Consistent estimation at person-level and household-level

Anne Konrad, Jan Pablo Burgard, Ralf Münnich

Conference of European Statistics Stakeholders 2016
Budapest

University Trier

October 21, 2016

Motivation

- Many household surveys are based on cluster sampling: at the first stage the households are sampled, at the second stage all persons within a household.
- Allows the simultaneous estimation at the person- and at the household-level.
- In practice, **integrated weighting**, which substitutes individual auxiliary variables with (aggregated or) mean values, is often used.
- Eurostat recommends integrated weighting for EU-SILC (European Commission, 2013).

Research questions

1) Is there a price to pay to enforce consistent estimates due to the restriction of unique weights?

2) Does an alternative weighting strategy exists which is capable of both, ensuring consistent estimates at both levels and allowing for different weights for persons within the same household?

Table of contents

Table of contents

Usual person-level GREG estimator

The GREG estimator for totals is given by:

$$\hat{T}_{Y,GREG} = \hat{T}_{Y,HT} + \hat{\mathbf{B}}^{\mathbf{T}}(\mathbf{T}_{\mathbf{x}} - \hat{\mathbf{T}}_{\mathbf{x},\mathbf{HT}}) \quad (1)$$

with $\hat{\mathbf{B}} = (\mathbf{X}^{\mathbf{T}}\mathbf{\Pi}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{\Pi}^{-\mathbf{1}}\mathbf{Y}$ $(p \times 1)$ as regression coefficient.

Notation:

$\mathbf{Y}$: variable of interest $(n \times 1)$

$\mathbf{X}$: auxiliary variables $(n \times p)$

$\mathbf{T}_{\mathbf{x}}$: known totals of the auxiliaries $(p \times 1)$

$\hat{\mathbf{T}}_{\mathbf{x},\mathbf{HT}}$: estimated totals of the auxiliaries $(p \times 1)$

$\mathbf{\Pi}$: diagonal matrix with inclusion probabilities π_i $(n \times n)$

Integrated GREG estimator

Lemaître, G., Dufour, J. (1987): Substitution of the individual auxiliaries with their **constructed mean values**

The integrated GREG estimator for totals is given by:

$$\hat{T}_{Y,int} = \hat{T}_{Y,HT} + \mathbf{B}_{\mathbf{int}}^{\mathbf{T}}(\mathbf{T_x} - \hat{\mathbf{T}}_{\mathbf{x,HT}}) \quad (2)$$

with $\hat{\mathbf{B}}_{\mathbf{int}} = (\mathbf{D}^{\mathbf{T}}\mathbf{\Pi}^{-\mathbf{1}}\mathbf{D})^{-1}\mathbf{D}^{\mathbf{T}}\mathbf{\Pi}^{-\mathbf{1}}\mathbf{Y}$ $(p \times 1)$ as regression coefficient.

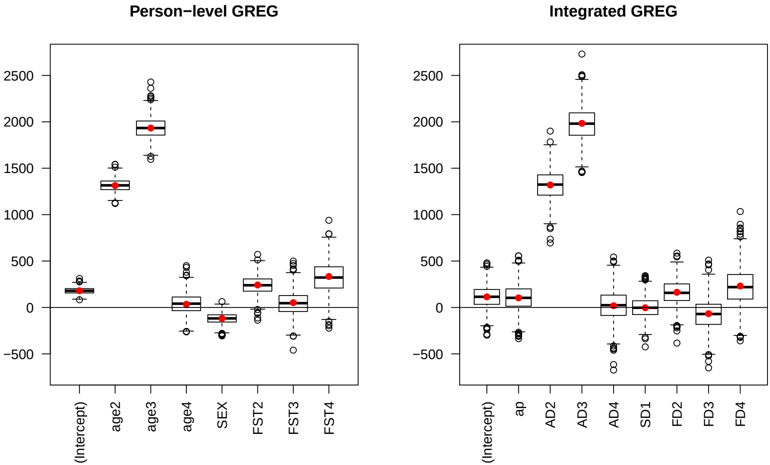
Further notation:

D: mean values of auxiliary variables $(n \times p)$

Simulation study: person-level vs. integrated GREG estimator

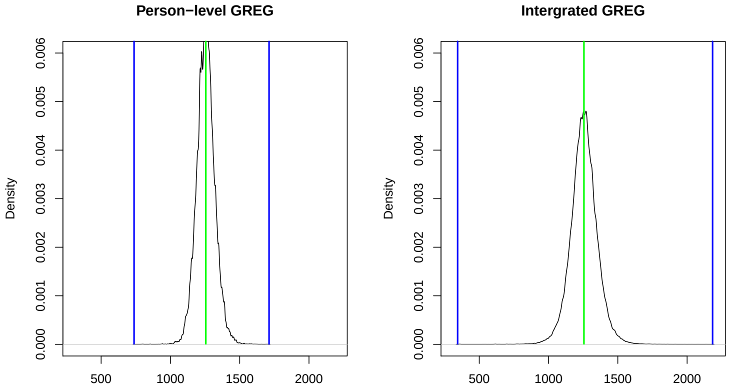
- Data: RIFOSS population of Rhineland-Palatinate (1,881,167 households and 4,225,729 persons)
- Sampling design: SRS of households of $n = 1500$
- Auxiliaries: sex, age classes, family status

Regression coefficients



$$\Rightarrow V(\mathbf{B_p}) < V(\mathbf{B_{int}})$$

Distribution of weights



⇒ Integrated weights have a significantly higher range!

Estimation results

	Person-level GREG	Integrated GREG
OCC_1	26,859	26,723
OCC_2	11,978	11,937
OCC_3	11,580	11,605
OCC_4	26,572	26,566
SELF	7,972	7,978
INC	121,242,544	120,915,970
UNEMP	7,179,708	7,181,217
PEN	39,823,873	39,970,062
PEK_HHG1	62,412,942	56,614,498
PEK_HHG2	101,730,314	99,704,938
PEK_HHG3	88,774,374	89,260,997
PEK_HHG4	87,359,338	85,215,552
PEK_HHG5	73,271,371	64,975,914
PEK_FST1	57,590,242	57,291,391
PEK_FST2	99,925,949	99,547,440
PEK_FST3	24,262,527	24,304,438
PEK_FST4	39,526,247	39,722,635

Table: MC standard errors

Table of contents

Alternative weighting approach

Idea: Intern consistency is solely required for common variables at the person- and household-level. Hence, utilize this common variables as additional auxiliaries in the calibration.

Modify the usual person-level GREG estimator and add the common variables matrix $\mathbf{C}$ $(n \times p)$:

$$\hat{T}_{y,Alternative} = \hat{T}_{y,HT} + \hat{\mathbf{B}}_{\mathbf{x}}^{\mathbf{T}}(\mathbf{T}_{\mathbf{x}} - \hat{\mathbf{T}}_{\mathbf{x,HT}}) + \hat{\mathbf{B}}_{\mathbf{c}}^{\mathbf{T}}(\hat{\mathbf{T}}_{\mathbf{c}} - \hat{\mathbf{T}}_{\mathbf{c,HT}})$$

Distribution of the weights

	Mean	SD	Min	Max	Range
Integrative GREG	66.69	4.90	21.58	116.98	95.04
Alternative GREG*	66.69	3.29	-37.00	172.07	209.07
Alternative GREG**	66.69	3.28	20.83	114.24	93.41

Table: Summary Statistics (3,365,765 observations)

* Improved model for common variables

** Stratification, improved model

Estimation results

	Integrated GREG	Alternative GREG
OCC_1	26,723	13,328
OCC_2	11,937	11,996
OCC_3	11.605	11,591
OCC_4	26.566	16,293
SELF	7,978	7,970
INC	120,915,970	91,355,871
UNEMP	7,181,217	7,085,061
PEN	39,970,062	39,048,784
PEK_HHG1	56,614,498	51,470,551
PEK_HHG2	99,704,938	76,115,807
PEK_HHG3	89,260,997	62,342,796
PEK_HHG4	85,215,552	56,442,895
PEK_HHG5	64,975,914	45,234,234
PEK_FST1	57,291,391	52,290,368
PEK_FST2	99,547,440	88,224,743
PEK_FST3	24,304,438	24,245,146
PEK_FST4	39,722,635	39,404,944

Table: MC standard errors

Table of contents

Conclusion

1) Yes, there is a price to pay for consistency in the integrated weighting approach due to unique weights:

- Higher variances of the auxiliaries and the regression coefficients.
- Higher deviation from sampling weights.

2) Yes, our alternative weighting approach ensures consistent estimates for the common variables without unique weights.

- The spread of the weights is comparable with the integrated weights, however the variation is significant smaller.
- More efficient estimation results.
- More flexible in model selection and independence of the household size.

Thank you for your attention!

This talk was developed within the project Research innovations for official and survey statistics (RIFOSS), funded by the German Statistical Office.

Literature

Bethlehem, J.G., Keller, W. (1987): Linear Weighting of Sample Survey data, Journal of official statistics, 3(2), 141-153.

European Commission (2013): Methodological guidelines and description of EU-SILC target variables, Eurostat, Directorate F: Social Statistics, Doc-SILC065 (2014 operation), zuletzt abgerufen am 28.04.2014.

Lemaître, G., Dufour, J. (1987): An integrated method for weighting persons and families, Survey Methodology, 13, 199-207.

Nieuwenbroek, N. (1993): An integrated method for weighting characteristics of persons and households using the linear regression estimator, Netherlands Central Bureau of Statistics.

Renssen, R.H., Nieuwenbroek, N.J. (1997): Aligning estimates for common variables in two or more sample surveys, Journal of the American Statistical Association, 92, 368-374.

Särndal, C. Swensson, B., Wretman, J. (1992): Model Assisted Survey Sampling, New York: Springer-Verlag.

Steel D.G., Clark, R.G. (2007): Person-level and household-level regression estimation in household surveys, Survey Methodology, 33(1), 51-60.

van den Brakel, J. (2013): Sampling and estimation techniques for household panels, Discussion Paper, 15, Statistics Netherlands.

Verma, V., Betti, G., Ghellini, G. (2006): Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC, Working Papers, 67, Dipartimento di Metodi Quantitativi, Università di Siena.