Regression with Imprecise Data:
A Robust Approach

Marco Cattaneo and Andrea Wiencierz
Department of Statistics, LMU Munich

ISIPTA '11, Innsbruck, Austria
25 July 2011

regression problem with imprecise data:

We investigate the relationship between the variables $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$ (where $\mathcal{X}$ can be any set).

We would like to describe this relationship by means of a function $f \in \mathcal{F}$, where $\mathcal{F}$ is a particular set of functions $f : \mathcal{X} \to \mathbb{R}$ (e.g., the set of all linear functions).

The regression problem consists in trying to identify the function $f \in \mathcal{F}$ minimizing in some sense the (absolute) residuals $R_{f,i} := |Y_i - f(X_i)|$.

However, instead of precise data $V_i := (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, we can often observe only imprecise data $V_i^* \subseteq \mathcal{X} \times \mathbb{R}$, and therefore (for each function $f \in \mathcal{F}$) the residuals $R_{f,i}$ are imprecisely observed as well.

regression as a decision problem:

We choose a nonparametric probabilistic model: $\mathcal{P}$ is the set of all probability measures such that $(V_1, V_1^*), \dots, (V_n, V_n^*)$ are independent and identically distributed and satisfy $P(V_i \in V_i^*) \geq 1 - \varepsilon$ (where $\varepsilon \in [0, 1]$ is fixed).

We try to identify the function $f \in \mathcal{F}$ minimizing the p-quantile of the distribution of $R_{f,i}$ (where $p \in (0, 1)$ is fixed): the main reason for the choice of the p-quantile (instead, e.g., of a moment of the distribution of $R_{f,i}$) is that it can be estimated even under the nonparametric model $\mathcal{P}$.

The regression problem can be expressed as a decision problem:

- the set of possible decisions is $\mathcal{F}$,
- the set of possible "states of the world" is $\mathcal{P}$, and
- the loss associated to $f \in \mathcal{F}$ and $P \in \mathcal{P}$ is the p-quantile $Q_f(P)$ of the distribution of $R_{f,i}$ under P.

likelihood-based imprecise regression:

The observed (imprecise) data $V_1^* = A_1, \ldots, V_n^* = A_n$ induce the (normalized) likelihood function $lik : \mathcal{P} \to [0, 1]$ with $lik(P) = \frac{P(V_1^*=A_1,\ldots,V_n^*=A_n)}{\sup_{P' \in \mathcal{P}} P'(V_1^*=A_1,\ldots,V_n^*=A_n)}$.

We use lik to reduce the model $\mathcal{P}$ to $\mathcal{P}_{>\beta} := \{P \in \mathcal{P} : lik(P) > \beta\}$ (where $\beta \in (0, 1)$ is fixed).

The imprecise value of the loss $Q_f(P)$ becomes $\mathcal{C}_f := \{Q_f(P) : lik(P) > \beta\}$, which has a simple geometrical interpretation:

- $\underline{B}_{f,q} := \{(x, y) \in \mathcal{X} \times \mathbb{R} : |y - f(x)| < q\}$ is the open band of width $2\,q$ around f,
- $\overline{B}_{f,q} := \{(x, y) \in \mathcal{X} \times \mathbb{R} : |y - f(x)| \leq q\}$ is the closed band of width $2\,q$ around f,
- $\mathcal{C}_f$ consists of all $q \in [0, +\infty)$ such that the closed band $\overline{B}_{f,q}$ is wide enough to intersect at least $\underline{k} + 1$ imprecise data, and the open band $\underline{B}_{f,q}$ is thin enough to contain at most $\overline{k} - 1$ imprecise data (where $\underline{k}, \overline{k}$ depend on n, ε, p, β).

The function $f_{LRM} \in \mathcal{F}$ minimizing $\sup \mathcal{C}_f$ is the (Γ-)minimax decision, and has a simple geometrical interpretation: $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ is the thinnest band of the form $\overline{B}_{f,q}$ containing at least $\overline{k}$ imprecise data.

Each function $f \in \mathcal{F}$ such that $\mathcal{C}_f$ intersects $\mathcal{C}_{f_{LRM}}$ is undominated with respect to interval dominance: geometrically, f is undominated when $\overline{B}_{f,\overline{q}_{LRM}}$ intersects at least $\underline{k} + 1$ imprecise data.

example with social survey data:

We use data from the "ALLBUS — German General Social Survey" of 2008 to investigate the relationship between two variables (with $n = 3247$): age $X_i \in \mathcal{X} := [18, 100)$ and personal income (on average per month) $Y_i \in [0, +\infty)$.

We consider the set $\mathcal{F} = \{f_{a,b_1,b_2} : a, b_1, b_2 \in \mathbb{R}\}$ of all quadratic functions $f_{a,b_1,b_2}(x) = a + b_1\,x + b_2\,x^2$, and choose $\varepsilon = 0$, $p = 0.5$, and $\beta = 0.15$ (implying $\underline{k} = 1568$ and $\overline{k} = 1679$).

In 4 different data situations, we compare 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) with the results of the ordinary least squares regression applied after reducing the imprecise data to their centers and choosing 15 000 (blue curve) or 10 000 (green curve) as the upper income limit.

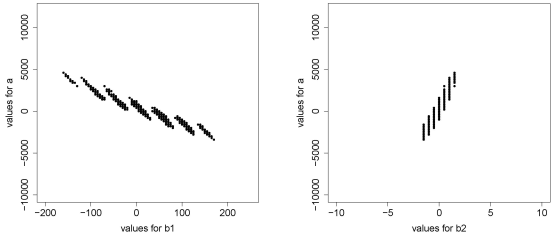
original data:

age data:

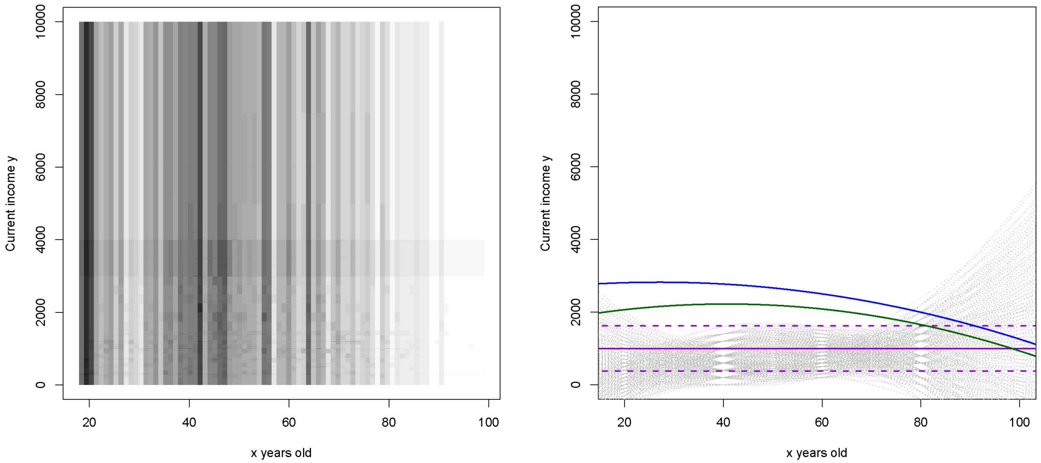
3236 “precise” (in years: 83 classes), 11 missing

income data:

2266 precise, 361 categorized (22 classes), 620 missing



The set of undominated parameter values.



Two-dimensional histogram of the data set and regression results.

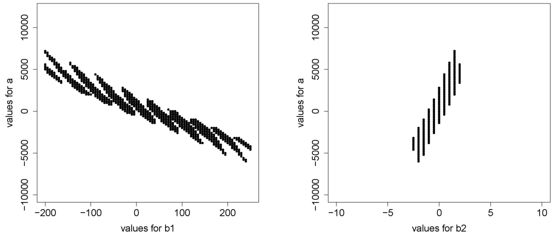
categorized income data:

age data:

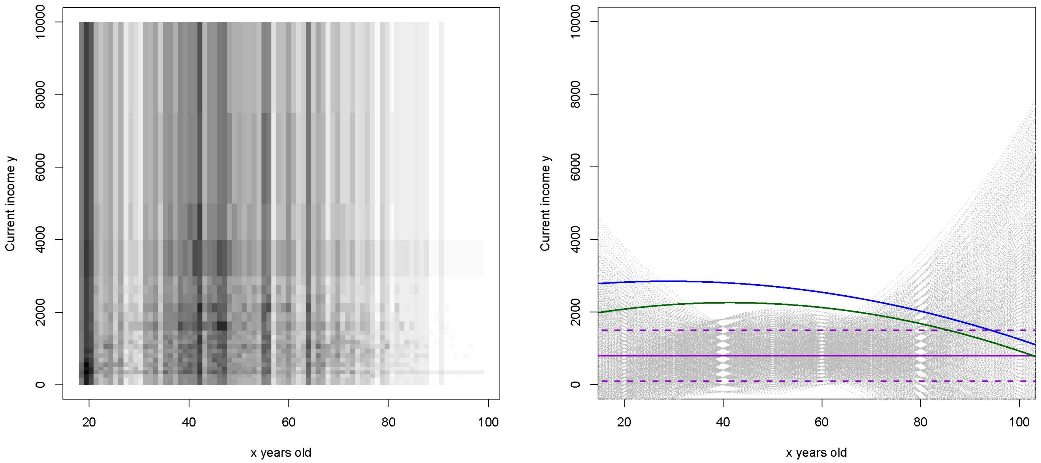
3236 “precise” (in years: 83 classes), 11 missing

income data:

2627 categorized (22 classes), 620 missing



The set of undominated parameter values.

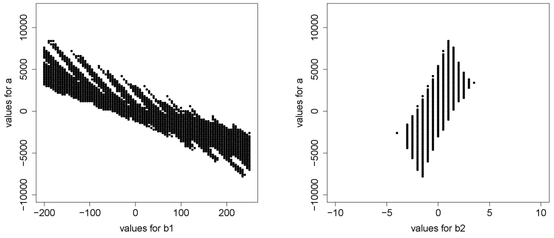


Two-dimensional histogram of the data set and regression results.

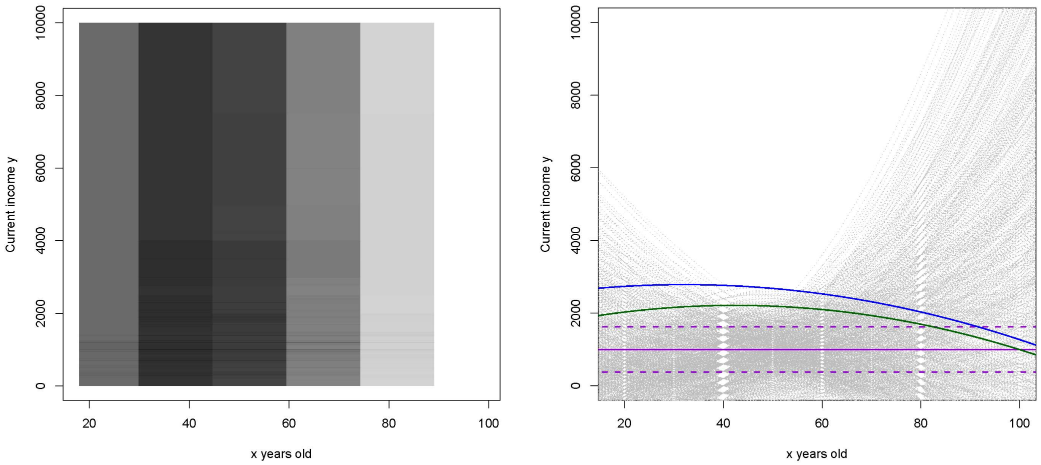
categorized age data:

age data:
3236 categorized (6 classes), 11 missing

income data:
2266 precise, 361 categorized (22 classes), 620 missing



The set of undominated parameter values.

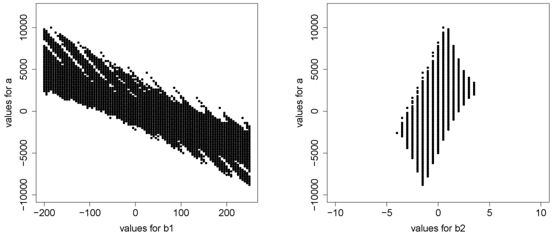


Two-dimensional histogram of the data set and regression results.

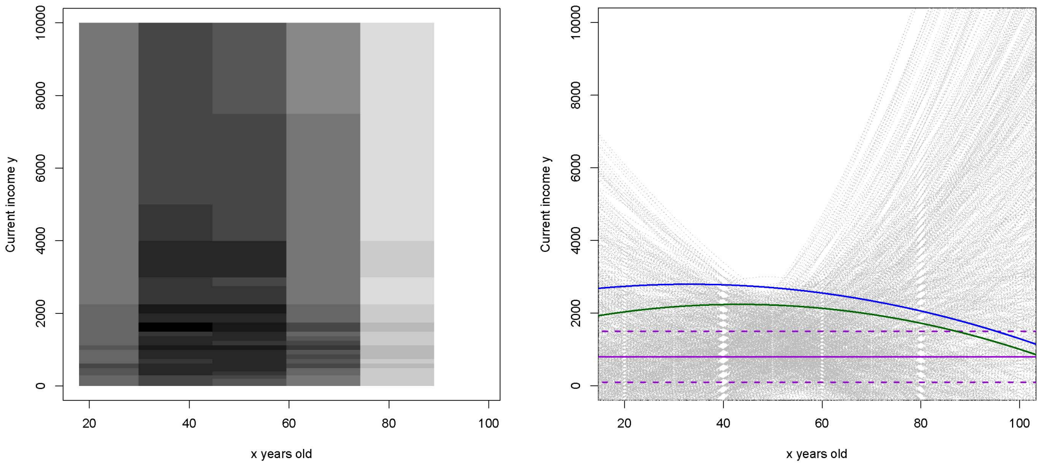
categorized age and income data:

age data:
3236 categorized (6 classes), 11 missing

income data:
2627 categorized (22 classes), 620 missing



The set of undominated parameter values.



Two-dimensional histogram of the data set and regression results.