Recent Developments in the Statistical Analysis of Interval Data

The Case of Regression

Ulrich Pötter[1] Georg Schollmeyer[2] Thomas Augustin[2]
Marco Cattaneo[2] Andrea Wiencierz[2]

[1] German Youth Institute (DJI) [2] Ludwig-Maximilians University (LMU)
Munich, Germany

Applied Statistics 2012
Ribno, September 23rd 2012

Interval Data

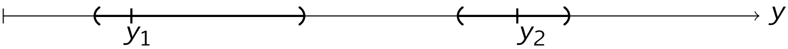
Data are often observed or recorded imprecisely. They may be grouped, censored, coarsened to some extend.

The situation may be represented by interval valued data in the form of

$$y^* := [\underline{y}, \overline{y}] = \{(y_1, \dots, y_n) | \underline{y}_1 \leq y_1 \leq \overline{y}_1, \dots, \underline{y}_n \leq y_n \leq \overline{y}_n\}.$$

where it is assumed that the intervals contain the actual data

$$y = (y_1, \dots, y_n), \quad y_i \in [\underline{y}_i, \overline{y}_i]$$



Interval Data

Consequence: An additional type of uncertainty apart from classical statistical uncertainty. This uncertainty can't be decreased by sampling more data.

Two Approaches

1. Likelihood inference based on a non-parametric model of interval-valued data.
2. All least-squares projections compatible with the interval-valued data.

Profile Likelihood

Probability Model

Joint distribution of exact and interval-valued random variables with marginal distributions P (exact data) and P^* (interval-valued data):

$$\begin{tikzcd}
\Omega \arrow[r] \arrow[d] & \mathcal{Y}^* \sim P^* \\
\mathcal{Y} \sim P &
\end{tikzcd}$$

with the consistency condition:

$$\Pr\big(Y \in Y^*\big) = 1$$

Consider all statistical models which are plausible enough in the light of the observed (interval-valued) data.

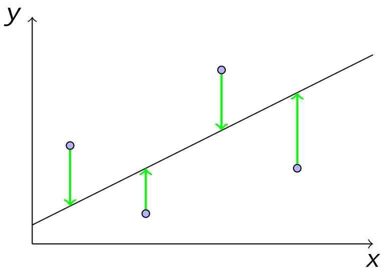
Profile Likelihood

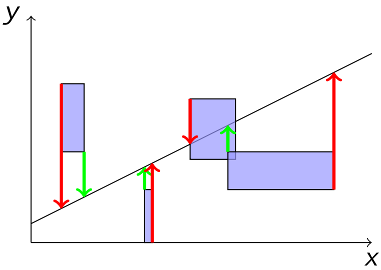
Likelihood

$$\mathcal{L}(P; y^*) = \sup_{\{P^* \text{ compatible with } P\}} P^*(y^*)$$

Profile Likelihood

Look at the residuals:





Since the data are interval-valued, the residuals are interval-valued as well.

Minimize the median (or another quantile) of the absolute residuals

Profile Likelihood

- Compute all linear models for which the median of the residuals is not dominated by the residuals of another linear model.
- The set of all undominated models is the final estimate.

This method is a generalization of the least median of squares method. It is implemented in the package linLIR available from CRAN.

Profile Likelihood

- Compute all linear models for which the median of the residuals is not dominated by the residuals of another linear model.
- The set of all undominated models is the final estimate.

This method is a generalization of the least median of squares method. It is implemented in the package linLIR available from CRAN.

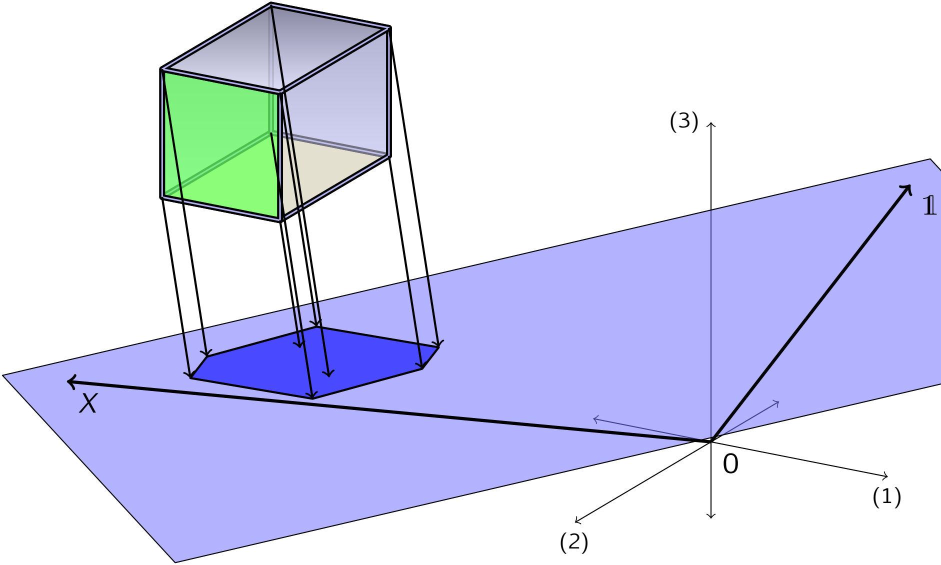
All Consistent Projections

All Consistent Least-Squares Solutions

General idea: Consider the set of all estimates obtained by applying the estimator to all exact observations compatible with the interval-valued data.

Linear regression: Apply the least-squares estimator to all possible y consistent with the interval-valued data $[\underline{y}, \overline{y}]$. I.e. take the set of all orthogonal projections of $y \in y^*$ on the space spanned by the covariates x as reasonable estimates.

All Consistent Projections



All Consistent Projections

- The least-squares estimator is linear in the dependent variables.
- Thus it is easy to compute the image of set-valued data $[\underline{y}, \overline{y}]$ under a linear mapping:

  It is essentially the computation of Minkowski-sums whose computational aspects are well studied in computational geometry.

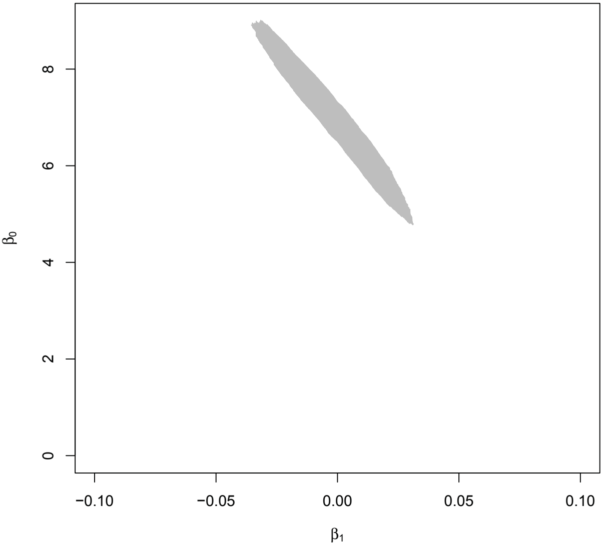
Example

German General Social Survey (ALLBUS) 2008:
y ... log of income (interval-valued)
x ... age (precise)

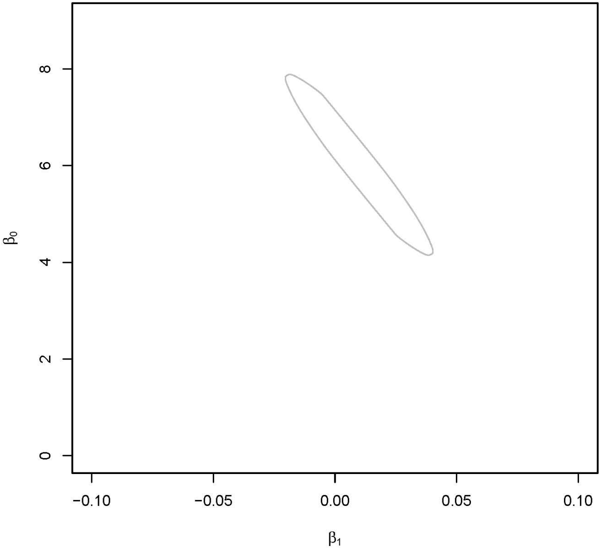
1067 observations from Eastern Germany with some information on income and age.

25% reported only income brackets.

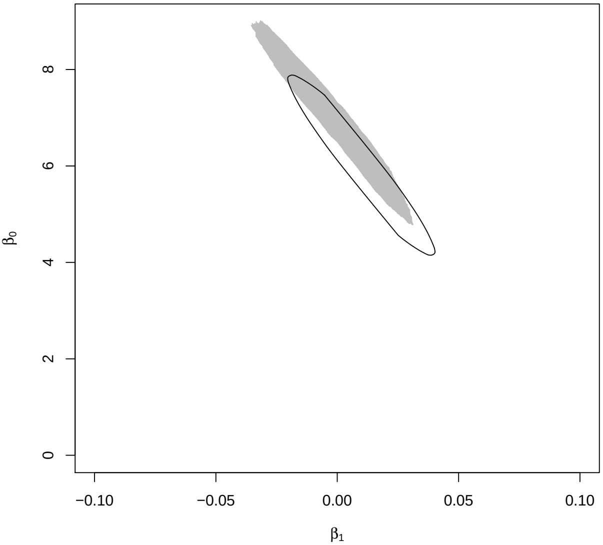
Profile Likelihood



All Consistent Projections



Comparison



Comparison

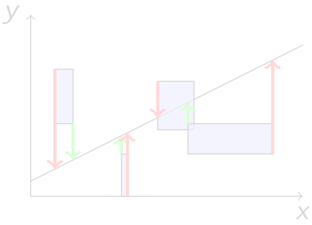
- The set of solutions in the all-projections approach is always convex, thus easy to describe and to handle.
  In contrast, the set of solutions in the profile likelihood approach need not be convex and may be hard to characterize completely.
- The computational complexity of the all-projections approach is of the order $O(n)$, the one of the profile likelihood approach is $O(n^3 \log(n))$. In terms of real computation time, the latter may take much longer than the former.
- The profile likelihood approach can easily be adapted to other forms of coarsened data including gross reporting errors and misclassifications. It can be used to estimate general regression functions and other parameters of interest.
  The all-projections approach is restricted to the situation of least-squares computations in linear models.

Comparison

- The all-projections approach inherits the non-robustness from the least-squares estimator.

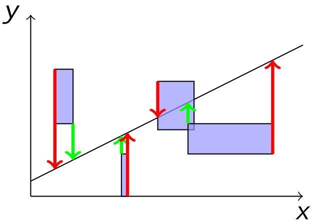
  In contrast, the profile likelihood approach uses (outlier) robust quantiles in its construction and can be expected to be much more robust.

  However, notions of robustness are not straight forwardly transferable to the coarsened data context.



Comparison

- The all-projections approach inherits the non-robustness from the least-squares estimator.
  In contrast, the profile likelihood approach uses (outlier) robust quantiles in its construction and can be expected to be much more robust.
  However, notions of robustness are not straight forwardly transferable to the coarsened data context.



References I

M. Cattaneo, A. Wiencierz (2012). Likelihood-based Imprecise Regression, International Journal of Approximate Reasoning, 53 (8), 1137–1154.

A. Beresteanu, F. Molinari (2008). Asymptotic Properties for a Class of Partially Identified Models, Econometrica, 76 (4), 763–814.

Profile Likelihood Regression

Observations $y_1^*,\dots,y_n^*$ induce a (normalized) profile likelihood function for the p-quantile of the distribution of residuals R_f for each set of regression coefficients β.

$$\underline{r}_{\beta,i} = \min_{(x,y)\in[\underline{x}_i,\overline{x}_i]\times[\underline{y}_i,\overline{y}_i]} \left|y - x\beta\right|, \quad \overline{r}_{\beta,i} = \sup_{(x,y)\in[\underline{x}_i,\overline{x}_i]\times[\underline{y}_i,\overline{y}_i]} \left|y - x\beta\right|$$

The result is

$$\mathcal{U} = \{\beta : \underline{r}_{\beta,(\underline{k}+1)} \leq \overline{q}_{LRM}\}$$

where $\overline{q}_{LRM} = \inf_\beta \overline{r}_{\beta,(\overline{k})}$ and where $\underline{k}$ and $\overline{k}$ depend on n, p and a cut-off point of the profile likelihood.

Further details in: M. Cattaneo, A. Wiencierz (2012). *Likelihood-based Imprecise Regression*. Int. J. Approx. Reasoning 53. 1137-1154.