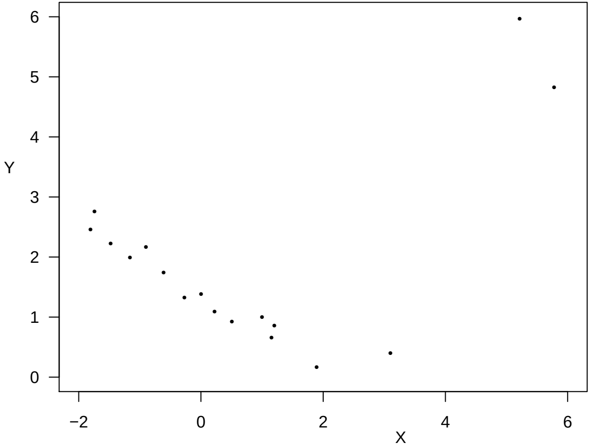
Linear Likelihood-based Imprecise Regression (LIR) with interval data

Andrea Wiencierz and Marco Cattaneo
Department of Statistics, LMU Munich

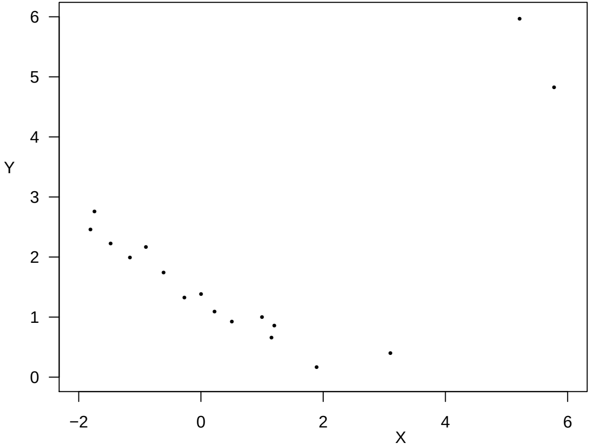
WPMSIIP 5, LMU Munich, Germany
September 11, 2012

(Simple) linear LIR with precise data



(Simple) linear LIR with precise data

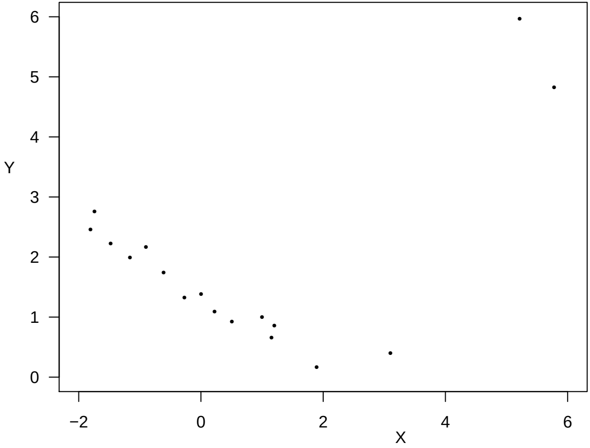
- $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P$



(Simple) linear LIR with precise data

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $(X_i, Y_i) \stackrel{\text{i.i.d.}}{\sim} P$
- simple linear regression:

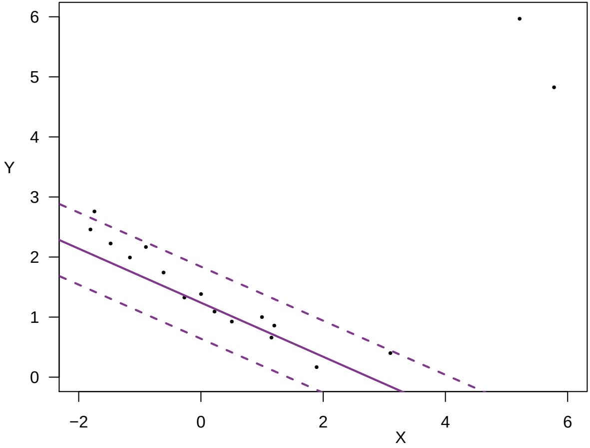
  $Y_i = f(X_i) = a + b\,X_i$



(Simple) linear LIR with precise data

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P$
- simple linear regression:

  $Y_i = f(X_i) = a + b\,X_i$

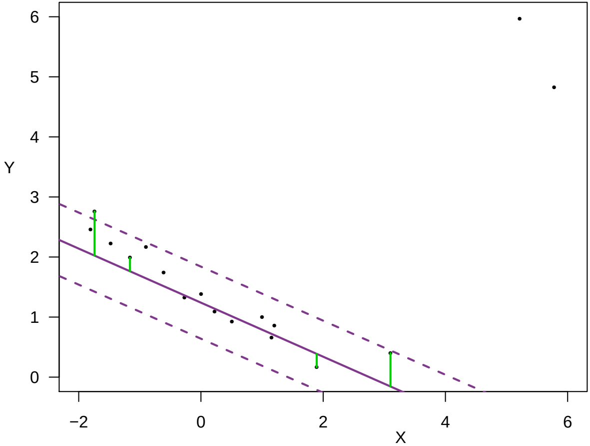


(Simple) linear LIR with precise data

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P$
- simple linear regression:

  $$Y_i = f(X_i) = a + b\,X_i$$
- residuals:

  $$R_{f,i} = |Y_i - f(X_i)|$$



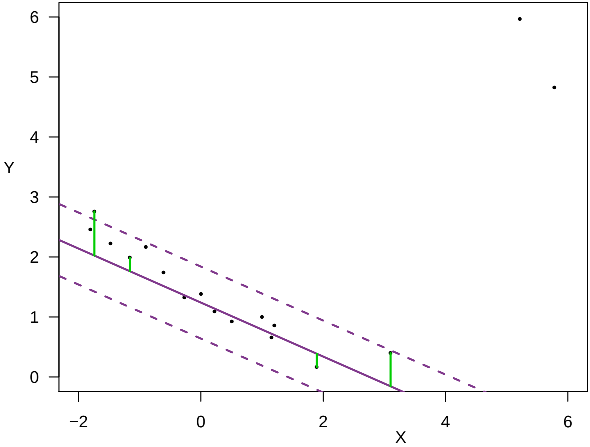
(Simple) linear LIR with precise data

- $(X_1, Y_1), \dots, (X_n, Y_n)$ with $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P$
- simple linear regression:

  $Y_i = f(X_i) = a + b\,X_i$
- residuals:

  $R_{f,i} = |Y_i - f(X_i)|$
- loss function p-quantile:

  $l_p(f) = Q_{R_f,p}, p \in (0,1)$



(Simple) linear LIR with precise data

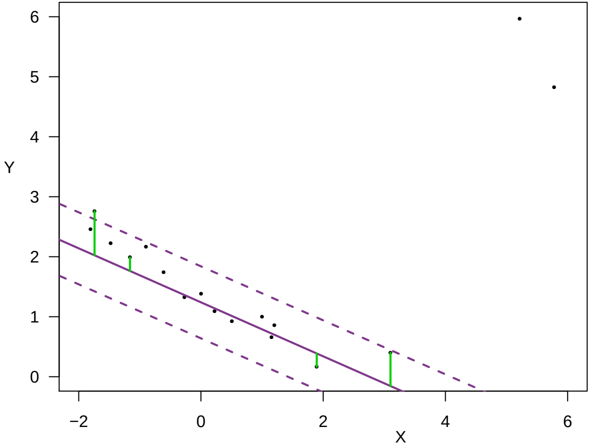
- $(X_1, Y_1), \dots, (X_n, Y_n)$ with $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P$
- simple linear regression:

  $$Y_i = f(X_i) = a + b\,X_i$$
- residuals:

  $$R_{f,i} = |Y_i - f(X_i)|$$
- loss function p-quantile:

  $$l_p(f) = Q_{R_f,p},\, p \in (0,1)$$
- account for statistical uncertainty:

  $\mathcal{C}_{f,p,\beta}$ likelihood-based confidence region with cutoff point $\beta \in (0,1)$



(Simple) linear LIR with precise data

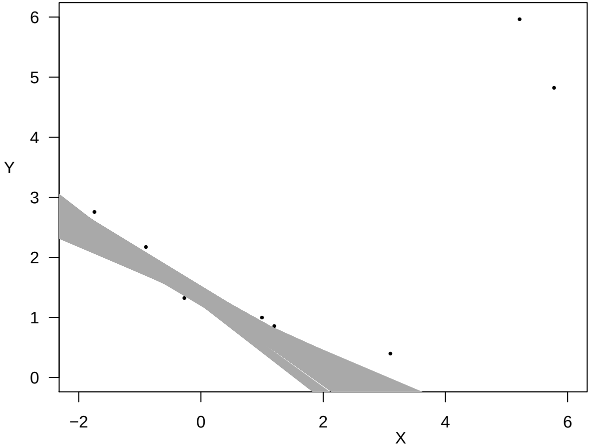
- $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P$
- simple linear regression:

  $Y_i = f(X_i) = a + b\,X_i$
- residuals:

  $R_{f,i} = |Y_i - f(X_i)|$
- loss function p-quantile:

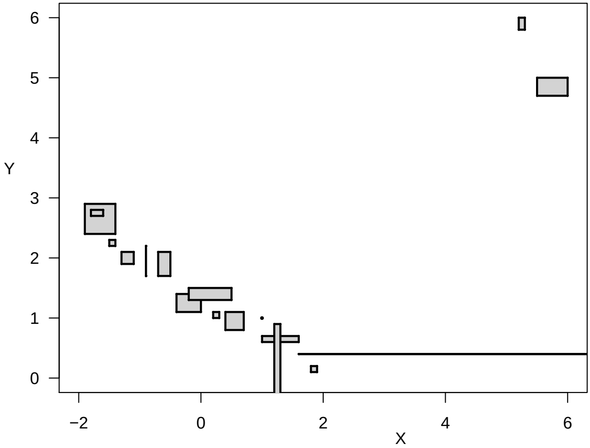
  $l_p(f) = Q_{R_f,p}, p \in (0,1)$
- account for statistical uncertainty:

  $\mathcal{C}_{f,p,\beta}$ likelihood-based confidence region with cutoff point $\beta \in (0,1)$
- result $\mathcal{U}$: set of plausible functions



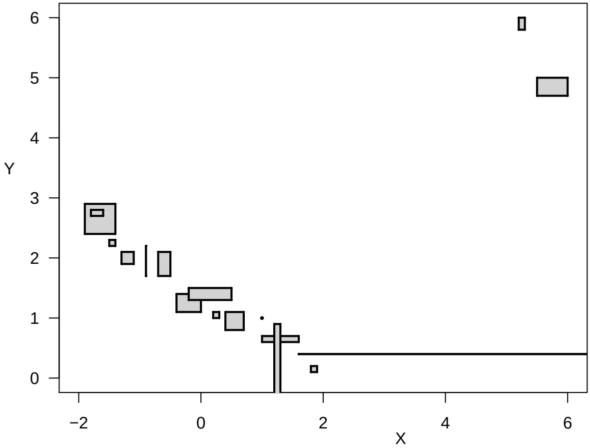
(Simple) linear LIR with interval data

- $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$
  where $X_i^* = [\underline{X}_i, \overline{X}_i]$
  and $Y_i^* = [\underline{Y}_i, \overline{Y}_i]$



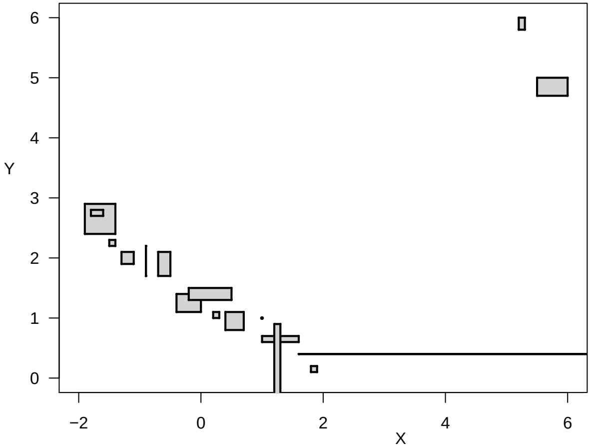
(Simple) linear LIR with interval data

- $(X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$
  where $X_i^* = [\underline{X}_i, \overline{X}_i]$
  and $Y_i^* = [\underline{Y}_i, \overline{Y}_i]$
- with $V_i^* = X_i^* \times Y_i^*$
  $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P$
  such that for $\varepsilon \in [0, 1]$
  $P((X_i, Y_i) \notin V_i^*) \leq \varepsilon$



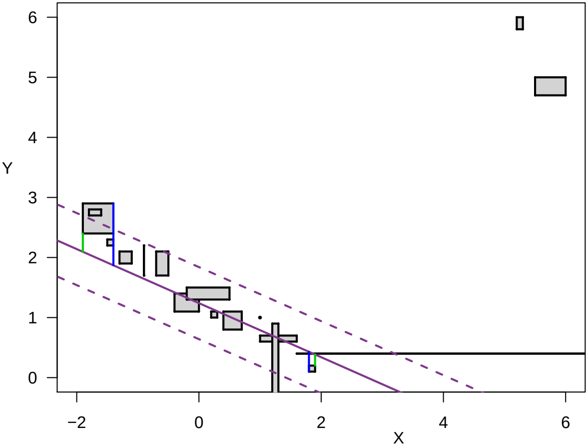
(Simple) linear LIR with interval data

- $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$
  where $X_i^* = [\underline{X}_i, \overline{X}_i]$
  and $Y_i^* = [\underline{Y}_i, \overline{Y}_i]$
- with $V_i^* = X_i^* \times Y_i^*$
  $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P$
  such that for $\varepsilon \in [0, 1]$
  $P((X_i, Y_i) \notin V_i^*) \leq \varepsilon$
- simple linear regression:
  $Y_i = f(X_i) = a + b\,X_i$



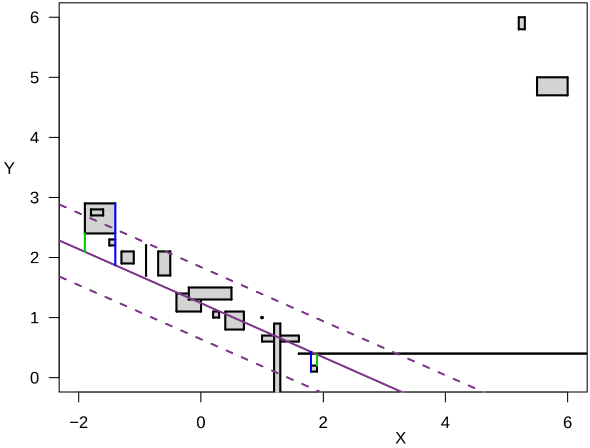
(Simple) linear LIR with interval data

- $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ where $X_i^* = [\underline{X}_i, \overline{X}_i]$ and $Y_i^* = [\underline{Y}_i, \overline{Y}_i]$
- with $V_i^* = X_i^* \times Y_i^*$ $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P$ such that for $\varepsilon \in [0, 1]$ $P((X_i, Y_i) \notin V_i^*) \leq \varepsilon$
- simple linear regression: $Y_i = f(X_i) = a + b\, X_i$



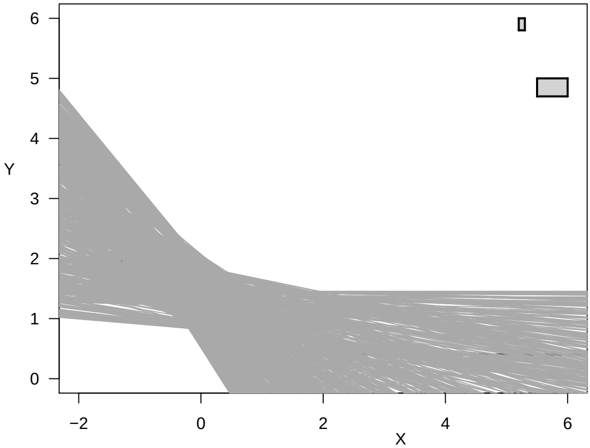
(Simple) linear LIR with interval data

- $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$
  where $X_i^* = [\underline{X}_i, \overline{X}_i]$
  and $Y_i^* = [\underline{Y}_i, \overline{Y}_i]$
- with $V_i^* = X_i^* \times Y_i^*$
  $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P$
  such that for $\varepsilon \in [0, 1]$
  $P((X_i, Y_i) \notin V_i^*) \leq \varepsilon$
- simple linear regression:
  $Y_i = f(X_i) = a + b\, X_i$
- $\mathcal{C}_{f,p,\beta,\varepsilon}$



(Simple) linear LIR with interval data

- $(X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$ where $X_i^* = [\underline{X}_i, \overline{X}_i]$ and $Y_i^* = [\underline{Y}_i, \overline{Y}_i]$
- with $V_i^* = X_i^* \times Y_i^*$ $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P$ such that for $\varepsilon \in [0,1]$ $P((X_i, Y_i) \notin V_i^*) \leq \varepsilon$
- simple linear regression: $Y_i = f(X_i) = a + b\, X_i$
- $\mathcal{C}_{f,p,\beta,\varepsilon}$
- result $\mathcal{U}$: set of plausible functions



(Simple) linear LIR with interval data

- $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P\,,\ P \in \mathcal{P}_\varepsilon = \{P : P((X_i, Y_i) \notin V_i^*) \leq \varepsilon\}\ ,\ \varepsilon \in [0, 1]$

(Simple) linear LIR with interval data

- $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P\,,\ P \in \mathcal{P}_\varepsilon = \{P : P((X_i, Y_i) \notin V_i^*) \leq \varepsilon\}\ ,\ \varepsilon \in [0, 1]$
- $Y_i = f(X_i)\,,\ f \in \mathcal{F} = \left\{ f_{a,b} : \begin{array}{ccc} \mathbb{R} & \to & \mathbb{R} \\ X & \mapsto & a + b\,X \end{array}\ ,\ a, b \in \mathbb{R} \right\}$

(Simple) linear LIR with interval data

- $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P\,,\ P \in \mathcal{P}_\varepsilon = \{P : P((X_i, Y_i) \notin V_i^*) \leq \varepsilon\}\,,\ \varepsilon \in [0, 1]$
- $Y_i = f(X_i)\,,\ f \in \mathcal{F} = \left\{ f_{a,b} : \begin{array}{ccc} \mathbb{R} & \to & \mathbb{R} \\ X & \mapsto & a + b\,X \end{array},\ a, b \in \mathbb{R} \right\}$
- observations $v_1^*, \ldots, v_n^*$ induce (normalized) profile likelihood function for the p-quantile of the distribution of R_f for each $f \in \mathcal{F}$

(Simple) linear LIR with interval data

- $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P\,,\ P \in \mathcal{P}_\varepsilon = \{P : P((X_i, Y_i) \notin V_i^*) \leq \varepsilon\}\ ,\ \varepsilon \in [0, 1]$
- $Y_i = f(X_i)\,,\ f \in \mathcal{F} = \left\{ f_{a,b} : \begin{array}{ccc} \mathbb{R} & \to & \mathbb{R} \\ X & \mapsto & a + b\,X \end{array}\ ,\ a, b \in \mathbb{R} \right\}$
- observations $v_1^*, \ldots, v_n^*$ induce (normalized) profile likelihood function for the p-quantile of the distribution of R_f for each $f \in \mathcal{F}$
- $\underline{r}_{f,i} = \min_{(x,y)\in[\underline{x}_i,\overline{x}_i]\times[\underline{y}_i,\overline{y}_i]} |y - f(x)|\ ,\quad \overline{r}_{f,i} = \sup_{(x,y)\in[\underline{x}_i,\overline{x}_i]\times[\underline{y}_i,\overline{y}_i]} |y - f(x)|$

(Simple) linear LIR with interval data

- $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P\,,\ P \in \mathcal{P}_\varepsilon = \{P : P((X_i, Y_i) \notin V_i^*) \leq \varepsilon\}\ ,\ \varepsilon \in [0, 1]$
- $Y_i = f(X_i)\,,\ f \in \mathcal{F} = \left\{ f_{a,b} : \begin{array}{ccc} \mathbb{R} & \to & \mathbb{R} \\ X & \mapsto & a + b\,X \end{array}\ ,\ a, b \in \mathbb{R} \right\}$
- observations $v_1^*, \ldots, v_n^*$ induce (normalized) profile likelihood function for the p-quantile of the distribution of R_f for each $f \in \mathcal{F}$
- $\underline{r}_{f,i} = \min_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} |y - f(x)|\ ,\quad \overline{r}_{f,i} = \sup_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} |y - f(x)|$
- $0 = \underline{r}_{f,(0)}, \ldots, \underline{r}_{f,(\lceil n(p-\varepsilon) \rceil)}, \overline{r}_{f,(\lfloor n(p+\varepsilon) \rfloor + 1)}, \ldots, \overline{r}_{f,(n+1)} = +\infty$

(Simple) linear LIR with interval data

- $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P\,,\; P \in \mathcal{P}_\varepsilon = \{P : P((X_i, Y_i) \notin V_i^*) \leq \varepsilon\}\,,\; \varepsilon \in [0, 1]$
- $Y_i = f(X_i)\,,\; f \in \mathcal{F} = \left\{ f_{a,b} : \begin{array}{ccc} \mathbb{R} & \to & \mathbb{R} \\ X & \mapsto & a + b\,X \end{array}\,,\; a, b \in \mathbb{R} \right\}$
- observations $v_1^*, \ldots, v_n^*$ induce (normalized) profile likelihood function for the p-quantile of the distribution of R_f for each $f \in \mathcal{F}$
- $\underline{r}_{f,i} = \min_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} |y - f(x)|\,, \quad \overline{r}_{f,i} = \sup_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} |y - f(x)|$
- $0 = \underline{r}_{f,(0)}, \ldots, \underline{r}_{f,(\lceil n(p-\varepsilon) \rceil)}, \overline{r}_{f,(\lfloor n(p+\varepsilon) \rfloor + 1)}, \ldots, \overline{r}_{f,(n+1)} = +\infty$
- $\mathcal{C}_f = \left[\underline{r}_{f,(\underline{k}+1)}, \overline{r}_{f,(\overline{k})}\right]$, values of $\underline{k}, \overline{k} \in \mathbb{N} \cup \{0\}$ depend on n, p, β, ε

(Simple) linear LIR with interval data

- $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P\,,\ P \in \mathcal{P}_\varepsilon = \{P : P((X_i, Y_i) \notin V_i^*) \leq \varepsilon\}\,,\ \varepsilon \in [0, 1]$
- $Y_i = f(X_i)\,,\ f \in \mathcal{F} = \left\{ f_{a,b} : \begin{array}{ccc} \mathbb{R} & \to & \mathbb{R} \\ X & \mapsto & a + b\,X \end{array}\,,\ a, b \in \mathbb{R} \right\}$
- observations $v_1^*, \ldots, v_n^*$ induce (normalized) profile likelihood function for the p-quantile of the distribution of R_f for each $f \in \mathcal{F}$
- $\underline{r}_{f,i} = \min_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} |y - f(x)|\,, \quad \overline{r}_{f,i} = \sup_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} |y - f(x)|$
- $0 = \underline{r}_{f,(0)}, \ldots, \underline{r}_{f,(\lceil n(p-\varepsilon) \rceil)}, \overline{r}_{f,(\lfloor n(p+\varepsilon) \rfloor + 1)}, \ldots, \overline{r}_{f,(n+1)} = +\infty$
- $\mathcal{C}_f = [\underline{r}_{f,(\underline{k}+1)}, \overline{r}_{f,(\overline{k})}]\,,$ values of $\underline{k}, \overline{k} \in \mathbb{N} \cup \{0\}$ depend on n, p, β, ε
- LIR result $\mathcal{U} = \{f \in \mathcal{F} : \underline{r}_{f,(\underline{k}+1)} \leq \overline{q}_{LRM}\}\,,$ where $\overline{q}_{LRM} = \inf_{f \in \mathcal{F}} \overline{r}_{f,(\overline{k})}$

(Simple) linear LIR with interval data

- $((X_i, Y_i), V_i^*) \overset{\text{i.i.d.}}{\sim} P\,,\ P \in \mathcal{P}_\varepsilon = \{P : P((X_i, Y_i) \notin V_i^*) \leq \varepsilon\}\,,\ \varepsilon \in [0,1]$
- $Y_i = f(X_i)\,,\ f \in \mathcal{F} = \left\{ f_{a,b} : \begin{array}{ccc} \mathbb{R} & \to & \mathbb{R} \\ X & \mapsto & a + b\,X \end{array},\ a, b \in \mathbb{R} \right\}$
- observations $v_1^*, \ldots, v_n^*$ induce (normalized) profile likelihood function for the p-quantile of the distribution of R_f for each $f \in \mathcal{F}$
- $\underline{r}_{f,i} = \min_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} |y - f(x)|\,, \quad \overline{r}_{f,i} = \sup_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} |y - f(x)|$
- $0 = \underline{r}_{f,(0)}, \ldots, \underline{r}_{f,(\lceil n(p-\varepsilon) \rceil)}, \overline{r}_{f,(\lfloor n(p+\varepsilon) \rfloor + 1)}, \ldots, \overline{r}_{f,(n+1)} = +\infty$
- $\mathcal{C}_f = [\underline{r}_{f,(\underline{k}+1)}, \overline{r}_{f,(\overline{k})}]$, values of $\underline{k}, \overline{k} \in \mathbb{N} \cup \{0\}$ depend on n, p, β, ε
- LIR result $\mathcal{U} = \{f \in \mathcal{F} : \underline{r}_{f,(\underline{k}+1)} \leq \overline{q}_{LRM}\}$, where $\overline{q}_{LRM} = \inf_{f \in \mathcal{F}} \overline{r}_{f,(\overline{k})}$
- further details in: M. Cattaneo, A. Wiencierz (2012). *Likelihood-based Imprecise Regression*. Int. J. Approx. Reasoning 53. 1137-1154.

Statistical properties of the LIR method

- robustness:

Statistical properties of the LIR method

- robustness: breakdown-point $\epsilon^* = \frac{\min\{\underline{k}, n-\overline{k}\}}{n} \stackrel{n\to\infty}{\longrightarrow} \min\{p, 1-p\} - \varepsilon$

Statistical properties of the LIR method

- robustness: breakdown-point $\epsilon^* = \frac{\min\{\underline{k}, n-\overline{k}\}}{n} \overset{n\to\infty}{\longrightarrow} \min\{p, 1-p\} - \varepsilon$
- exact confidence level of $\mathcal{C}_f$:

Statistical properties of the LIR method

- robustness: breakdown-point $\epsilon^* = \frac{\min\{\underline{k}, n-\overline{k}\}}{n} \overset{n\to\infty}{\longrightarrow} \min\{p, 1-p\} - \varepsilon$
- exact confidence level of $\mathcal{C}_f$:

$$P(\mathcal{C}_f \ni Q_{R_f}) \geq \begin{cases} \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} p^k (1-p)^{n-k} & \varepsilon = 0 \\ \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} (p+\varepsilon)^k (1-(p+\varepsilon))^{n-k} & \varepsilon > 0, p \leq 0.5 \\ \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} (p-\varepsilon)^k (1-(p-\varepsilon))^{n-k} & \varepsilon > 0, p > 0.5 \end{cases}$$

Statistical properties of the LIR method

- robustness: breakdown-point $\epsilon^* = \frac{\min\{\underline{k}, n-\overline{k}\}}{n} \stackrel{n\to\infty}{\longrightarrow} \min\{p, 1-p\} - \varepsilon$
- exact confidence level of $\mathcal{C}_f$:

$$P(\mathcal{C}_f \ni Q_{R_f}) \geq \begin{cases} \displaystyle\sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} p^k (1-p)^{n-k} & \varepsilon = 0 \\ \displaystyle\sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} (p+\varepsilon)^k (1-(p+\varepsilon))^{n-k} & \varepsilon > 0, p \leq 0.5 \\ \displaystyle\sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} (p-\varepsilon)^k (1-(p-\varepsilon))^{n-k} & \varepsilon > 0, p > 0.5 \end{cases}$$

- confidence level of the result $\mathcal{U}$:

Statistical properties of the LIR method

- robustness: breakdown-point $\epsilon^* = \frac{\min\{\underline{k}, n-\overline{k}\}}{n} \overset{n\to\infty}{\longrightarrow} \min\{p, 1-p\} - \varepsilon$
- exact confidence level of $\mathcal{C}_f$:

$$P(\mathcal{C}_f \ni Q_{R_f}) \geq \begin{cases} \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} p^k (1-p)^{n-k} & \varepsilon = 0 \\ \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} (p+\varepsilon)^k (1-(p+\varepsilon))^{n-k} & \varepsilon > 0, p \leq 0.5 \\ \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} (p-\varepsilon)^k (1-(p-\varepsilon))^{n-k} & \varepsilon > 0, p > 0.5 \end{cases}$$

- confidence level of the result $\mathcal{U}$: We don't know yet.

Statistical properties of the LIR method

- robustness: breakdown-point $\epsilon^* = \frac{\min\{\underline{k}, n-\overline{k}\}}{n} \overset{n\to\infty}{\longrightarrow} \min\{p, 1-p\} - \varepsilon$
- exact confidence level of $\mathcal{C}_f$:

$$P(\mathcal{C}_f \ni Q_{R_f}) \geq \begin{cases} \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} p^k (1-p)^{n-k} & \varepsilon = 0 \\ \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} (p+\varepsilon)^k (1-(p+\varepsilon))^{n-k} & \varepsilon > 0, p \leq 0.5 \\ \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} (p-\varepsilon)^k (1-(p-\varepsilon))^{n-k} & \varepsilon > 0, p > 0.5 \end{cases}$$

- confidence level of the result $\mathcal{U}$: We don't know yet.
- consistency of $\mathcal{U}$:

Statistical properties of the LIR method

- robustness: breakdown-point $\epsilon^* = \frac{\min\{\underline{k}, n-\overline{k}\}}{n} \stackrel{n\to\infty}{\longrightarrow} \min\{p, 1-p\} - \varepsilon$
- exact confidence level of $\mathcal{C}_f$:

$$P(\mathcal{C}_f \ni Q_{R_f}) \geq \begin{cases} \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} p^k (1-p)^{n-k} & \varepsilon = 0 \\ \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} (p+\varepsilon)^k (1-(p+\varepsilon))^{n-k} & \varepsilon > 0, p \leq 0.5 \\ \sum_{k=\underline{k}+1}^{\overline{k}} \binom{n}{k} (p-\varepsilon)^k (1-(p-\varepsilon))^{n-k} & \varepsilon > 0, p > 0.5 \end{cases}$$

- confidence level of the result $\mathcal{U}$: We don't know yet.
- consistency of $\mathcal{U}$: What does that mean? $\rightarrow$ tomorrow

Implementation: Exact algorithm for simple linear LIR

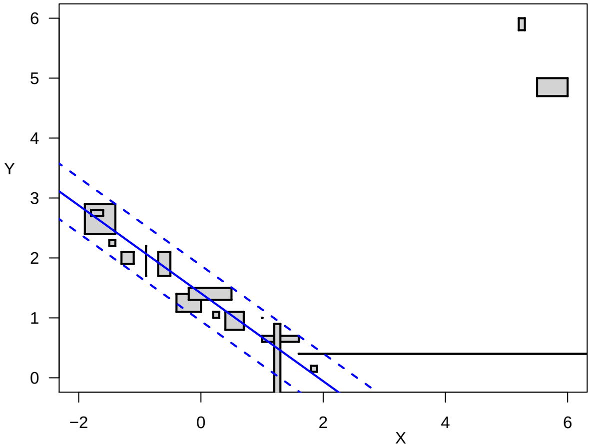
- aim: determine the set of undominated functions $\mathcal{U} = \{f \in \mathcal{F} : \underline{r}_{f,(\underline{k}+1)} \leq \overline{q}_{LRM}\}$

Implementation: Exact algorithm for simple linear LIR

- aim: determine the set of undominated functions $\mathcal{U} = \{f \in \mathcal{F} : \underline{r}_{f,(\underline{k}+1)} \leq \overline{q}_{LRM}\}$
- 1st step: find $\overline{q}_{LRM}$

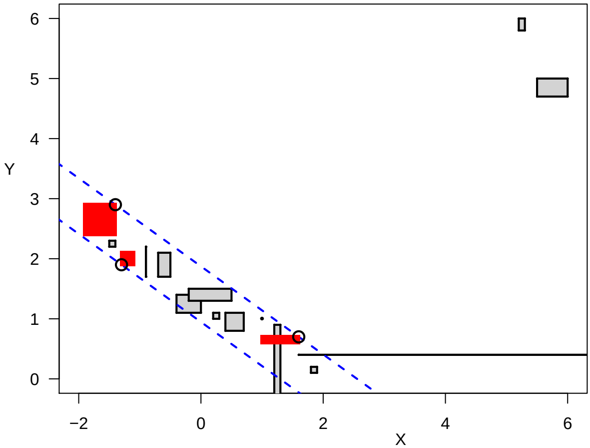
Implementation: Exact algorithm for simple linear LIR

- aim: determine the set of undominated functions $\mathcal{U} = \{f \in \mathcal{F} : \underline{r}_{f,(\underline{k}+1)} \leq \overline{q}_{LRM}\}$
- 1st step: find $\overline{q}_{LRM}$
- $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ (blue dashed lines) is the thinnest band containing at least $\overline{k}$ imprecise data
- here $\beta = 0.8, \underline{p} = 0.6,$ $n = 17$, and $\overline{k} = 12$



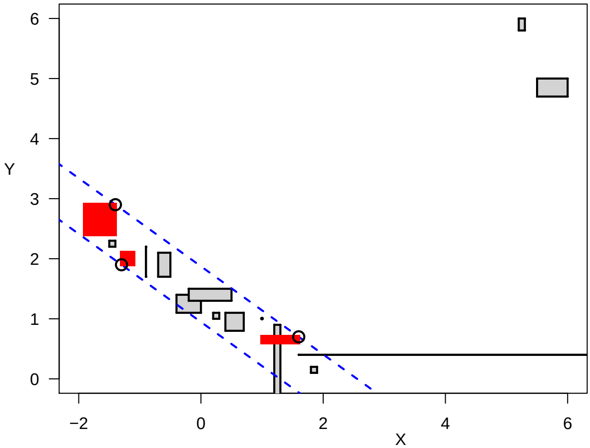
Implementation: Exact algorithm - Part 1

- some of the included $\overline{k}$ imprecise observations touch the borders of $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ in 3 different points



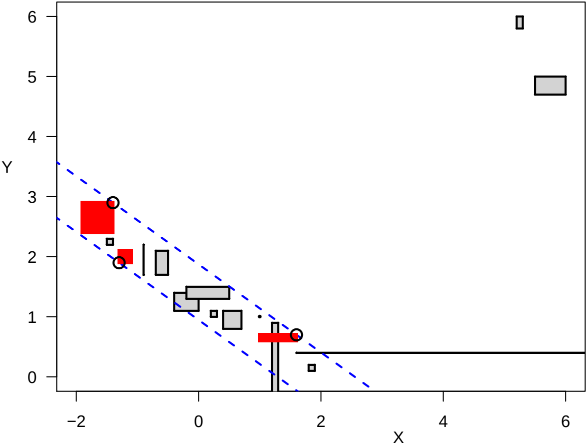
Implementation: Exact algorithm - Part 1

- some of the included $\overline{k}$ imprecise observations touch the borders of $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ in 3 different points
- b_{LRM} can be any slope determined by the corresponding corner points of 2 imprecise data or 0



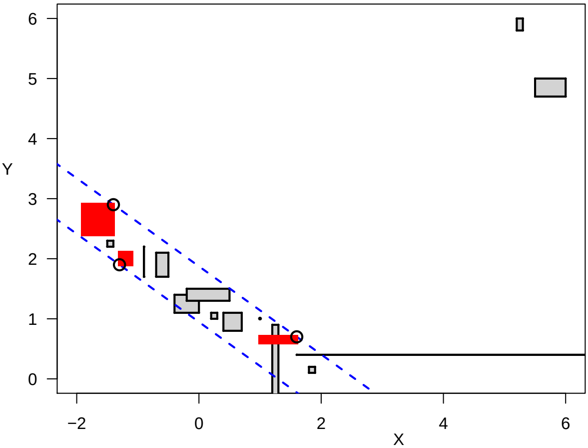
Implementation: Exact algorithm - Part 1

- some of the included $\overline{k}$ imprecise observations touch the borders of $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ in 3 different points
- b_{LRM} can be any slope determined by the corresponding corner points of 2 imprecise data or 0
- $\mathcal{B}$ is the set of $4\binom{n}{2}+1$ possible values for b_{LRM}



Implementation: Exact algorithm - Part 1

- some of the included $\overline{k}$ imprecise observations touch the borders of $\overline{B}_{f_{LRM}, \overline{q}_{LRM}}$ in 3 different points
- b_{LRM} can be any slope determined by the corresponding corner points of 2 imprecise data or 0
- $\mathcal{B}$ is the set of $4\binom{n}{2} + 1$ possible values for b_{LRM}
- for each $b \in \mathcal{B}$ determine $a_b \in \mathbb{R}$ such that $\overline{r}_{f_{a_b,b},(\overline{k})}$ is minimal



Implementation: Exact algorithm - Part 1

for each $b \in \mathcal{B}$

- consider transformed data $z_i^* = [\underline{z}_i, \overline{z}_i]$ with

$$\underline{z}_i = \begin{cases} \underline{y}_i - b\,\overline{x}_i\,, & b > 0 \\ \underline{y}_i - b\,\underline{x}_i\,, & b \leq 0 \end{cases} \quad \text{and} \quad \overline{z}_i = \begin{cases} \overline{y}_i - b\,\underline{x}_i\,, & b > 0 \\ \overline{y}_i - b\,\overline{x}_i\,, & b \leq 0 \end{cases}$$

Implementation: Exact algorithm - Part 1

for each $b \in \mathcal{B}$

- consider transformed data $z_i^* = [\underline{z}_i, \overline{z}_i]$ with

$$\underline{z}_i = \begin{cases} \underline{y}_i - b\,\overline{x}_i\,, & b > 0 \\ \underline{y}_i - b\,\underline{x}_i\,, & b \leq 0 \end{cases} \quad \text{and} \quad \overline{z}_i = \begin{cases} \overline{y}_i - b\,\underline{x}_i\,, & b > 0 \\ \overline{y}_i - b\,\overline{x}_i\,, & b \leq 0 \end{cases}$$

- determine the shortest of the $n - \overline{k} + 1$ intervals of the form $\left(\overline{z}_{[j]} - \underline{z}_{(j)}\right)$,

  where $\overline{z}_{[j]}$ is the $\overline{k}$th smallest value among the $\overline{z}_{b,i}$ such that $\underline{z}_{b,i} \geq \overline{z}_{b,(j)}$

Implementation: Exact algorithm - Part 1

for each $b \in \mathcal{B}$

- consider transformed data $z_i^* = [\underline{z}_i, \overline{z}_i]$ with

$$\underline{z}_i = \begin{cases} \underline{y}_i - b\,\overline{x}_i\,, & b > 0 \\ \underline{y}_i - b\,\underline{x}_i\,, & b \leq 0 \end{cases} \quad \text{and} \quad \overline{z}_i = \begin{cases} \overline{y}_i - b\,\underline{x}_i\,, & b > 0 \\ \overline{y}_i - b\,\overline{x}_i\,, & b \leq 0 \end{cases}$$

- determine the shortest of the $n - \overline{k} + 1$ intervals of the form $\left(\overline{z}_{[j]} - \underline{z}_{(j)}\right)$, where $\overline{z}_{[j]}$ is the $\overline{k}$th smallest value among the $\overline{z}_{b,i}$ such that $\underline{z}_{b,i} \geq \overline{z}_{b,(j)}$
- the length of the shortest interval corresponds to the bandwidth

Implementation: Exact algorithm - Part 1

for each $b \in \mathcal{B}$

- consider transformed data $z_i^* = [\underline{z}_i, \overline{z}_i]$ with

$$\underline{z}_i = \begin{cases} \underline{y}_i - b\,\overline{x}_i\,, & b > 0 \\ \underline{y}_i - b\,\underline{x}_i\,, & b \leq 0 \end{cases} \quad \text{and} \quad \overline{z}_i = \begin{cases} \overline{y}_i - b\,\underline{x}_i\,, & b > 0 \\ \overline{y}_i - b\,\overline{x}_i\,, & b \leq 0 \end{cases}$$

- determine the shortest of the $n - \overline{k} + 1$ intervals of the form $\left(\overline{z}_{[j]} - \underline{z}_{(j)}\right)$,

  where $\overline{z}_{[j]}$ is the $\overline{k}$th smallest value among the $\overline{z}_{b,i}$ such that $\underline{z}_{b,i} \geq \overline{z}_{b,(j)}$
- the length of the shortest interval corresponds to the bandwidth
- the corresponding intercept a_b is given by the midpoint of this interval

Implementation: Exact algorithm - Part 1

for each $b \in \mathcal{B}$

- consider transformed data $z_i^* = [\underline{z}_i, \overline{z}_i]$ with

$$\underline{z}_i = \begin{cases} \underline{y}_i - b\,\overline{x}_i\,, & b > 0 \\ \underline{y}_i - b\,\underline{x}_i\,, & b \leq 0 \end{cases} \quad \text{and} \quad \overline{z}_i = \begin{cases} \overline{y}_i - b\,\underline{x}_i\,, & b > 0 \\ \overline{y}_i - b\,\overline{x}_i\,, & b \leq 0 \end{cases}$$

- determine the shortest of the $n - \overline{k} + 1$ intervals of the form $\left(\overline{z}_{[j]} - \underline{z}_{(j)}\right)$, where $\overline{z}_{[j]}$ is the $\overline{k}$th smallest value among the $\overline{z}_{b,i}$ such that $\underline{z}_{b,i} \geq \overline{z}_{b,(j)}$
- the length of the shortest interval corresponds to the bandwidth
- the corresponding intercept a_b is given by the midpoint of this interval

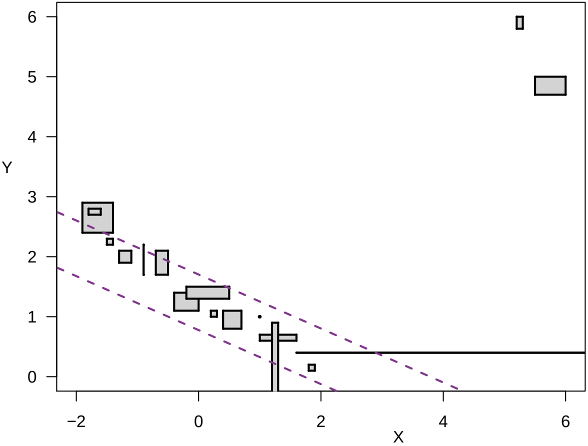
$$\Rightarrow \quad \overline{q}_{LRM} = \frac{1}{2} \min_{(b,j) \in \mathcal{B} \times \{1,\ldots,n-\overline{k}+1\}} \left(\overline{z}_{b,[j]} - \underline{z}_{b,(j)}\right)$$

Implementation: Exact algorithm - Part 2

- step 2: determine $\mathcal{U}$

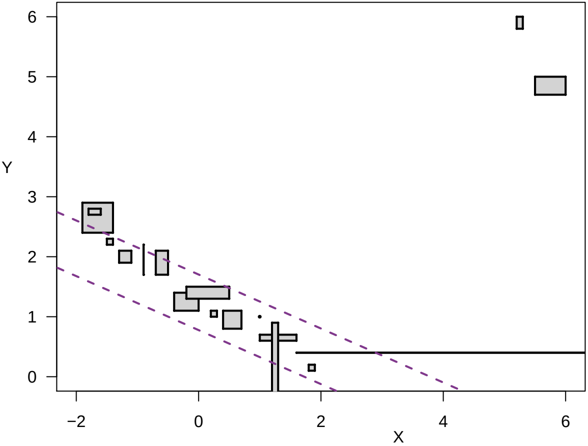
Implementation: Exact algorithm - Part 2

- step 2: determine $\mathcal{U}$
- if $f \in \mathcal{U}$, then $\overline{B}_{f,\overline{q}_{LRM}}$ intersects at least $\underline{k} + 1$ imprecise data
- here $\underline{k} = 8$



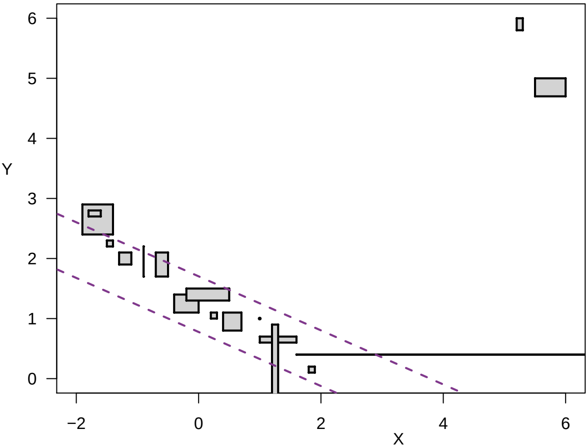
Implementation: Exact algorithm - Part 2

- step 2: determine $\mathcal{U}$
- if $f \in \mathcal{U}$, then $\overline{B}_{f,\overline{q}_{LRM}}$ intersects at least $\underline{k} + 1$ imprecise data
- here $\underline{k} = 8$
- for each $b \in \mathcal{B}$ determine set $A_b \subset \mathbb{R}$ such that $\{f_{a,b} : a \in A_b\} \subset \mathcal{U}$



Implementation: Exact algorithm - Part 2

- step 2: determine $\mathcal{U}$
- if $f \in \mathcal{U}$, then $\overline{B}_{f,\overline{q}_{LRM}}$ intersects at least $\underline{k}+1$ imprecise data
- here $\underline{k} = 8$
- for each $b \in \mathcal{B}$ determine set $A_b \subset \mathbb{R}$ such that $\{f_{a,b} : a \in A_b\} \subset \mathcal{U}$
- $A_b = \bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{(\underline{k}+j)} - \overline{q}_{LRM}, \overline{z}_{(j)} + \overline{q}_{LRM}]$



Implementation: Exact algorithm - Result

$$\Rightarrow \quad \mathcal{U} = \left\{ f_{a,b} : b \in \mathbb{R} \text{ and } a \in \bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}, \overline{z}_{b,(j)} + \overline{q}_{LRM}] \right\}$$

Implementation: Exact algorithm - Result

$$\Rightarrow \quad \mathcal{U} = \left\{ f_{a,b} : b \in \mathbb{R} \text{ and } a \in \bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM},\, \overline{z}_{b,(j)} + \overline{q}_{LRM}] \right\}$$

- $\mathcal{U}$ contains all functions that are plausible relations of X and Y in the light of the imprecise observations

Implementation: Exact algorithm - Result

$$\Rightarrow \quad \mathcal{U} = \left\{ f_{a,b} : b \in \mathbb{R} \text{ and } a \in \bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}, \, \overline{z}_{b,(j)} + \overline{q}_{LRM}] \right\}$$

- $\mathcal{U}$ contains all functions that are plausible relations of X and Y in the light of the imprecise observations
- exact algorithm for the simple linear LIR analysis with interval data

Implementation: Exact algorithm - Result

$$\Rightarrow \quad \mathcal{U} = \left\{ f_{a,b} : b \in \mathbb{R} \text{ and } a \in \bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}, \, \overline{z}_{b,(j)} + \overline{q}_{LRM}] \right\}$$

- $\mathcal{U}$ contains all functions that are plausible relations of X and Y in the light of the imprecise observations
- exact algorithm for the simple linear LIR analysis with interval data
- the presented algorithm has computational complexity $O(n^3 \log n)$

Implementation: Exact algorithm - Result

$$\Rightarrow \quad \mathcal{U} = \left\{ f_{a,b} : b \in \mathbb{R} \text{ and } a \in \bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}, \overline{z}_{b,(j)} + \overline{q}_{LRM}] \right\}$$

- $\mathcal{U}$ contains all functions that are plausible relations of X and Y in the light of the imprecise observations
- exact algorithm for the simple linear LIR analysis with interval data
- the presented algorithm has computational complexity $O(n^3 \log n)$
- further details in: M. Cattaneo, A. Wiencierz (2012). *On the implementation of LIR: the case of simple linear regression with interval data*. Technical Report 127. Department of Statistics. LMU Munich.

R package `linLIR`

- `linLIR`: linear Likelihood-based Imprecise Regression
- R package version 1.0-2 is available at CRAN:
  `http://cran.r-project.org/`

R package `linLIR`

- `linLIR`: linear Likelihood-based Imprecise Regression
- R package version 1.0-2 is available at CRAN:
  `http://cran.r-project.org/`
- plot 2-dimensional interval data set

R package `linLIR`

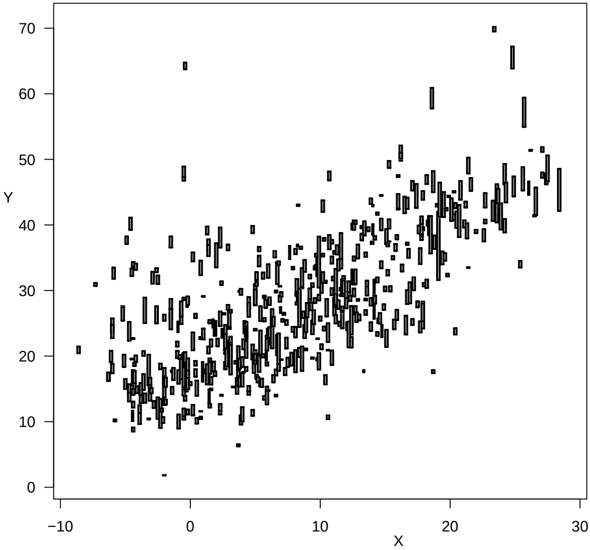
- `linLIR`: linear Likelihood-based Imprecise Regression
- R package version 1.0-2 is available at CRAN:
  `http://cran.r-project.org/`
- plot 2-dimensional interval data set
- `s.linlir` function implements the exact algorithm

R package `linLIR`

- `linLIR`: linear Likelihood-based Imprecise Regression
- R package version 1.0-2 is available at CRAN: `http://cran.r-project.org/`
- plot 2-dimensional interval data set
- `s.linlir` function implements the exact algorithm
- further tools to summarize and visualize results

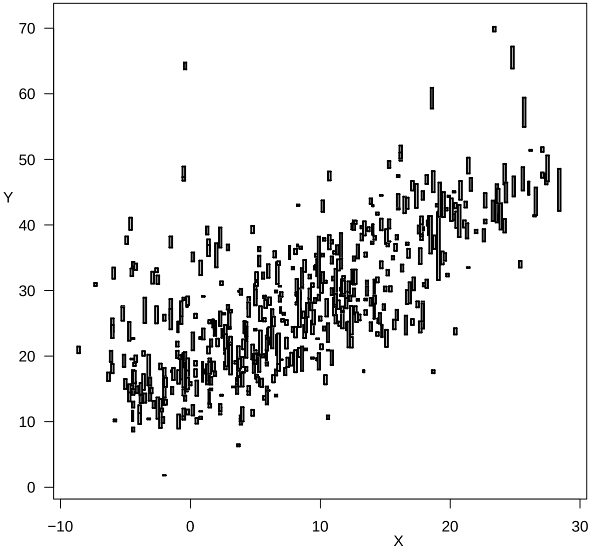
Example

- 2-dimensional interval data set of $n = 514$ observations



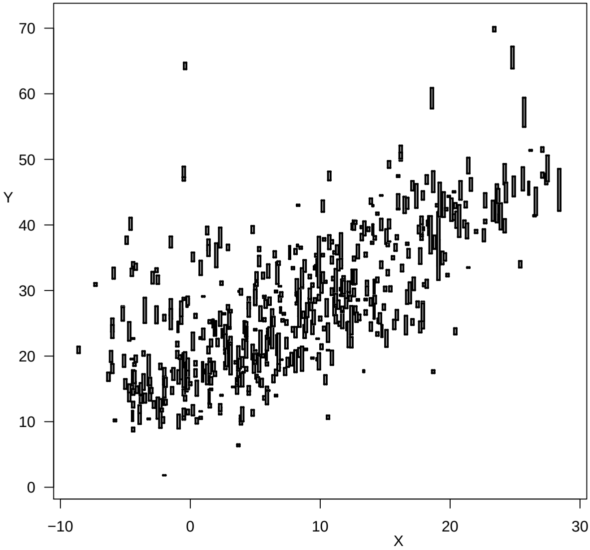
Example

- 2-dimensional interval data set of $n = 514$ observations
- LIR analysis with $p = 0.5, \beta = 0.26, \varepsilon = 0$



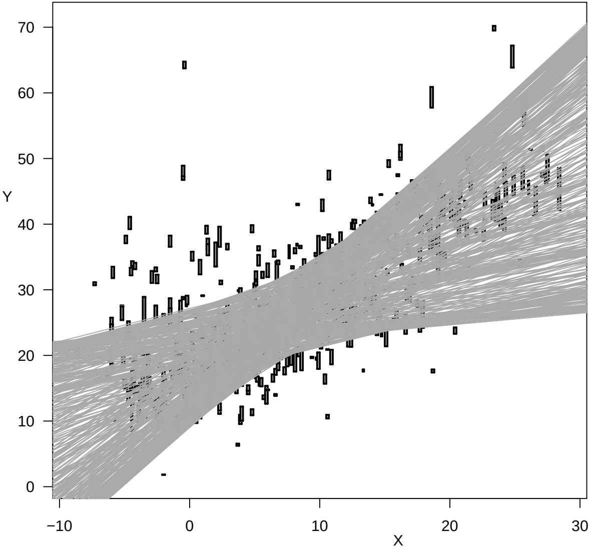
Example

- 2-dimensional interval data set of $n = 514$ observations
- LIR analysis with $p = 0.5, \beta = 0.26, \varepsilon = 0$
- $\underline{k} = 238\,,\ \overline{k} = 276$



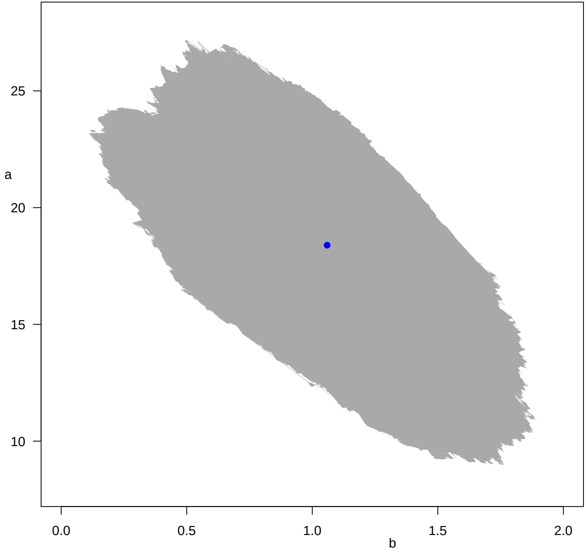
Example

- 2-dimensional interval data set of $n = 514$ observations
- LIR analysis with $p = 0.5, \beta = 0.26, \varepsilon = 0$
- $\underline{k} = 238\,,\ \overline{k} = 276$
- obtained set of undominated functions



Example

- 2-dimensional interval data set of $n = 514$ observations
- LIR analysis with $p = 0.5, \beta = 0.26, \varepsilon = 0$
- $\underline{k} = 238$, $\overline{k} = 276$
- obtained set of undominated functions
- obtained set of parameters



Future work

Future work

- generalize algorithm to multiple linear regression

Future work

- generalize algorithm to multiple linear regression
- supplement R package

Future work

- generalize algorithm to multiple linear regression
- supplement R package
- further investigate statistical properties

Future work

- generalize algorithm to multiple linear regression
- supplement R package
- further investigate statistical properties
- study LIR for more general regression functions