Reliable likelihood inference under coarse data and the (non-)testability of coarsening assumptions

Leibniz-Institut für Bildungsverläufe
Arbeitsbereich: Methoden der Survey-Statistik

Julia Plass*, Marco Cattaneo**, Thomas Augustin*
Georg Schollmeyer*, Christian Heumann*

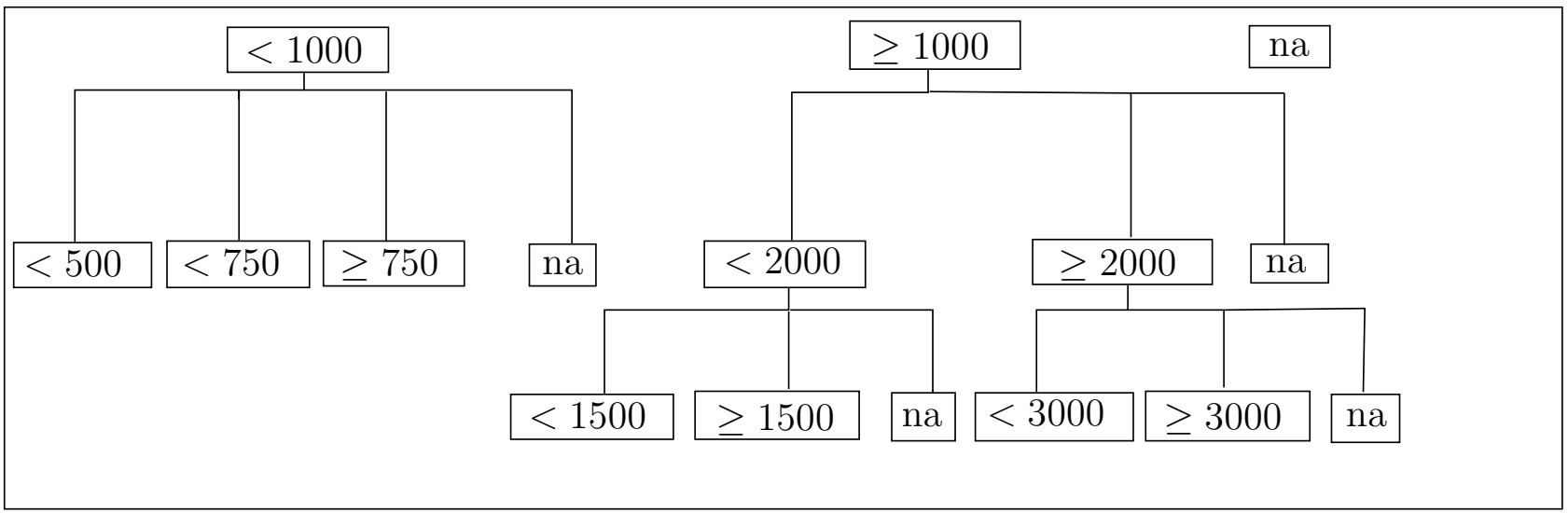
*Department of Statistics, Ludwig-Maximilians University and
**School of Mathematics and Physical Sciences, University of Hull

&

Bamberg, 29th of January 2018

Example for coarse categorical data

- coarse data: Data that can not be observed in the required accuracy originally intended in the substance matter context
- German Panel Study "Labour Market and Social Security" conducted by the IAB (PASS study)
- sensitive income question precisely asked, follow-up questions are directed to nonresponders:



- missing data as a special case

What's the problem?

Common dealing with incomplete data: assumptions

$\Rightarrow$ Missing at random (MAR) / coarsening at random (CAR)

$\Rightarrow$ Frequently: assumptions only for pragmatic reasons

What's the problem?

Common dealing with incomplete data: assumptions

$\Rightarrow$ Missing at random (MAR) / coarsening at random (CAR)

$\Rightarrow$ Frequently: assumptions only for pragmatic reasons

What's the problem?

Common dealing with incomplete data: assumptions

$\Rightarrow$ Missing at random (MAR) / coarsening at random (CAR)

$\Rightarrow$ Frequently: assumptions only for pragmatic reasons

Here:

1.) Reliable likelihood inference

2.) Testability of coarsening assumptions

3.) Reliable regression estimators

Coarse data (categorical setting)

LATENT

i.i.d. random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of categ. variables (X,Y) with $\Omega_X \times \Omega_Y$

coarsening

$y_i \in \mathcal{Y}_i$, i=1,...,n

OBSERVABLE

i.i.d. random sample $(x_1, \mathcal{Y}_1), \ldots, (x_n, \mathcal{Y}_n)$ of categ. variables (X,$\mathcal{Y}$) with $\Omega_X \times \Omega_{\mathcal{Y}}$, $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$

Coarse data (categorical setting)

LATENT

i.i.d. random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of categ. variables (X,Y) with $\Omega_X \times \Omega_Y$

coarsening

$y_i \in \mathcal{Y}_i$, i=1,...,n

OBSERVABLE

i.i.d. random sample $(x_1, \mathcal{Y}_1), \ldots, (x_n, \mathcal{Y}_n)$ of categ. variables (X,$\mathcal{Y}$) with $\Omega_X \times \Omega_{\mathcal{Y}}$, $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$

		$\mathcal{Y}$							
		{a}	{b}	{c}	{a,b}	{a,c}	{b,c}	{a,b,c}	sum
X	0	$n_{0\{a\}}$	$n_{0\{b\}}$	$n_{0\{c\}}$	$n_{0\{a,b\}}$	$n_{0\{a,c\}}$	$n_{0\{b,c\}}$	$n_{0\{a,b,c\}}$	n_0
	1	$n_{1\{a\}}$	$n_{1\{b\}}$	$n_{1\{c\}}$	$n_{1\{a,b\}}$	$n_{1\{a,c\}}$	$n_{1\{b,c\}}$	$n_{1\{a,b,c\}}$	n_1

Contingency table for coarse data

Coarse data (categorical setting)

LATENT

i.i.d. random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of categ. variables (X,Y) with $\Omega_X \times \Omega_Y$

coarsening

$y_i \in \mathcal{y}_i$, i=1,...,n

OBSERVABLE

i.i.d. random sample $(x_1, \mathcal{y}_1), \ldots, (x_n, \mathcal{y}_n)$ of categ. variables (X,$\mathcal{Y}$) with $\Omega_X \times \Omega_{\mathcal{Y}}$, $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$

		$\mathcal{Y}$			
		{a}	{b}	{a,b}	sum
X	0	$n_{0\{a\}}$	$n_{0\{b\}}$	$n_{0\{a,b\}}$	n_0
	1	$n_{1\{a\}}$	$n_{1\{b\}}$	$n_{1\{a,b\}}$	n_1

Contingency table for missing data

Coarse data (categorical setting)

LATENT

i.i.d. random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of categ. variables (X,Y) with $\Omega_X \times \Omega_Y$

coarsening

$y_i \in \mathcal{y}_i$, i=1,...,n

OBSERVABLE

i.i.d. random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of categ. variables $(X,\mathcal{Y})$ with $\Omega_X \times \Omega_{\mathcal{Y}}$, $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$

- **UBII**: receipt of Unemployment Benefit II
- $\mathcal{Y}$: categorical income
  - $\{<\}$: $< 1000€$
  - $\{\geq\}$: $\geq 1000€$
  - $\{<, \geq\}$: $<$ or $\geq$

UBII

	$\mathcal{Y}$			
	$\{<\}$	$\{\geq\}$	$\{<, \geq\}$	sum
0	38	385	95	518
1	36	42	9	87

PASS, w5 (Trappmann, 2010)

Coarse data (categorical setting)

LATENT

i.i.d. random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of categ. variables (X,Y) with $\Omega_X \times \Omega_Y$

coarsening

$y_i \in \mathcal{Y}_i$, i=1,...,n

OBSERVABLE

i.i.d. random sample $(x_1, \mathcal{Y}_1), \ldots, (x_n, \mathcal{Y}_n)$ of categ. variables (X,$\mathcal{Y}$) with $\Omega_X \times \Omega_{\mathcal{Y}}$, $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$

		$\mathcal{Y}$		
		$<$	$\geq$	sum
UBII	0	38+**95**	385	518
	1	36+**9**	42	87

potential true table

		$\mathcal{Y}$			
		$\{<\}$	$\{\geq\}$	$\{<, \geq\}$	sum
UBII	0	38	385	**95**	518
	1	36	42	**9**	87

observed table

Coarse data (categorical setting)

LATENT

i.i.d. random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of categ. variables (X,Y) with $\Omega_X \times \Omega_Y$

coarsening

$y_i \in \mathcal{Y}_i$, i=1,...,n

OBSERVABLE

i.i.d. random sample $(x_1, \mathcal{Y}_1), \ldots, (x_n, \mathcal{Y}_n)$ of categ. variables (X,$\mathcal{Y}$) with $\Omega_X \times \Omega_{\mathcal{Y}}$, $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$

		$\mathcal{Y}$		
		$<$	$\geq$	sum
UBII	0	38	385+**95**	518
	1	36	42+**9**	87

potential true table

		$\mathcal{Y}$			
		$\{<\}$	$\{\geq\}$	$\{<, \geq\}$	sum
UBII	0	38	385	**95**	518
	1	36	42	**9**	87

observed table

Coarse data (categorical setting)

LATENT

i.i.d. random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of categ. variables (X,Y) with $\Omega_X \times \Omega_Y$

coarsening

$y_i \in \mathcal{Y}_i$, i=1,...,n

OBSERVABLE

i.i.d. random sample $(x_1, \mathcal{Y}_1), \ldots, (x_n, \mathcal{Y}_n)$ of categ. variables (X,$\mathcal{Y}$) with $\Omega_X \times \Omega_{\mathcal{Y}}$, $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$

		$\mathcal{Y}$		
		$<$	$\geq$	sum
UBII	0	38+**20**	385+**75**	518
	1	36+**5**	42+**4**	87

potential true table

		$\mathcal{Y}$			
		$\{<\}$	$\{\geq\}$	$\{<, \geq\}$	sum
UBII	0	38	385	**95**	518
	1	36	42	**9**	87

observed table

General ML estimation (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

LATENT

$\pi_{xy} := P(Y=y|X=x)$

(error-freeness)

Observation model

$q_{y|xy} := P(\mathcal{Y} = y|X=x, Y=y)$

OBSERVABLE

$p_{xy} := P(\mathcal{Y} = y|X=x)$

General ML estimation (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

LATENT

$$\vartheta = (\pi_{xy}^{\mathsf{T}},\ \mathsf{q}_{y|xy}^{\mathsf{T}})^{\mathsf{T}}$$

OBSERVABLE

$$\mathsf{p}_{xy} := \mathsf{P}(\mathcal{Y} = y|\mathsf{X}{=}\mathsf{x})$$

General ML estimation (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

LATENT

$$\vartheta = (\pi_{xy}^{\mathsf{T}},\ \mathsf{q}_{y|xy}^{\mathsf{T}})^{\mathsf{T}}$$

OBSERVABLE

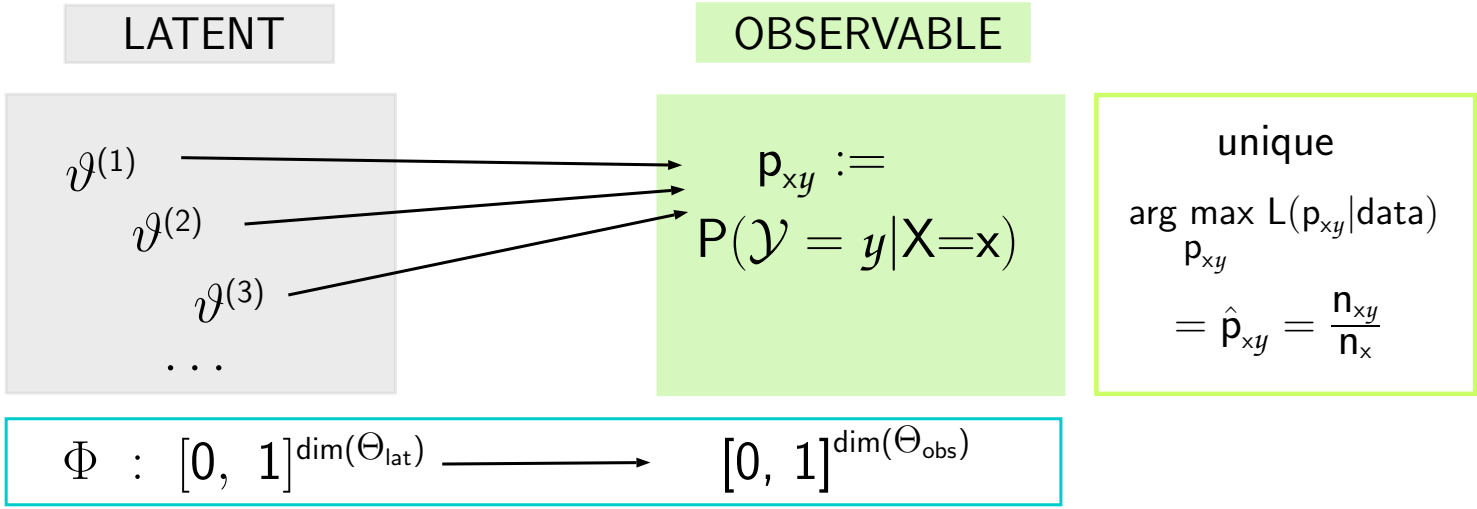
$$\mathsf{p}_{xy} := \mathsf{P}(\mathcal{Y} = y|\mathsf{X}{=}\mathsf{x})$$

unique

$$\underset{\mathsf{p}_{xy}}{\arg\max}\ \mathsf{L}(\mathsf{p}_{xy}|\mathsf{data}) = \hat{\mathsf{p}}_{xy} = \frac{\mathsf{n}_{xy}}{\mathsf{n}_{x}}$$

1.) Determine MLE of observed variable distribution

General ML estimation (Plass, Augustin, Cattaneo, Schollmeyer, 2015)

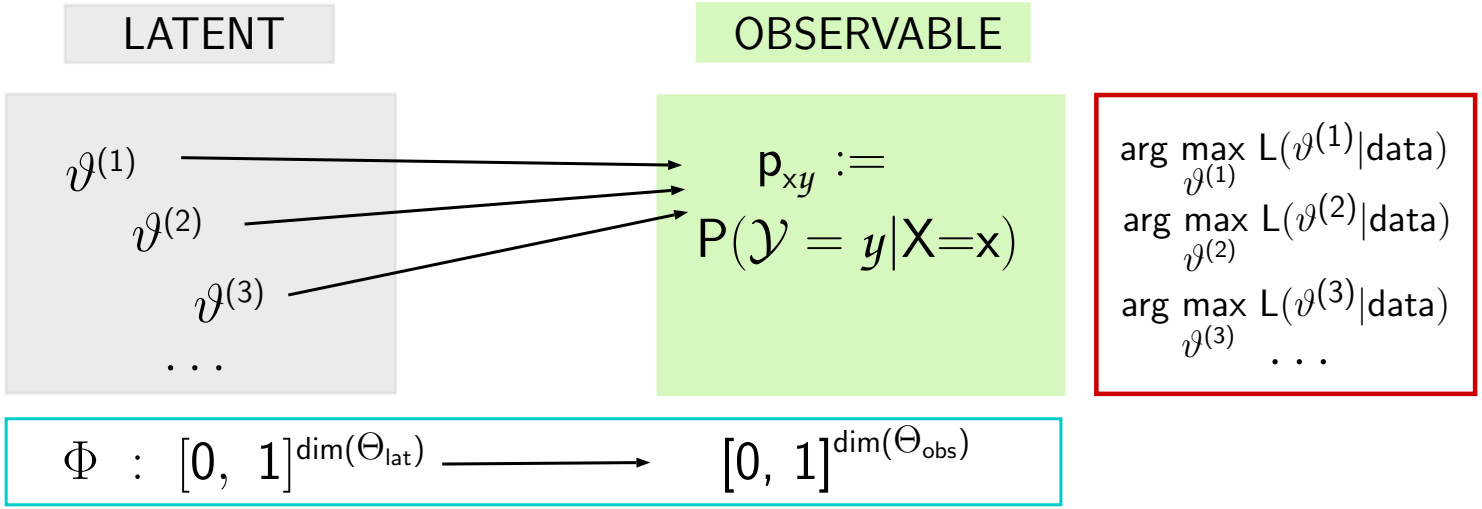


1.) Determine MLE of observed variable distribution

2.) Use connection between both worlds

$$p_{xy} = \sum_{y \in y} \left(\pi_{xy} \cdot q_{y|xy} \right).$$

General ML estimation (Plass, Augustin, Cattaneo, Schollmeyer, 2015)



1.) Determine MLE of observed variable distribution

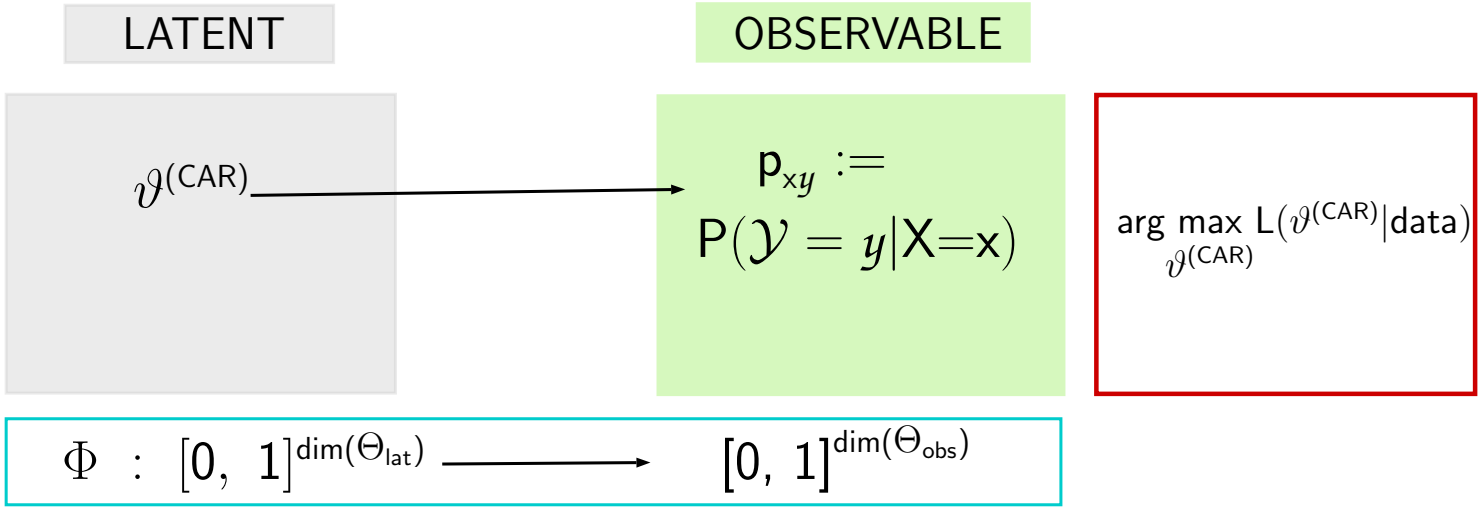
2.) Use connection between both worlds

$$p_{xy} = \sum_{y \in y} \left(\pi_{xy} \cdot q_{y|xy} \right).$$

3.) Use invariance of the likelihood

$$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x}, \frac{\sum_{y \ni y} n_{xy}}{n_x} \right], \quad \hat{q}_{y|xy} \in \left[0, \frac{n_{xy}}{n_{x\{y\}} + n_{xy}} \right].$$

General ML estimation (Plass, Augustin, Cattaneo, Schollmeyer, 2015)



1.) Determine MLE of observed variable distribution

2.) Use connection between both worlds

$$p_{xy} = \sum_{y \in y} \left(\pi_{xy} \cdot q_{y|xy} \right).$$

3.) Use invariance of the likelihood

$$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x}, \frac{\sum_{y \ni y} n_{xy}}{n_x} \right], \quad \hat{q}_{y|xy} \in \left[0, \frac{n_{xy}}{n_{x\{y\}} + n_{xy}} \right].$$

Coarsening at random (CAR)

Definition of CAR (Heitjan, Rubin, 1991):

For each fixed y and x, $q_{y|xy}$ takes the same value $\forall y \in y$.

Coarsening at random (CAR)

Definition of CAR (Heitjan, Rubin, 1991):
For each fixed y and x, $q_{y|xy}$ takes the same value $\forall y \in y$.

Illustrated by the example:

- The probability of $\{<, \geq\}$ is taken to be independent of the true income category in both subgroups split by UBII:

$$q_{\{<,\geq\}|0<} = q_{\{<,\geq\}|0\geq} \quad \text{and} \quad q_{\{<,\geq\}|1<} = q_{\{<,\geq\}|1\geq}$$

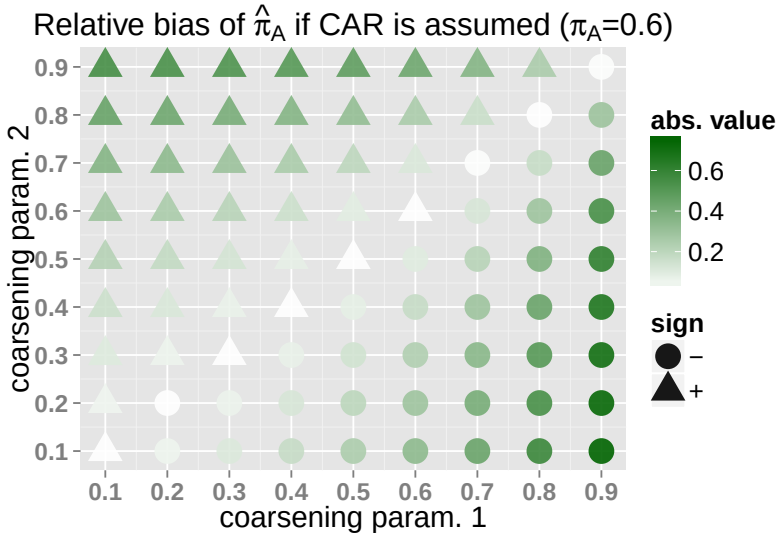
Coarsening at random (CAR)

Definition of CAR (Heitjan, Rubin, 1991):
For each fixed y and x, $q_{y|xy}$ takes the same value $\forall y \in y$.

Illustrated by the example:

- The probability of $\{<, \geq\}$ is taken to be independent of the true income category in both subgroups split by UBII:

$$q_{\{<,\geq\}|0<} = q_{\{<,\geq\}|0\geq} \quad \text{and} \quad q_{\{<,\geq\}|1<} = q_{\{<,\geq\}|1\geq}$$



Coarsening at random (CAR)

Definition of CAR (Heitjan, Rubin, 1991):
For each fixed y and x, $q_{y|xy}$ takes the same value $\forall y \in y$.

Illustrated by the example:

- The probability of $\{<, \geq\}$ is taken to be independent of the true income category in both subgroups split by UBII:

$$q_{\{<,\geq\}|0<} = q_{\{<,\geq\}|0\geq} \quad \text{and} \quad q_{\{<,\geq\}|1<} = q_{\{<,\geq\}|1\geq}$$

- Resulting estimators:

$$\hat{\pi}^{(CAR)}_{x<} = \frac{n_{x\{<\}}}{n_{x\{<\}} + n_{x\{\geq\}}}, \quad \hat{q}^{(CAR)}_{\{<,\geq\}|x<} = \hat{q}^{(CAR)}_{\{<,\geq\}|x\geq} = \frac{n_{x\{<,\geq\}}}{n_x}$$

Summary: Reliable likelihood inference

Estimators for subgroup 0 (PASS data example) under ...

- ... no assumptions

$$\hat{\pi}_{0<} \in [0.07, 0.26], \quad \hat{q}_{\{<,\geq\}|0<} \in [0, 0.71], \quad \hat{q}_{\{<,\geq\}|0\geq} \in [0, 0.20]$$

- ... CAR

$$\hat{\pi}^{(CAR)}_{0<} = 0.09, \qquad \hat{q}^{(CAR)}_{\{<,\geq\}|0<} = 0.18, \qquad \hat{q}^{(CAR)}_{\{<,\geq\}|0\geq} = 0.18$$

- all tenable auxiliary information formalized via (Nordheim, 1984)

$$R_{0,<,\geq,\{<,\geq\}} = \frac{q_{\{<,\geq\}|0<}}{q_{\{<,\geq\}|0\geq}} \quad \in [0, 1] \ :$$

$$\hat{\pi}_{0<} \in [0.07, 0.09], \quad \hat{q}_{\{<,\geq\}|0<} \in [0, 0.18], \quad \hat{q}_{\{<,\geq\}|0\geq} \in [0, 0.18]$$

What's the problem?

Common dealing with incomplete data: assumptions

$\Rightarrow$ Missing at random (MAR) / coarsening at random (CAR)

$\Rightarrow$ Frequently: assumptions only for pragmatic reasons

Here:

1.) Reliable likelihood inference

2.) Testability of coarsening assumptions

3.) Reliable regression estimators

Subgroup independence (SI)

Subgroup independence (Plass, Augustin, Cattaneo, Schollmeyer, 2015):
For each fixed y and y $\in y$, $q_{y|xy}$ takes the same value $\forall x \in \Omega_X$.

Subgroup independence (SI)

Subgroup independence (Plass, Augustin, Cattaneo, Schollmeyer, 2015):
For each fixed y and $y \in \mathcal{y}$, $q_{y|xy}$ takes the same value $\forall x \in \Omega_X$.

Illustrated by the example:

- The probability of $\{<, \geq\}$ is taken to be independent of the receipt of the UBII given y:

$$q_{\{<,\geq\}|\mathbf{0}<} = q_{\{<,\geq\}|\mathbf{1}<} \quad \text{and} \quad q_{\{<,\geq\}|\mathbf{0}\geq} = q_{\{<,\geq\}|\mathbf{1}\geq}$$

Subgroup independence (SI)

Subgroup independence (Plass, Augustin, Cattaneo, Schollmeyer, 2015):
For each fixed y and $y \in y$, $q_{y|xy}$ takes the same value $\forall x \in \Omega_X$.

Illustrated by the example:

- The probability of $\{<, \geq\}$ is taken to be independent of the receipt of the UBII given y:

  $q_{\{<,\geq\}|\mathbf{0}<} = q_{\{<,\geq\}|\mathbf{1}<}$ and $q_{\{<,\geq\}|\mathbf{0}\geq} = q_{\{<,\geq\}|\mathbf{1}\geq}$

- Resulting estimators (if well-defined and inside [0,1]):

$$\hat{\pi}^{(SI)}_{x<} = \frac{n_{x\{<\}}}{n_x}\frac{v}{w}, \quad \hat{q}^{(SI)}_{\{<,\geq\}|x<} = 1 - \frac{w}{v}, \quad \hat{q}^{(SI)}_{\{<,\geq\}|x\geq} = 1 - \frac{w}{z}$$

with $v = n_0 n_{1\{\geq\}} - n_{0\{\geq\}} n_1$, $w = n_{0\{<\}} n_{1\{\geq\}} - n_{0\{\geq\}} n_{1\{<\}}$
and $z = n_{0\{<\}} n_1 - n_{1\{<\}} n_0$

Testing SI? $\Rightarrow$:-)

Estimators for subgroup 0:

$$\hat{\pi}_{0<} \in [0.07, 0.26], \quad \hat{q}_{\{<,\geq\}|0<} \in [0,\ 0.71], \quad \hat{q}_{\{<,\geq\}|0\geq} \in [0,\ 0.198]$$

$$\hat{\pi}_{0<}^{(SI)} = 0.42, \qquad \hat{q}_{\{<,\geq\}|0<}^{(SI)} = -0.04, \qquad \hat{q}_{\{<,\geq\}|0\geq}^{(SI)} = 0.20$$

Testing SI? $\Rightarrow$:-)

Estimators for subgroup 0:

$$\hat{\pi}_{0<} \in [0.07, 0.26], \quad \hat{q}_{\{<,\geq\}|0<} \in [0,\ 0.71], \quad \hat{q}_{\{<,\geq\}|0\geq} \in [0,\ 0.198]$$

$$\hat{\pi}_{0<}^{(SI)} = 0.42, \qquad \hat{q}_{\{<,\geq\}|0<}^{(SI)} = -0.04, \qquad \hat{q}_{\{<,\geq\}|0\geq}^{(SI)} = 0.20$$

$\Rightarrow$ There are data situations that might hint to (partial) incompatibility with SI

$\Rightarrow$ SI is testable in our setting

Testability of SI

- Number of degrees of freedom under SI:

$$\begin{aligned} \mathrm{df}^{\mathrm{SI}} &= \dim(\Theta_{\mathrm{obs}}) - \dim(\Theta_{\mathrm{lat,\ SI}}) \\ &= k \cdot (2^m - 2) - [k \cdot (m-1) + m \cdot (2^{m-1} - 1)] \end{aligned}$$

  with k as the number of subgroups and $m = |\Omega_Y|$

- Point-identification and testability are only valid if sufficient subgroups are available inducing $\mathrm{df} \geq 0$

m	2	3,4,5	6,7	8,9	...	20	...	50	...
minimum k	2	3	4	5	...	11	...	26	...

LR-Test for SI: Hypotheses and test statistic

- Hypotheses:

$$
\begin{aligned}
H_0 &: \quad q_{y|xy} = q_{y|x'y} \text{ for all } y \in \Omega_{\mathcal{Y}},\ x,\ x' \in \Omega_X,\ y \in \Omega_Y, \\
H_1 &: \quad q_{y|xy} \neq q_{y|x'y} \quad \text{for some } y \in \Omega_{\mathcal{Y}}\ x,\ x' \in \Omega_X,\ y \in \Omega_Y.
\end{aligned}
$$

LR-Test for SI: Hypotheses and test statistic

- Hypotheses:

$$\begin{aligned} H_0 &: q_{y|xy} = q_{y|x'y} \text{ for all } y \in \Omega_Y,\ x,\ x' \in \Omega_X,\ y \in \Omega_Y, \\ H_1 &: q_{y|xy} \neq q_{y|x'y} \text{ for some } y \in \Omega_Y\ x,\ x' \in \Omega_X,\ y \in \Omega_Y. \end{aligned}$$

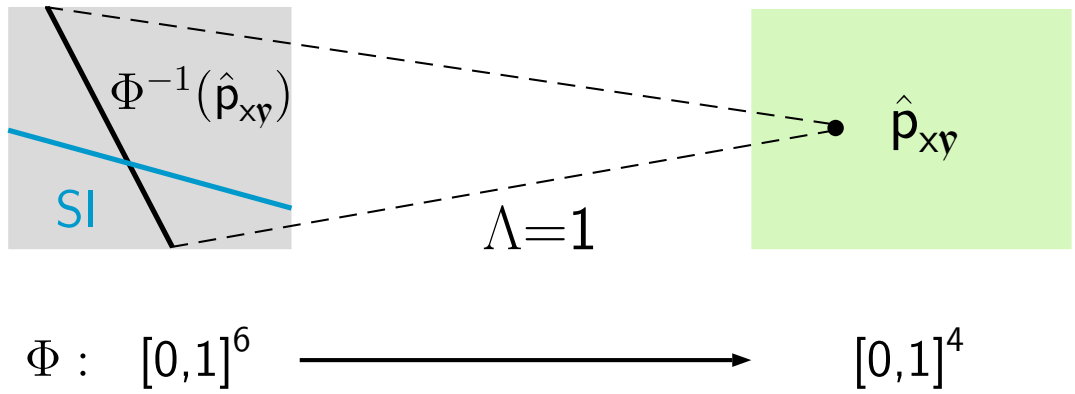
- Test based on test statistic (Wilks, 1938)

$$T = -2 \cdot \ln(\Lambda(y_1, \ldots, y_n, x_1, \ldots, x_n)) \quad \text{with}$$

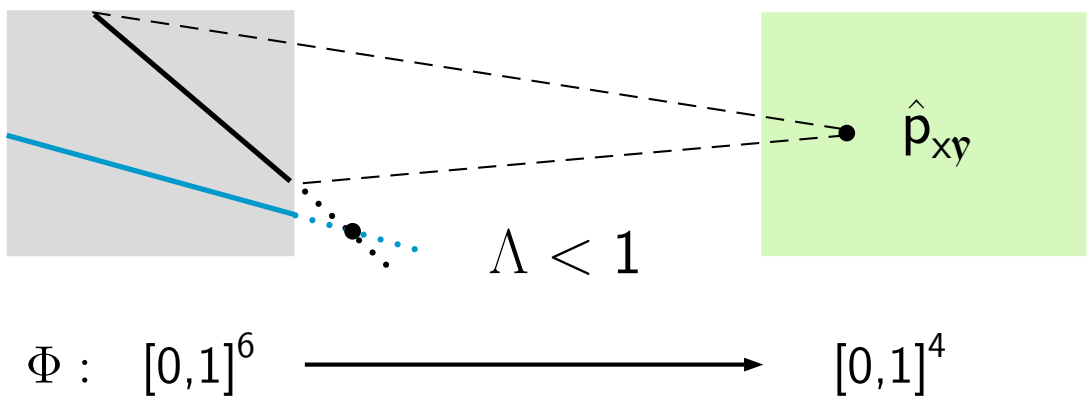
$$\Lambda(y_1, \ldots, y_n, x_1, \ldots, x_n) = \frac{\sup_{H_0} L(\vartheta | y_1, \ldots, y_n, x_1, \ldots, x_n)}{\sup_{H_0 \cup H_1} L(\vartheta | y_1, \ldots, y_n, x_1, \ldots, x_n)} .$$

LR-Test for SI: Sensitivity of Λ

- No evidence to reject SI



- Some evidence to reject SI



LR-Test for SI: Distribution of the test statistic under H_0...

- Asymptotic distribution of T under H_0:
  - $\frac{1}{2}\delta_0 + \frac{1}{2}\chi^2_1$ if $df^{SI} = 0$
  - $\chi^2_{df^{SI}}$ if $df^{SI} > 0$,

  where δ_0 is the Dirac distribution at 0 (Chernoff, 1954)
- Critical values for test decision:
  - $\chi^2_{1,1-2\cdot\alpha}$ if $df^{SI} = 0$
  - $\chi^2_{df^{SI},1-\alpha}$ if $df^{SI} > 0$,

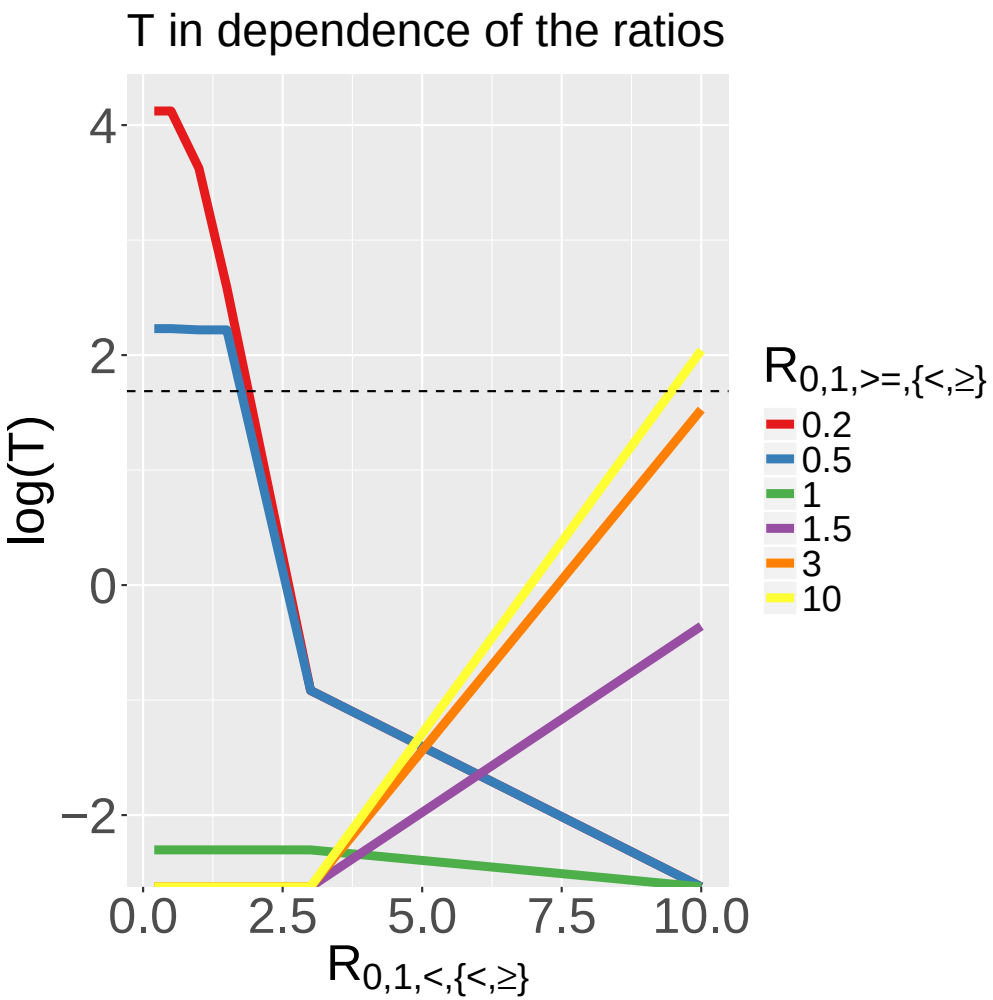
LR-Test for SI: Distribution of the test statistic under H_0...

- Asymptotic distribution of T under H_0:
  - $\frac{1}{2}\delta_0 + \frac{1}{2}\chi^2_1$ if $df^{SI} = 0$
  - $\chi^2_{df^{SI}}$ if $df^{SI} > 0$,

  where δ_0 is the Dirac distribution at 0 (Chernoff, 1954)
- Critical values for test decision:
  - $\chi^2_{1,1-2\cdot\alpha}$ if $df^{SI} = 0$
  - $\chi^2_{df^{SI},1-\alpha}$ if $df^{SI} > 0$,
- In data example (with $\alpha = 0.01$):

  $$T = 0.14 < 5.4 = \chi^2_{1,1-2\cdot 0.01} \quad \Rightarrow H_0 \text{ can not be rejected}$$

LR-Test for generalized SI



with ratios

$$R_{0,1,<,\{<,\geq\}} = \frac{q_{\{<,\geq\}|0<}}{q_{\{<,\geq\}|1<}}$$

and

$$R_{0,1,\geq,\{<,\geq\}} = \frac{q_{\{<,\geq\}|0\geq}}{q_{\{<,\geq\}|1\geq}}$$

Summary: CAR versus SI

	CAR	SI
Point-identifying?	always	in specific settings
Testability	generally impossible	possible in specific settings

Construction of a hypothesis test:

- H_0: (g)SI, H_1: no (g)SI
- Test statistic based on the Likelihood Ratio

(Relative) profile log-likelihood

- global log-likelihood

$$l(\Phi^{-1}(p_{00<}, p_{00\geq}, \ldots)) = l(\pi_{00<}, q_{na|00<}, q_{na|00\geq}, \ldots)$$

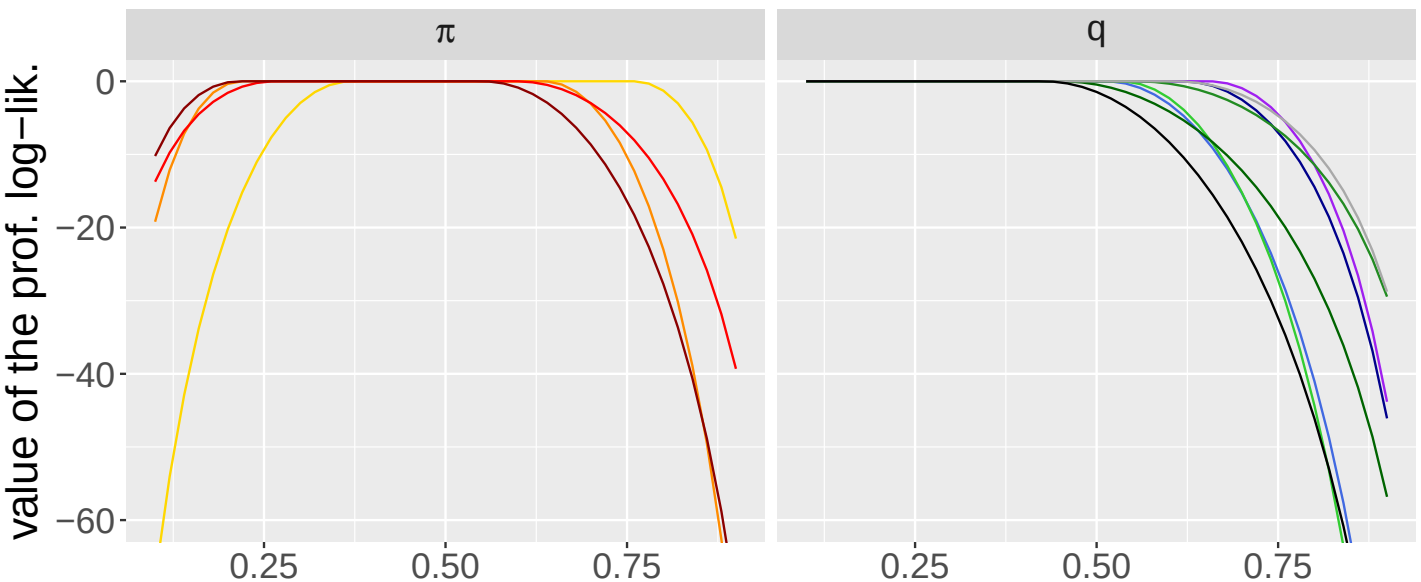
(Relative) profile log-likelihood

- global log-likelihood

$$l(\Phi^{-1}(p_{00<}, p_{00\geq}, \ldots)) = l(\pi_{00<}, q_{na|00<}, q_{na|00\geq}, \ldots)$$

- profile log-likelihood with nuissance param. $\xi = \vartheta \setminus \pi_{00<}$ or $\xi = \vartheta \setminus q_{na|00<}$

$$l(\pi_{00<}) = \max_{\xi} l(\pi_{00<}, \xi) \quad \text{or} \quad l(q_{y|\mathbf{x}y}) = \max_{\xi} l(q_{na|00<}, \xi)$$



What's the problem?

Common dealing with incomplete data: assumptions

$\Rightarrow$ Missing at random (MAR) / coarsening at random (CAR)

$\Rightarrow$ Frequently: assumptions only for pragmatic reasons

Here:

1.) Reliable likelihood inference

2.) Testability of coarsening assumptions

3.) Reliable regression estimators

Reliable regression estimators (Direct method)

Logit model with linear predictor
$\eta_{\mathbf{x}} = \beta_0 + \beta_1 \cdot x_{Abitur} + \beta_2 \cdot x_{age} + \beta_{12} \cdot x_{Abitur} x_{age}$:

Response function h *Link function $g = h^{-1}$*

$$\pi_{\mathbf{x}<} = h(\eta_{\mathbf{x}}) = \frac{\exp(\eta_{\mathbf{x}})}{1 + \exp(\eta_{\mathbf{x}})} \quad \text{and} \quad \eta_{\mathbf{x}} = g(\pi_{\mathbf{x}<}) = \ln\left(\frac{\pi_{\mathbf{x}<}}{1 - \pi_{\mathbf{x}<}}\right).$$

Reliable regression estimators (Direct method)

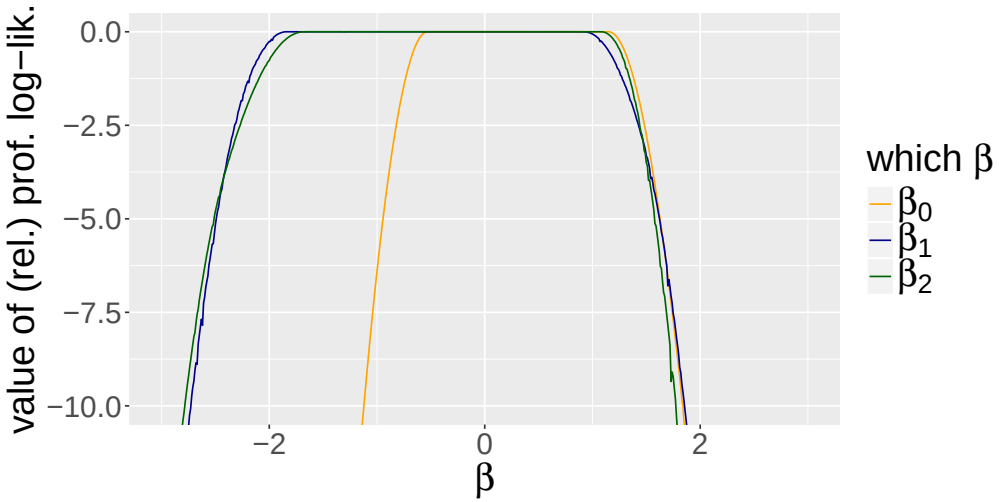
Logit model with linear predictor
$\eta_{\mathbf{x}} = \beta_0 + \beta_1 \cdot x_{Abitur} + \beta_2 \cdot x_{age} + \beta_{12} \cdot x_{Abitur} x_{age}$:

Response function h *Link function* $g = h^{-1}$

$$\pi_{\mathbf{x}<} = h(\eta_{\mathbf{x}}) = \frac{\exp(\eta_{\mathbf{x}})}{1 + \exp(\eta_{\mathbf{x}})} \quad \text{and} \quad \eta_{\mathbf{x}} = g(\pi_{\mathbf{x}<}) = \ln\left(\frac{\pi_{\mathbf{x}<}}{1 - \pi_{\mathbf{x}<}}\right).$$

$\Rightarrow$ global likelihood
$l(\beta_0, \beta_1, \ldots, q_{na|00<}, \ldots)$

$\Rightarrow$ Consider (rel.)
profile log-lik. for
each β:



Reliable regression estimators (Two-step method)

No parametric assumption on the regression model:

- Saturated model, all interactions included
- $\ln\left(\frac{\pi_{\mathbf{x}<}}{1-\pi_{\mathbf{x}<}}\right) = \beta_0 + \beta_1 \cdot x_{Abitur} + \beta_2 \cdot x_{age} + \beta_{12} x_{Abitur} x_{age}$
- $\pi_{\mathbf{x}<}$ and regression coefficients basically represent same information

Parametric assumption on the regression model:

- (some) interactions are equal to 0 (e.g. $\beta_{12} = 0$)
- $\ln\left(\frac{\pi^*_{\mathbf{x}<}}{1-\pi^*_{\mathbf{x}<}}\right) = \beta_0 + \beta_1 \cdot x_{Abitur} + \beta_2 \cdot x_{age}$
- $\pi_{\mathbf{x}<}$ and regression coefficients do not represent same information

Aim:
We want to study the impact of the parametric assumption on the regression model in the coarse data problem

Reliable regression estimators (Two-step method)

Two-step method:

1.) Estimate the bounds $\underline{\hat{\pi}}_{\mathbf{x}<}$ and $\overline{\hat{\pi}}_{\mathbf{x}<}$ (no param. assump.)

2.) Use these bounds to obtain the reliable regression estimators

No parametric assumption:

Function h is bijective $\Rightarrow$ transform $\underline{\hat{\pi}}_{\mathbf{x}<}$ and $\overline{\hat{\pi}}_{\mathbf{x}<}$ via link function

Parametric assumption: solve an optimization problem

Reliable regression estimators (Two-step method)

$\beta_1 \to \min/\max$ given

$$\underline{\hat{\pi}}_{00<} \leq \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \leq \overline{\hat{\pi}}_{00<}, \qquad \underline{\hat{\pi}}_{10<} \leq \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} \leq \overline{\hat{\pi}}_{10<},$$

$$\underline{\hat{\pi}}_{01<} \leq \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)} \leq \overline{\hat{\pi}}_{01<}, \qquad \underline{\hat{\pi}}_{11<} \leq \frac{\exp(\beta_0+\beta_1+\beta_2)}{1+\exp(\beta_0+\beta_1+\beta_2)} \leq \overline{\hat{\pi}}_{11<}.$$

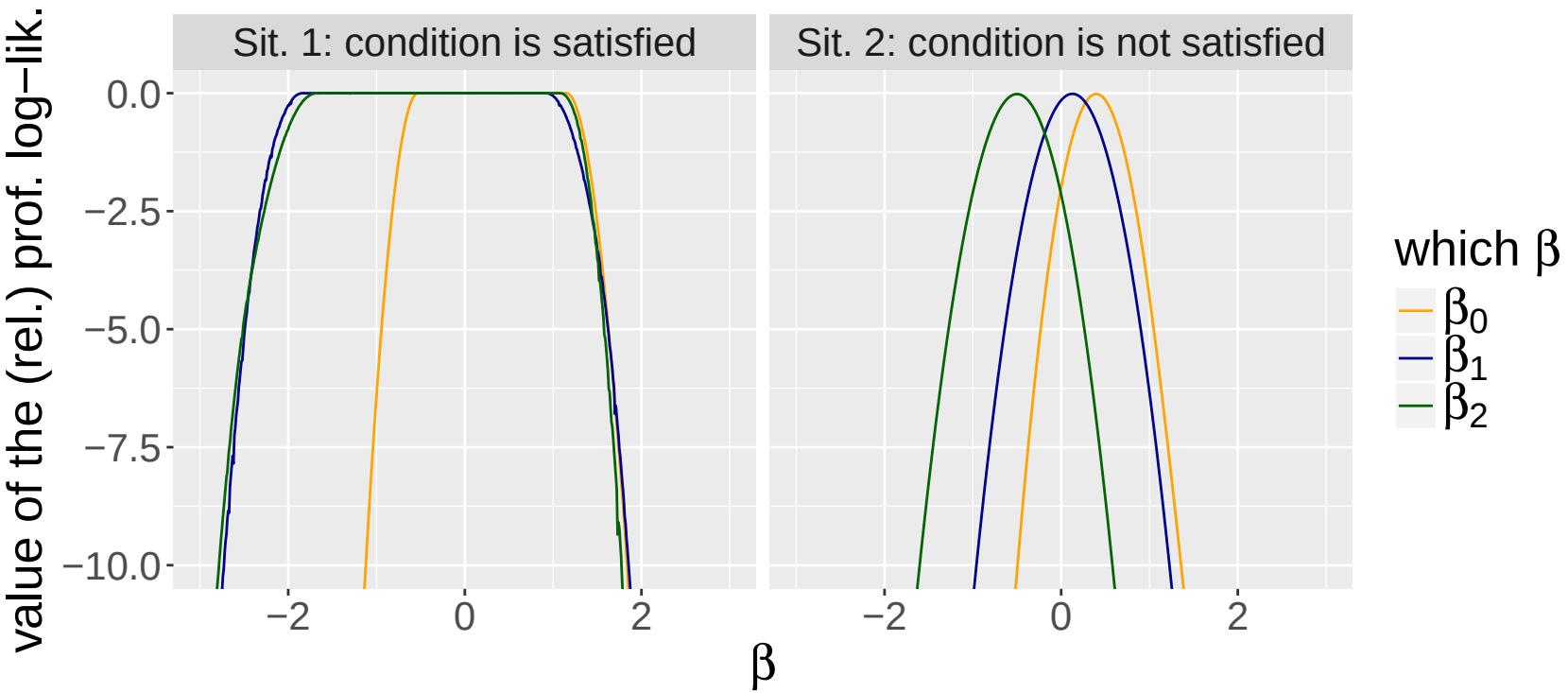
Different types of results

1.) There is a solution.
   a) Regression estimators obtainable that produce $\underline{\hat{\pi}}_{\mathbf{x}<}$ and $\overline{\hat{\pi}}_{\mathbf{x}<}$
   $\Rightarrow$ No impact on coarsening (same $\underline{\hat{q}}_{\{<,\geq\}|\mathbf{x}<}$ and $\overline{\hat{q}}_{\{<,\geq\}|\mathbf{x}<}$)
   b) The resulting regression estimators can only represent tighter bounds of the estimated latent variable distribution
   $\Rightarrow$ Tighter bounds compared to $\underline{\hat{q}}_{\{<,\geq\}|\mathbf{x}<}$ and $\overline{\hat{q}}_{\{<,\geq\}|\mathbf{x}<}$

2.) There is no solution.

Impact on the compatible coarsening scenarios?

Relative profile log-likelihood for the regression coefficients:



Results and conclusion

Reliable estimation and estimation under CAR (MAR)

- No parametric assumption on the regression model

$\hat{\beta}_0 \in [-0.53,\ 1.15]$	$\hat{\beta}_1 \in [-2.16,\ 0.92]$	$\hat{\beta}_2 \in [-2.42,\ 1.08]$	$\hat{\beta}_{12} \in [-2.76,\ 3.64]$
$\hat{\beta}_0 = 0.43$	$\hat{\beta}_1 = -0.85$	$\hat{\beta}_2 = -0.94$	$\hat{\beta}_{12} = 0.63$

- Parametric assumption on the regression model

$\hat{\beta}_0 \in [-0.53,\ 1.15]$	$\hat{\beta}_1 \in [-1.84,\ 0.92]$	$\hat{\beta}_2 \in [-1.68,\ 1.08]$
$\hat{\beta}_0 = 0.35$	$\hat{\beta}_1 = 0.05$	$\hat{\beta}_2 = 0.00$

Conclusions:

- Two methods for a reliable regression estimators have been studied
- Principally different impact on the estimated coarsening parameters
- Compatibility with observed data can be “repaired” in the presence of coarse data

Some important references


Couso, Dubois.
Statistical Reasoning with Set-Valued Information: Ontic vs. Epistemic Views, IJAR, 2014.


Chernoff.
On the distribution of the likelihood ratio, Ann. Stat. Math., 1954.


Heitjan, Rubin.
Ignorability and Coarse Data, Annals of Statistics, 1991.


Jaeger.
On testing the missing at random assumption, ECML, 2006.


Manski.
Partial Identification of Probability Distributions, Springer, 2003.


Nordheim.
Inference from nonrandomly missing categorical data: an example from a genetic study on Turner's syndrome, JASA, 1984.


Trappmann, Gundert, Wenzig, Gebhardt.
PASS: a household panel survey for research on unemployment and poverty, Schmollers Jahrbuch, 2010.


Wilks.
The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Stat., 1938.

Own references


Plass, Augustin, Cattaneo, Schollmeyer.
Statistical modelling under epistemic data imprecision, ISIPTA, 2015.


Plass, Cattaneo, Schollmeyer, Augustin.
On the testability of coarsening assumptions: A hypothesis test for subgroup independence, IJAR, 2017.


Plass, Cattaneo, Schollmeyer, Augustin.
Towards a reliable categorical regression analysis for nonrandomly coarsened observations: An analysis with German labour market data, under Review.


Plass, Omar, Augustin.
Towards a cautious modelling of missing data in Small Area Estimation, ISIPTA, 2017.


Plass, Fink, Schöning, Augustin.
Statistical modelling in surveys without neglecting The Undecided: Multinomial logistic regression models and imprecise classification trees under ontic data imprecision, ISIPTA, 2015.