M-Estimation with Imprecise Data

Marco E. G. V. Cattaneo
Department of Mathematics
University of Hull
m.cattaneo@hull.ac.uk

Real data often do not have the level of precision required by conventional statistical methods. In particular, a data point can be incompletely observed, in the sense that the only available observation is a set known to contain the data point. An important problem is then how to perform statistical estimation, and in particular regression, when some (or all) data points are incompletely observed. This problem has recently attracted much attention in the statistical literature in general, and at ISIPTAs in particular: see for example Cattaneo and Wiencierz (2012); Liu and Vandal (2011); Schollmeyer and Augustin (2015); Utkin and Coolen (2011).

The typical setting in these works is that instead of the precise data points $x_i \in \mathcal{X}$, only the sets $s_i \subseteq \mathcal{X}$ are observed. It is assumed that $x_i \in s_i$, but no other information about x_i is available. In particular, precisely observed data points x_i can be represented by singletons $s_i = \{x_i\}$, while missing data points x_i can be represented by observations $s_i = \mathcal{X}$. The statistical problem consists in estimating a quantity of interest $\theta \in \Theta$ on the basis of the data.

In the case of precisely observed data, most statistical estimation methods can be expressed as M-estimators (or slight generalizations thereof):

$$\hat{\theta}(x_1, \ldots, x_n) = \arg\min_{\theta \in \Theta} \textstyle\sum_{i=1}^{n} \rho(x_i, \theta), \qquad (1)$$

where $\rho : \mathcal{X} \times \Theta \to \mathbb{R}$ describes some kind of estimation error. For example, when $\mathcal{X} = \Theta = \mathbb{R}$, the squared error $\rho(x_i, \theta) = (x_i - \theta)^2$ leads to the least squares estimation of location.

An apparently very intuitive idea for generalizing an estimator $\hat{\theta}$ to the case of incompletely observed data is to interpret

$$\{\hat{\theta}(x_1, \ldots, x_n) : x_i \in s_i\} \qquad (2)$$

as the set-valued estimate based on the observations s_i. However, for M-estimators an alternative approach is possible: replacing $\sum_{i=1}^{n} \rho(x_i, \theta)$ with

$$\{\textstyle\sum_{i=1}^{n} \rho(x_i, \theta) : x_i \in s_i\} \qquad (3)$$

(or its convex hull) in the minimization task (1). Since the quantity (3) to be minimized is set-valued, several definitions of minimum are possible and can lead to different kinds of estimators.

The present work investigates the imprecise minimization approach (3) and compares it with the set of estimates approach (2). Both approaches have interesting connections with the statistical method of estimating equations, and face some difficulties in parametric models. An important advantage of the former is the possibility, if desired, of easily obtaining a precise estimate, for example by interpreting the minimization as a minimax problem. By contrast, the interpretation of the set-valued estimates intrinsically tied to the latter approach is difficult, because they mix aspects of the different statistical concepts of point estimate and confidence region.

Keywords. M-estimator, regression, imprecise data, interval data, coarse data, missing data, estimating equations, robust statistics, set-valued estimates.

References

Cattaneo, M., and Wiencierz, A. (2012). Likelihood-based Imprecise Regression. *Int. J. Approx. Reasoning* 53, 1137–1154. [based on an ISIPTA '11 paper]

Liu, X., and Vandal, A. C. (2011). Bounds for self-consistent CDF estimators for univariate and multivariate censored data. In *ISIPTA '11*, 267–276.

Schollmeyer, G., and Augustin, T. (2015). Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *Int. J. Approx. Reasoning* 56, 224–248. [based on an ISIPTA '13 paper]

Utkin, L. V., and Coolen, F. P. A. (2011). Interval-valued regression and classification models in the framework of machine learning. In *ISIPTA '11*, 371–380.