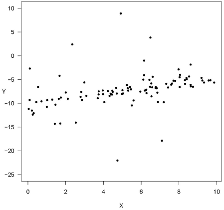
Robust Regression with Coarse Data

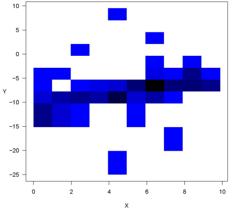
Marco Cattaneo and Andrea Wiencierz
Department of Statistics, LMU Munich

Statistische Woche 2011, Leipzig, Germany
21 September 2011

coarse data

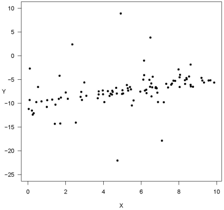


unobserved precise data

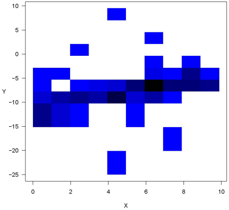


observed coarse data

coarse data



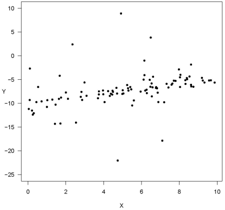
unobserved precise data



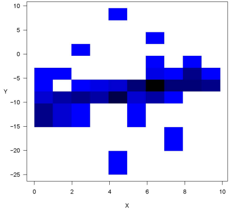
observed coarse data

- in the literature, two kinds of general approaches to regression with coarse data:

coarse data



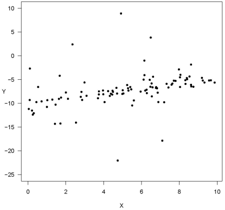
unobserved precise data



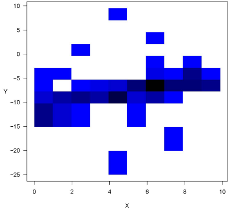
observed coarse data

- in the literature, two kinds of general approaches to regression with coarse data:
  - represent the observed coarse data by *few precise values* (e.g., intervals by center and width), and apply standard regression methods to those values: see for instance Domingues et al. (2010)

coarse data



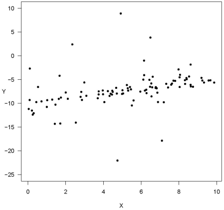
unobserved precise data



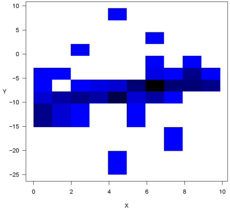
observed coarse data

- in the literature, two kinds of general approaches to regression with coarse data:
  - represent the observed coarse data by *few precise values* (e.g., intervals by center and width), and apply standard regression methods to those values: see for instance Domingues et al. (2010)
  - apply standard regression methods to *all possible precise data* compatible with the observed coarse data, and consider the range of outcomes as the imprecise result: see for example Ferson et al. (2007)

coarse data



unobserved precise data



observed coarse data

- in the literature, two kinds of general approaches to regression with coarse data:
  - represent the observed coarse data by *few precise values* (e.g., intervals by center and width), and apply standard regression methods to those values: see for instance Domingues et al. (2010)
  - apply standard regression methods to *all possible precise data* compatible with the observed coarse data, and consider the range of outcomes as the imprecise result: see for example Ferson et al. (2007)
- LIR (Likelihood-based Imprecise Regression): new regression method directly applicable to coarse data (Cattaneo and Wiencierz, 2011)

nonparametric likelihood

- precise data (unobserved): random variables $V_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$

nonparametric likelihood

- precise data (unobserved): random variables $V_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$
- coarse data (observed): random sets $V_i^* \subseteq \mathcal{X} \times \mathbb{R}$

nonparametric likelihood

- precise data (unobserved): random variables $V_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$
- coarse data (observed): random sets $V_i^* \subseteq \mathcal{X} \times \mathbb{R}$
- nonparametric model: $\mathcal{P}$ is the set of all probability measures such that

nonparametric likelihood

- precise data (unobserved): random variables $V_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$
- coarse data (observed): random sets $V_i^* \subseteq \mathcal{X} \times \mathbb{R}$
- nonparametric model: $\mathcal{P}$ is the set of all probability measures such that
  - $(V_1, V_1^*), \ldots, (V_n, V_n^*)$ i.i.d.

nonparametric likelihood

- precise data (unobserved): random variables $V_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$
- coarse data (observed): random sets $V_i^* \subseteq \mathcal{X} \times \mathbb{R}$
- nonparametric model: $\mathcal{P}$ is the set of all probability measures such that
  - $(V_1, V_1^*), \ldots, (V_n, V_n^*)$ i.i.d.
  - $P(V_i \in V_i^*) \geq 1 - \varepsilon$ (where $\varepsilon \in [0, 1]$ is fixed)

nonparametric likelihood

- precise data (unobserved): random variables $V_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$
- coarse data (observed): random sets $V_i^* \subseteq \mathcal{X} \times \mathbb{R}$
- nonparametric model: $\mathcal{P}$ is the set of all probability measures such that
  - $(V_1, V_1^*), \ldots, (V_n, V_n^*)$ i.i.d.
  - $P(V_i \in V_i^*) \geq 1 - \varepsilon$ (where $\varepsilon \in [0, 1]$ is fixed)
- the observed (coarse) data $V_1^* = A_1, \ldots, V_n^* = A_n$ induce the (normalized) likelihood function $lik : \mathcal{P} \to [0, 1]$ with

$$lik(P) = \frac{P(V_1^* = A_1, \ldots, V_n^* = A_n)}{\max_{P' \in \mathcal{P}} P'(V_1^* = A_1, \ldots, V_n^* = A_n)}$$

regression problem

- regression functions: $\mathcal{F}$ is a certain set of functions $f : \mathcal{X} \to \mathbb{R}$

regression problem

- regression functions: $\mathcal{F}$ is a certain set of functions $f : \mathcal{X} \to \mathbb{R}$
- absolute residuals: $R_{f,i} = |Y_i - f(X_i)|$

regression problem

- regression functions: $\mathcal{F}$ is a certain set of functions $f : \mathcal{X} \to \mathbb{R}$
- absolute residuals: $R_{f,i} = |Y_i - f(X_i)|$
- for each function $f \in \mathcal{F}$, the quantiles of the distribution of the absolute residuals $R_{f,i}$ can be estimated even under the nonparametric model $\mathcal{P}$

regression problem

- regression functions: $\mathcal{F}$ is a certain set of functions $f : \mathcal{X} \to \mathbb{R}$
- absolute residuals: $R_{f,i} = |Y_i - f(X_i)|$
- for each function $f \in \mathcal{F}$, the quantiles of the distribution of the absolute residuals $R_{f,i}$ can be estimated even under the nonparametric model $\mathcal{P}$
- the regression problem can be interpreted as the *minimization of the p-quantile* of the distribution of the absolute residuals $R_{f,i}$ (where $p \in (0, 1)$ is fixed)

generalized LQS regression

- likelihood-based confidence interval for the p-quantile of the distribution of the absolute residuals $R_{f,i}$ (where $Q_f(P)$ is the interval of all p-quantiles of $R_{f,i}$ under P, and $\beta \in (0,1)$ is fixed):

$$\mathcal{C}_f = \bigcup_{P \in \mathcal{P} \,:\, lik(P) > \beta} Q_f(P)$$

generalized LQS regression

- likelihood-based confidence interval for the p-quantile of the distribution of the absolute residuals $R_{f,i}$ (where $Q_f(P)$ is the interval of all p-quantiles of $R_{f,i}$ under P, and $\beta \in (0, 1)$ is fixed):

$$\mathcal{C}_f = \bigcup_{P \in \mathcal{P} \,:\, lik(P) > \beta} Q_f(P)$$

- *point estimate*: f_{LRM} is the function in $\mathcal{F}$ minimizing $\sup \mathcal{C}_f$ (Likelihood-based Region Minimax: see Cattaneo, 2007)

generalized LQS regression

- likelihood-based confidence interval for the p-quantile of the distribution of the absolute residuals $R_{f,i}$ (where $Q_f(P)$ is the interval of all p-quantiles of $R_{f,i}$ under P, and $\beta \in (0,1)$ is fixed):

$$\mathcal{C}_f = \bigcup_{P \in \mathcal{P} \,:\, lik(P) > \beta} Q_f(P)$$

- *point estimate*: f_{LRM} is the function in $\mathcal{F}$ minimizing $\sup \mathcal{C}_f$ (Likelihood-based Region Minimax: see Cattaneo, 2007)
- f_{LRM} has a simple geometrical interpretation: $\overline{B}_{f_{LRM}, \overline{q}_{LRM}}$ is the thinnest band of the form $\overline{B}_{f,q} = \{(x,y) \in \mathcal{X} \times \mathbb{R} : |y - f(x)| \leq q\}$ *containing* at least $\overline{k}$ coarse data (where $\overline{k} > (p + \varepsilon)\, n$ depends on n, ε, p, β), for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$

generalized LQS regression

- likelihood-based confidence interval for the p-quantile of the distribution of the absolute residuals $R_{f,i}$ (where $Q_f(P)$ is the interval of all p-quantiles of $R_{f,i}$ under P, and $\beta \in (0,1)$ is fixed):

$$\mathcal{C}_f = \bigcup_{P \in \mathcal{P} \,:\, lik(P) > \beta} Q_f(P)$$

- *point estimate*: f_{LRM} is the function in $\mathcal{F}$ minimizing $\sup \mathcal{C}_f$ (Likelihood-based Region Minimax: see Cattaneo, 2007)
- f_{LRM} has a simple geometrical interpretation: $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ is the thinnest band of the form $\overline{B}_{f,q} = \{(\underline{x}, y) \in \mathcal{X} \times \mathbb{R} : |y - f(x)| \leq q\}$ *containing* at least $\overline{k}$ coarse data (where $\overline{k} > (p + \varepsilon)\, n$ depends on n, ε, p, β), for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$
- when the observed data are in fact precise, f_{LRM} corresponds to the LQS (Least Quantile of Squares) estimate with quantile $\frac{\overline{k}}{n}$

generalized LQS regression

- likelihood-based confidence interval for the p-quantile of the distribution of the absolute residuals $R_{f,i}$ (where $Q_f(P)$ is the interval of all p-quantiles of $R_{f,i}$ under P, and $\beta \in (0,1)$ is fixed):

$$\mathcal{C}_f = \bigcup_{P \in \mathcal{P} \,:\, lik(P) > \beta} Q_f(P)$$

- *point estimate*: f_{LRM} is the function in $\mathcal{F}$ minimizing $\sup \mathcal{C}_f$ (Likelihood-based Region Minimax: see Cattaneo, 2007)
- f_{LRM} has a simple geometrical interpretation: $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ is the thinnest band of the form $\overline{B}_{f,q} = \{(x,y) \in \mathcal{X} \times \mathbb{R} : |y - f(x)| \leq q\}$ *containing* at least $\overline{k}$ coarse data (where $\overline{k} > (p + \varepsilon)\, n$ depends on n, ε, p, β), for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$
- when the observed data are in fact precise, f_{LRM} corresponds to the LQS (Least Quantile of Squares) estimate with quantile $\frac{\overline{k}}{n}$
- in the case of linear regression with interval data, f_{LRM} can be computed by generalizing the algorithm of Rousseeuw and Leroy (1987)

nonparametric LIR

- interval dominance: interval I (strictly) dominates interval J iff $x < y$ for all $x \in I$ and all $y \in J$

nonparametric LIR

- interval dominance: interval I (strictly) dominates interval J iff $x < y$ for all $x \in I$ and all $y \in J$
- imprecise regression: set of *all undominated functions* (that is, all $f \in \mathcal{F}$ such that $\overline{q}_{LRM} \in \mathcal{C}_f$)

nonparametric LIR

- interval dominance: interval I (strictly) dominates interval J iff $x < y$ for all $x \in I$ and all $y \in J$
- imprecise regression: set of *all undominated functions* (that is, all $f \in \mathcal{F}$ such that $\overline{q}_{LRM} \in \mathcal{C}_f$)
- the undominated functions have a simple geometrical interpretation: f is undominated iff $\overline{B}_{f,\overline{q}_{LRM}}$ *intersects* at least $\underline{k} + 1$ coarse data (where $\underline{k} < (p - \varepsilon)\, n$ depends on n, ε, p, β)

nonparametric LIR

- interval dominance: interval I (strictly) dominates interval J iff $x < y$ for all $x \in I$ and all $y \in J$

- imprecise regression: set of *all undominated functions* (that is, all $f \in \mathcal{F}$ such that $\overline{q}_{LRM} \in \mathcal{C}_f$)

- the undominated functions have a simple geometrical interpretation: f is undominated iff $\overline{B}_{f,\overline{q}_{LRM}}$ *intersects* at least $\underline{k} + 1$ coarse data (where $\underline{k} < (p - \varepsilon)\, n$ depends on n, ε, p, β)

- complex uncertainty, consisting of two kinds of uncertainty:

nonparametric LIR

- interval dominance: interval I (strictly) dominates interval J iff $x < y$ for all $x \in I$ and all $y \in J$

- imprecise regression: set of *all undominated functions* (that is, all $f \in \mathcal{F}$ such that $\overline{q}_{LRM} \in \mathcal{C}_f$)

- the undominated functions have a simple geometrical interpretation: f is undominated iff $\overline{B}_{f,\overline{q}_{LRM}}$ *intersects* at least $\underline{k} + 1$ coarse data (where $\underline{k} < (p - \varepsilon)\, n$ depends on n, ε, p, β)

- complex uncertainty, consisting of two kinds of uncertainty:
  - *sample uncertainty*: decreases when n increases (reflected by the spread between $\frac{\underline{k}+1}{n}$ and $\frac{\overline{k}}{n}$)

nonparametric LIR

- interval dominance: interval I (strictly) dominates interval J iff $x < y$ for all $x \in I$ and all $y \in J$

- imprecise regression: set of *all undominated functions* (that is, all $f \in \mathcal{F}$ such that $\overline{q}_{LRM} \in \mathcal{C}_f$)

- the undominated functions have a simple geometrical interpretation: f is undominated iff $\overline{B}_{f,\overline{q}_{LRM}}$ *intersects* at least $\underline{k} + 1$ coarse data (where $\underline{k} < (p - \varepsilon)\, n$ depends on n, ε, p, β)

- complex uncertainty, consisting of two kinds of uncertainty:
  - *sample uncertainty*: decreases when n increases (reflected by the spread between $\frac{\underline{k}+1}{n}$ and $\frac{\overline{k}}{n}$)
  - *coarseness uncertainty*: unavoidable under such weak assumptions (reflected by the difference between *containing* and *intersecting* coarse data)

example with social survey data

- data from the "ALLBUS — German General Social Survey" of 2008 (provided by GESIS — Leibniz Institute for the Social Sciences)

example with social survey data

- data from the "ALLBUS — German General Social Survey" of 2008 (provided by GESIS — Leibniz Institute for the Social Sciences)
- relationship between age $X_i \in \mathcal{X} = [18, 100)$ and personal income (on average per month) $Y_i \in [0, +\infty)$, with $n = 3247$

example with social survey data

- data from the "ALLBUS — German General Social Survey" of 2008 (provided by GESIS — Leibniz Institute for the Social Sciences)
- relationship between age $X_i \in \mathcal{X} = [18, 100)$ and personal income (on average per month) $Y_i \in [0, +\infty)$, with $n = 3247$
- choice of regression functions: $\mathcal{F} = \{f_{a,b_1,b_2} : a, b_1, b_2 \in \mathbb{R}\}$ is the set of all quadratic functions $f_{a,b_1,b_2}(x) = a + b_1\, x + b_2\, x^2$

example with social survey data

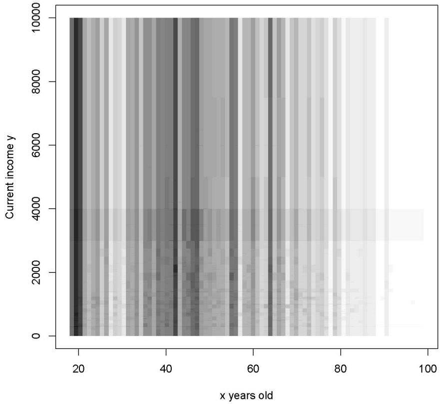
- data from the "ALLBUS — German General Social Survey" of 2008 (provided by GESIS — Leibniz Institute for the Social Sciences)
- relationship between age $X_i \in \mathcal{X} = [18, 100)$ and personal income (on average per month) $Y_i \in [0, +\infty)$, with $n = 3247$
- choice of regression functions: $\mathcal{F} = \{f_{a,b_1,b_2} : a, b_1, b_2 \in \mathbb{R}\}$ is the set of all quadratic functions $f_{a,b_1,b_2}(x) = a + b_1\, x + b_2\, x^2$
- choice of parameters: $\varepsilon = 0$ (no error in the coarsening process), $p = 0.5$ (median), and $\beta = 0.15$ (each $\mathcal{C}_f$ is approximately a conservative 95% confidence interval), implying $\underline{k} = 1568$ and $\overline{k} = 1679$

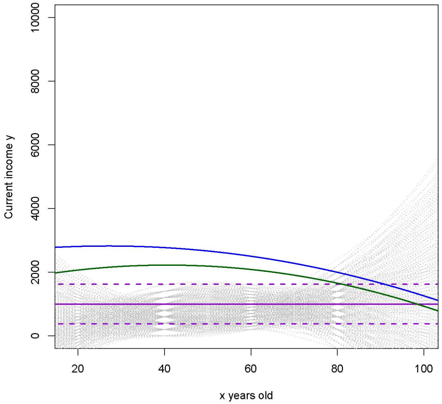
example with social survey data

- data from the "ALLBUS — German General Social Survey" of 2008 (provided by GESIS — Leibniz Institute for the Social Sciences)
- relationship between age $X_i \in \mathcal{X} = [18, 100)$ and personal income (on average per month) $Y_i \in [0, +\infty)$, with $n = 3247$
- choice of regression functions: $\mathcal{F} = \{f_{a,b_1,b_2} : a, b_1, b_2 \in \mathbb{R}\}$ is the set of all quadratic functions $f_{a,b_1,b_2}(x) = a + b_1\,x + b_2\,x^2$
- choice of parameters: $\varepsilon = 0$ (no error in the coarsening process), $p = 0.5$ (median), and $\beta = 0.15$ (each $\mathcal{C}_f$ is approximately a conservative 95% confidence interval), implying $\underline{k} = 1568$ and $\overline{k} = 1679$
- in 4 different data situations, f_{LRM} (violet solid line, with $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ represented by the violet dashed lines) and the undominated functions (gray dotted curves) are compared with the results of the ordinary least squares regression applied after reducing the interval data to their centers and choosing 15 000 (blue curve) or 10 000 (green curve) as the upper income limit

original data

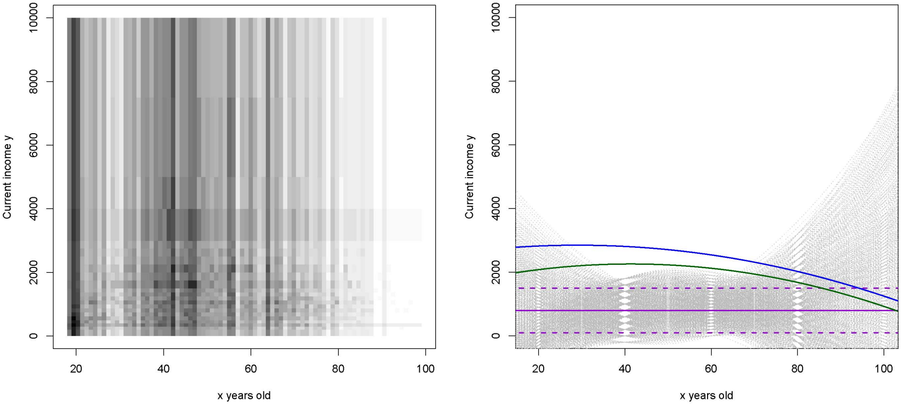
- age data: 3236 "precise" (in years: 83 classes), 11 missing
- income data: 2266 precise, 361 categorized (22 classes), 620 missing





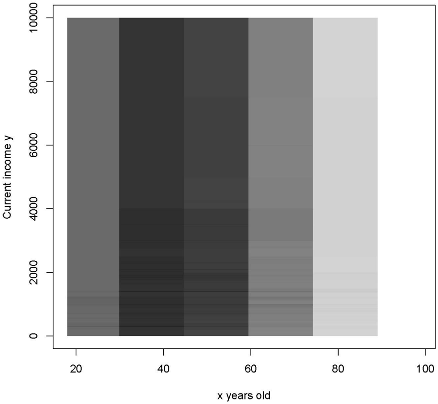
categorized income data

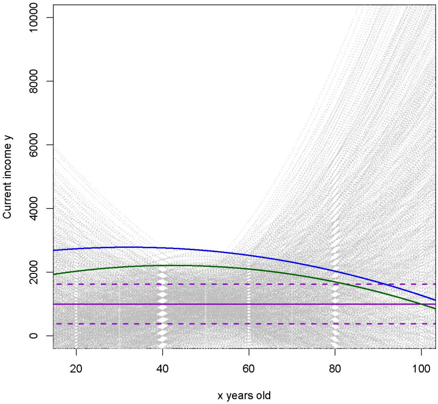
- age data: 3236 "precise" (in years: 83 classes), 11 missing
- income data: 2627 categorized (22 classes), 620 missing



categorized age data

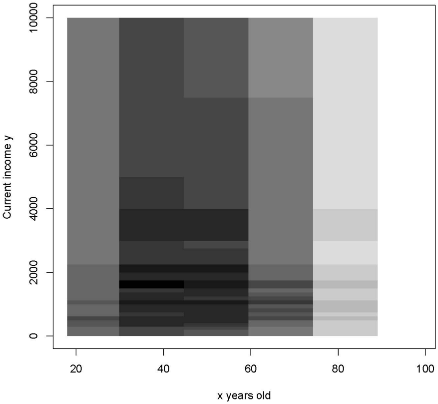
- age data: 3236 categorized (6 classes), 11 missing
- income data: 2266 precise, 361 categorized (22 classes), 620 missing

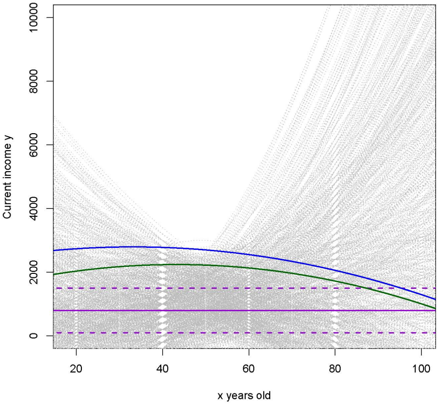




categorized age and income data

- age data: 3236 categorized (6 classes), 11 missing
- income data: 2627 categorized (22 classes), 620 missing





conclusion and outlook

- LIR: new line of approach to regression with coarse data

conclusion and outlook

- LIR: new line of approach to regression with coarse data
- LIR is directly applicable to any kind of coarse data (with precise data as a special case), where the coarsening process can be informative (and even wrong with a certain probability)

conclusion and outlook

- LIR: new line of approach to regression with coarse data
- LIR is directly applicable to any kind of coarse data (with precise data as a special case), where the coarsening process can be informative (and even wrong with a certain probability)
- the result of the regression is imprecise, reflecting the total uncertainty in the data

conclusion and outlook

- LIR: new line of approach to regression with coarse data
- LIR is directly applicable to any kind of coarse data (with precise data as a special case), where the coarsening process can be informative (and even wrong with a certain probability)
- the result of the regression is imprecise, reflecting the total uncertainty in the data
- nonparametric LIR: extremely weak assumptions, leading to very robust results (generalized LQS regression)

conclusion and outlook

- LIR: new line of approach to regression with coarse data
- LIR is directly applicable to any kind of coarse data (with precise data as a special case), where the coarsening process can be informative (and even wrong with a certain probability)
- the result of the regression is imprecise, reflecting the total uncertainty in the data
- nonparametric LIR: extremely weak assumptions, leading to very robust results (generalized LQS regression)
- future work:

conclusion and outlook

- LIR: new line of approach to regression with coarse data
- LIR is directly applicable to any kind of coarse data (with precise data as a special case), where the coarsening process can be informative (and even wrong with a certain probability)
- the result of the regression is imprecise, reflecting the total uncertainty in the data
- nonparametric LIR: extremely weak assumptions, leading to very robust results (generalized LQS regression)
- future work:
  - improve the *implementation* of LIR

conclusion and outlook

- LIR: new line of approach to regression with coarse data
- LIR is directly applicable to any kind of coarse data (with precise data as a special case), where the coarsening process can be informative (and even wrong with a certain probability)
- the result of the regression is imprecise, reflecting the total uncertainty in the data
- nonparametric LIR: extremely weak assumptions, leading to very robust results (generalized LQS regression)
- future work:
  - improve the *implementation* of LIR
  - study in more detail the *statistical properties* of the method (e.g., the coverage probability of the imprecise result), even though the repeated sampling evaluation is problematic with coarse data

conclusion and outlook

- LIR: new line of approach to regression with coarse data
- LIR is directly applicable to any kind of coarse data (with precise data as a special case), where the coarsening process can be informative (and even wrong with a certain probability)
- the result of the regression is imprecise, reflecting the total uncertainty in the data
- nonparametric LIR: extremely weak assumptions, leading to very robust results (generalized LQS regression)
- future work:
  - improve the *implementation* of LIR
  - study in more detail the *statistical properties* of the method (e.g., the coverage probability of the imprecise result), even though the repeated sampling evaluation is problematic with coarse data
  - investigate the consequences of *stronger assumptions* (e.g., the existence of a true regression function with an additive, homoscedastic, normal error)

conclusion and outlook

- LIR: new line of approach to regression with coarse data
- LIR is directly applicable to any kind of coarse data (with precise data as a special case), where the coarsening process can be informative (and even wrong with a certain probability)
- the result of the regression is imprecise, reflecting the total uncertainty in the data
- nonparametric LIR: extremely weak assumptions, leading to very robust results (generalized LQS regression)
- future work:
  - improve the *implementation* of LIR
  - study in more detail the *statistical properties* of the method (e.g., the coverage probability of the imprecise result), even though the repeated sampling evaluation is problematic with coarse data
  - investigate the consequences of *stronger assumptions* (e.g., the existence of a true regression function with an additive, homoscedastic, normal error)
  - consider the minimization of other properties of the distribution of the absolute residuals (besides the quantiles), in order to increase the *efficiency* of the method (e.g., generalized LTS regression)

references

Cattaneo, M. (2007). *Statistical Decisions Based Directly on the Likelihood Function*. PhD thesis, ETH Zurich. doi:10.3929/ethz-a-005463829.

Cattaneo, M., and Wiencierz, A. (2011). Regression with Imprecise Data: A Robust Approach. In *ISIPTA '11, Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, eds. F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. SIPTA, 119–128.

Domingues, M. A. O., de Souza, R. M. C. R., and Cysneiros, F. J. A. (2010). A robust method for linear regression of symbolic interval data. *Pattern Recognit. Lett.* 31, 1991–1996.

Ferson, S., Kreinovich, V., Hajagos, J., Oberkampf, W., and Ginzburg, L. (2007). *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. Technical Report SAND2007-0939. Sandia National Laboratories.

Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley.