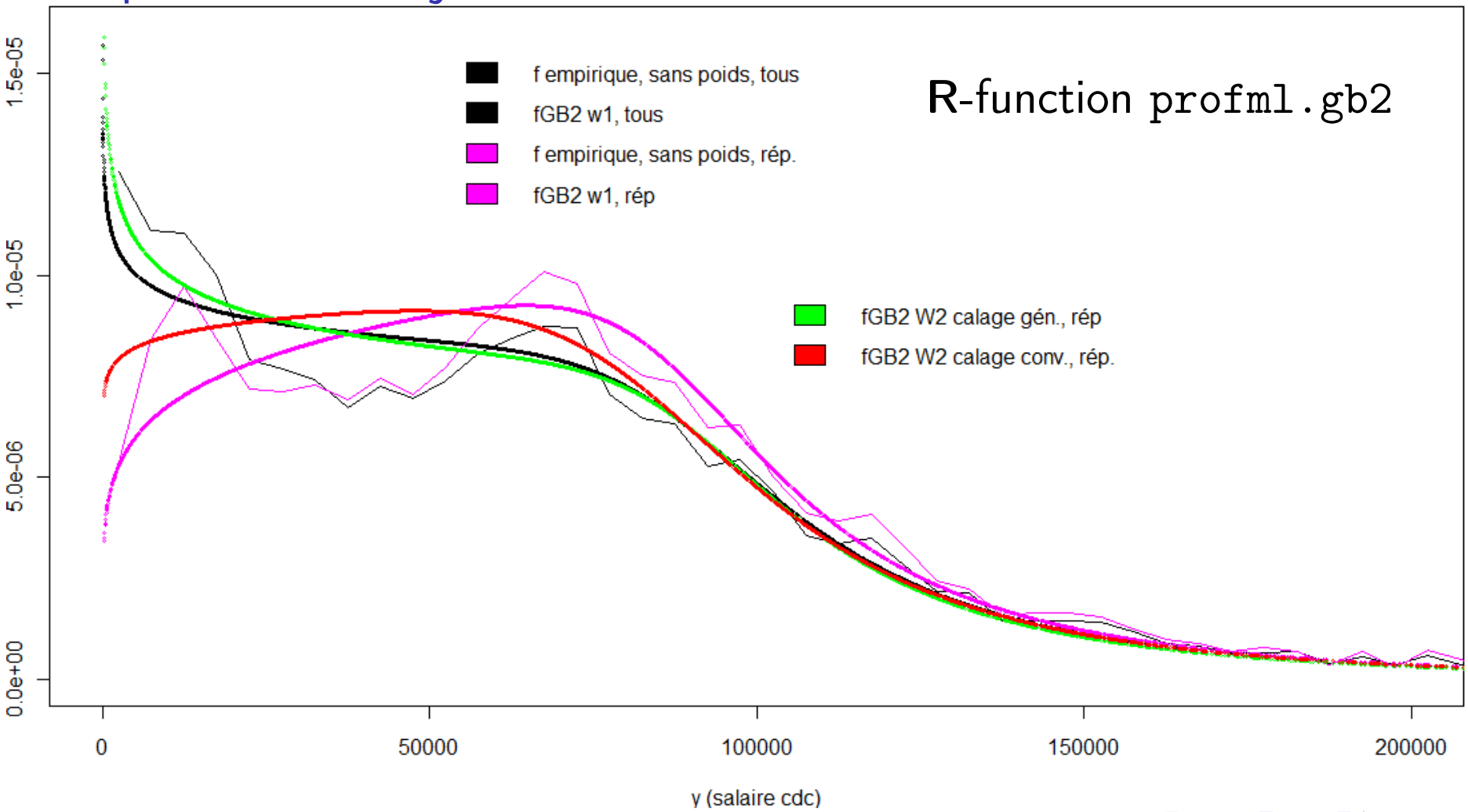
Example : GB2 adjustments on SILC09, salaried income



Example : GB2 adjustments on SILC09, salaried income

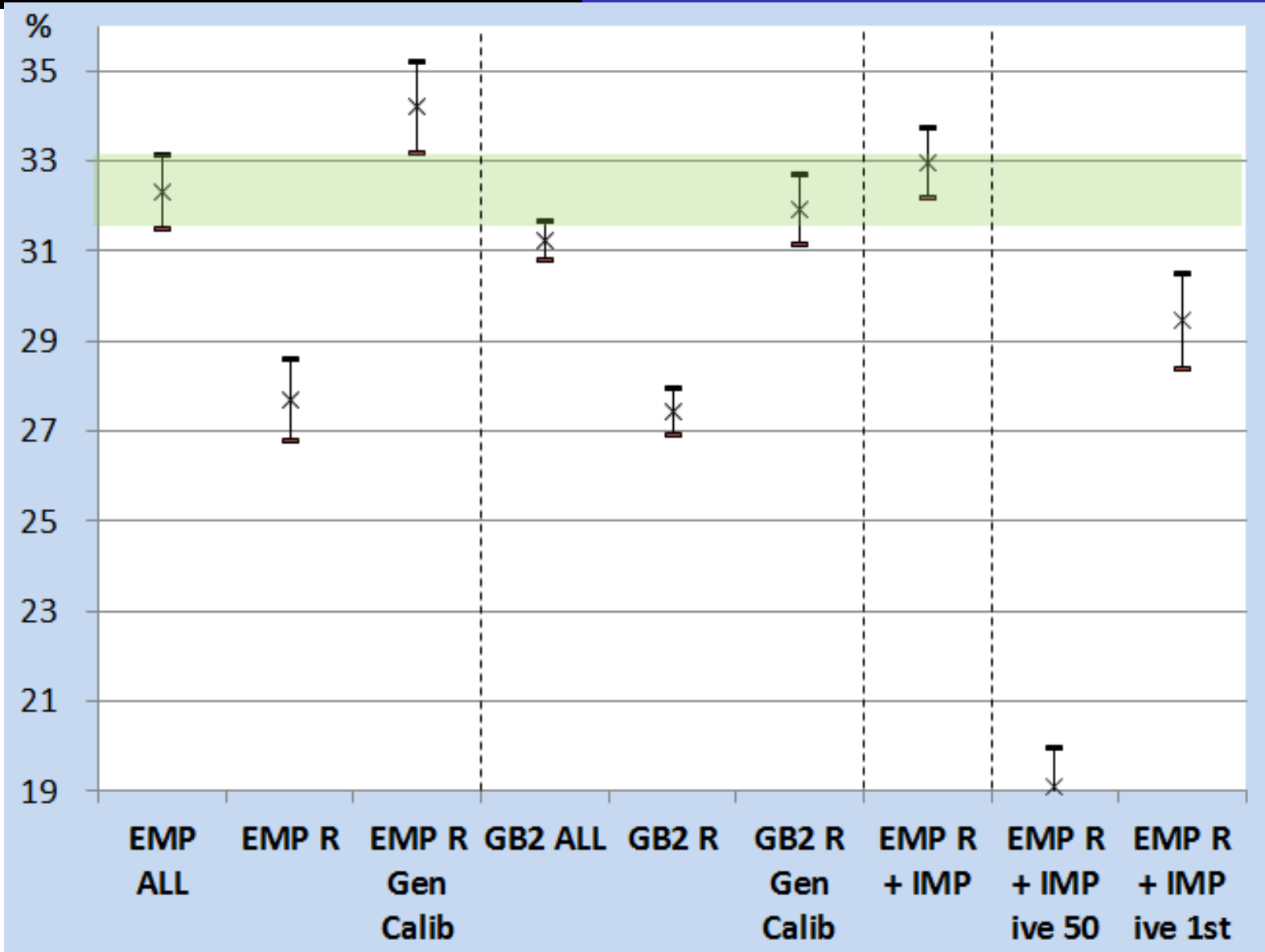


Example : GB2 adjustments on SILC09, salaried income



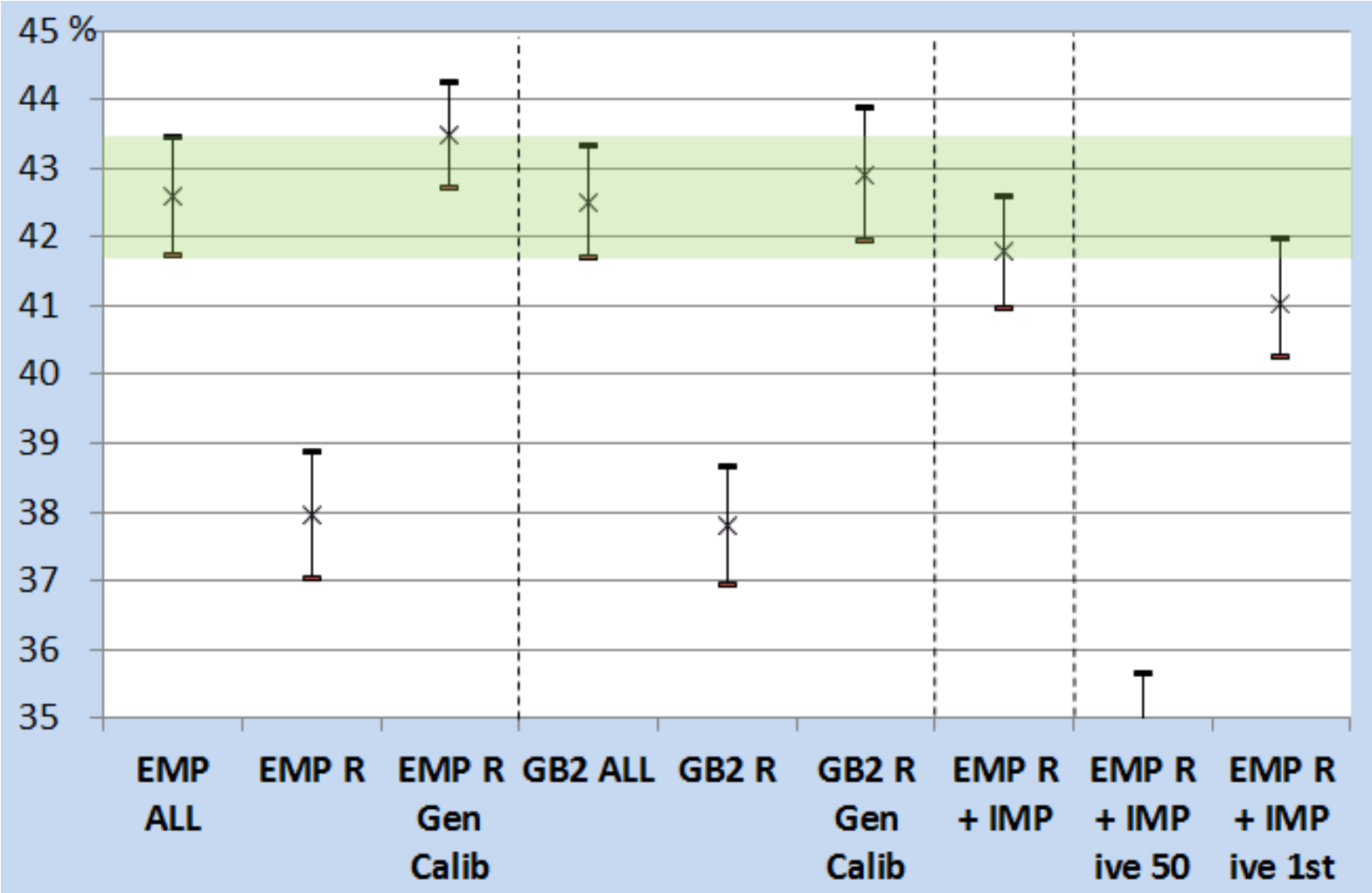
Results SILC09, salaried income

ARPR

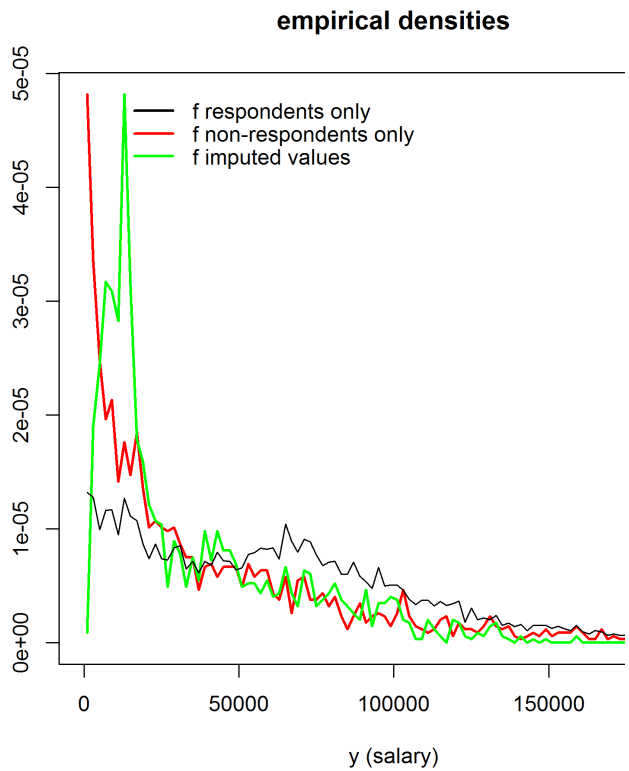


Results SILC09, salaried income

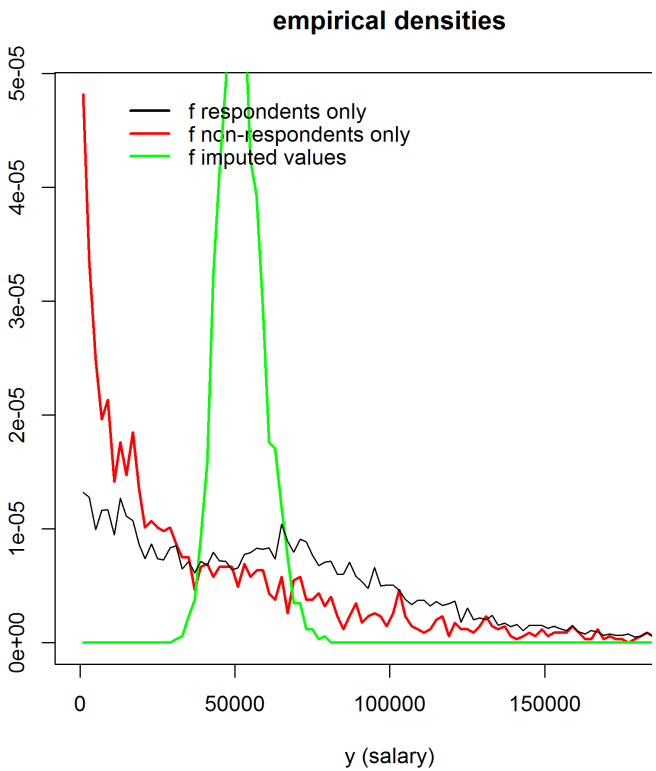
GINI



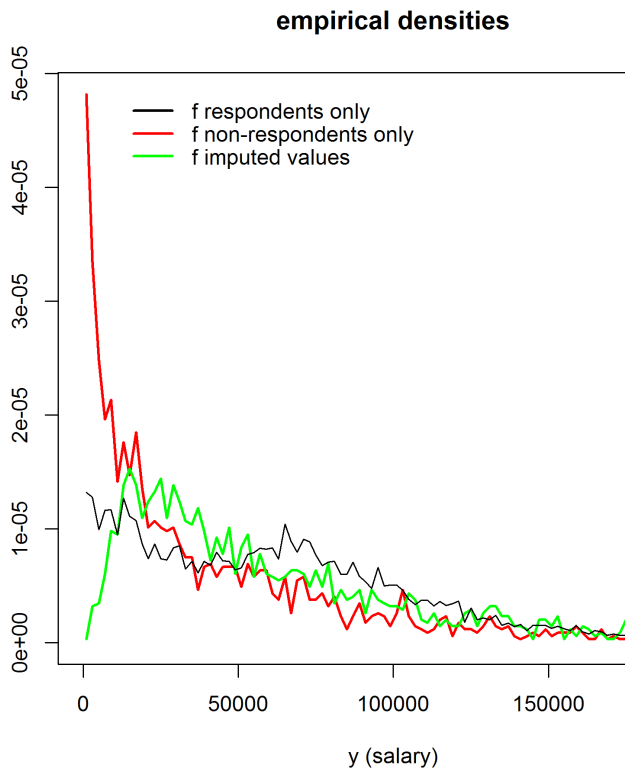
Results



Our method.



IVEware, mean on 50 imp.



IVEware, 1 imp.

Conclusions - Imputation method

- We used the GB2 adjustment as a graphical test to identify the best combination of auxiliary/instrumental variables defining the gen. calib. $\Rightarrow$ weights compensating for NMAR NR.
- the choice of good combination of aux. and instr. variables is more important than the choice of the calibration function.
- Imputations fit and respect the natural distribution of income variables.
- Significatively better results than IVEware.

Conclusions - Imputation method

- We used the GB2 adjustment as a graphical test to identify the best combination of auxiliary/instrumental variables defining the gen. calib. $\Rightarrow$ weights compensating for NMAR NR.
- the choice of good combination of aux. and instr. variables is more important than the choice of the calibration function.
- Imputations fit and respect the natural distribution of income variables.
- Significatively better results than IVEware.

Conclusions - Imputation method

- We used the GB2 adjustment as a graphical test to identify the best combination of auxiliary/instrumental variables defining the gen. calib. $\Rightarrow$ weights compensating for NMAR NR.
- the choice of good combination of aux. and instr. variables is more important than the choice of the calibration function.
- Imputations fit and respect the natural distribution of income variables.
- Significatively better results than IVEware.

Conclusions - Imputation method

- We used the GB2 adjustment as a graphical test to identify the best combination of auxiliary/instrumental variables defining the gen. calib. $\Rightarrow$ weights compensating for NMAR NR.
- the choice of good combination of aux. and instr. variables is more important than the choice of the calibration function.
- Imputations fit and respect the natural distribution of income variables.
- Significatively better results than IVEware.

Outline

Motivation & Introduction

Imputation of Income Data with Generalized Calibration and GB2 Distribution : Illustration with SILC Data

Variance Estimation Using Linearization for Laeken Indicators

- Generalized linearization
- Empirical influence functions for poverty indicators
- Estimation of income density function
- Results from simulations
- Conclusion for generalized linearization

Generalized linearization technique

- Usable for estimating the variance of sample-based non linear statistics (Deville, 1999 ; Demnati and Rao, 2004)
- Relies on the concept of influence functions proposed in robust statistic (Hampel, 1974)
- Applied by Osier (2009) for several EU-SILC indicators

Key result : Under asymptotic conditions, that are in principle satisfied if the sample is "large enough" : the variance of the estimated total of *estimated linearized variable* or *empirical influence function* $\hat{z}_k$ is an approximation of the variance of the (complex) statistic $\widehat{A}$.

Generalized linearization technique

- Usable for estimating the variance of sample-based non linear statistics (Deville, 1999 ; Demnati and Rao, 2004)
- Relies on the concept of influence functions proposed in robust statistic (Hampel, 1974)
- Applied by Osier (2009) for several EU-SILC indicators

Key result : Under asymptotic conditions, that are in principle satisfied if the sample is "large enough" : the variance of the estimated total of *estimated linearized variable* or *empirical influence function* $\hat{z}_k$ is an approximation of the variance of the (complex) statistic $\widehat{A}$.

Influence functions for poverty and inequality measure

See Osier (2009) ; Langel and Tillé (2011, 2013) ; Graf and Tillé (2014)

Example : $A = \mathrm{ARPT}$ At risk of poverty threshold

$$\hat{z}_k^{\mathrm{ARPT}} = -\frac{0.6}{f(\hat{q}_{50})}\frac{1}{\hat{N}}[\mathbf{1}_{[y_k \leq \hat{q}_{\mathbf{50}}]} - 0.5]$$

$\hat{z}_k^{\mathrm{ARPR}}$, $\hat{z}_k^{m_p}$ and $\hat{z}_k^{\mathrm{RMPG}}$ also require to estimate $f(\cdot)$ in several points of the income distribution.

Estimation of income density function f

Deville (1999) ; Osier (2009) propose to proceed through Gaussian kernel density estimation

(1) $$\widehat{f_1}(x) = \frac{1}{h\sqrt{2\pi}} \frac{1}{\hat{N}} \sum_{k \in S} w_k \exp\left[-\frac{(x-y_k)^2}{2h^2}\right]$$

- h is the *bandwidth* that Osier estimates by $\hat{h} = \hat{\sigma}\hat{N}^{-0.2}$
- $\hat{\sigma}$ the empirical standard deviation of the income variable $\mathcal{Y}$.
- The estimation of σ is non robust (sensitive to extreme values)
- In surveys, one encounter often some concentration of observations around some values what may cause trouble with a fixed bandwidth.

(2) Estimate through the logarithm

$$\widehat{f}_2(x) = \frac{\widehat{f}_1(log(x))}{x}$$

Diminishes problems due to extreme values in the GK density estimation

(3) Nearest neighbours with minimal bandwidth $\widehat{f}_{NNMB}(x) = \frac{p(x)}{nh(x)}$

At each point of the distribution, use at least p neighbours and a minimal bandwidth $h(x) \geq h_{opt}$ where h_{opt} is the *rule of thumb* of Silverman (1986) to determine the bandwidth.
This solution is more robust and avoids problems arising from the concentration of observations around some values.

$$\widehat{f}_3(x) = \frac{\widehat{f}_{NNMB}(log(x))}{x}$$

(2) Estimate through the logarithm

$$\widehat{f}_2(x) = \frac{\widehat{f}_1(log(x))}{x}$$

Diminishes problems due to extreme values in the GK density estimation

(3) Nearest neighbours with minimal bandwidth $\widehat{f}_{NNMB}(x) = \frac{p(x)}{nh(x)}$
At each point of the distribution, use at least p neighbours and a minimal bandwidth $h(x) \geq h_{opt}$ where h_{opt} is the *rule of thumb* of Silverman (1986) to determine the bandwidth.
This solution is more robust and avoids problems arising from the concentration of observations around some values.

$$\widehat{f}_3(x) = \frac{\widehat{f}_{NNMB}(log(x))}{x}$$

Results from simulations

Relative bias of the variance

10,000 SRSWR from Swiss SILC09 data

y = salaried income, $N = 7,922$

Indicator	Sample size (sampling rate)								
	$n = 500(6.3\%)$			$n = 750(9.5\%)$			$n = 1000(12.6\%)$		
GINI	-0.03			-0.03			-0.02		
QSR	-0.00			0.00			0.00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	0.07	0.05	**0.13**	0.06	0.04	**0.10**	0.06	0.03	0.08
ARPR	-0.05	-0.04	-0.02	-0.05	-0.04	-0.01	-0.06	-0.05	-0.02
RMPG	**0.61**	**0.12**	**0.15**	**0.60**	**0.11**	0.08	**0.59**	0.09	0.05
MEDP	**0.73**	**0.17**	**0.18**	**0.72**	**0.16**	**0.10**	**0.72**	**0.15**	0.07
MED	0.07	0.04	**0.13**	0.06	0.04	**0.10**	0.05	0.03	0.07

Conclusion for generalized linearization I

Generalized linearization permits for large samples ($\geq$ 1000 units) to estimate variances for widely used complex indicators, but one must be careful how the income density is estimated in such computations :

- The Gaussian kernel density estimation method currently implemented in most cases is not recommended without at least using the logarithm.

Conclusion for generalized linearization II

- The nearest neighbour method which also imposes a minimum bandwidth, may yield even better results, especially if there are agglomerations of observations with certain values in the given data. However, this method requires setting a minimum number p of neighbours.
- Attention : In presence of nonresponse, we have conducted other studies showing that ignoring the stage of imputation in the variance estimation may lead to severe underestimation of the total variance.

Acknowledgements

Special gratitude goes to :

- Prof. Yves Tillé, University of Neuchâtel,
- Dr. Philippe Eichenberger and Dr. Jean-Pierre Renfer from section METH at the Swiss Federal Statistical Office (SFSO),
- SFSO : this work was carried out within the framework of a cooperation agreement between the Institute of Statistics of the University of Neuchâtel and the SFSO.
- Thesis evaluation committee : Prof. Yves Tillé, Prof. Starica Catalin, Prof. Louis-Paul Rivest, Dr. Camelia Goga and Dr. Jean-Pierre Renfer.

Acknowledgements

Special gratitude goes to :

- Prof. Yves Tillé, University of Neuchâtel,
- Dr. Philippe Eichenberger and Dr. Jean-Pierre Renfer from section METH at the Swiss Federal Statistical Office (SFSO),
- SFSO : this work was carried out within the framework of a cooperation agreement between the Institute of Statistics of the University of Neuchâtel and the SFSO.
- Thesis evaluation committee : Prof. Yves Tillé, Prof. Starica Catalin, Prof. Louis-Paul Rivest, Dr. Camelia Goga and Dr. Jean-Pierre Renfer.

Thank you for your attention !

References I

Antal, E., Langel, M., and Tilllé, Y. (2011). Variance estimation of inequality indices in complex sampling designs. In *Proc. 58th World Statistical Congress*, Dublin. Session IPS056.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81 :1063–1069.

Conover, W. J. (1999). *Practical nonparametric statistics*. New York : J. Wiley, cop., 3rd ed. edition.

Croux, C. (1998). Limit behaviour of the empirical influence function of the median. *Statistics & Probability Letters*, 37 :331–340.

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). Linearization variance estimators for survey data. *Survey Methodology*, 30(1) :17–26.

Deng, L.-Y. and Chhikara, R. S. (1991). On asymptotically design-unbiased estimators of a finite population mean. *Biometrika*, 78 :189–195.

References II

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

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. In *Compstat - Proceedings in Computational Statistics : 14th Symposium Held in Utrecht, The Netherlands*, pages 65–76, New York. Springer.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. In *Actes des Journées de Méthodologie Statistique*, Paris. 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. and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10(4) :381–394.

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.

European Union (2011). Project AMELI : Advanced Methodology for European Laeken Indicators. FP7.

References III

Eurostat (2003). Laeken indicators - detailed calculation methodology. Doc.e2/ipse/2003, Directorate E : Social Statistics, Unit E-2 : Living Conditions, Luxembourg.

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.

Graf, E. and Tillé, Y. (2014). Variance estimation through linearization for poverty and social exclusion indicators. *Survey Methodology*, 40(1) :61–79.

Graf, M. (2009). An efficient algorithm for the computation of the gini coefficient of the generalized beta distribution of the second kind. In *JMS Proceedings*, pages 4835–4843. American Statistical Association.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69 :383–393.

Hulliger, B. (1999). Simple and robust estimators for sampling. In *Survey Research Methods Section*, pages 54–63. American Statistical Association.

References IV

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50 :361–365.

Jenkins, S. P. (2007). Inequality and the GB2 income distribution. *Discussion Paper IZA No. 2831*.

Jones, A., Lomas, J., and Rice, N. (2011). Applying beta-type size distributions to healthcare cost regressions. Technical report, University of York. Health, Econometrics and Data Group.

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

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2) :133–142.

Langel, M. and Tillé, Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference*, 141 :2976–2985.

Langel, M. and Tillé, Y. (2013). Variance estimation of the Gini index : Revisiting a result several times published. *Journal of the Royal Statistical Society*, A176(2) :521–540.

References V

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

Lee, H., Rancourt, E., and Särndal, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10(3) :231–243.

Lesage, E. and Haziza, D. (2013a). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*. To appear.

Lesage, E. and Haziza, D. (2013b). On the problem of bias and variance amplification of the instrumental calibration estimator in the presence of unit nonresponse. *Journal of Survey Statistics and Methodology*. To appear.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. Wiley, New York, 2nd edition.

Massiani, A. (2013). Estimation of the variance of cross-sectional indicators for the SILC survey in Switzerland. *Survey Methodology*, 39(1) :121–148.

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

References VI

McDonald, J. B., Sorensen, J., and Turley, P. A. (2013). Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth*, 2 :360–374.

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 :1–25.

Park, S. and Kim, J. K. (2014). Instrumental-variable calibration estimation in survey sampling. *Statistica Sinica*, 24 :1001–1015.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modles de régression. *Techniques d'enquête*, 27(1) :91–103.

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

References VII

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

SFSO (2014). Swiss pendent of the European Union statistics on Income and Living Conditions (EU-SILC).

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.