Statistical modelling under epistemic data imprecision

Some results, next steps, questions,...

Julia Plass, Thomas Augustin
Georg Schollmeyer, Marco Cattaneo

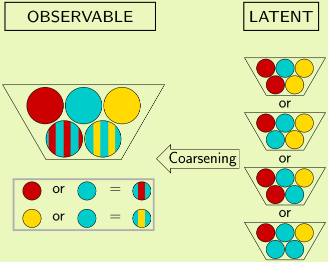
Department of Statistics, Ludwigs-Maximilians University

WPMSIIP 2015, 2nd of September 2015

Epistemic vs. ontic interpretation (Couso, Dubois, Sánchez, 2014)

Epistemic imprecision:

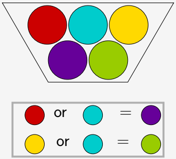
"Imprecise observation of something precise"



⇒ Truth is hidden due to the underlying coarsening mechanism

Ontic imprecision:

"Precise observation of something imprecise"

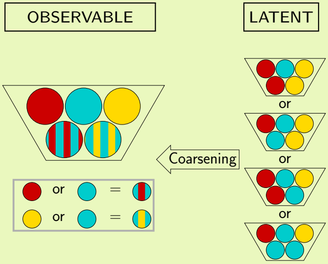


⇒ Truth is represented by coarse observation

Examples of data under epistemic imprecision

Epistemic imprecision:

"Imprecise observation of something precise"



$\Rightarrow$ Truth is hidden due to the underlying coarsening mechanism

Examples:

- Matched data sets with partially overlapping variables
- Coarsening as anonymization technique
- Missing data as special case

Here: PASS-data
$\mathcal{Y}$: income, X: UBII

$$\Omega_{\mathcal{Y}} = \{<, \geq, \text{na}\}$$
$$\Omega_X = \{0 \text{ (no)}, 1 \text{ (yes)}\}$$

Basic problem

OBSERVABLE		LATENT
$\mathcal{Y}$ coarse data		Y latent variable
		<u>Main goal:</u>
	coarsening mechanism	Estimation of
$p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i),\ i = 1, \ldots, n$	$q_{\mathscr{y}_i \vert y_i} = P(\mathcal{Y}_i = \mathscr{y}_i \vert Y_i = y_i)$	$\pi_{ij} = P(Y_i = j)$

Basic problem

OBSERVABLE

$\mathcal{Y}$ coarse data

$p_{y_i} = P(\mathcal{Y}_i = y_i),\ i = 1, \ldots, n$

Observation model $\mathcal{Q}$

error-freeness

coarsening mechanism

$q_{y_i|y_i} = P(\mathcal{Y}_i = y_i | Y_i = y_i)$

LATENT

Y latent variable

Main goal:

Estimation of

$\pi_{ij} = P(Y_i = j)$

1. Frequentist approach:

   Find maximum-likelihood estimator of $\boldsymbol{\gamma} = (\boldsymbol{q}^T_{y_i|\boldsymbol{y_i}},\ \boldsymbol{\pi}^T_{\boldsymbol{y}})^T$

   (i.i.d. case , regression case)

2. Bayesian approach:

   Find posterior mean

   (Next step)

Basic problem

OBSERVABLE

$\mathcal{Y}$ coarse data

$p_{\mathscr{Y}_i} = P(\mathcal{Y}_i = \mathscr{Y}_i), \; i = 1, \ldots, n$

Observation model $\mathcal{Q}$

error-freeness

coarsening mechanism

$q_{\mathscr{Y}_i|y_i} = P(\mathcal{Y}_i = \mathscr{Y}_i | Y_i = y_i)$

LATENT

Y latent variable

Main goal:

Estimation of

$\pi_{ij} = P(Y_i = j)$

1. Frequentist approach:

   Find maximum-likelihood estimator of $\boldsymbol{\gamma} = (\boldsymbol{q}^T_{\mathscr{Y}_i|\boldsymbol{y_i}}, \; \boldsymbol{\pi}^T_{\boldsymbol{y}})^T$

   (i.i.d. case , regression case)

2. Bayesian approach:

   Find posterior mean

   (Next step)

Basic idea for the i.i.d. case

OBSERVABLE

$\mathcal{Y}$ coarse data

$p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i),\ i = 1, \ldots, n$

Observation model $\mathcal{Q}$

error-freeness

coarsening mechanism

$q_{\mathscr{y}|y} = P(\mathcal{Y} = \mathscr{y} | Y = y)$

LATENT

Y latent variable

Main goal:

Estimation of $\pi_{ij} = P(Y_i = j)$

$\pi_{i1} = \pi_1, \ldots, \pi_{iK} = \pi_K$

$\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{y}|\boldsymbol{y}}^T,\ \boldsymbol{\pi}_{\boldsymbol{y}}^T)^T$

Use random-set perspective and determine maximum-likelihood estimator $\hat{p}_{\mathscr{y}}$

Likelihood for parameters $\boldsymbol{p} = (p_1, \ldots, p_{|\Omega_{\mathcal{Y}}|-1})^T$

$L(\boldsymbol{p}) \propto \prod_{\mathscr{y} \in \Omega_{\mathcal{Y}}} p_{\mathscr{y}}^{n_{\mathscr{y}}}$ is *uniquely* maximized by

$\boxed{\hat{p}_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n},}\qquad \mathscr{y} \in \{1, \ldots, |\Omega_{\mathcal{Y}}| - 1\}$

and thus $\hat{p}_{|\Omega_{\mathcal{Y}}|} = 1 - \sum_{m=1}^{|\Omega_{\mathcal{Y}}|-1} \hat{p}_m$.

Use the **connection between $\boldsymbol{p}$ and $\boldsymbol{\gamma}$**

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

and the **invariance of the likelihood** under parameter transformations, i.e.:

$$\boxed{\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}}$$

$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\mathscr{y} \ni y} n_{\mathscr{y}}}{n}\right]$

$\hat{q}_{\mathscr{y}|y} \in \left[0,\ \frac{n_{\mathscr{y}}}{n_{\{y\}} + n_{\mathscr{y}}}\right]$

Basic idea for the i.i.d. case

OBSERVABLE

$\mathcal{Y}$ coarse data

$p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i),\ i = 1, \dots, n$

Observation model $\mathcal{Q}$

error-freeness

coarsening mechanism

$q_{\mathscr{y}|y} = P(\mathcal{Y} = \mathscr{y} | Y = y)$

LATENT

Y latent variable

Main goal:

Estimation of $\pi_{ij} = P(Y_i = j)$

$\pi_{i1} = \pi_1, \dots, \pi_{iK} = \pi_K$

$\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{y}|\boldsymbol{y}}^T,\ \boldsymbol{\pi}_{\boldsymbol{y}}^T)^T$

Use the **connection between $\boldsymbol{p}$ and $\boldsymbol{\gamma}$**

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

Use random-set perspective and determine maximum-likelihood estimator $\hat{p}_{\mathscr{y}}$

Likelihood for parameters $\boldsymbol{p} = (p_1, \dots, p_{|\Omega_\mathcal{Y}|-1})^T$

$L(\boldsymbol{p}) \propto \prod_{\mathscr{y} \in \Omega_\mathcal{Y}} p_{\mathscr{y}}^{n_{\mathscr{y}}}$ is *uniquely* maximized by

$\hat{p}_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n}$, $\quad \mathscr{y} \in \{1, \dots, |\Omega_\mathcal{Y}| - 1\}$

and thus $\hat{p}_{|\Omega_\mathcal{Y}|} = 1 - \sum_{m=1}^{|\Omega_\mathcal{Y}|-1} \hat{p}_m$.

and the **invariance of the likelihood** under parameter transformations, i.e.:

$$\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}$$

$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\mathscr{y} \ni y} n_{\mathscr{y}}}{n}\right]$

$\hat{q}_{\mathscr{y}|y} \in \left[0,\ \frac{n_{\mathscr{y}}}{n_{\{y\}} + n_{\mathscr{y}}}\right]$

Basic idea for the i.i.d. case

OBSERVABLE

$\mathcal{Y}$ coarse data

Observation model $\mathcal{Q}$

error-freeness

LATENT

Y latent variable

Main goal:

Estimation of $\pi_{ij} = P(Y_i = j)$

$\pi_{i1} = \pi_1, \ldots, \pi_{iK} = \pi_K$

$p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i),\ i = 1, \ldots, n$

coarsening mechanism

$q_{\mathscr{y}|y} = P(\mathcal{Y} = \mathscr{y} | Y = y)$

$\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{y}|\boldsymbol{y}}^T,\ \boldsymbol{\pi}_{\boldsymbol{y}}^T)^T$

Use random-set perspective and determine maximum-likelihood estimator $\hat{p}_{\mathscr{y}}$

Likelihood for parameters $\boldsymbol{p} = (p_1, \ldots, p_{|\Omega_{\mathcal{Y}}|-1})^T$

$L(\boldsymbol{p}) \propto \prod_{\mathscr{y} \in \Omega_{\mathcal{Y}}} p_{\mathscr{y}}^{n_{\mathscr{y}}}$ is *uniquely* maximized by

$\boxed{\hat{p}_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n},}$ $\quad \mathscr{y} \in \{1, \ldots, |\Omega_{\mathcal{Y}}| - 1\}$

and thus $\hat{p}_{|\Omega_{\mathcal{Y}}|} = 1 - \sum_{m=1}^{|\Omega_{\mathcal{Y}}|-1} \hat{p}_m$.

Use the **connection between $\boldsymbol{p}$ and $\boldsymbol{\gamma}$**

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

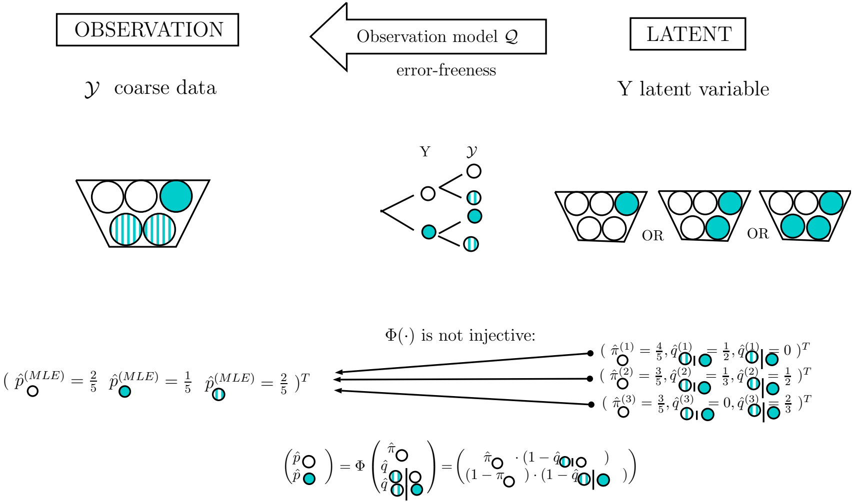
and the **invariance of the likelihood** under parameter transformations, i.e.:

$$\boxed{\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}}$$

$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\mathscr{y} \ni y} n_{\mathscr{y}}}{n}\right]$

$\hat{q}_{\mathscr{y}|y} \in \left[0,\ \frac{n_{\mathscr{y}}}{n_{\{y\}} + n_{\mathscr{y}}}\right]$

Basic idea for the i.i.d. case



Basic idea for the i.i.d. case

OBSERVABLE

$\mathcal{Y}$ coarse data

$p_{\mathscr{Y}_i} = P(\mathcal{Y}_i = \mathscr{Y}_i),\ i = 1, \dots, n$

Observation model $\mathcal{Q}$

error-freeness

coarsening mechanism

$q_{\mathscr{Y}|y} = P(\mathcal{Y} = \mathscr{Y} | Y = y)$

LATENT

Y latent variable

Main goal:

Estimation of $\pi_{ij} = P(Y_i = j)$

$\pi_{i1} = \pi_1, \dots, \pi_{iK} = \pi_K$

$\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{Y}|\boldsymbol{y}}^T,\ \boldsymbol{\pi}_{\boldsymbol{y}}^T)^T$

Use random-set perspective and determine maximum-likelihood estimator $\hat{p}_{\mathscr{Y}}$

Likelihood for parameters $\boldsymbol{p} = (p_1, \dots, p_{|\Omega_\mathcal{Y}|-1})^T$

$L(\boldsymbol{p}) \propto \prod_{\mathscr{Y} \in \Omega_\mathcal{Y}} p_{\mathscr{Y}}^{n_{\mathscr{Y}}}$ is *uniquely* maximized by

$\boxed{\hat{p}_{\mathscr{Y}} = \frac{n_{\mathscr{Y}}}{n},} \qquad \mathscr{Y} \in \{1, \dots, |\Omega_\mathcal{Y}| - 1\}$

and thus $\hat{p}_{|\Omega_\mathcal{Y}|} = 1 - \sum_{m=1}^{|\Omega_\mathcal{Y}|-1} \hat{p}_m$.

Use the **connection between $\boldsymbol{p}$ and $\boldsymbol{\gamma}$**

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

and the **invariance of the likelihood** under parameter transformations, i.e.:

$$\boxed{\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}}$$

$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\mathscr{Y} \ni y} n_{\mathscr{Y}}}{n}\right]$

$\hat{q}_{\mathscr{Y}|y} \in \left[0,\ \frac{n_{\mathscr{Y}}}{n_{\{y\}} + n_{\mathscr{Y}}}\right]$

Basic idea for the i.i.d. case

OBSERVABLE

$\mathcal{Y}$ coarse data

$p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i),\ i = 1, \dots, n$

Observation model $\mathcal{Q}$

error-freeness

coarsening mechanism

$q_{\mathscr{y}|y} = P(\mathcal{Y} = \mathscr{y} | Y = y)$

LATENT

Y latent variable

Main goal:

Estimation of $\pi_{ij} = P(Y_i = j)$

$\pi_{i1} = \pi_1, \dots, \pi_{iK} = \pi_K$

$\boldsymbol{\gamma} = (\boldsymbol{q}^T_{\mathscr{y}|\boldsymbol{y}},\ \boldsymbol{\pi}^T_{\boldsymbol{y}})^T$

Use random-set perspective and determine maximum-likelihood estimator $\hat{p}_{\mathscr{y}}$

Likelihood for parameters $\boldsymbol{p} = (p_1, \dots, p_{|\Omega_\mathcal{Y}|-1})^T$

$L(\boldsymbol{p}) \propto \prod_{\mathscr{y} \in \Omega_\mathcal{Y}} p_{\mathscr{y}}^{n_{\mathscr{y}}}$ is *uniquely* maximized by

$\boxed{\hat{p}_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n},}$ $\quad \mathscr{y} \in \{1, \dots, |\Omega_\mathcal{Y}| - 1\}$

and thus $\hat{p}_{|\Omega_\mathcal{Y}|} = 1 - \sum_{m=1}^{|\Omega_\mathcal{Y}|-1} \hat{p}_m$.

Use the **connection between $\boldsymbol{p}$ and $\boldsymbol{\gamma}$**

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

and the **invariance of the likelihood** under parameter transformations, i.e.:

$$\boxed{\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}}$$

$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n}, \frac{\sum_{\mathscr{y} \ni y} n_{\mathscr{y}}}{n}\right]$

$\hat{q}_{\mathscr{y}|y} \in \left[0, \frac{n_{\mathscr{y}}}{n_{\{y\}} + n_{\mathscr{y}}}\right]$

Basic idea for the i.i.d. case

OBSERVABLE

$\mathcal{Y}$ coarse data

$p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i),\ i = 1, \ldots, n$

Observation model $\mathcal{Q}$

error-freeness

coarsening mechanism

$q_{\mathscr{y}|y} = P(\mathcal{Y} = \mathscr{y} | Y = y)$

LATENT

Y latent variable

Main goal:

Estimation of $\pi_{ij} = P(Y_i = j)$

$\pi_{i1} = \pi_1, \ldots, \pi_{iK} = \pi_K$

$\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{y}|\boldsymbol{y}}^T,\ \boldsymbol{\pi}_{\boldsymbol{y}}^T)^T$

Use the **connection between $\boldsymbol{p}$ and $\boldsymbol{\gamma}$**

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

and the **invariance of the likelihood** under parameter transformations, i.e.:

$$\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}$$

Use random-set perspective and determine maximum-likelihood estimator $\hat{p}_{\mathscr{y}}$

Likelihood for parameters $\boldsymbol{p} = (p_1, \ldots, p_{|\Omega_{\mathcal{Y}}|-1})^T$

$L(\boldsymbol{p}) \propto \prod_{\mathscr{y} \in \Omega_{\mathcal{Y}}} p_{\mathscr{y}}^{n_{\mathscr{y}}}$ is *uniquely* maximized by

$\hat{p}_{\mathscr{y}} = \frac{n_{\mathscr{y}}}{n}$, $\quad \mathscr{y} \in \{1, \ldots, |\Omega_{\mathcal{Y}}| - 1\}$

and thus $\hat{p}_{|\Omega_{\mathcal{Y}}|} = 1 - \sum_{m=1}^{|\Omega_{\mathcal{Y}}|-1} \hat{p}_m$.

$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\mathscr{y} \ni y} n_{\mathscr{y}}}{n}\right]$

$\hat{q}_{\mathscr{y}|y} \in \left[0,\ \frac{n_{\mathscr{y}}}{n_{\{y\}} + n_{\mathscr{y}}}\right]$

Illustration (PASS data)

$n_< = 238,\ n_\geq = 835,\ n_{\text{na}} = 338$

$\hat{\pi}_< \in \left[\frac{238}{1411},\ \frac{238+338}{1411}\right]$

Basic idea for the regression case

OBSERVABLE

$\mathcal{Y}$ coarse data

Observation model $\mathcal{Q}$

error-freeness

LATENT

Y latent variable

Main goal:

Estimation of $\pi_{ij} = P(Y_i = j)$

coarsening mechanism

$p_{x\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i | X = x),\ i = 1, \dots, n$
$x \in \{0,\ 1\}$

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

$\pi_{ij} = P(Y_i = j | \mathbf{x}_i) = \frac{\exp(\beta_{j0} + \mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)}$

$\pi_{iK} = \left(1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)\right)^{-1}$

(multinomial logit model)

Use the **connection between $\boldsymbol{p}$** and $\boldsymbol{\gamma}$

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

$$\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{y}|\boldsymbol{xy}}^T,\ \boldsymbol{\pi}_{\boldsymbol{y}}^T)^T$$

Use random-set perspective and determine maximum-likelihood estimator $\hat{p}_{x\mathscr{y}}$

For fixed x, $(n_{x1}, \dots n_{x|\Omega_\mathcal{Y}|}) \sim M(n_x, (p_{x1}, \dots, p_{x|\Omega_\mathcal{Y}|}))$ with conditional probabilities $p_{x\mathscr{y}} = P(\mathcal{Y} = \mathscr{y} | X = x)$.

Likelihood for parameters $\boldsymbol{p} = (p_{x1}, \dots, p_{x|\Omega_\mathcal{Y}|-1})^T$
$L(\boldsymbol{p}) \propto \prod_{\mathscr{y} \in \Omega_\mathcal{Y}} p_{0\mathscr{y}}^{n_{0\mathscr{y}}} \prod_{\mathscr{y} \in \Omega_\mathcal{Y}} p_{1\mathscr{y}}^{n_{1\mathscr{y}}}$ is *uniquely* maximized by

$\hat{p}_{x\mathscr{y}} = \frac{n_{x\mathscr{y}}}{n_x}$, for $x \in \{0,\ 1\}$.

and the **invariance of the likelihood** under parameter transformations, i.e.:

$$\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}$$

$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x},\ \frac{\sum_{\mathfrak{y} \ni y} n_{x\mathfrak{y}}}{n_x}\right]$

$\hat{q}_{\mathfrak{y}|xy} \in \left[0,\ \frac{n_{x\mathfrak{y}}}{n_{x\{y\}} + n_{x\mathfrak{y}}}\right]$

Basic idea for the regression case

OBSERVABLE

$\mathcal{Y}$ coarse data

Observation model $\mathcal{Q}$

error-freeness

LATENT

Y latent variable

Main goal:

Estimation of $\pi_{ij} = P(Y_i = j)$

coarsening mechanism

$p_{x\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i | X = x),\ i = 1, \dots, n$

$x \in \{0,\ 1\}$

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

$\pi_{ij} = P(Y_i = j|\mathbf{x}_i) = \frac{\exp(\beta_{j0} + \mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)}$

$\pi_{iK} = \left(1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)\right)^{-1}$

(multinomial logit model)

Use the **connection between $\boldsymbol{p}$ and $\boldsymbol{\gamma}$**

$$\Phi(\boldsymbol{\gamma}) = \mathbf{p}$$

$$\boldsymbol{\gamma} = (\boldsymbol{q}_{\mathscr{y}|\boldsymbol{xy}}^T,\ \boldsymbol{\pi}_{\boldsymbol{y}}^T)^T$$

Use random-set perspective and determine maximum-likelihood estimator $\hat{p}_{x\mathscr{y}}$

For fixed x, $(n_{x1}, \dots n_{x|\Omega_\mathcal{Y}|}) \sim M(n_x, (p_{x1}, \dots, p_{x|\Omega_\mathcal{Y}|}))$ with conditional probabilities $p_{x\mathscr{y}} = P(\mathcal{Y} = \mathscr{y}|X = x)$.

Likelihood for parameters $\boldsymbol{p} = (p_{x1}, \dots, p_{x|\Omega_\mathcal{Y}|-1})^T$

$L(\boldsymbol{p}) \propto \prod_{\mathscr{y} \in \Omega_\mathcal{Y}} p_{0\mathscr{y}}^{n_{0\mathscr{y}}} \prod_{\mathscr{y} \in \Omega_\mathcal{Y}} p_{1\mathscr{y}}^{n_{1\mathscr{y}}}$ is *uniquely* maximized by

$\boxed{\hat{p}_{x\mathscr{y}} = \frac{n_{x\mathscr{y}}}{n_x},}$ for $x \in \{0,\ 1\}$.

and the **invariance of the likelihood** under parameter transformations, i.e.:

$$\hat{\Gamma} = \{\boldsymbol{\gamma} \mid \Phi(\boldsymbol{\gamma}) = \hat{\boldsymbol{p}}\}$$

$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x},\ \frac{\sum_{\mathfrak{y} \ni y} n_{x\mathfrak{y}}}{n_x}\right]$

$\hat{q}_{\mathfrak{y}|xy} \in \left[0,\ \frac{n_{x\mathfrak{y}}}{n_{x\{y\}} + n_{x\mathfrak{y}}}\right]$

Illustration (PASS data)

$\hat{\pi}_{0<} \in [0.41,\ 0.64]$ $\hat{\pi}_{1<} \in [0.10,\ 0.34]$

$\hat{\beta}_{<0} \in [-0.37,\ 0.59]$ $\hat{\beta}_{<} \in [-1.83,\ -1.25]$

Reliable incorporation of auxiliary information

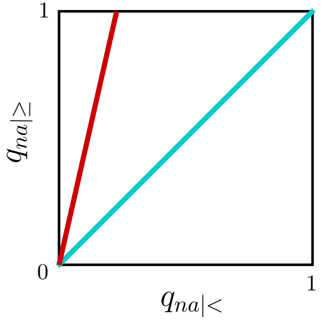
Starting from point-identifying assumptions, we use sensitivity parameters to allow inclusion of partial knowledge.

Assumption about exact value of $R = \frac{q_{na|\geq}}{q_{na|<}}$ (Nordheim, 1984):

e.g. $\mathcal{Q}$ specified by R=1, R=4

where R=1 corresponds to CAR

(Heitjan, Rubin, 1991).



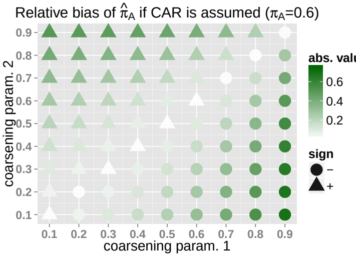
Starting from point-identifying assumptions, we use sensitivity parameters to allow inclusion of partial knowledge.

Assumption about exact value of $R = \frac{q_{na|\geq}}{q_{na|<}}$ (Nordheim, 1984):

e.g. $\mathcal{Q}$ specified by R=1, R=4

where R=1 corresponds to CAR

(Heitjan, Rubin, 1991).



Starting from point-identifying assumptions, we use sensitivity parameters to allow inclusion of partial knowledge.

Assumption about exact value of $R = \frac{q_{na|\geq}}{q_{na|<}}$ (Nordheim, 1984):

e.g. $\mathcal{Q}$ specified by R=1, R=4

where R=1 corresponds to CAR

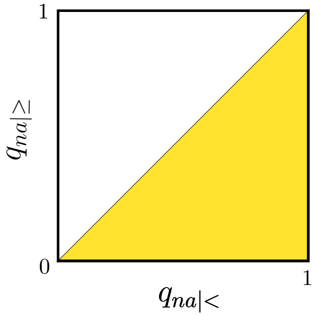
(Heitjan, Rubin, 1991).

Rough evaluation of R:

e.g. $\mathcal{Q}$ specified by R $\leq$ 1:

low income group has a higher tendency

to report "na"



Sensitivity analysis (cf. Vansteelandt, Goetghebeur, Kenward, Molenberghs, 2006)

- $R = \frac{q_{na|\geq}}{q_{na|<}}$: sensitivity parameter $\delta \in \Delta$
- Honestly Estimated Ignorance Region (HEIR) $\hat{ir}(\pi_Y, \Delta)$
  - Example 1: No knowledge about R
    $\Rightarrow \hat{ir}(\pi_<, [0, \infty)) = \left[\frac{n_<}{n_< + n_\geq + n_{na}}, \ \frac{n_< + n_{na}}{n_< + n_\geq + n_{na}} \right]$
  - Example 2: $R \leq 1$
    $\Rightarrow \hat{ir}(\pi_<, [0, 1]) = \left[\frac{n_<}{n_< + n_\geq + n_{na}}, \ \frac{n_<}{n_< + n_\geq} \right]$

How to adopt for uncertainty induced by finite sampling?

Uncertainty regions...

... for the ignorance region?

... for the parameter of interest?

Likelihood

$$L(\pi_<, q_{na|<}, q_{na|\geq} || \mathbf{y}_1, \ldots, \mathbf{y}_n)$$
$$\propto ((1 - q_{na|<}) \cdot \pi_<)^{n_<} \cdot ((1 - q_{na|\geq}) \cdot (1 - \pi_<))^{n_\geq}$$
$$\cdot (q_{na|<} \cdot \pi_A + q_{na|\geq} \cdot (1 - \pi_<))^{n_{na}}$$

Likelihood ratio test:

- $H_0 : \boldsymbol{\gamma} = \gamma_0 \qquad H_1 : \boldsymbol{\gamma} \neq \gamma_0$
- $LQ = 2 log \frac{L(\hat{\gamma}_{ML})}{L(\hat{\gamma}_{0,ML})} = 2 \log(l(\hat{\gamma}_{ML}) - l(\hat{\gamma}_{0,ML})$
- Reject H_0 if $LQ \geq \chi^2_{dim(\gamma)}$

Basic problem, now: Bayesian approach

OBSERVABLE

$\mathcal{Y}$ coarse data

LATENT

Y latent variable

Main goal:

Estimation of

coarsening mechanism

$p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i), \ i = 1, \dots, n$

$q_{\mathscr{y}_i|y_i} = P(\mathcal{Y}_i = \mathscr{y}_i | Y_i = y_i)$

$\pi_{ij} = P(Y_i = j)$

1. Frequentist approach:

   Find maximum-likelihood estimator of $\boldsymbol{\gamma} = (\boldsymbol{q}^T_{\mathscr{y}_i|\boldsymbol{y_i}}, \ \boldsymbol{\pi}^T_{\boldsymbol{y}})^T$

   (i.i.d. case , regression case)

2. Bayesian approach:

   Find posterior mean

   (Next step)

Summary and outlook

- Via the observation model $\mathcal{Q}$ maximum-likelihood estimators referring to the latent variable may be obtained for both cases
  - ... the homogeneous case
  - ... the case with categorical covariates
- Proper inclusion of auxiliary information via further restrictions on $\mathcal{Q}$

Next steps:

- Likelihood-based hypothesis tests and uncertainty regions for coarse categorical data
- Inclusion of auxiliary information via sets of priors
- Consideration of other "deficiency" processes

References

- Couso, I., Dubois, D., Sánchez, L.
  Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables, Springer, 2014.
- Heitjan, D., Rubin, D.
  Ignorability and Coarse Data, Annals of Statistics, 1991.
- Manski, C.
  Partial Identification of Probability Distributions, Springer, 2003.
- E. Nordheim.
  Inference from nonrandomly missing categorical data:An example from a genetic study on Turner's syndrome, J. Am. Stat. Assoc., 1984.
- Vansteelandt, S., Goetghebeur, E., Kenward, M., Molenberghs, G.
  Ignorance and uncertainty regions as inferential tools in a sensitivity analysis, Stat. Sin., 2006.