Towards a Cautious Modelling of Missing Data in Small Area Estimation

ISIPTA '17, Lugano

Julia Plass[1], Aziz Omar[1,2], Thomas Augustin[1]

[1] Department of Statistics, Ludwig-Maximilians University and
[2] Department of Mathematics, Insurance and Appl. Statistics, Helwan University

11th of July 2017

Our team and aim

Aziz Omar

"Towards a Cautious Modelling of Missing Data in Small Area Estimation"

Thomas Augustin
Julia Plass

Our team and aim

Aziz Omar

"Towards a Cautious Modelling of Missing Data in Small Area Estimation"

Thomas Augustin
Julia Plass

Our team and aim

Aziz Omar

"Towards a Cautious Modelling of Missing Data in Small Area Estimation"

Thomas Augustin
Julia Plass

Our team and aim

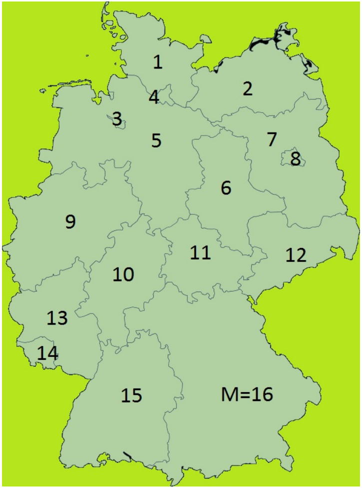
Aziz Omar

"Towards a Cautious Modelling of Missing Data in Small Area Estimation"

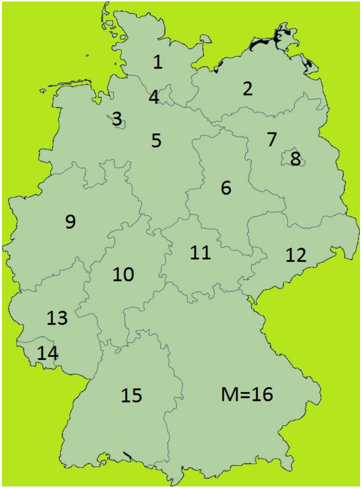
Thomas Augustin
Julia Plass

- Existing approaches for dealing with nonresponse in SAE are based on strong assumptions on the missingness process
- Such assumptions are usually **not testable**, and wrongly imposing them may lead to **biased** results.
  (Manski, 2003, Partial Identification of Probability Distributions, Jaeger, 2006, ECML,...)

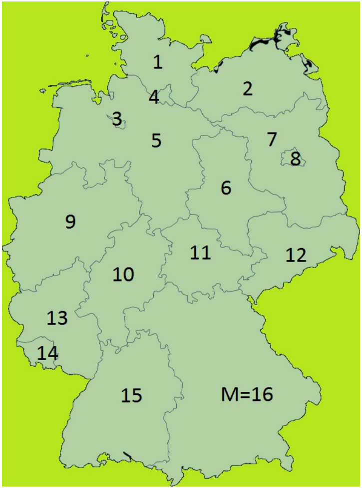
- Population with N individuals



- Population with N individuals
- M areas, each contains N_i individuals, $i = 1, \ldots, M$



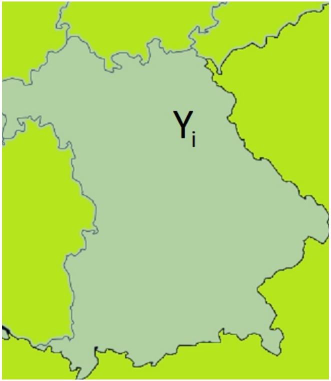
- Population with N individuals
- M areas, each contains N_i individuals, $i = 1, \ldots, M$
- **Of interest:**
  Area-specific mean $\bar{Y}_i$



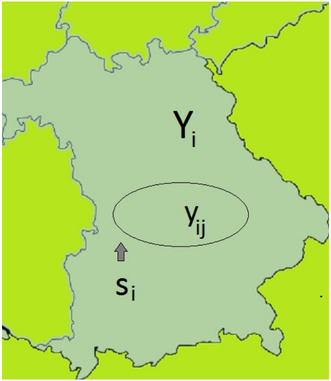
- Population with N individuals
- M areas, each contains N_i individuals, $i = 1, \ldots, M$
- **Of interest:** Area-specific mean $\bar{Y}_i$
- **Problem:** For each area, only sample s_i with small sample size n_i available

⇒ Using auxiliary variables (covariates) $X_1, \ldots, X_k$

⇒ "borrowing strength"

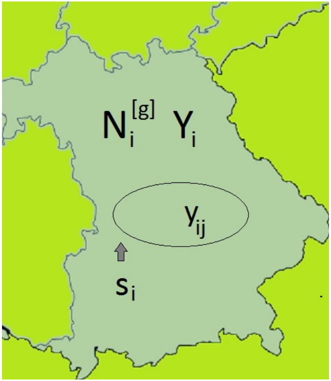


- Binary variable of interest
  ⇒ probability that Y_i is equal to 1
  $:= \pi_i$ (poverty rate)



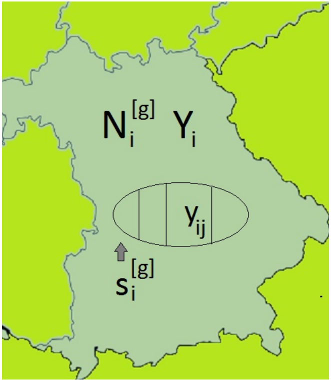
- Binary variable of interest
  ⇒ probability that Y_i is equal to 1
  $:= \pi_i$ (poverty rate)
- $1/w_{ij}$ is the probability that individual j in area i is selected in s_i
- Sample values y_{ij} known for $j \in s_i$
- Sample data from German General Social Survey (GESIS Leibniz Institute for the Social Sciences, 2016), $y_{ij} = 1$: 'poor', $y_{ij} = 0$: 'rich'

What's the problem? ⇒ 1. Small Area Estimation (SAE)



- Binary covariates (Abitur, sex)
- Cross classifications of the covariates
  ⇒ subgroup g, $g = 1, \ldots, v$
- Known absolute frequencies $N_i^{[g]}$
  Federal Statistical Office's data report:

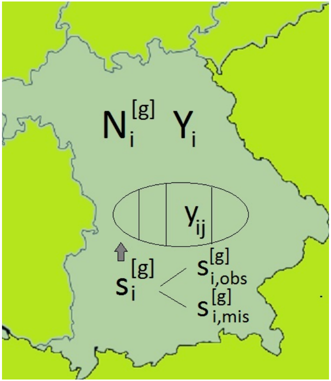
		Abitur	
		no	yes
sex	male	$N_i^{[1]}$	$N_i^{[2]}$
	female	$N_i^{[3]}$	$N_i^{[4]}$



- Binary covariates (Abitur, sex)
- Cross classifications of the covariates
  ⇒ subgroup g, $g = 1, \ldots, v$
- Known absolute frequencies $N_i^{[g]}$
  Federal Statistical Office's data report:

		Abitur	
		no	yes
sex	male	$N_i^{[1]}$	$N_i^{[2]}$
	female	$N_i^{[3]}$	$N_i^{[4]}$

- Joint information about x_{ij} and y_{ij}
  ⇒ We know y_{ij} for $j \in s_i^{[g]}$



- some sample values y_{ij} are missing
- $s_i^{[g]}$ is partitioned into $s_{i,obs}^{[g]}$ and $s_{i,mis}^{[g]}$

Cautious Approach for Dealing with Nonresponse

(ISIPTA '15, Plass, Augustin, Cattaneo, Schollmeyer)

- An observation model is determined by the missingness parameters $q^{[g]}_{na|y}$ (:= probability to refuse the answer ("na"), given subgroup g and the true value y)

- Maximizing the log-likelihood

$$\ell(\pi^{[g]},\ q^{[g]}_{na|0},\ q^{[g]}_{na|1}) = n^{[g]}_1\Big(\ln(\pi^{[g]}) + \ln(1 - q^{[g]}_{na|1})\Big)$$
$$+\, n^{[g]}_0\Big(\ln(1 - \pi^{[g]}) + \ln(1 - q^{[g]}_{na|0})\Big) + n^{[g]}_{na}\Big(\ln(\pi^{[g]} q^{[g]}_{na|1} + (1 - \pi^{[g]}) q^{[g]}_{na|0})\Big)$$

gives set-valued estimator.

- Resulting bounds of $\hat{\pi}^{[g]}$ under **no assumptions about $q^{[g]}_{na|y}$:**

$$\underline{\hat{\pi}}^{[g]} = \frac{n^{[g]}_1}{n^{[g]}_{na} + n^{[g]}_1 + n^{[g]}_0} \quad \text{and} \quad \overline{\hat{\pi}}^{[g]} = \frac{n^{[g]}_1 + n^{[g]}_{na}}{n^{[g]}_{na} + n^{[g]}_1 + n^{[g]}_0}\ .$$

Cautious Approach for Dealing with Nonresponse

(ISIPTA '15, Plass, Augustin, Cattaneo, Schollmeyer)

- **Incorporate assumptions** by missingness ratio (Nordheim, 1984)

$$R = q^{[g]}_{na|1} / q^{[g]}_{na|0} \ , \quad \text{with } R \in \mathcal{R} \subseteq \mathbb{R}^+_0$$

- Specific values of R point-identify $\pi^{[g]}$

- Partial assumptions, expressed by $\mathcal{R} = [\underline{R}, \overline{R}]$, refine the result without any missingness assumptions $(R \in [0,\ 1])$

  $\Rightarrow$ Bounds for $\hat{\pi}^{[g],\mathcal{R}}$, $\hat{q}^{[g],\mathcal{R}}_{na|0}$ and $\hat{q}^{[g],\mathcal{R}}_{na|1}$ obtained under $\underline{R}$ and $\overline{R}$

The synthetic estimator (without nonresponse)

- Horvitz-Thompson (HT) estimator
  (Horvitz and Thompson, 1952, JASA)

$$\hat{\pi}_{HT,i} = \frac{1}{N_i} \sum_{j \in s_i} w_{ij} y_{ij}$$

- The synthetic estimator (González, 1973, JASA)

$$\hat{\pi}_{SYN} \equiv \hat{\pi}_{SYN,i} = \frac{1}{N} \sum_{i=1}^{M} \sum_{j \in s_i} w_{ij} y_{ij} = \frac{1}{N} \sum_{i=1}^{M} N_i \cdot \hat{\pi}_{HT,i}$$

Cautious synthetic estimator

- **No assumptions:**

$$\hat{\pi}_{SYN} = \frac{1}{N}\sum_{i=1}^{M}\Big(\sum_{j\in s_{i,obs}} w_{ij}y_{ij} + \sum_{j\in s_{i,mis}} w_{ij}\cdot y_{ij}\Big)$$

$$\underline{\hat{\pi}}_{SYN} = \ldots\Big(\ldots + \sum_{j\in s_{i,mis}} w_{ij}\cdot 0\Big)\ ,\ \overline{\hat{\pi}}_{SYN} = \ldots\Big(\ldots + \sum_{j\in s_{i,mis}} w_{ij}\cdot 1\Big)$$

Cautious synthetic estimator

- **No assumptions:**

$$\hat{\pi}_{SYN} = \frac{1}{N} \sum_{i=1}^{M} \Big(\sum_{j \in s_{i,obs}} w_{ij} y_{ij} + \sum_{j \in s_{i,mis}} w_{ij} \cdot y_{ij} \Big)$$

$$\underline{\hat{\pi}}_{SYN} = \dots \Big(\dots + \sum_{j \in s_{i,mis}} w_{ij} \cdot 0 \Big) \;, \; \overline{\hat{\pi}}_{SYN} = \dots \Big(\dots + \sum_{j \in s_{i,mis}} w_{ij} \cdot 1 \Big)$$

- **Partial assumptions:**

$$\underline{\hat{\pi}}_{SYN}^{\mathcal{R}} = \frac{1}{N} \sum_{i=1}^{M} \Big(\sum_{j \in s_{i,obs}} w_{ij} y_{ij} + \underbrace{\underline{\hat{q}}_{na|1i}^{\mathcal{R}} \cdot \underline{\hat{\pi}}_{i}^{\mathcal{R}} \cdot \sum_{j \in s_i} w_{ij}} \Big)$$

smallest est. weighted number of nonrespondents with $y_{ij} = 1$, under the assumption in focus.

Analogously, $\overline{\hat{\pi}}_{SYN}^{\mathcal{R}}$ is achieved by using $\overline{\hat{q}}_{na|1i}^{\mathcal{R}}$ and $\overline{\hat{\pi}}_{i}^{\mathcal{R}}$.

The LGREG estimator (without nonresponse)...

(Lehtonen and Veijanen, 1998, Surv. Methodol.)

- ... in its representation how we need it:

$$\hat{\pi}_{LGREG,i} = \sum_{g=1}^{v} \Bigg(\overbrace{\sum_{j \in s_i^{[g]}} w_{ij} y_{ij}}^{\text{HT-part}} + \overbrace{\hat{\pi}^{[g]} \Big(N_i^{[g]} - \sum_{j \in s_i^{[g]}} w_{ij} \Big)}^{\text{correction term}} \Bigg) \Big/ N_i$$

$$\text{with} \quad \hat{\pi}^{[g]} = \sum_{i=1}^{M} \sum_{j \in s_i^{[g]}} \frac{y_{ij}}{n^{[g]}}$$

- The correction term accounts for under/overrepresentation of certain constellations of covariates in the sample
- In most cases: $w_{ij} = w_i, \forall j = 1, \ldots, n_i, i = 1, \ldots, M$.

No assumptions: Cautious LGREG estimator

Breaking the summation over all areas into a term for area i^* of interest and areas $i \neq i^*$ leads to

$$\sum_{g=1}^{v} \Bigg(\Big(\frac{1}{n^{[g]}} \sum_{\substack{i=1 \\ i \neq i^*}}^{M} \Big(\sum_{j \in s_{i,obs}^{[g]}} y_{ij} + \sum_{j \in s_{i,mis}^{[g]}} y_{ij} \Big) \Big) \Big(N_{i^*}^{[g]} - n_{i^*}^{[g]} w_{i^*} \Big)$$

$$+ \frac{1}{n^{[g]}} \Big(\sum_{j \in s_{i^*,obs}^{[g]}} y_{i^* j} + \sum_{j \in s_{i^*,mis}^{[g]}} y_{i^* j} \Big) \Big(N_{i^*}^{[g]} - w_{i^*} (n_{i^*}^{[g]} + n^{[g]}) \Big) \Bigg) \Big/ N_{i^*}$$

No assumptions: Cautious LGREG estimator

Breaking the summation over all areas into a term for area i^* of interest and areas $i \neq i^*$ leads to

$$\sum_{g=1}^{v} \Bigg(\Big(\frac{1}{n^{[g]}} \sum_{\substack{i=1 \\ i \neq i^*}}^{M} \Big(\sum_{j \in s_{i,obs}^{[g]}} y_{ij} + \sum_{j \in s_{i,mis}^{[g]}} y_{ij} \Big) \Big) \Big(N_{i^*}^{[g]} - n_{i^*}^{[g]} w_{i^*} \Big)$$

$$+ \frac{1}{n^{[g]}} \Big(\sum_{j \in s_{i^*,obs}^{[g]}} y_{i^*j} + \sum_{j \in s_{i^*,mis}^{[g]}} y_{i^*j} \Big) \Big(N_{i^*}^{[g]} - w_{i^*} (n_{i^*}^{[g]} + n^{[g]}) \Big) \Bigg) / N_{i^*}$$

To determine $\hat{\underline{\pi}}_{LGREG,i^*}$:

	$N_{i^*}^{[g]} \geq w_{i^*}(n_{i^*}^{[g]} + n^{[g]})$	$N_{i^*}^{[g]} < w_{i^*}(n_{i^*}^{[g]} + n^{[g]})$
$N_{i^*}^{[g]} \geq n_{i^*}^{[g]} w_{i^*}$	$y_{ij} = 0,\ \forall j \in s_{i,mis}$	$y_{ij} = \begin{cases} 0 & \forall j \in s_{i,mis},\, i \neq i^* \\ 1 & \forall j \in s_{i,mis},\, i = i^* \end{cases}$
$N_{i^*}^{[g]} < n_{i^*}^{[g]} w_{i^*}$	$y_{ij} = \begin{cases} 1 & \forall j \in s_{i,mis},\, i \neq i^* \\ 0 & \forall j \in s_{i,mis},\, i = i^* \end{cases}$	$y_{ij} = 1,\ \forall j \in s_{i,mis}$

Partial assumptions: Cautious LGREG estimator

1.) Regard $\hat{\pi}_{LGREG,i^*}$ as a combination of two estimators:
$\Rightarrow$ a global one that borrows strength and
$\Rightarrow$ a specific one associated to area i^*.

2.) Maximize the two log-likelihoods under $\underline{R}$ and $\overline{R}$:

- $\ell(\pi^{[g],\mathcal{R}},\ q^{[g],\mathcal{R}}_{na|0},\ q^{[g],\mathcal{R}}_{na|1})$ and
- $\ell(\pi^{[g],\mathcal{R}}_{i^*},\ q^{[g],\mathcal{R}}_{na|0i^*},\ q^{[g],\mathcal{R}}_{na|1i^*})$

3.) Include the estimators that minimize

$$\sum_{g=1}^{v}\Bigg(\overbrace{\sum_{j\in s^{[g]}_{i^*,obs}} w_{i^*}y_{i^*j} + \hat{q}^{[g],\mathcal{R}}_{na|1i^*}\hat{\pi}^{[g],\mathcal{R}}_{i^*}\sum_{j\in s^{[g]}_{i^*}} w_{i^*j}}^{\text{HT-part}} + \overbrace{\hat{\pi}^{[g],\mathcal{R}}(N^{[g]}_{i^*} - n^{[g]}_{i^*}w_{i^*})}^{\text{correction term}}\Bigg)\Big/ N_{i^*}$$

$\Rightarrow$ Since $\pi^{[g]}$ and $\pi^{[g]}_{i^*}$ are estimated distinctively, interrelation between them should be considered.

Some results (example)

- Intervals for the synthetic estimator

no assumption	$\mathcal{R} = [0, 1]$
$[0.167,\ 0.300]$	$[0.167,\ 0.193]$

- Intervals for the LGREG estimator

Federal state	no assumption	$\mathcal{R} = [0, 1]$
BW	$[0.129,\ 0.366]$	$[0.129,\ 0.210]$
BY	$[0.088,\ 0.233]$	$[0.088,\ 0.133]$
HB	$[0.077,\ 0.405]$	$[0.115,\ 0.193]$
...	...	...

Further work

- Optimization of one overall likelihood, instead of two, to obtain the cautious LGREG-estimator
- Comparison of the magnitude of both principally differing kinds of uncertainty induced by the two problems in focus