Towards a Cautious Modelling of Missing Data in Small Area Estimation

Statistische Woche 2017, Rostock

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

21st of September 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

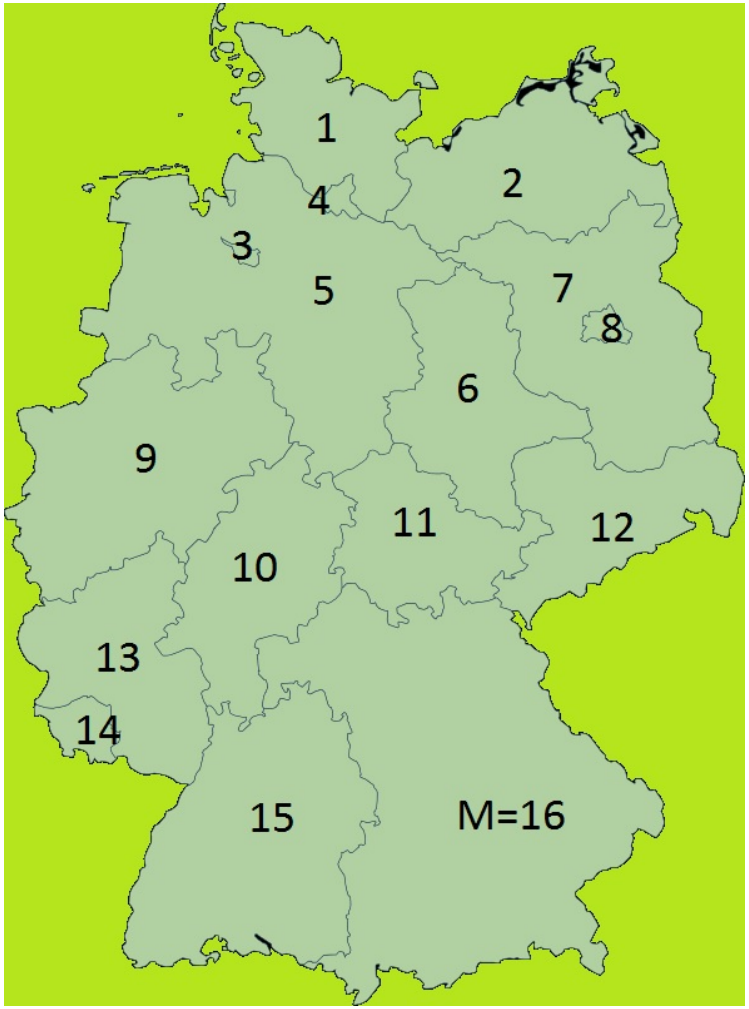
- Existing approaches for dealing with nonresponse in SAE are based on strong assumptions on the missingness process
- Such assumptions are usually **not testable**, and wrongly imposing them may lead to **biased** results.

  (Manski, 2003, Partial Identification of Probability Distributions, Jaeger, 2006, ECML,...)

What's the problem? $\Rightarrow$ 1. Small Area Estimation (SAE)

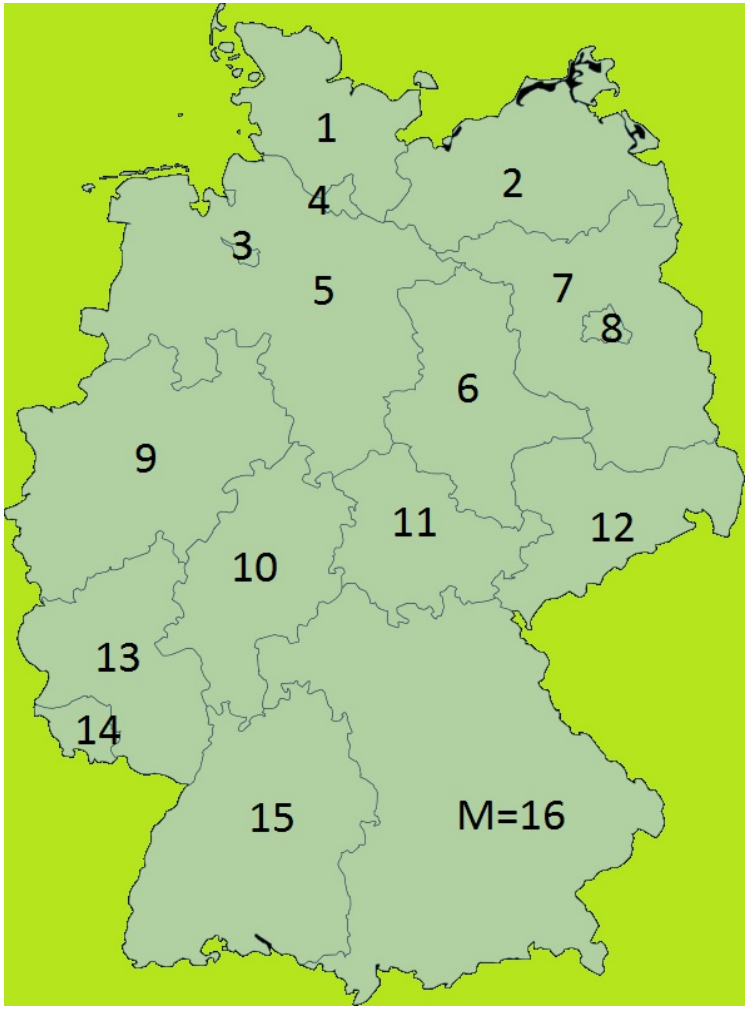
- Population with N individuals

What's the problem? $\Rightarrow$ 1. Small Area Estimation (SAE)



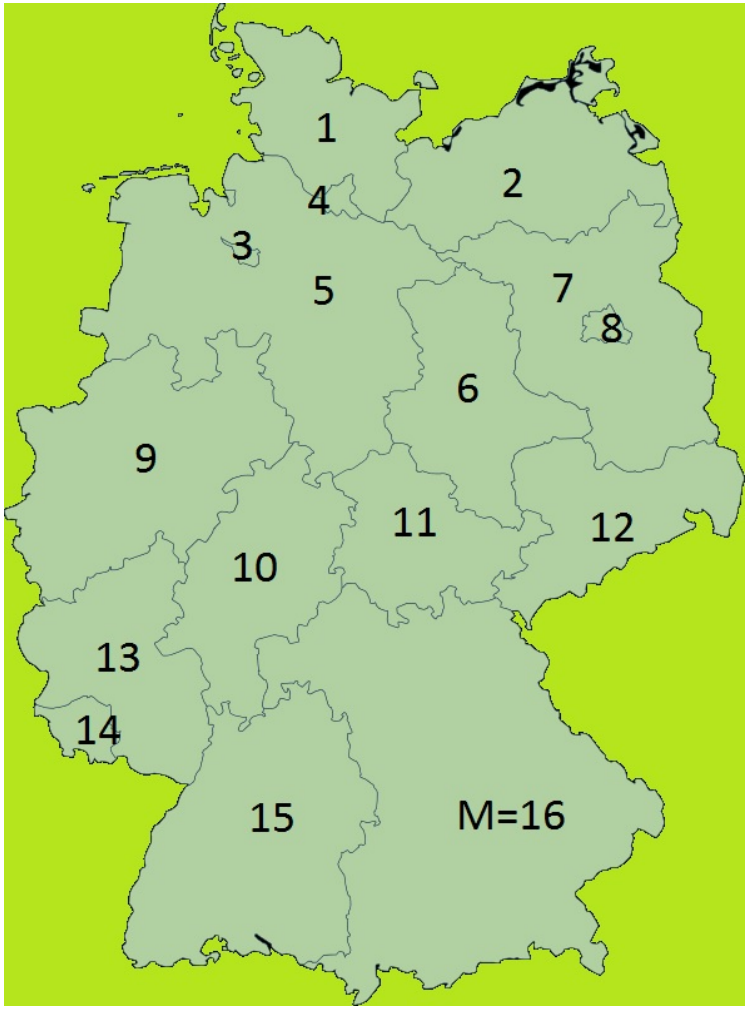
- Population with N individuals
- M areas, each contains N_i individuals, $i = 1, \ldots, M$

What's the problem? $\Rightarrow$ 1. Small Area Estimation (SAE)



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

What's the problem? $\Rightarrow$ 1. Small Area Estimation (SAE)

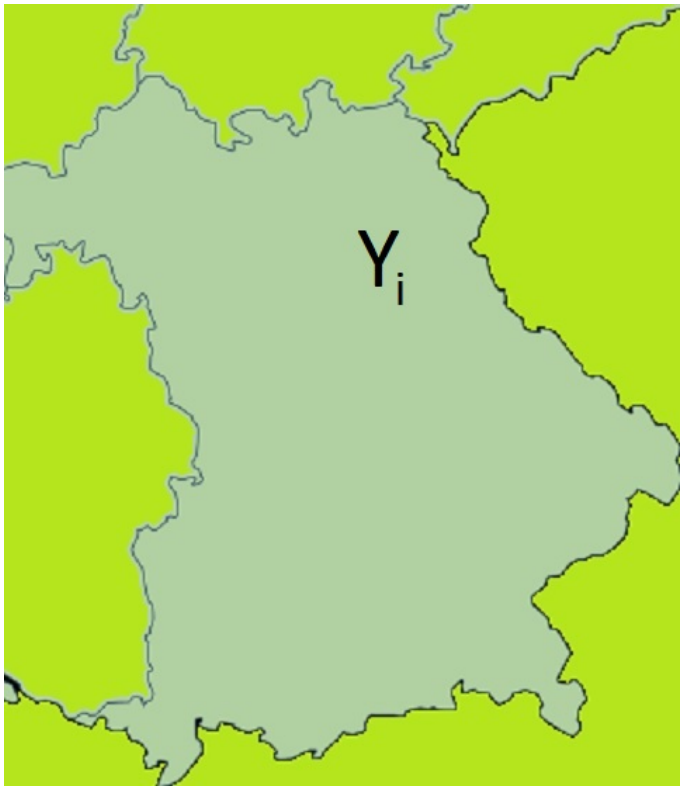


- Population with N individuals
- M areas, each contains N_i individuals, $i = 1, \ldots, M$
- **Of interest:**
  Area-specific mean $\bar{Y}_i$
- **Problem:**
  For each area, only sample s_i with small sample size n_i available

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

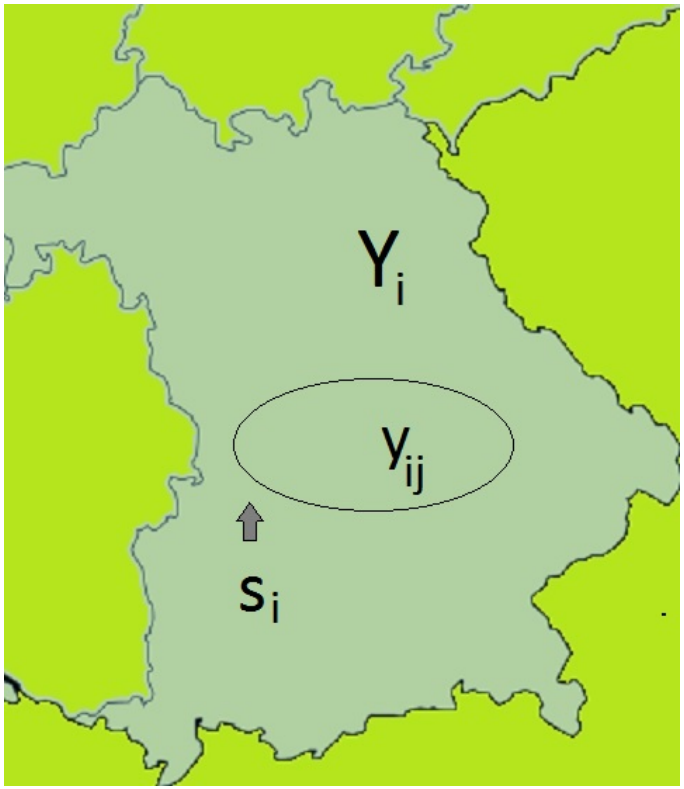
$\Rightarrow$ "borrowing strength"

What's the problem? $\Rightarrow$ 1. Small Area Estimation (SAE)



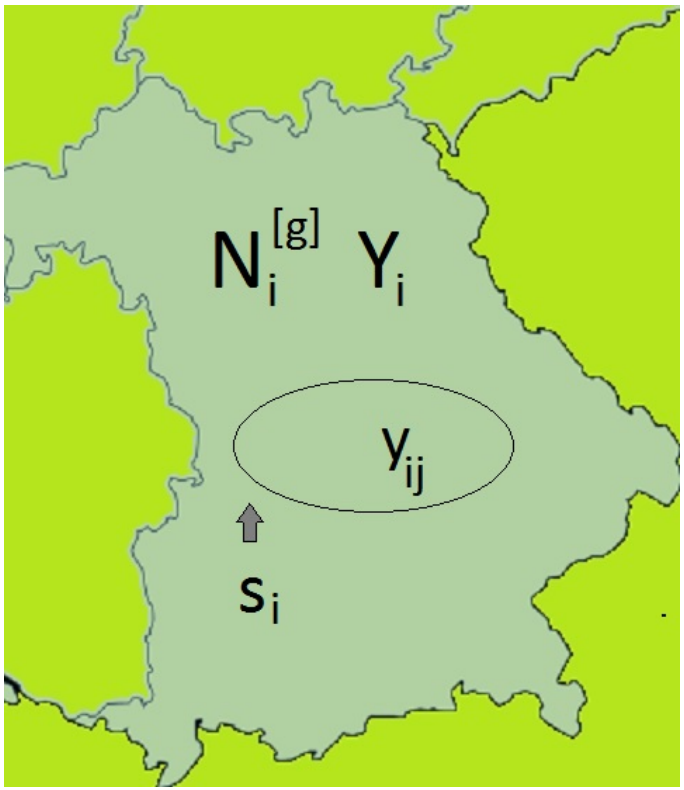
- Binary variable of interest
  $\Rightarrow$ probability that Y_i is equal to 1
  $:= \pi_i$ (poverty rate)

What's the problem? $\Rightarrow$ 1. Small Area Estimation (SAE)



- Binary variable of interest
  $\Rightarrow$ 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'

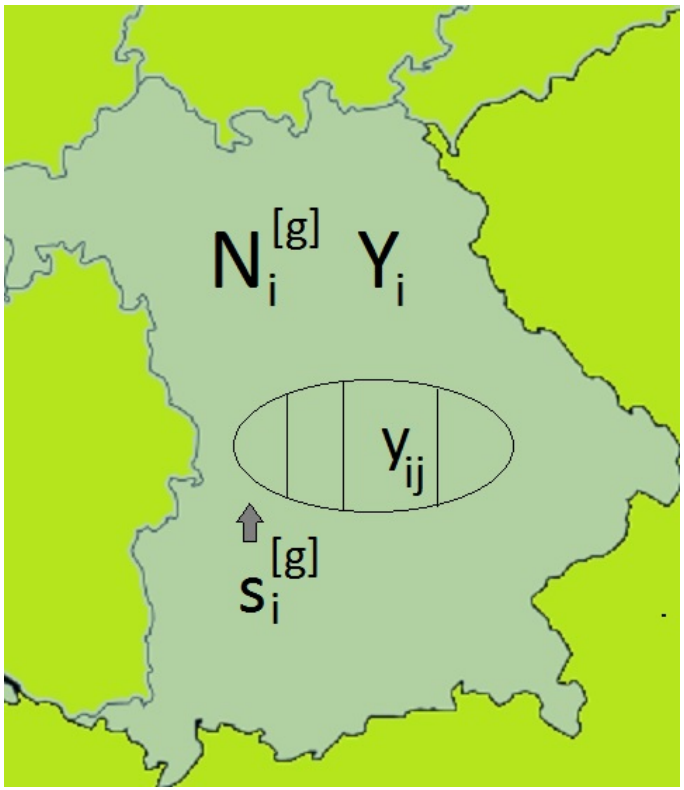
What's the problem? $\Rightarrow$ 1. Small Area Estimation (SAE)



- Binary covariates (Abitur, sex)
- Cross classifications of the covariates $\Rightarrow$ 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	$N_i^{[3]}$	$N_i^{[4]}$

What's the problem? $\Rightarrow$ 1. Small Area Estimation (SAE)

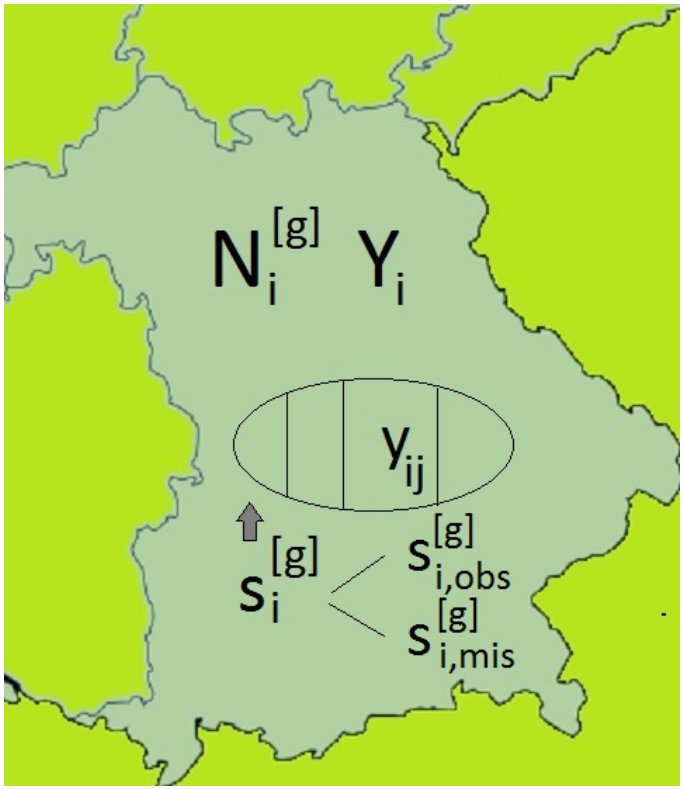


- Binary covariates (Abitur, sex)
- Cross classifications of the covariates
  $\Rightarrow$ subgroup g, $g = 1, \ldots, v$
- Known absolute frequencies $N_i^{[g]}$
  Federal Statistical Office's data report:

		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}
  $\Rightarrow$ We know y_{ij} for $j \in s_i^{[g]}$

What's the problem? $\Rightarrow$ 2. Missing data



- 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 $q_{na|y}^{[g]}$ (:= probability to refuse the answer ("na"), given subgroup g and the true value y)
- Maximizing the log-likelihood

$$\ell(\pi^{[g]},\ q_{na|0}^{[g]},\ q_{na|1}^{[g]}) = n_1^{[g]}\Big(\ln(\pi^{[g]}) + \ln(1-q_{na|1}^{[g]})\Big) + n_0^{[g]}\Big(\ln(1-\pi^{[g]}) + \ln(1-q_{na|0}^{[g]})\Big) + n_{na}^{[g]}\Big(\ln(\pi^{[g]}q_{na|1}^{[g]} + (1-\pi^{[g]})q_{na|0}^{[g]})\Big)$$

  gives set-valued estimator.

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

$$\underline{\hat{\pi}}^{[g]} = \frac{n_1^{[g]}}{n_{na}^{[g]} + n_1^{[g]} + n_0^{[g]}} \quad \text{and} \quad \overline{\hat{\pi}}^{[g]} = \frac{n_1^{[g]} + n_{na}^{[g]}}{n_{na}^{[g]} + n_1^{[g]} + n_0^{[g]}}\ .$$

Cautious Approach for Dealing with Nonresponse

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

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

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

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

Illustration of the cautious approach (survey data only)

observation

		poverty class		
sex	**Abitur**	rich (0)	poor (1)	na
male (0)	no (0)	777	217	131
	yes (1)	492	61	81
female (1)	no (0)	668	284	144
	yes (1)	441	72	98

Illustration of the cautious approach (survey data only)

observation

		poverty class		
sex	**Abitur**	rich (0)	poor (1)	na
male (0)	no (0)	777	217	131
	yes (1)	492	61	81
female (1)	no (0)	668	284	144
	yes (1)	441	72	98

underlying truth

		poverty class	
sex	**Abitur**	0	1
0	0	777+□	217+□
	1	492+□	61+□
1	0	668+□	284+□
	1	441+□	72+□

no asmpt. about $q^{[g]}_{na|y}$

Illustration of the cautious approach (survey data only)

observation

		poverty class		
sex	**Abitur**	rich (0)	poor (1)	na
male (0)	no (0)	777	217	131
	yes (1)	492	61	81
female (1)	no (0)	668	284	144
	yes (1)	441	72	98

underlying truth

		poverty class	
sex	**Abitur**	0	1
0	0	777+131	217
	1	492+81	61
1	0	668+144	284
	1	441+98	72

no asmpt. about $q_{na|y}^{[g]}$

$\underline{\hat{\pi}}^{[1]} \approx 0.19$

$\underline{\hat{\pi}}^{[2]} \approx 0.10$

$\underline{\hat{\pi}}^{[3]} \approx 0.27$

$\underline{\hat{\pi}}^{[4]} \approx 0.12$

Illustration of the cautious approach (survey data only)

observation

		poverty class		
sex	**Abitur**	rich (0)	poor (1)	na
male (0)	no (0)	777	217	131
	yes (1)	492	61	81
female (1)	no (0)	668	284	144
	yes (1)	441	72	98

underlying truth

		poverty class	
sex	**Abitur**	0	1
0	0	777	217+131
	1	492	61+81
1	0	668	284+144
	1	441	72+98

no asmpt. about $q^{[g]}_{na|y}$

$\underline{\hat{\pi}}^{[1]} \approx 0.19$, $\overline{\hat{\pi}}^{[1]} \approx 0.31$

$\underline{\hat{\pi}}^{[2]} \approx 0.10$, $\overline{\hat{\pi}}^{[2]} \approx 0.22$

$\underline{\hat{\pi}}^{[3]} \approx 0.27$, $\overline{\hat{\pi}}^{[3]} \approx 0.37$

$\underline{\hat{\pi}}^{[4]} \approx 0.12$, $\overline{\hat{\pi}}^{[4]} \approx 0.28$

Illustration of the cautious approach (survey data only)

observation

		poverty class		
sex	**Abitur**	rich (0)	poor (1)	na
male (0)	no (0)	777	217	131
	yes (1)	492	61	81
female (1)	no (0)	668	284	144
	yes (1)	441	72	98

underlying truth

		poverty class	
sex	**Abitur**	0	1
0	0	777+102	217+29
	1	492+72	61+9
1	0	668+101	284+43
	1	441+84	72+14

no asmpt. about $q^{[g]}_{na|y}$

$\underline{\hat{\pi}}^{[1]} \approx 0.19,\ \overline{\hat{\pi}}^{[1]} \approx 0.31$

$\underline{\hat{\pi}}^{[2]} \approx 0.10,\ \overline{\hat{\pi}}^{[2]} \approx 0.22$

$\underline{\hat{\pi}}^{[3]} \approx 0.27,\ \overline{\hat{\pi}}^{[3]} \approx 0.37$

$\underline{\hat{\pi}}^{[4]} \approx 0.12,\ \overline{\hat{\pi}}^{[4]} \approx 0.28$

Specific value about R, here: $R = 1$ (MAR):

$\hat{\pi}^{[1]} \approx 0.22,\ \hat{\pi}^{[2]} \approx 0.11,\ \hat{\pi}^{[3]} \approx 0.30, \hat{\pi}^{[4]} \approx 0.14$

Illustration of the cautious approach (survey data only)

observation

		poverty class		
sex	**Abitur**	rich (0)	poor (1)	na
male (0)	no (0)	777	217	131
	yes (1)	492	61	81
female (1)	no (0)	668	284	144
	yes (1)	441	72	98

underlying truth

		poverty class	
sex	**Abitur**	0	1
0	0	777+□	217+□
	1	492+□	61+□
1	0	668+□	284+□
	1	441+□	72+□

weak info $R \in [0, 1]$

$\underline{\hat{\pi}}^{[1]} \approx 0.19,\ \overline{\hat{\pi}}^{[1]} \approx 0.22$

$\underline{\hat{\pi}}^{[2]} \approx 0.10,\ \overline{\hat{\pi}}^{[2]} \approx 0.11$

$\underline{\hat{\pi}}^{[3]} \approx 0.27,\ \overline{\hat{\pi}}^{[3]} \approx 0.30$

$\underline{\hat{\pi}}^{[4]} \approx 0.12,\ \overline{\hat{\pi}}^{[4]} \approx 0.14$

Specific value about R, here: $R = 1$ (MAR):

$\hat{\pi}^{[1]} \approx 0.22,\quad \hat{\pi}^{[2]} \approx 0.11,\quad \hat{\pi}^{[3]} \approx 0.30, \hat{\pi}^{[4]} \approx 0.14$

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:**

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

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

Details: Plass, Omar, Augustin, 2017, PMLR

Cautious synthetic estimator

- **No assumptions:**

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

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

- **Partial assumptions:**

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

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

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

Details: Plass, Omar, Augustin, 2017, PMLR

The LGREG estimator (without nonresponse)...

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

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

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

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

- 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

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

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

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 $\hat{\bar{\pi}}_{LGREG,i^*}$:

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

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 $\overline{R}$:

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

3.) Include the estimators that minimize

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

$\Rightarrow$ Since $\pi^{[g]}$ and $\pi_{i^*}^{[g]}$ 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]
...	...	...

Outlook and summary

Next steps:

- 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

Conclusion:

	no assumptions	weak auxiliary information
synthetic estimator	✓	✓
LGREG estimator	✓	$\sim$

Furthermore (not in this talk):
First sensitivity analysis for model-based SAE estimators