Computing Simple Bounds for Regression Estimates for Linear Regression with Interval-valued Covariates

Basic Situation

- Simple linear model under interval-valued covariates:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n \tag{1}$$

$$x_i \in [\underline{x}_i, \overline{x}_i] \quad a.s., \qquad i = 1, \ldots, n. \tag{2}$$

- $(\varepsilon_1, \ldots, \varepsilon_n)$ assumed i.i.d. with expectation 0 and variance σ^2 (but can be relaxed).
- y_i precisely observed.
- x_i only observed in intervals (epistemic data imprecision).
- Because the x_i's are not precisely observed, the model is generally only partially identified.

Partial Identification

- Set-valued estimator for the best linear predictor:

$$OLS = \bigcup \left\{ \underset{\beta}{\operatorname{argmin}} \{||X\beta - y||_2\} \mid X \in [\underline{X}, \overline{X}] \right\}. \tag{3}$$

- Under certain assumptions, this set-valued estimator converges to the sharp identification region for the best linear predictor. However, computing *OLS* is very difficult. (Already computing exact bounds for $\hat{\sigma}^2$ is **NP**-hard.)
- Here, we are only interested in the slope-parameter β_1.

Approach 1: Interval-arithmetic

- $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - mean(x))(y_i - mean(y))}{\sum_{i=1}^{n} (x_i - mean(x))^2}. \tag{4}$$
- Then, apply interval-arithmetic (for simplicity separately for the nominator and the denominator).

Approach 2: Reverse Regression and Analytical Bounds

- Firstly, regress x on y: $\beta_{xy} = \left[(Y'Y)^{-1}Y'x\right]_{21}$.
- Only x is interval-valued and β_{xy} is linear in x.
- $OLS_{xy} = \{\beta_{xy} \mid x \in [\underline{x}, \overline{x}]\}$ and especially the minimal slope parameter for the reverse regression $\underline{\beta_{xy}}$ is easy to compute.
- Since $|\beta_{yx}| \leq \frac{1}{|\beta_{xy}|}$ (Cauchy-Schwarz inequality), we have $\overline{\beta_{yx}} \leq \frac{1}{\underline{\beta_{xy}}}$ (for positive slope parameters).
- This gives an upper bound for $\hat{\beta}_1$.

Approach 3: Replacing OLS by Another Estimator

- Use another estimator that is linear in y:
- $\hat{\beta}_1 = \sum_{j>i} \alpha_{ji} \cdot \frac{y_j - y_i}{x_j - x_i}$ with coefficients $\alpha_{ji} \geq 0$ and $\sum_{j>i} \alpha_{ji} = 1$.
- This is a convex combination of all the simple estimates $\frac{y_j - y_i}{x_j - x_i}$ for the slope based on pairs of two data points.
- This estimator is unbiased and the variance can be minimized by optimizing the variance in dependence on the coefficients α_{ji}.
- **Theorem 1:** For precise x, this estimator is exactly the OLS-estimator.
- For interval-valued x, simply apply interval-arithmetic to all the estimates $\frac{y_j - y_i}{x_j - x_i}$.
- Conservative confidence intervals are also attainable by estimating an upper bound for $\hat{\sigma}^2$ and by analyzing the coefficients α_{ji}.

Results and Outlook

- Approach 3 usually gives the sharpest bounds.
- Further possible modifications of approach 3:
- Replace weighted mean by weighted median to obtain more robust estimates.
- Also for confidence intervals, more robust estimates for the scale parameter are thinkable.
- One can also adjust for possible heteroscedasticity.
- Does also work for imprecise y.
- Also applicable for multiple linear regression. Open question: Is there a generalization of Theorem 1 for the case of multiple linear regression?

Georg Schollmeyer, LMU Munich

georg.schollmeyer@stat.uni-muenchen.de