On Sharp Identification Regions for Regression Under Interval Data

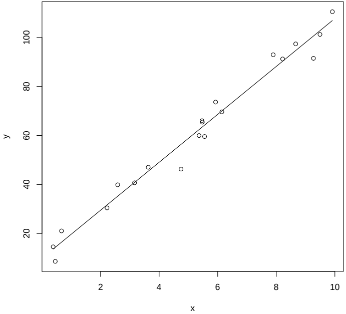
Georg Schollmeyer Thomas Augustin

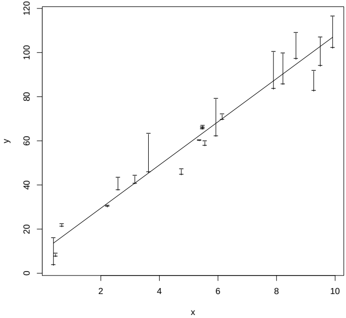
Department of Statistics LMU Munich

Situation

Situation: Standard simple linear model $Y = \beta_0 + \beta_1 X + \varepsilon$ with interval-censored outcomes.

- Y unobserved, only $\underline{Y}$ and $\overline{Y}$ observed and all we know is $Y \in [\underline{Y}, \overline{Y}]$.
- model partially identified, there are many parameters (β_0, β_1) leading to the same distribution of the observable variables.

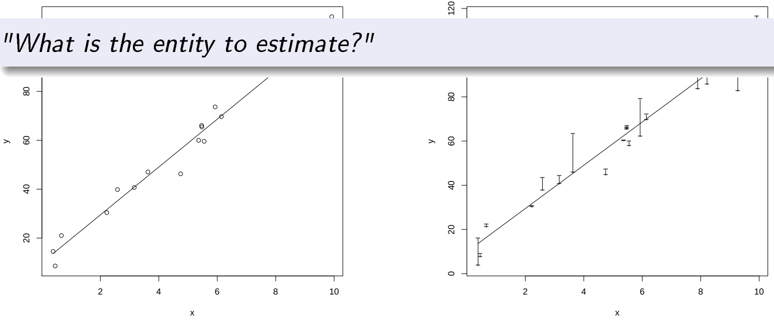




Situation

Situation: Standard simple linear model $Y = \beta_0 + \beta_1 X + \varepsilon$ with interval-censored outcomes.

- Y unobserved, only $\underline{Y}$ and $\overline{Y}$ observed and all we know is $Y \in [\underline{Y}, \overline{Y}]$.
- model partially identified, there are many parameters (β_0, β_1) leading to the same distribution of the observable variables.



"What is the entity to estimate?"

Identification regions for the simple linear model

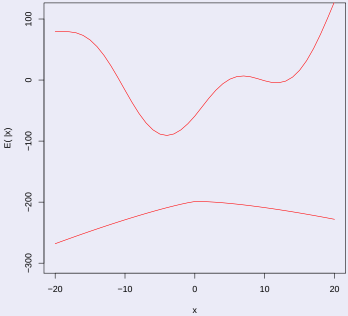
a) Sharp Marrow Region:

$$SMR(\underline{Y}, \overline{Y}) = \{\beta \mid \mathbb{E}(\underline{Y} \mid X) \leq \beta_0 + \beta_1 X \leq \mathbb{E}(\overline{Y} \mid X)\}$$

Identification regions for the simple linear model

a) Sharp Marrow Region:

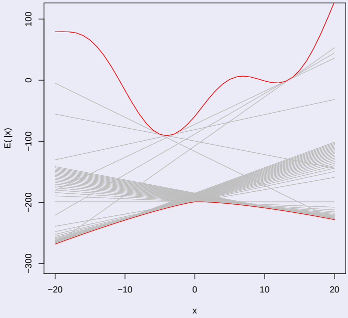
$$SMR(\underline{Y},\overline{Y}) = \{\beta \mid \mathbb{E}(\underline{Y} \mid X) \le \beta_0 + \beta_1 X \le \mathbb{E}(\overline{Y} \mid X)\}$$



Identification regions for the simple linear model

a) Sharp Marrow Region:

$$SMR(\underline{Y}, \overline{Y}) = \{\beta \mid \mathbb{E}(\underline{Y} \mid X) \leq \beta_0 + \beta_1 X \leq \mathbb{E}(\overline{Y} \mid X)\}$$



Identification regions for the simple linear model

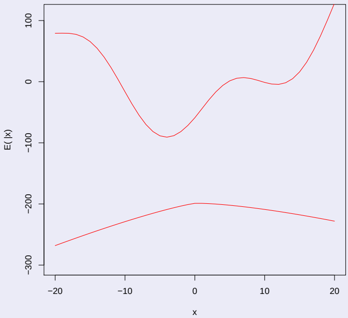
b) Sharp Collection Region:

$$SCR(\underline{\mathrm{Y}}, \overline{\mathrm{Y}}) = \{\operatorname{argmin} \mathbb{E}((\beta_0 + \beta_1 X - Y)^2) \mid Y \in [\underline{\mathrm{Y}}, \overline{\mathrm{Y}}]\}$$

Identification regions for the simple linear model

b) Sharp Collection Region:

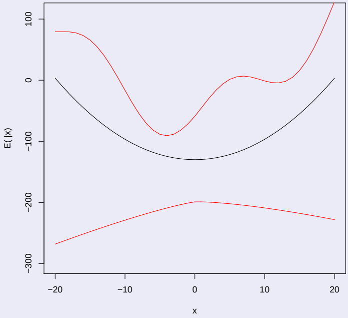
$$SCR(\underline{Y}, \overline{Y}) = \{\operatorname{argmin} \mathbb{E}((\beta_0 + \beta_1 X - Y)^2) \mid Y \in [\underline{Y}, \overline{Y}]\}$$



Identification regions for the simple linear model

b) Sharp Collection Region:

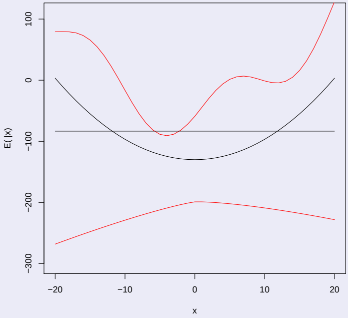
$$SCR(\underline{Y}, \overline{Y}) = \{\text{argmin}\, \mathbb{E}((\beta_0 + \beta_1 X - Y)^2) \mid Y \in [\underline{Y}, \overline{Y}]\}$$



Identification regions for the simple linear model

b) Sharp Collection Region:

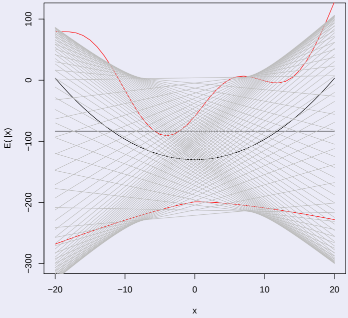
$$SCR(\underline{Y}, \overline{Y}) = \{\operatorname{argmin} \mathbb{E}((\beta_0 + \beta_1 X - Y)^2) \mid Y \in [\underline{Y}, \overline{Y}]\}$$



Identification regions for the simple linear model

b) Sharp Collection Region:

$$SCR(\underline{Y}, \overline{Y}) = \{\operatorname{argmin} \mathbb{E}((\beta_0 + \beta_1 X - Y)^2) \mid Y \in [\underline{Y}, \overline{Y}]\}$$



Sharp Setloss Region

c) Sharp Setloss Region:

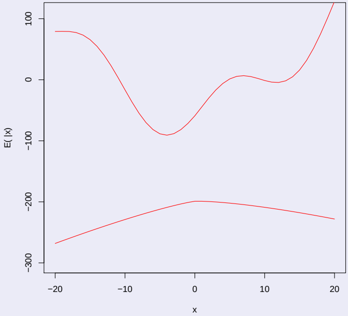
$$L_S(\underline{Y}, \overline{Y}, \Gamma) = \int \left[\mathbb{E}(\overline{Y} \mid x) - \sup_{\beta \in \Gamma} (\beta_0 + \beta_1 x) \right]^2 + \left[\mathbb{E}(\underline{Y} \mid x) - \inf_{\beta \in \Gamma} (\beta_0 + \beta_1 x) \right]^2 d\mathbb{P}(x)$$

Sharp Setloss Region

c) Sharp Setloss Region:

$$L_S(\underline{Y}, \overline{Y}, \Gamma) = \int \left[\mathbb{E}(\overline{Y} \mid x) - \sup_{\beta \in \Gamma} (\beta_0 + \beta_1 x) \right]^2 + \left[\mathbb{E}(\underline{Y} \mid x) - \inf_{\beta \in \Gamma} (\beta_0 + \beta_1 x) \right]^2 d\mathbb{P}(x)$$

$$SSR(\underline{Y}, \overline{Y}) = \bigcup_{\Gamma \subseteq \mathbb{R}^2} \operatorname{argmin} L_S(\underline{Y}, \overline{Y}, \Gamma)$$

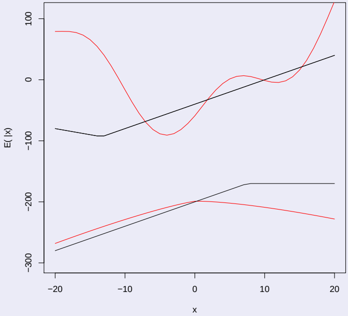


Sharp Setloss Region

c) Sharp Setloss Region:

$$L_S(\underline{Y}, \overline{Y}, \Gamma) = \int \left[\mathbb{E}(\overline{Y} \mid x) - \sup_{\beta \in \Gamma} (\beta_0 + \beta_1 x) \right]^2 + \left[\mathbb{E}(\underline{Y} \mid x) - \inf_{\beta \in \Gamma} (\beta_0 + \beta_1 x) \right]^2 d\mathbb{P}(x)$$

$$SSR(\underline{Y}, \overline{Y}) = \bigcup_{\Gamma \subseteq \mathbb{R}^2} \operatorname{argmin} L_S(\underline{Y}, \overline{Y}, \Gamma)$$

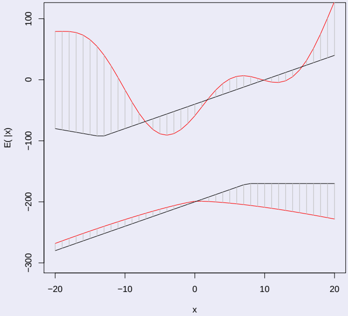


Sharp Setloss Region

c) Sharp Setloss Region:

$$L_S(\underline{Y}, \overline{Y}, \Gamma) = \int \left[\mathbb{E}(\overline{Y} \mid x) - \sup_{\beta \in \Gamma} (\beta_0 + \beta_1 x)\right]^2 + \left[\mathbb{E}(\underline{Y} \mid x) - \inf_{\beta \in \Gamma} (\beta_0 + \beta_1 x)\right]^2 d\mathbb{P}(x)$$

$$SSR(\underline{Y}, \overline{Y}) = \bigcup_{\Gamma \subseteq \mathbb{R}^2} \operatorname{argmin} L_S(\underline{Y}, \overline{Y}, \Gamma)$$

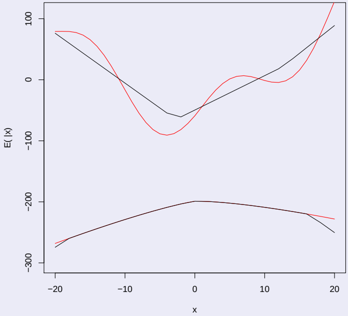


Sharp Setloss Region

c) Sharp Setloss Region:

$$L_S(\underline{Y}, \overline{Y}, \Gamma) = \int \left[\mathbb{E}(\overline{Y} \mid x) - \sup_{\beta \in \Gamma} (\beta_0 + \beta_1 x) \right]^2 + \left[\mathbb{E}(\underline{Y} \mid x) - \inf_{\beta \in \Gamma} (\beta_0 + \beta_1 x) \right]^2 d\mathbb{P}(x)$$

$$SSR(\underline{Y}, \overline{Y}) = \bigcup_{\Gamma \subseteq \mathbb{R}^2} \operatorname{argmin} L_S(\underline{Y}, \overline{Y}, \Gamma)$$



Relations between SMR, SCR and SSR

- $SSR \supseteq SMR \subseteq SCR$

Relations between SMR, SCR and SSR

- $SSR \supseteq SMR \subseteq SCR$
- $SSR \subseteq SCR$?

Relations between SMR, SCR and SSR

- $SSR \supseteq SMR \subseteq SCR$
- $SSR \subseteq SCR$? *"often approximately"*

region	well-specified case	misspecified case
SMR	• true parameters $(\beta_0, \beta_1) \in SMR$ • not "well-estimable"	• ?
SSR	• true parameters $(\beta_0, \beta_1) \in SSR$ • "well-estimable" (e.g. replace expectations by means)	• good description of the bounds $\underline{Y}, \overline{Y}$ and thus good description of Y
SCR	• true parameters $(\beta_0, \beta_1) \in SCR$ • "well-estimable" (e.g. replace expectations by means)	• set of all possible good linear descriptions for every $Y \in [\underline{Y}, \overline{Y}]$

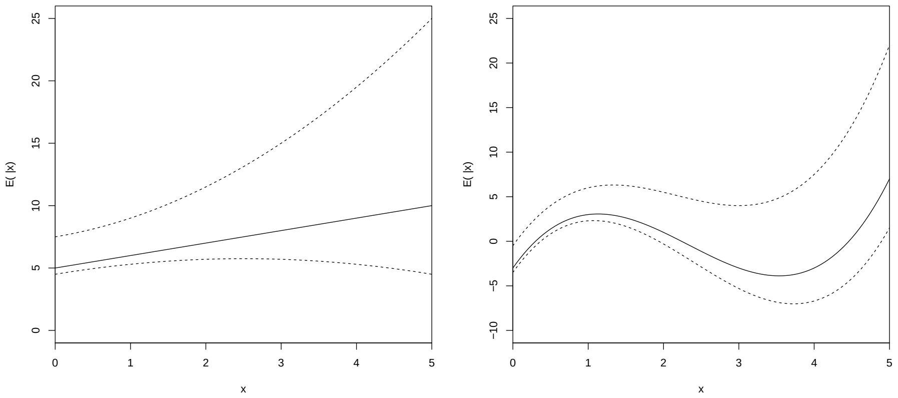


Figure: The conditional expectations $\mathbb{E}(Y \mid x)$ (black), $\mathbb{E}(\underline{Y} \mid x)$ (dashed) and $\mathbb{E}(\overline{Y} \mid x)$ (dashed) for a well-specified (left) and a misspecified (right) situation.

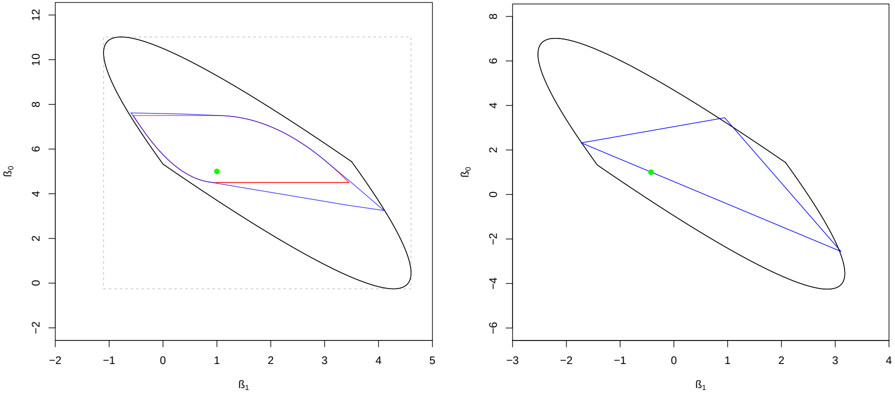


Figure: The identification regions SMR (red), SCR (black) and SSR (blue) for a well-specified (left) and a misspecified (right) situation. The true Parameter is dotted green (for the misspecified situation the green point is the best linear predictor for y). In the misspecified case SMR is empty.

- Beresteanu, A., & Molinari, F. (2008): Asymptotic properties for a class of partially identified models, Econometrica, 76, 763–814.
- Černý, M. & Rada M. (2011): On the possibilistic approach to linear regression with rounded or interval-censored data. *Measurement Science Review*, 11, 34–40.
- Ponomareva, M., & Tamer, E. (2011): Misspecification in moment inequality models: back to moment equalities?. *Econometrics Journal*, 14, 186-203.
- Stoye, J. (2007): Bounds on generalized linear predictors with incomplete outcome data. *Reliable Computing*, 13, 293–302.

region	well-specified case	misspecified case
SMR	• truth $\in SMR$ • not "well-estimable" • monotone • idempotent	• ?
SSR	• truth $\in SSR$ • "well-estimable" • not monotone • idempotent	• good description of the bounds $\underline{Y}, \overline{Y}$ and thus good description of Y
SCR	• truth $\in SCR$ • "well-estimable" • monotone • not idempotent and thus maybe too rough	• set of all possible good linear descriptions for every $Y \in [\underline{Y}, \overline{Y}]$