High breakdown point robust regression with censored data

Matías Salibian-Barrera[1] and Víctor J. Yohai[2]

[1] Carleton University, School of Mathematics and Statistics,1125 Colonel By Drive, Room 4302HP, Ottawa, ON, Canada, K1S 5B6

[2] University of Buenos Aires, Departamento de Matemáticas, Pabellón 1, Ciudad Universitaria 1428 Buenos Aires, Argentina

Keywords: linear regression, robust estimates, censored data, high breakdown point

Abstract

We consider the linear regression model

$$y_i = \beta' \mathbf{x}_i + u_i$$

whith right censoring. Then the observed sample is $\mathbf{z}_i = (y_i^*, \mathbf{x}_i, \delta_i), 1 \leq i \leq n$, where

$$y_i^* = \min(y_i, c_i)$$

and $\delta_i = I_{\{y_i \leq c_i\}}$.

M-estimates in the non censoring case are defined by

$$\sum_{i=1}^{n} \rho(r_i(\beta)) = E_{F_{n\beta}}(\rho(u)) \ = \min!, \tag{1}$$

where $F_{n\beta}$ is the empirical distribution of

$$r_i(\beta) = y_i - \beta' \mathbf{x}_i.$$

M-estimates also satisfy

$$\sum_{i=1}^{n} \psi(r_i(\beta)) \mathbf{x}_i = E_{H_{n\beta}}(\psi(u) \mathbf{x}) \ = \mathbf{0}, \tag{2}$$

where $\psi = \rho'$ and $H_{n\beta}$ is the empirical distribution of $(r_1(\beta), \mathbf{x}_1), (r_2(\beta), \mathbf{x}_2), ..., (r_n(\beta), \mathbf{x}_n)$.

Let F_β be the distribution of $r_i(\beta)$. Since the y_i are not available, one way to generalize (1) and (2) is replacing these equations by

$$\sum_{i=1}^{n} E(\rho(r_i(\beta)|\mathbf{z}_i)) = \sum_{i=1}^{n} E_{F_\beta}(\rho(u)|\mathbf{z}_i)) \ = \ min!$$

and

$$\sum_{i=1}^{n} E(\psi(r_i(\beta)_i|\mathbf{z}_i))\mathbf{x}_i = \sum_{i=1}^{n} E_{F_\beta}(\psi(u)|\mathbf{z}_i))\mathbf{x}_i = \mathbf{0}.$$

Since F_β is unknown, we can replace this distribution by a nonparametric estimate based on the censored residuals $r_i^*(\beta) = y_i^* - \beta' \mathbf{x}_i$. A natural choice is the Kaplan-Meyer estimate $F_{n\beta}^*$. However, $r_i^*(\beta)$ is independent of the corresponding censoring time $c_i - \beta' \mathbf{x}_i$only when $\beta = \beta_0$. Therefore the consistency of $F_{n,\beta}^*$ to F_β is only guaranteed under the true value. Then, the estimate defined by

$$\sum_{i=1}^{n} \ E_{F_{n\beta}^*}(\rho(u)|\mathbf{z}_i)) \ = \ min! \tag{3}$$

is not consistent.

On the other hand, the estimate defined by

$$\sum_{i=1}^{n} E_{F^*_{n\beta}}(\psi(u)|\mathbf{z}_i))\mathbf{x}_i = \mathbf{0} \tag{4}$$

is Fisher consistent. M-estimates defined by (4) were first proposed by Ritov (1990) and further studied by Li and Ying (1994).

The solution of (4) is well defined only when ψ is non decreasing. However, it is well know that M-estimates with nondecreasing ψ are only robust against low leverage outliers and high leverage outliers may have a large influence on these estimates. Therefore, it is desirable to define M-estimates with a redescending ψ. Unfortunately, for redescending ψ (4) has, in general, several solutions and not all of them correspond to consistent estimates.

For this reason we must modify (3) to get consistent estimates with high breakdown point. Define

$$C_n(\beta, \gamma) = \sum_{i=1}^{n} E_{F^*_{n\beta}}(\rho(u - \gamma'\mathbf{x}_i)|\mathbf{z}_i))$$

and

$$\gamma_n(\beta) = \arg\min C_n(\beta, \gamma).$$

Observe that since $F^*_{n\beta_0}$ is a consistent estimate of F, the distribution of the error u, we have that

$$\gamma_n(\beta_0) \to \mathbf{0}.$$

Then a Fisher consistent estimate of β_0 is defined by the equation

$$\gamma_n(\widehat{\beta}) = \mathbf{0} \tag{5}$$

The estimate defined by (5) may be considered an extension of the Ritov M-estimates for bounded ρ functions. Using the same ideas we can also extend other high breakdown point estimates as the least median of squares estimate (Rousseeuw, 1984), S-estimates (Rousseeuw and Yohai, 1984), MM-estimates (Yohai, 1987) and τ-estimates (Yohai and Zamar, 1988) to the case of censored data.

References

Li, T. L. and Ying, Z. (1994). A missing information principle and M-estimators in regression analysis with censored and truncated data. Annals of Statistics, **22**, 1222-1255.

Ritov, Y. (1990). Estimation in the linear model with censored data. Annals of Statistics, **18**, 303-328/

Rousseeuw, P. J. (1984). Least median of squares regression, J. Amer. Statist. Assoc., **79**, 871–880.

Rousseeuw P. J. and Yohai, V. J. (1984). Robust regression by means of S–estimators, in *Robust and Nonlinear Time Series Analysis*, J. Franke, W. Hardle, and R. D. Martin (eds.), Lecture Notes in Statistics, **26**, Springer, New York, 256–272.

Yohai, V. J. (1987). High breakdown-point and high efficiency M-estimates for regression. *Annals of Statistics*, **15**, 642-656.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown–point estimates of regression by means of the minimization of an efficient scale, *J. Amer. Statist. Assoc.*, **83**, 406–413.

Please fill in this form and mail it together with your abstract.

My abstract fits best to topic number 20 (**"PLEASE INDICATE HERE THE NUMBER FROM THE LIST OF TOPICS BELOW WHICH BEST FITS TO YOUR ABSTRACT"**.).

List of Topics:

1. Algorithms
2. Applications
3. Biostatistics
4. Computing and graphics
5. Data analysis
6. Data mining
7. Economics, finance
8. Efficiency and robustness
9. Functionals and bias
10. Fuzzy statistics
11. Geostatistics
12. Inference for robust methods, model testing
13. Location depth and regression depth
14. Multivariate methods
15. Neural networks
16. Rank-based methods
17. Regression quantiles, trimming
18. Robust covariance
19. Robust designs
20. Robust regression
21. Time series analysis
22. Wavelets
23. Other (please specify)