Likelihood based analyses concerning coarsened categorical data

Julia Plaß

18^{th} of June 2014

Outline

Introduction to the problem of coarse data

Reasons for coarse categorical data:

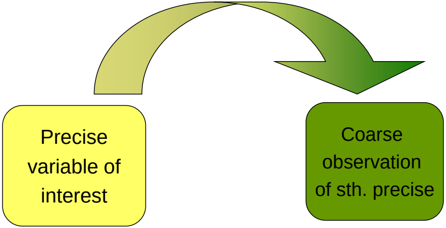
- Guarantee of anonymization, prevention of refusals

  Example:
  "Which kind of party did you elect?"
  ☐ rather left ☐ center ☐ rather right

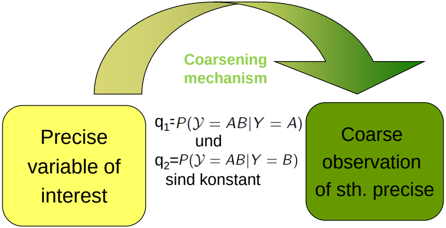
- Different levels of reporting accuracy
  (lack of knowledge, vague question formulation)

  Examples:
  "To which electoral district do you belong to?"
  "Which car do you drive?"

Initial situation (iid case)



Initial situation (iid case)



Y	Y_coarse
A	A
B	B
B	AB
A	AB

10 000 Obs.

100 Datasets

The general log-likelihood

Likelihood :

$$
\begin{aligned}
L(\pi_A, q_1, q_2) &= \prod_{\mathcal{Y}_i} P(\mathcal{Y} = \mathfrak{y}) \\
&= \prod_{i:\mathcal{Y}_i = A} \underbrace{P(\mathcal{Y} = A | Y = A)}_{(1-q_1)} \pi_{iA} \prod_{i:\mathcal{Y}_i = B} \underbrace{P(\mathcal{Y} = B | Y = B)}_{(1-q_2)} (1 - \pi_{iA}) \\
&\quad \prod_{i:\mathcal{Y}_i = AB} \underbrace{P(\mathcal{Y} = AB | Y = A)}_{q_1} \pi_{iA} + \underbrace{P(\mathcal{Y} = AB | Y = B)}_{q_2} (1 - \pi_{iA})
\end{aligned}
$$

log-Likelihood under the iid assumption :

$$
\begin{aligned}
l(\pi_A, q_1, q_2) &= n_A \cdot [ln(1 - q_1) + ln(\pi_A)] + n_B \cdot [ln(1 - q_2) + ln(1 - \pi_A)] \\
&+ n_{AB} \cdot [q_1 \pi_A + q_2(1 - \pi_A))]
\end{aligned}
$$

The general log-likelihood

FOC :

$$
\begin{aligned}
\text{I.)}\ \frac{\partial}{\partial \pi_A} &= \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}(q_1 - q_2) + \frac{n_A}{\pi_A} - \frac{n_B}{1-\pi_A} \overset{!}{=} 0 \\
\text{II.)}\ \frac{\partial}{\partial q_1} &= \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}\pi_A - \frac{n_A}{1-q_1} \overset{!}{=} 0 \\
\text{III.)}\ \frac{\partial}{\partial q_2} &= \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}(1-\pi_A) - \frac{n_B}{1-q_2} \overset{!}{=} 0
\end{aligned}
$$

Neccessary and sufficient solutions:

Estimators $(\hat{\pi}_A, \hat{q}_1, \hat{q}_2)$ are solutions of the estimation problem if and only if

$$\frac{n_{AB}}{n} = \hat{q}_1 \cdot \hat{\pi}_A + \hat{q}_2 \cdot (1 - \hat{\pi}_A)$$

is fulfilled with $\hat{\pi}_A$, $\hat{q}_1$ and $\hat{q}_1 > 0$ and < 1.

Distinguishing different cases

Estimation of parameter of interest ....

... implying point-identifying assumptions

- known coarsening mechanism
- $q_1 = q_2$: data are coarsened at random (CAR)

$$\hat{\pi}_A = \frac{n_A}{n_A + n_B}$$
$$\hat{q}_1 = \hat{q}_2 = \frac{n_{AB}}{n_A + n_B + n_{AB}}$$

- relation between coarsening parameters $R = \frac{q_1}{q_2}$ is known
  $\Rightarrow$ Generalization of CAR

... without any assumptions

$\Rightarrow$ Find lower and upper bounds of parameter estimators

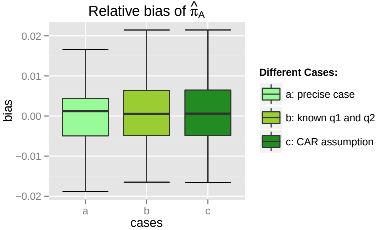
Implying assumptions: CAR

Evaluation of $\hat{\pi}_A$

by means of comparing relative bias

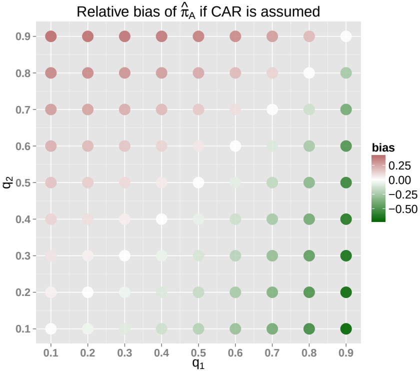
$$Bias_{\text{rel}} = \frac{\hat{\pi}_A - \pi_A}{|\pi_A|}$$

in the following three situations:



Analysis if CAR is wrongly assumed

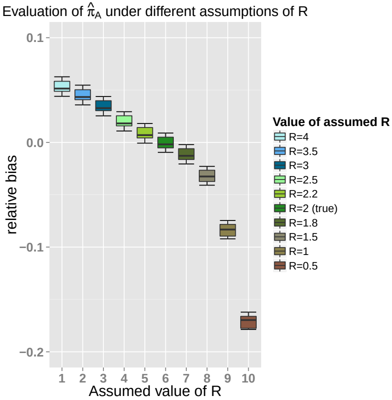
Median relative bias of $\hat{\pi}_A$ for different combinations of true q_1- and q_2 values is considered:

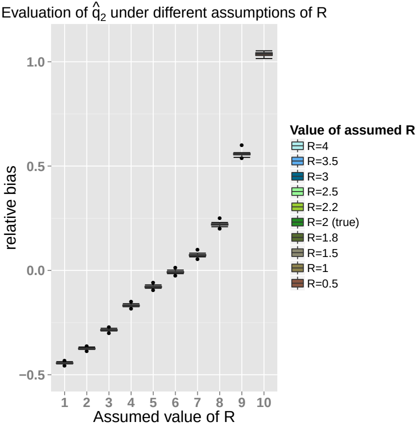


Implying assumptions: relation R (two true categories)

Regarded situation:

$q_1 = 0.3,\ q_2 = 0.15 \Rightarrow R_{\text{true}} = \frac{q_1}{q_2} = 2$



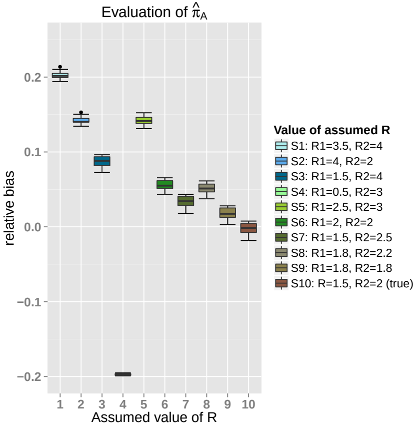


Implying assumptions: relation R (three true categories)

Now three true categories $(A,\ B,\ C)$ and coarse categories AB and AC

$$q_{AB|A} = 0.4,\ q_{AB|B} = 0.2\overline{6} \Rightarrow R_{1,\text{true}} = \frac{q_{AB|A}}{q_{AB|B}} = 1.5$$

$$q_{AC|A} = 0.4,\ q_{AC|C} = 0.2 \Rightarrow R_{2,\text{true}} = \frac{q_{AC|A}}{q_{AC|C}} = 2$$



Different assumptions of R:

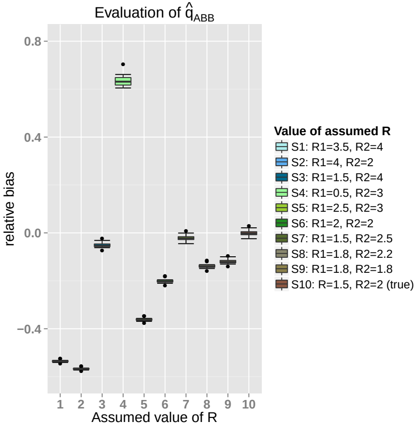
- large...
  ... deviation of R1 & R2 (S1)
  ... deviation of R1 only (S2)
  ... deviationof R2 only (S3)
- medium...
  ... reverse deviation of R1 & R2 (S4)
  ... concordant deviation of R1 & R2 (S5)
  ... deviation of R1 only (S6)
  ... deviation of R2 only (S7)
- small...
  ... concordant deviation of R1 & R2 (S8)
  ... reverse deviation of R1 & R2 (S9)
- true assumptions (S10)

Implying assumptions: relation R (three true categories)

Now three true categories $(A, \ B, \ C)$ and coarse categories AB and AC

$q_{AB|A} = 0.4, \ q_{AB|B} = 0.2\overline{6} \Rightarrow R_{1,\text{true}} = \frac{q_{AB|A}}{q_{AB|B}} = 1.5$

$q_{AC|A} = 0.4, \ q_{AC|C} = 0.2 \Rightarrow R_{2,\text{true}} = \frac{q_{AC|A}}{q_{AC|C}} = 2$



Different assumptions of R:

- large...
  ... deviation of R1 & R2 (S1)
  ... deviation of R1 only (S2)
  ... deviationof R2 only (S3)
- medium...
  ... reverse deviation of R1 & R2 (S4)
  ... concordant deviation of R1 & R2 (S5)
  ... deviation of R1 only (S6)
  ... deviation of R2 only (S7)
- small...
  ... concordant deviation of R1 & R2 (S8)
  ... reverse deviation of R1 & R2 (S9)
- true assumptions (S10)

Implying no assumptions: Approaches

Main goal:

Find bounds for π_A

Different approaches:

- Take Dempster-Shafer estimators: $\underline{\hat{\pi}_A} = \frac{n_A}{n}$ and $\overline{\hat{\pi}_A} = \frac{n_A + n_{AB}}{n}$
- Consider $l(\pi_A, \overline{\hat{q}_1}, \underline{\hat{q}_2})$ ($\Rightarrow \overline{\hat{\pi}_A}$) and $l(\pi_A, \underline{\hat{q}_1}, \overline{\hat{q}_2})$ ($\Rightarrow \underline{\hat{\pi}_A}$),

$$\text{where} \qquad \hat{q}_1 = \frac{n_{AB}}{n_{AB} + n_A}, \quad \hat{q}_2 = 0$$

$$\text{and} \qquad \hat{q}_1 = 0, \quad \hat{q}_2 = \frac{n_{AB}}{n_B + n_{AB}} \quad \text{resp.}$$

- Solution by optimization problem

The empirical evidence only: Optimization problems

Optimization problem considering $l(\pi_A, q_1, q_2)$

Original problem:

$$\pi_A \to \max$$
$$\pi_A \to \min$$

Constraints:

$$\begin{array}{ccccc} & S(\pi_A, q_1, q_2) & & = & \mathbf{0} \\ 0 & \leq & \pi_A & \leq & 1 \\ 0 & \leq & q_1 & \leq & 1 \\ 0 & \leq & q_2 & \leq & 1 \end{array}$$

Rearranged problem:

$$\left.\begin{array}{l} \pi_A \to \max \\ \pi_A \to \min \\ S(\pi_A, q_1, q_2) = \mathbf{0} \end{array}\right\} \begin{array}{l} \pi_A - \lambda \cdot S(\pi_A, q_1, q_2)^2 \to \max \\ \pi_A + \lambda \cdot S(\pi_A, q_1, q_2)^2 \to \min \end{array}$$

Constraints:

$$\begin{array}{cccc} 0 & \leq & \pi_A & \leq 1 \\ 0 & \leq & q_1 & \leq 1 \\ 0 & \leq & q_2 & \leq 1 \end{array}$$

$\Rightarrow$ One can obtain the Dempster-Shafer estimators

The empirical evidence only: Optimization problems

Optimization problem considering $L(\pi_A)$

- Regarding the Likelihood $L(\pi_A)$:

$$L(\pi_A) = (\pi_A)^{\sum_{i=1}^n y_{iA}} \cdot (1-\pi_A)^{\sum_{i=1}^n y_{iB}}$$
$$\text{where} \quad y_{iA} \text{ and } y_{iB} \in \{0,1\}$$

- Corresponding Scorefunction

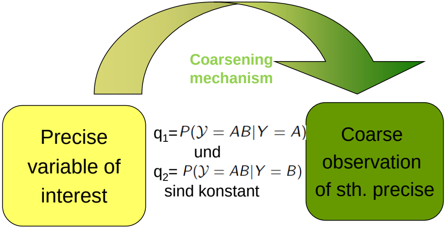
$$\begin{aligned} S(\pi_A) &= \frac{\sum_{i=1}^n y_{iA}}{\pi_A} - \frac{\sum_{i=1}^n y_{iB}}{1-\pi_A} \\ &= \frac{n_A + n_{AB|A}}{\pi_A} - \frac{n_B + n_{AB} - n_{AB|A}}{1-\pi_A} \end{aligned}$$

- Possible optimization problem:

$$\begin{aligned} \text{Objective function:} \quad & \pi_A - \lambda \cdot (S(\pi_A))^2 \to \max \\ & \pi_A + \lambda \cdot (S(\pi_A))^2 \to \min \\ \text{Constraints:} \quad & 0 \leq \pi_A \leq 1 \\ & 0 \leq n_{AB|A} \leq n_{AB} \end{aligned}$$

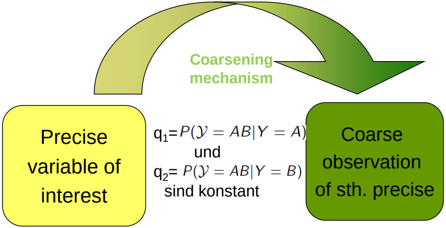
$\Rightarrow$ One can obtain the Dempster-Shafer estimators

Initial situation (with covariates)



Y	Y_coarse
A	A
B	B
B	AB
A	AB

Initial situation (with covariates)



Y	Y_coarse	X
A	A	1
B	B	1
B	AB	0
A	AB	1

10 000 Obs.

100 Datasets

The general likelihood (with covariates)

Now:

π_{iA} are dependent on the values $\mathbf{x}_i$ by

$$\pi_{iA} = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_A)}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta}_A)}$$

Resulting log-likelihood:

$$\begin{aligned} l(\boldsymbol{\beta}_A, q_1, q_2) \quad = \quad & \sum_{i=1}^{N_1} \ln\left((1-q_1)\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_A)}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_A)}\right) + \sum_{i=N_1+1}^{N_2} \ln\left((1-q_2)\frac{1}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_A)}\right) + \\ & \sum_{i=N_2+1}^{N} \ln\left(q_1\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_A)}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_A)} + \frac{q_2}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_A)}\right) \end{aligned}$$

Addressed cases:

- implying CAR-assumption
- general investigation

Implying CAR

Addressed situation:

model with two covariates

assumption of CAR: $q_1 = q_2$

parameters of main interest: $\boldsymbol{\beta}$

- Evaluation of $\hat{\boldsymbol{\beta}}$ by means of the relative bias

$$Bias_{\text{rel}} = \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{|\boldsymbol{\beta}|}$$

if CAR is involved into the estimation

Implying CAR

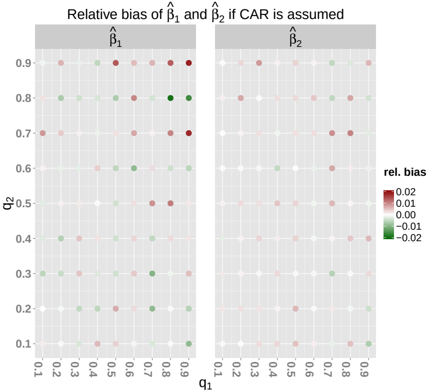
Addressed situation:
model with two covariates
assumption of CAR: $q_1 = q_2$
parameters of main interest: $\boldsymbol{\beta}$

- Evaluation of $\hat{\boldsymbol{\beta}}$ by means of the relative bias

$$Bias_{\text{rel}} = \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{|\boldsymbol{\beta}|}.$$

if CAR is involved into the estimation

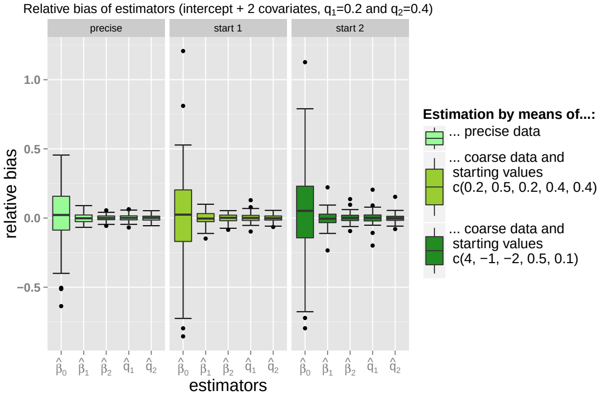
- Nearly unbiased estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ in all situations that have been considered



No assumptions (with covariates)

Addressed model:

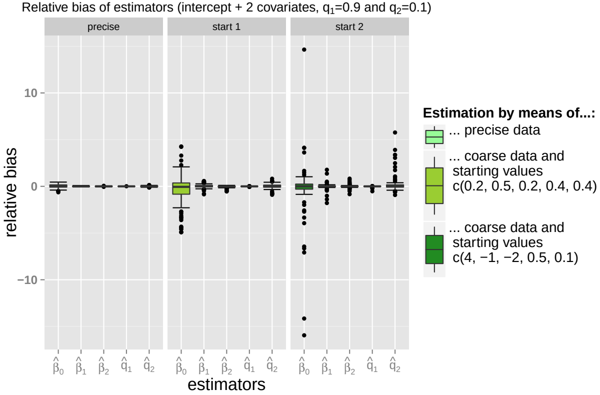
Two covariates - $X_1 \sim Po(3)$ and $X_2 \sim N(mean = 0, sd = 2)$



No assumptions (with covariates)

Addressed model:

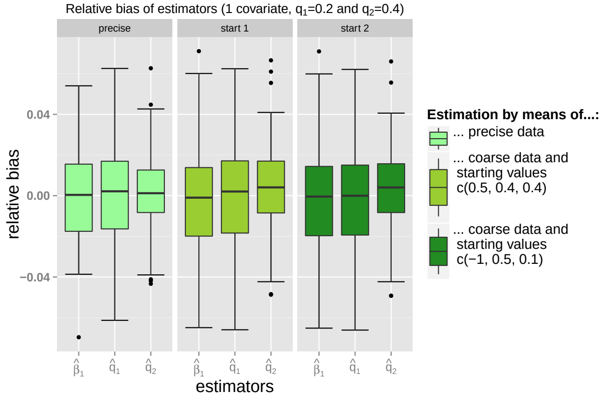
Two covariates - $X_1 \sim Po(3)$ and $X_2 \sim N(mean = 0, sd = 2)$



No assumptions (with covariates)

Addressed model:

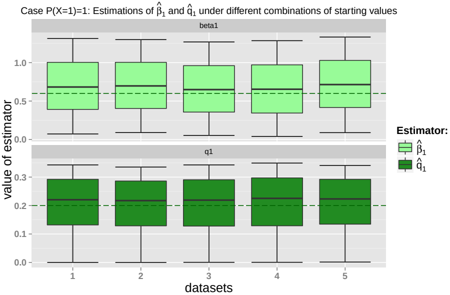
One covariates - $X_1 \sim N(mean = 3, sd = 7)$



Limiting case

Investigation of the transition from iid-case to case with one covariate

Cases:
P(X=1)=1
P(X=1)=0.99
P(X=1)=0.9



- Calculate corresponding $\hat{\pi}_{iA}$ for $X_i = 1$ by

$$\hat{\pi}_{iA} = \frac{\exp(x_i \hat{\beta}_A)}{1 + \exp(x_i \hat{\beta}_A)}$$

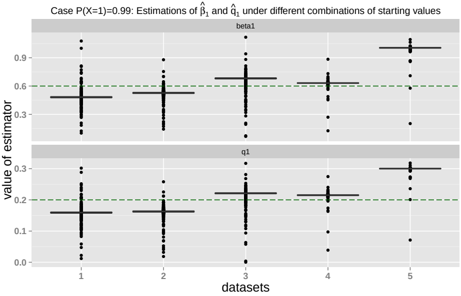
- Compare...
  - ... $\min(\hat{\pi}_A)$ and $\max(\hat{\pi}_A)$ with bounds from DST interval
  - ... $\min(\hat{q}_1)$ and $\max(\hat{q}_2)$ with $\underline{\hat{q}_1} = 0$ and $\overline{\hat{q}_1} = \frac{n_{AB}}{n_{AB}+n_A}$

$\Rightarrow$ nearly same values in both cases

Limiting case

Investigation of the transition from iid-case to case with one covariate

Cases:
P(X=1)=1
P(X=1)=0.99
P(X=1)=0.9



Result of comparison:

Results for $\hat{\pi}_A$ ($\hat{\pi}_{A,prec} \approx 0.63$):

Data	DST interval	bounds from estimation
1	[0.5188, 0.7885]	[0.5254, 0.7459] (0.64)
2	[0.5253, 0.7879]	[0.5358, 0.7066] (0.65)
3	[0.5161, 0.7876]	[0.5166, 0.7535] (0.65)
4	[0.5105 0.7872]	[0.5315, 0.7076] (0.64)
5	[0.5104, 0.7887]	[0.5505, 0.7489] (0.64)

Results for $\hat{q}_1$ ($\hat{q}_{1,prec} \approx 0.20$):

Data	DST interval	bounds from estimation
1	[0, 0.3420]	[0.0120, 0.3017] (0.20)
2	[0, 0.3332]	[0.0186, 0.2580] (0.19)
3	[0, 0.3447]	[0.0001, 0.3170] (0.20)
4	[0, 0.3515]	[0.0388, 0.2747] (0.21)
5	[0, 0.3529]	[0.0712, 0.3181] (0.21)

Limiting case

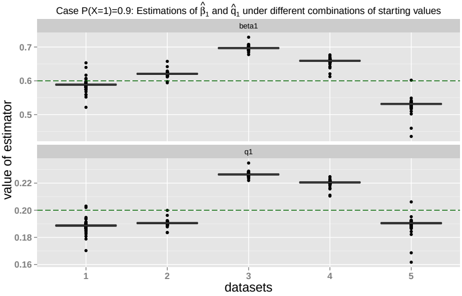
Investigation of the transition from iid-case to case with one covariate

Cases:

P(X=1)=1

P(X=1)=0.99

P(X=1)=0.9



Result of comparison:

Results for $\hat{\pi}_A$ ($\hat{\pi}_{A,prec} \approx 0.63$):

Data	DST interval	bounds from estimation
1	[0.5104, 0.7828]	[0.6275, 0.6577] (0.63)
2	[0.5141, 0.7795]	[0.6443, 0.6587] (0.63)
3	[0.5035, 0.7775]	[0.6632, 0.6746] (0.63)
4	[0.5009, 0.7800]	[0.6485, 0.7076] (0.63)
5	[0.4985, 0.7793]	[0.6072, 0.6461] (0.63)

Results for $\hat{q}_1$ ($\hat{q}_{1,prec} \approx 0.20$):

Data	DST interval	bounds from estimation
1	[0, 0.3480]	[0.1702, 0.2030] (0.19)
2	[0, 0.3405]	[0.1835, 0.1999] (0.19)
3	[0, 0.3524]	[0.2219, 0.2348] (0.20)
4	[0, 0.3578]	[0.2105, 0.2247] (0.21)
5	[0, 0.3603]	[0.1617, 0.2061] (0.21)

Summary and outlook

Summary

- In case of *iid* variables
  - ... generally a set of estimators results whose bounds can be obtained by nonlinear optimization
  - ... using correctly the assumptions of *CAR* leads to identified and nearly unbiased estimators
- In case of incorporating covariates
  - ... generally identified and nearly unbiased estimators seem to result
  - ... using CAR even if it is not valid at least results in nearly unbiased estimators for $\hat{\beta}_1$ and $\hat{\beta}_2$ (in model with 2 covariates)

Outlook

- Further investigation of transition from non-identifiable to identifiable model
- Think about reasons for identifiability in case of involving covariates
- Including other true coarsening mechanisms