THE TEACHING OF ROBUST STATISTICS FOR REGRESSION

Víctor J. Yohai
Universidad de Buenos Aires and CONICET, Argentina
vyohai@dm.uba.ar

We present some ideas on how to teach Robust Regression. We motivate the importance of robust estimation using a real data set and briefly discuss why diagnostic procedures based on the least squares estimates do not guarantee the detection of outlier observations. We also introduced regression M-estimates and discuss why they require an estimate of the error scale. Finally we introduce the class of S-estimates to obtain a robust estimate of the error scale.

WHY ROBUST ESTIMATES?

To motivate the introduction of robust estimates for regression we start by pointing out the shortcoming of the least squares estimates

Let us consider the linear regression model

$$y \;=\theta' x_i + u_i , i = 1,...,n$$

$\boldsymbol{x}_i$=(x_{i1},...,x_{ip}) ∈ R^p, where x_i and u_i are independent.

Least squares (LS) estimates are defined by

$$\sum_{i=1}^{n} r_i^2(\theta) \text{=minimum}$$

where

$$r_i(\theta) = y_i \theta' x_i \, .$$

The students probably know that these estimates are optimal when the u_i's are normal. It is well known that in this case the LS estimate is the maximum likelihood estimate (MLE). However, a few atypical observations (outliers) can have a large influence on the LS-estimate. Even one outlier can take the LS-estimate beyond any bound. The sensitivity to outliers of the LS estimate is illustrated with the following example.

We consider an experiment on the speed of learning of rats. Times were recorded for a rat to go through a shuttlebox in successive attempts. If the time exceeded 5 seconds, the rat received an electric shock for the duration of the next attempt. The data consist of the number of shocks received by the rat and the average time for all attempts between shocks.

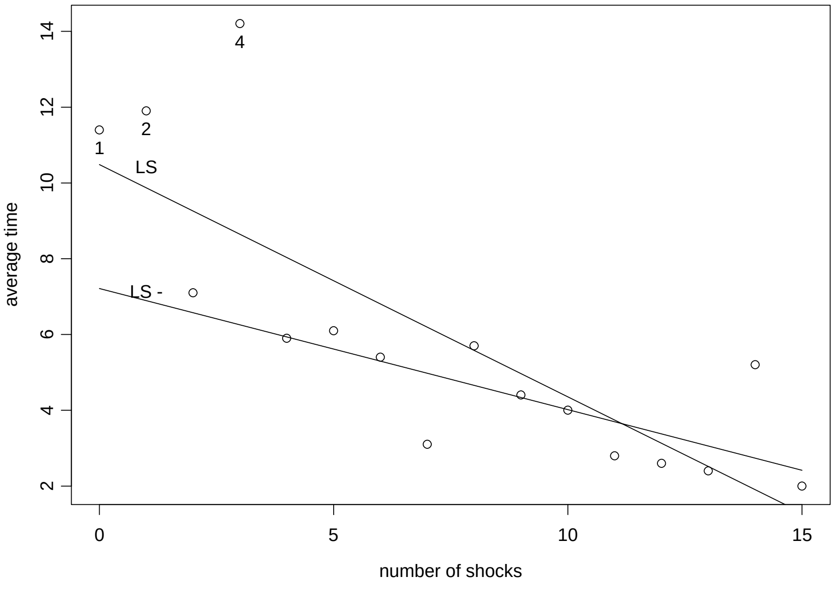

Figure 1: Shock data- LS fits with all data (LS) and omitting points 1-2-4. (LS-)

Figure 1 show two lines fitted using the LS estimate, one using the full set of data and the second the line obtained after deleting the outlying observations 1, 2 and 4. We observe that that after deleting these three observations the LS line fits much better the bulk of the data.

One common belief is that the outliers have large LS residuals, and that therefore they can be detected looking at residuals plots. The next Figure shows the QQ plot of the standardized residuals for the learning experiment. We observe that the observations 1 and 2, which are outliers, do not have significant large residuals.

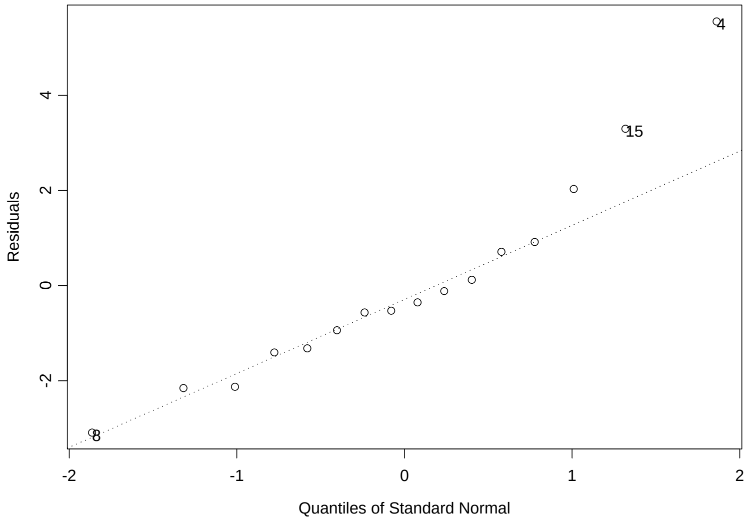

Figure 2: Residuals QQ-plot for the shock data

As we have seen in the shock example, a small number of outliers may have a large influence on the LS estimate. This motivates the introduction of the concept of robust estimates. A robust estimate is an estimate that is not much influenced by a small percentage of outliers. That is, a robust estimate does not try to fit all the observations; it only tries to fit the bulk of the data.

M-ESTIMATES

The first class of robust estimates that we present here is the class of *M*-estimates, introduced by Huber (1964) for the location model and Huber (1973) for the regression model. *M*-estimates of regression are defined by

$$M(\theta)=\sum_{i=1}^{n}\rho(r_i(\theta))=\text{minimum}$$

where $\rho(u)$ is an even function, non-decreasing for positive u.

To have robustness properties, the function ρ should increase slower than u^2 and in this way give less weight to outliers. For example

$$\rho(u)=|u|$$

This is least absolute value (LAV) or L1-estimate. However a shortcoming of this estimate is that the asymptotic efficiency of this estimate for normal errors is $(2/\pi)=0.637$ which is a rather low value. It would be convenient to combine robustness with efficiency under the regression model with normal errors. For this purpose Huber (1964) proposed to take ρ in the following family

$$\rho^H(u)=\begin{cases} u^2 & \mathit{if} \quad |u|\le c \\ 2c|u|-c^2 & \mathit{if} \quad |u|>c \end{cases}$$

which is quadratic as the square function used for the LS estimate when $|u|$ is small and is linear like the absolute value function for large $|u|$.

These functions may be considered as intermediate functions between $|u|$ and u^2. One common feature of the ρ functions corresponding to the LAV estimate and those of the family $\rho^H(u)$ is that they are convex and that $\psi=\rho'$ is bounded and non-decreasing.

However we will show in the class with simulated and real examples that convex ρ-functions with bounded ψ define *M*-estimates, which are only robust, when the outliers have low leverage. Low leverage outliers are those outliers where the vector of independent variables is not an outlier.

To obtain estimates which are robust against low and high leverage outliers it is required that ρ be bounded. One family of bounded ρ functions is the bisquare family proposed by Tukey

$$\rho^B(u)=\begin{cases}1-(1-(u/c)^2)^3 & if \quad |u|\le c\\ 1 & if \quad |u|>c\end{cases}$$

Differentiating $S(\boldsymbol{\theta})$ we obtain the estimating equation of *M*-estimates

$$\sum_{i=1}^{n}\psi(r_i(\theta))x_i=0$$

where $\psi=\rho'$. For ρ convex, ψ is non-decreasing and when ρ is bounded ψ is redescending.

The problem of solving the estimating equation is equivalent to the minimization problem when ρ is convex. However, if ρ is bounded, the estimating equation may have many solutions corresponding to local minima and maxima. This will considerably complicate the computation of these estimates. Yohai (1987) study a class of robust *M*-estimates based on a bounded ρ function, which are simultaneously highly robust and highly efficient under normal errors.

The *M*-estimates that we have presented are not scale equivariant, i.e., they are not independent of the system of units. In order to get scale equivariance, the definition of *M*-estimates should be modified. Then we define the *M*-estimates by

$$M(\theta)=\sum_{i=1}^{n}\rho\left(\frac{r_i(\theta)}{s_n}\right)=\text{ minimum}$$

where s_n is an estimate of the scale of the residuals and if we want to keep the *M*-estimate robust, s_n should be a robust estimate too.

The next problem that we are going to address is how to obtain a robust residual scale estimate. One way of doing this is using a robust estimate that does not require a previous scale. In the next section we introduce a class of estimates with this property

ESTIMATES BASED ON A ROBUST SCALE

One class of estimates that do not require a previous scale is the class of estimates based on the minimization of a residual scale. Given a sample $u_1,...,u_n$, a scale estimate measures the largeness of the sample independent of the sign of their elements. They should satisfy the following properties:

$$S(u_1,...,u_n)\geq 0,\ S(u_1,...,u_n)=S(|u_1|,...,|u_n|),\ S(\lambda u_1,...,\lambda u_n)=|\lambda|\ S(u_1,...,u_n)$$

The square root mean squares scale is defined by

$$S_n^{\ 0}(u_1,...u_n)=\left(\frac{1}{n}\sum_{i=1}^{n}u_i^{\ 2}\right)^{1/2}$$

Observe that The LS- estimate can also be defined by

$$S_n^{\ 0}(r_1(\theta),...r_n(\theta))=\text{minimum}$$

Estimates defined by

$$S_n(r_1(\theta),...r_n(\theta))=\text{ minimum}$$

where S_n is an arbitrary scale are called estimated based on a scale. To obtain a robust regression estimate we should use a robust scale.

Given a sample $u_1,...,u_n$, an *M*-estimate of scale $S_n(u_1,...,u_n)$ is defined by the value s_n satisfying

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{u_i}{s_n}\right)=b$$

where $\rho(u)$ is even, non-decreasing for $u>0$ and bounded. Generally b is defined by $b=E_{\varphi}(\rho(u))$, where φ is the N(0,1) distribution. That makes s consistent to the standard deviation for normal data. Observe that if $\rho(u)=u^2$, then the *M*-scale coincides with S_n^0.

Rousseeuw and Yohai (1984) defined *S*-estimates of regression by

$$S_n(r_1(\theta),...r_n(\theta)) = \text{minimum},$$

where S_n is an *M*-scale corresponding to a function ρ.

One important fact is that an *S*-estimate is also an *M*-estimate. In fact, let $\boldsymbol{\theta}$ be the *S*-estimate corresponding to ρ and b. Put

$$s_n = S_n(r_1(\theta),...r_n(\theta))$$

then the *S*-estimate has the property of minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i(\theta)}{s_n}\right)$$

Therefore the *S*-estimates can be thought as self-scaled *M*-estimates whose scale is estimated simultaneously with the regression parameters..

Figure 3 show four lines for the shock data: The LS estimate based on all the observations, the LS after omitting observations 2 and 4. An L1 estimate and an *M*-estimate based on a ρ function belonging to the bisquare family. We observe that the L1 and the *M* estimates are close to the LS estimate after deleting the outlier observations

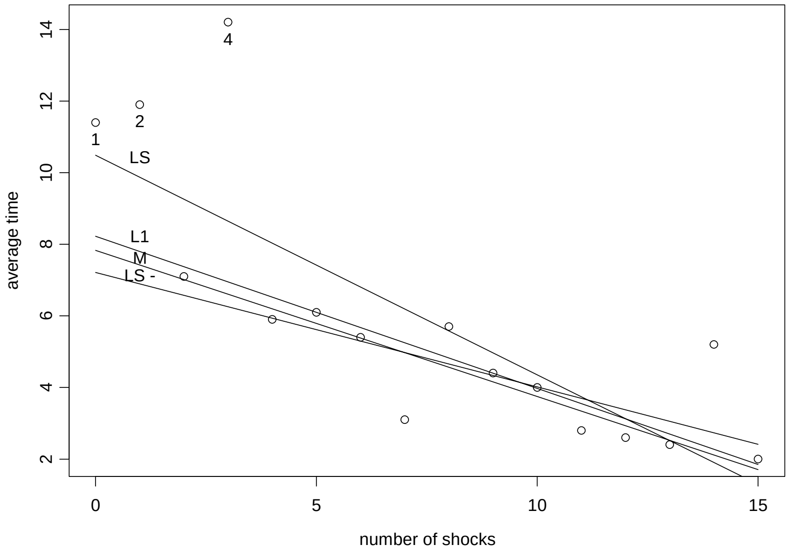

Figure 3: Shock data - The LS estimate with all observations (LS), the LS estimate after deleting observations 1, 2 and 4 (LS-), the L1 estimate (L1) and the M estimate with ρ in the bisquare family and efficiency 0.95 (M)

REFERENCES

Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 53, 73-101.

Huber, P. J. (1973) Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, 1, 799–821.

Rousseeuw, P. and Yohai, V. J. (1984). Robust regression by means of S-estimators. Robust and Nonlinear Time Series Analysis. *Lecture Notes in Statistics,* 26, 256-272. New York: Springer.

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