Robust regression: the influence of ψ functions, improving scale and covariance matrix estimates for tests, confidence and prediction intervals

Presenter: **Manuel Koller**, ETH Zürich, SfS, Switzerland,
`koller@stat.math.ethz.ch`

Co-authors: Prof. Werner A. Stahel

1 Goals. A goal of robust statistics is to perform linear regression with a high breakdown point as well as high efficiency of the estimate. For inference, the scale parameter and covariance matrix of the estimated coefficients must be estimated. The resulting tests or confidence intervals should adhere to their nominal levels, and prediction intervals should have correct coverage. More precisely, we consider ϵ-contamination neighborhoods V_ϵ around a central model H_0, the former is defined as $V_\epsilon(H_0) = \{H = (1-\epsilon)H_0 + \epsilon H^* : H^* \text{ arbitrary distribution}\}$. We measure the performance of the estimates at the central model. For example, prediction intervals are supposed to cover "the good part" of the data.

2 Program. In this talk we focus on MM-estimators, consisting of an S-estimator producing the scale $\hat\sigma_S$ and an M-estimator defined by the ψ-function ψ.

The estimator of the covariance matrix of the regression parameters β can be split into three parts: a scale, a correction factor based on the asymptotic normality of M-estimates, and a matrix part:

$$\widehat{\mathrm{cov}}(\hat\beta) = \hat\sigma^2 \gamma \mathbf{V}_{\mathbf{X}}^{-1}. \tag{1}$$

It is plausible that the above goals can be fulfilled if the procedures have the following properties: The scale estimate is unbiased and has high efficiency; the correction factor is correct at the central model; and the matrix part excludes "discarded" observations. All estimators should have the rejection property: the estimate should be invariant w.r.t. the inclusion of a clear outlier in the dataset.

3 $\hat\sigma$ Part. The scale estimate $\hat\sigma_S$ of the initial S-estimate does not have the rejection property described above. We therefore estimate the scale again, based on the final residuals $(r_1, \ldots, r_n)$, by solving

$$\sum_i \tau_i^2 \left[\chi\left(\frac{r_i}{\tau_i \hat\sigma_D}\right) - \kappa\nu\left(\frac{r_i}{\tau_i \hat\sigma_D}\right)\right] = 0. \tag{2}$$

The τ_i are used to standardize the residuals. By replacing sums with integrals over the approximated distribution of r_i we get a defining equation for τ_i (this was covered in last year's talk, see also [2]).

For the functions χ and ν we choose $\nu(r) = \psi(r)/r$ and $\chi(r) = \nu(r)r^2$. Note that ν gives the robustness weights w_i of the final regression M-estimate when applied to the residuals scaled by the initial scale estimate.

4 γ Part. Asymptotic normality of M-estimates gives us a correction factor $\gamma = \mathrm{E}[\psi(u/\sigma)^2]/(\mathrm{E}[\psi'(u/\sigma)])^2$, where $u \sim H_0$. This is usually estimated using the empirical distribution of the residuals and is multiplied by a factor $n/(n-p)$. For small samples this estimate requires an additional correction factor as outlined in [1].

We propose to use either the empirical distribution of the τ standardized residuals (without the small sample correction) or just the expected values evaluated at the central model.

5 $\mathbf{V}_{\mathbf{X}}^{-1}$ Part. We second the proposal of [4], to use

$$\mathbf{V}_{\mathbf{X}} = \sum_i w_i \mathbf{x}_i \mathbf{x}_i^T / (\sum_i w_i/n). \tag{3}$$

6 Results. It turns out that the choice of ψ-function is crucial for the tests and intervals to maintain the specified levels. While keeping the maximum asymptotic bias reasonably small, we stress the need to use slowly redescending ψ-functions such as Hampel (decent rate 1/2 or lower) or a modification of the Welsch. We define the latter as

$$\psi(x, c) = \begin{cases} x & |x| \le c \\ \exp\left(-\frac{1}{2}\frac{(|x|-c)^b}{a}\right) x & |x| > c \end{cases} \tag{4}$$

Using such a ψ-function and the proposed covariance matrix estimate showed in simulations to keep the specified levels even for n/p ratios as low as 5.

References

[1] P. J. Huber, E. M. Ronchetti. *Robust Statistics, Second Edition.* Wiley and Sons Inc., New York, 2009.

[2] M. Koller. *Robust Statistics: Tests for Robust Linear Regression.* Master's Thesis, 2008. http://stat.ethz.ch/research/dipl_arb/2008/koller.pdf

[3] Maronna, R.A. and Yohai, V.J. Correcting MM estimates for fat data sets Computational Statistics & Data Analysis, Elsevier, 2009.

[4] V. J. Yohai, W. A. Stahel, R. H. Zamar. A procedure for robust estimation and inference in linear regression. *Directions in Robust Statistics and Diagnostics (Part II).* W. A. Stahel and S. Weisberg (eds.). The IMA Volumes in Mathematics and its Applications, 365-374, Springer New York, 1991.

[5] Yohai, V. and Zamar, R. Optimal locally robust M-estimates of regression Journal of Statistical Planning and Inference 64(2): 309-323, Elsevier, 1997.