The institutionalized population in survey research

Assessing the peril of coverage bias with cross-national survey data

Jan-Lucas Schanze & Stefan Zins
ESRA conference, Lisbon (PT), 21st of July 2017

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 654221.
Share of population aged 85 years and over living in an institutional household

(% at NUTS 2)

- < 2
- 2 - < 10
- 10 - < 15
- 15 - < 20
- ≥ 20
- Data not available

Source: Eurostat 2015. „People in the EU“
Research question

If social surveys infer their results from a sample of private households to the general population, ...

... to what extend to they suffer from coverage error due to the size and distinctiveness of the institutionalized population?
Total Survey Error Framework

Source: Groves et al. 2004
Research findings: Significant differences

• Health-related variables
  • State of health, dementia, self-rated health

• Coping with daily living & housekeeping
  • Functional and/or cognitive impairments

• Demographic variables
  • Gender, age, ethnicity

• Social ties
  • Networks, marital status, parenthood, availability of informal caregiver

• Socio economic variables
  • Income, economic activity, housing, education

• Contextual variables
  • Social policies, demography of countries/regions
Hypotheses

Assuming bias in case of noncoverage or undercoverage of the institutionalized population

• H1: Increasing the coverage rate of institutionalized respondents will increase the aggregate mean of the number of limitations in activities of daily living (ADL)

• H2: Increasing the coverage rate of institutionalized respondents will decrease the aggregate mean of the self-perceived health perceptions
Data – Compiling our statistical population

• Five waves of the Survey of Health, Ageing and Retirement in Europe (SHARE)
  • Cross-national panel survey in Europe, beginning in 2004
  • The target population comprises the population aged 50 years or older living in private households and institutions for the elderly

• Pooling the data: Extracting the most recent observation for each SHARE respondent
  • Cross-sectional dataset with 91,671 cases without missing values in 15 countries
    • Among these: 2.0% (N = 1,840) identified as institutionalized residents
## Our statistical population

<table>
<thead>
<tr>
<th>(%)</th>
<th>Private hh.</th>
<th>Institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>45.2</td>
<td>35.5</td>
</tr>
<tr>
<td>Female</td>
<td>54.8</td>
<td>64.5</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 to 59</td>
<td>26.2</td>
<td>8.6</td>
</tr>
<tr>
<td>60 to 69</td>
<td>34.4</td>
<td>14.7</td>
</tr>
<tr>
<td>70 to 79</td>
<td>25.3</td>
<td>23.7</td>
</tr>
<tr>
<td>80 to 89</td>
<td>12.6</td>
<td>37.0</td>
</tr>
<tr>
<td>Older than 90</td>
<td>1.6</td>
<td>16.1</td>
</tr>
<tr>
<td><strong>Marital status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>71.3</td>
<td>35.6</td>
</tr>
<tr>
<td>Never married</td>
<td>5.6</td>
<td>11.2</td>
</tr>
<tr>
<td>Divorced</td>
<td>8.7</td>
<td>10.7</td>
</tr>
<tr>
<td>Widowed</td>
<td>14.4</td>
<td>42.5</td>
</tr>
<tr>
<td><strong>Number of observations</strong></td>
<td>89,831 (98%)</td>
<td>1,840 (2%)</td>
</tr>
</tbody>
</table>
Our statistical population (cont.)

<table>
<thead>
<tr>
<th>(%)</th>
<th>Private hh.</th>
<th>Institutions</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Limitations with ADL</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No limitations</td>
<td>87.2</td>
<td>51.1</td>
<td>86.5</td>
</tr>
<tr>
<td>1-3 limitations</td>
<td>10.3</td>
<td>27.5</td>
<td>10.7</td>
</tr>
<tr>
<td>&gt;3 limitations</td>
<td>2.5</td>
<td>21.4</td>
<td>2.9</td>
</tr>
<tr>
<td><strong>Self-perceived health</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>10.1</td>
<td>19.6</td>
<td>10.3</td>
</tr>
<tr>
<td>Fair</td>
<td>27.6</td>
<td>38.3</td>
<td>27.9</td>
</tr>
<tr>
<td>Good</td>
<td>36.2</td>
<td>30.2</td>
<td>36.1</td>
</tr>
<tr>
<td>Very good</td>
<td>18.5</td>
<td>7.9</td>
<td>18.2</td>
</tr>
<tr>
<td>Excellent</td>
<td>7.6</td>
<td>4.1</td>
<td>7.6</td>
</tr>
<tr>
<td><strong>Number of observations</strong></td>
<td>89,831</td>
<td>1,840</td>
<td>96,671</td>
</tr>
</tbody>
</table>
Monte Carlo Simulation

- 5,000 times simple random sampling with replacement of 3,000 observations with 5 different coverage rates

\[
\frac{N_i}{N} (Y_p - Y_i)
\]

- \( N \) = Total number of residents
- \( N_i \) = Number of institutionalized residents
- \( Y_p \) = Mean of residents living in private households
- \( Y_i \) = Mean of institutionalized residents

<table>
<thead>
<tr>
<th>Degree of inclusion of ( N_i ) in target population</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected elements from ( N_i ) in samples of 3,000 obs</td>
<td>0</td>
<td>15</td>
<td>30</td>
<td>45</td>
<td>60</td>
</tr>
</tbody>
</table>
Weighting the samples

- Applying different weighting schemes
  1. Unweighted sample
  2. Traditional survey weights
     - Age, gender, country
  3. Multivariate control weights
     - Age, gender, country
     - Marital status, partner in household, origin, number of children, household size
     - Country-specific income percentiles
     - Limited in health, long-term illness
Mean of limitations in ADL

Noncoverage (N_i = 0)
1. Unweighted
2. Survey weights
3. Multivariate control weights

1. 37.7% of all samples
2. 24.5% of all samples
3. 20.0% of all samples

N = 1000 samples of 3000 obs. each
Mean of ADL: Biased samples

- Noncoverage (0%)
- Undercoverage (25%)
- Undercoverage (50%)
- Undercoverage (75%)
- Equal coverage (100%)

Unweighted, Survey weights, Multivariate weights
Deviation of means in ADL

1. Unweighted

3. Multivariate weights

Noncoverage (0%)
Undercoverage (25%)
Undercoverage (50%)
Undercoverage (75%)
Equal coverage (100%)

excludes outside values
True value in statistical population = 100%
N of each category = 1000 samples of 3000 obs. each
Mean of self-perceived health

Noncoverage: 0% of institutionalized units in the target population

8.4% of all samples

Undercoverage: 25% of institutionalized units in the target population

5.7% of all samples

Undercoverage: 75% of institutionalized units in the target population

5.6% of all samples

Equal coverage: 100% of institutionalized units in the target population

4.9% of all samples

N = 1000 samples of 3000 obs. each; survey weights (gender, age)
Mean of health perception: Biased samples

Unweighted
Survey weights
Multivariate weights

- Noncoverage (0%)
- Undercoverage (25%)
- Undercoverage (50%)
- Undercoverage (75%)
- Equal coverage (100%)
Limitations

• Generalization of our simulation-based results for the universe of social surveys in Europe

• Survey data used for compiling our statistical population
  • Issue of undercoverage of the institutionalized population in some countries in the baseline wave
  • Institutionalization as a shock likely to increase noncontact and refusal of this group
  • Assumption of stability of variables over time (2004 to 2015) could be proved wrong in reality

• Analysis only comprises a small subset of variables and should be extended to further variables
Conclusion

• Noncoverage and undercoverage of the small group of institutionalized residents (2.0%) causes bias in some variables due to their statistical distinctiveness
  • Significant bias in health-related variables (ADL and self-perceived health)

• Weighting the samples for age and gender reduces the bias, but cannot eliminate it completely
  • In the ADL variable the bias even remains under control of multiple variables in case of noncoverage and undercoverage

• Undercoverage already improves the estimates compared to noncoverage, especially if weights are used!
Contact

Jan-Lucas Schanze

GESIS – Leibniz Institute for the Social Sciences
Survey Design and Methodology (SDM)

P.O. Box 12 21 55
68072 Mannheim (GER)

phone: 0049 (0)621 1246-405
mail: jan-lucas.schanze@gesis.org
References


De Luca, G. et al. 2015. “Sample design and weighting strategies in SHARE Wave 5”. In: SHARE Wave 5: Innovations & Methodology MEA, ed. by Börsch-Supan and Malter, Munich: pp.75–84.


Eurostat. 2015. “People in the EU - Who are we and how do we live” European Union: Luxembourg.


References (cont.)


Institutionalized population in EU-26*

Source: Eurostat 2016, CensusHub
* Lithuania and Bulgaria missing
Variables

- Institutionalization of respondents
  - Interviewer observations (Living in a nursing home)
  - Self-reports of respondents
    - Living permanently in a nursing home
    - Information about the housing type
- Additive index of limitations with activities in daily living (ADL)
  - Count variable of 9 dichotomous variables (scale 0 to 9)
- Self-perceived health (U.S. version)
  - Ordered 5-level scale
Deviation of means in self-perceived health

1. Unweighted

Equal coverage (100%)
Undercoverage (75%)
Undercoverage (50%)
Undercoverage (25%)
Noncoverage (0%)

excludes outside values
True value in statistical population = 100%
N of each category = 1000 samples of 3000 obs. each

3. Multivariate weights

Equal coverage (100%)
Undercoverage (75%)
Undercoverage (50%)
Undercoverage (25%)
Noncoverage (0%)

excludes outside values
True value in statistical population = 100%
N of each category = 1000 samples of 3000 obs. each
## Mean square error within categories (1)

<table>
<thead>
<tr>
<th>Limitations in ADL (%)</th>
<th>Inclusion of institutionalized units</th>
<th>True value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
<td>25%</td>
</tr>
<tr>
<td>No limitation</td>
<td>86.94</td>
<td>86.81</td>
</tr>
<tr>
<td>1-3 limitations</td>
<td>10.47</td>
<td>10.54</td>
</tr>
<tr>
<td>&gt;3 limitations</td>
<td>2.59</td>
<td>2.66</td>
</tr>
</tbody>
</table>

### Diagram

MSE (No limitation)  | MSE (1-3 limitations)  | MSE (>3 limitations)
Mean square error within categories (2)

<table>
<thead>
<tr>
<th>Self-perceived health (%, weighted with survey weights)</th>
<th>Inclusion of institutionalized units</th>
<th>True value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
<td>25%</td>
</tr>
<tr>
<td>Poor</td>
<td>10.16</td>
<td>10.21</td>
</tr>
<tr>
<td>Fair</td>
<td>27.71</td>
<td>27.76</td>
</tr>
<tr>
<td>Good</td>
<td>36.14</td>
<td>36.11</td>
</tr>
<tr>
<td>Very good</td>
<td>18.37</td>
<td>18.32</td>
</tr>
<tr>
<td>Excellent</td>
<td>7.62</td>
<td>7.60</td>
</tr>
</tbody>
</table>

The graph shows the mean square error (MSE) for different levels of self-perceived health across various inclusion percentages of institutionalized units.