Consistent estimation at person-level and household-level

Anne Konrad, Jan Pablo Burgard, Ralf Münich

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

Motivation

Research question 1
  Simulation study I

Research question 2
  Simulation study II

Conclusion
Table of contents

Motivation

Research question 1
  Simulation study I

Research question 2
  Simulation study II

Conclusion
Usual person-level GREG estimator

The GREG estimator for totals is given by:

\[ \hat{T}_{Y, GREG} = \hat{T}_{Y, HT} + \hat{B}^T (T_x - \hat{T}_{x, HT}) \]  (1)

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

Notation:
\( Y \) : variable of interest (\( n \times 1 \))
\( X \) : auxiliary variables (\( n \times p \))
\( T_x \) : known totals of the auxiliaries (\( p \times 1 \))
\( \hat{T}_{x, HT} \) : estimated totals of the auxiliaries (\( p \times 1 \))
\( \Pi \) : diagonal matrix with inclusion probabilities \( \pi_i \) (\( n \times n \))
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:

\[
\hat{T}_{Y, \text{int}} = \hat{T}_{Y, HT} + \hat{B}_{\text{int}}^T (T_x - \hat{T}_{x, HT})
\] (2)

with \( \hat{B}_{\text{int}} = (D^T \Pi^{-1} D)^{-1} D^T \Pi^{-1} Y \) \((p \times 1)\) 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

Person–level GREG

Integrated GREG

\[ V(B_p) < V(B_{int}) \]
Distribution of weights

\[ \Rightarrow \text{Integrated weights have a significantly higher range!} \]
## Estimation results

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

**Table:** MC standard errors
Table of contents

Motivation

Research question 1
  Simulation study I

Research question 2
  Simulation study II

Conclusion
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 $\mathbf{C}$ ($n \times p$):

$$\hat{T}_{y, \text{Alternative}} = \hat{T}_{y, HT} + \hat{\mathbf{B}}_x^T (T_x - \hat{T}_{x, HT}) + \hat{\mathbf{B}}_c^T (\hat{T}_c - \hat{T}_{c, HT})$$
Distribution of the weights

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrative GREG</td>
<td>66.69</td>
<td>4.90</td>
<td>21.58</td>
<td>116.98</td>
<td>95.04</td>
</tr>
<tr>
<td>Alternative GREG*</td>
<td>66.69</td>
<td>3.29</td>
<td>-37.00</td>
<td>172.07</td>
<td>209.07</td>
</tr>
<tr>
<td>Alternative GREG**</td>
<td>66.69</td>
<td>3.28</td>
<td>20.83</td>
<td>114.24</td>
<td>93.41</td>
</tr>
</tbody>
</table>

* Improved model for common variables
** Stratification, improved model

Table: Summary Statistics (3,365,765 observations)
## Estimation results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Integrated GREG</th>
<th>Alternative GREG</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCC_1</td>
<td>26,723</td>
<td>13,328</td>
</tr>
<tr>
<td>OCC_2</td>
<td>11,937</td>
<td>11,996</td>
</tr>
<tr>
<td>OCC_3</td>
<td>11,605</td>
<td>11,591</td>
</tr>
<tr>
<td>OCC_4</td>
<td>26,566</td>
<td>16,293</td>
</tr>
<tr>
<td>SELF</td>
<td>7,978</td>
<td>7,970</td>
</tr>
<tr>
<td>INC</td>
<td>120,915,970</td>
<td>91,355,871</td>
</tr>
<tr>
<td>UNEMP</td>
<td>7,181,217</td>
<td>7,085,061</td>
</tr>
<tr>
<td>PEN</td>
<td>39,970,062</td>
<td>39,048,784</td>
</tr>
<tr>
<td>PEK_HHG1</td>
<td>56,614,498</td>
<td>51,470,551</td>
</tr>
<tr>
<td>PEK_HHG2</td>
<td>99,704,938</td>
<td>76,115,807</td>
</tr>
<tr>
<td>PEK_HHG3</td>
<td>89,260,997</td>
<td>62,342,796</td>
</tr>
<tr>
<td>PEK_HHG4</td>
<td>85,215,552</td>
<td>56,442,895</td>
</tr>
<tr>
<td>PEK_HHG5</td>
<td>64,975,914</td>
<td>45,234,234</td>
</tr>
<tr>
<td>PEK_FST1</td>
<td>57,291,391</td>
<td>52,290,368</td>
</tr>
<tr>
<td>PEK_FST2</td>
<td>99,547,440</td>
<td>88,224,743</td>
</tr>
<tr>
<td>PEK_FST3</td>
<td>24,304,438</td>
<td>24,245,146</td>
</tr>
<tr>
<td>PEK_FST4</td>
<td>39,722,635</td>
<td>39,404,944</td>
</tr>
</tbody>
</table>

**Table:** MC standard errors
Table of contents

Motivation

Research question 1
Simulation study I

Research question 2
Simulation study II

Conclusion
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 significantly 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.


