Robust Inference

October 5, 2015

Robust Regression

In the early 1980's robustness was beginning to develop and much work needed to be done. Most research was focussed on the location-dispersion model and there were no computable robust regression estimates available. Although some modest progress had been made, serious theoretical and computational roadblocks needed to be removed.

In the mid 70's Ricardo Maronna had shown that generalized M-estimates of regression, the most promising regression estimates at that time, loose robustness as the dimension p of the problem increases. He showed that GM-estimates are in fact rather flimsy for $p > 5$. Therefore, an important open question at that time was whether robust regression - in terms of *breakdown point* (BP) was at all possible for high dimensional problems.

Repeated Medians

The first breakthrough in high BP regression is due to Siegel (1982) who introduced the repeated medians regression estimate (RM). Siegel showed that his estimate has $BP = 1/2$, regardless of the dimension of the problem.

The definition and computation of the RM is rather simple in the case of $p = 1$. Given n 2-dimesional points $(y_1, x_1), ..., (y_1, x_1)$ the RM estimate of slope is defined as

$$\hat{\beta}_1 \quad = \quad Med_i \; Med_j \; \left\{ \frac{y_i - y_j}{x_i - x_j} \right\}.$$

The corresponding intercept estimate is defined as

$$\hat{\beta}_0 \quad = \quad Med_i \; \left\{ y_i - \hat{\beta}_1 x_i \right\}.$$

An equivalent definition, which can be easily extended to cases with $p > 1$ (number of covariates) is as follows. For each pair of points (x_i, y_i) and (x_j, y_j) with $x_i \neq x_j$ let $\left(\widehat{\alpha}_{ij}, \widehat{\beta}_{ij} \right)$ be the corresponding LS fit. Now

$$\hat{\beta}_1 \quad = \quad Med_i \; Med_j \; \left\{ \widehat{\beta}_{ij} \right\}$$

$$\hat{\beta}_0 \quad = \quad Med_i \; \left\{ y_i - \hat{\beta}_1 x_i \right\}.$$

Unfortunately, the definition and computation of the repeated median estimate face severe limitations for $p > 2$. One particularly serious problem is that RM is no longer affine-regression equivariant for $p > 2$. This limitation opened the question of whether high BP regression and affine-regression equivariance was at all possible for high dimensional problems.

Least Median of Squares(LMS)

Rousseeuw (1984) broke new ground by showing that the least median of squares (LMS) regression estimate - originally defined by Hampel - answered this question in a positive way. Hampel-Rousseuw's LMS is based on a remarkable idea which we explain in the next two paragraphs.

We have seen before that robust location M-estimates are defined as the solution of an estimating equation with bounded score function $\psi(y)$. This is equivalent to replacing the non-robust quadratic loss function of LS estimates by a robust loss function $\rho(y)$ that ultimately grows linearly as $|t| \to \infty$. Robustness in the location case is therefore attained by replacing the non-robust quadratic loss function (and corresponding non-robust linear score function) by a robust loss function.

Hampel's idea was different and rather ingenious. He reasoned as follows: The LS estimate is not robust because it minimizes the non-robust average of the squared residuals. But we can perhaps achieve robustness by minimizing the robust median of the squared residuals. That is, instead of robustifying the traditional square residuals we could robustify the type of operation we perform over the squared residuals.

Given n $(p+1)$-dimensional points $(y_1, \mathbf{x}_1), ..., (y_1, \mathbf{x}_1)$ the regression LMS estimate is defined as

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1\right) \quad = \quad \arg\min_{(b_0, \mathbf{b})} Med_i \left\{ (y_i - b_0 - \mathbf{x}_i' \mathbf{b})^2 \right\},$$

where b_0 ranges over R and $\mathbf{b}$ ranges over R^p. Notice that we still work with the squared regression residuals $(y_i - b_0 - \mathbf{x}_i' \mathbf{b})^2$ but instead minimizing their average (or sum) we minimize their median.

Peter Rousseuw proved that LMS has $BP = 1/2$, independent from the dimension p of the explanatory variables. Since LMS is also affine-regression equivariant, it became the first affine-regression equivariant regression estimate which is also robust.

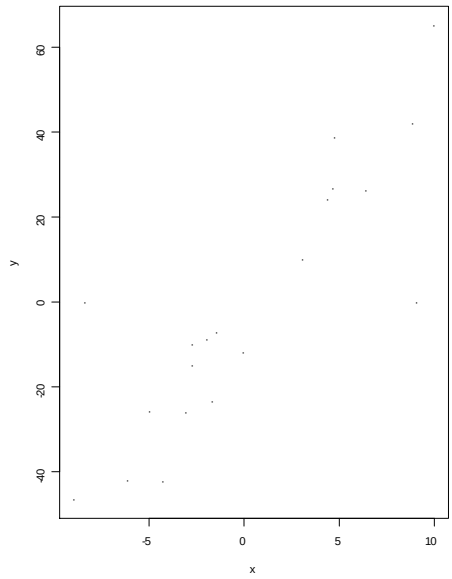

Figure 1: Simulated dataset with 20 points and two outliers.

Resampling Algorithm

The next problem faced by Rousseeuw was the actual computation of LMS. This is a highly non-trivial challenge because the median of the squared residuals is non-convex and non-smooth function. To begin to understand the difficulties here let's consider the following simulated data in Figure 1 and the regression-through-the-origin model

$$y_i \quad = \quad \beta x_i + \varepsilon_i, \quad i = 1, ..., 20$$

In Figure 2 we plot β versus $Med_i \left\{ (y_i - \beta x_i)^2 \right\}$ for $\beta = 2(0.01)8$. Notice that the plot is very jagged and the curve exhibits several local minimum and maximum. The situation is similar (usually worse) in the case of high dimensional data.

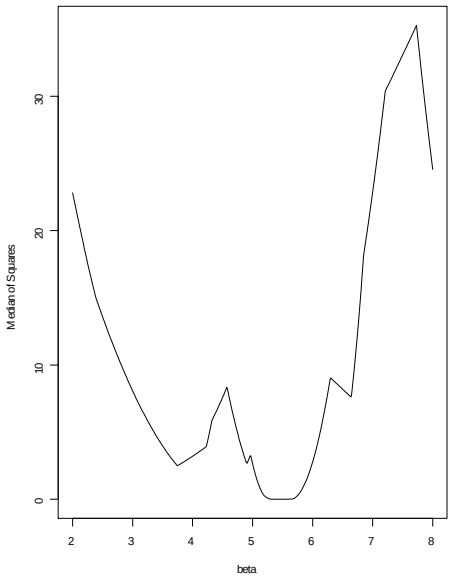

Figure 2: Median of squared residuals for the dataset displayed on Figure 1.

Werner Stahel - another former student of Hampel and contemporary of Rousseeuw - had recently proposed a very smart resampling algorithm to compute a robust multivariate location and covariance estimate now called Donoho-Stahel estimate. Rousseeuw decided to adapt this algorithm to compute LMS.

We will first focus on simple linear regression and use Forbes' dataset, to introduce the main ideas of the resampling algorithm. Forbes' dataset is similar to Hooker's and relates boiling point of water to atmospheric pressure at different altitude. The data is presented in Table 1

Resampling Algorithm When $p = 1$

The LMS resampling algorithm is based on the reasonable assumption (rather hope) that the fit $L(\hat{\alpha}, \hat{\beta}) = \left(x, \hat{\alpha} + \hat{\beta}x\right)$ that minimizes the median of squared residuals will be close to (or even equal to) one of the $m = \begin{pmatrix} n \\ 2 \end{pmatrix}$ lines connecting pairs of data points. Under this assumption, one can get a good (initial) evaluation of LMS by restricting attention to the m lines $L(\alpha_{ij}, \beta_{ij}) = \left(x, \alpha_{ij} + \beta_{ij}x\right)$, $i < j$ with slope and intercept coefficients

$$\beta_{ij} = \frac{y_i - y_j}{x_i - x_j} \quad \text{and} \quad \alpha_{ij} = \frac{y_j x_i - y_i x_j}{x_i - x_j} \tag{1}$$

respectively. For each pair of points $\{(y_i, x_i), (y_j, x_j)\}$ with $i < j$ satisfying $x_i - x_j \neq 0$, we compute the coefficients $\left(\alpha_{ij}, \beta_{ij}\right)$ given by (1) and the corresponding residuals

$$r_k^{ij} = y_k - \alpha_{ij} - \beta_{ij}x_k, \qquad k = 1, ..., n$$

The next step is to calculate the median of the squared residuals

$$MSR(i,j) = Med_k \left\{ \left(r_k^{ij}\right)^2 \right\} \qquad \text{for all } i < j \text{ such that } x_i - x_j \neq 0.$$

The final step is to determine

$$(i_0, j_0) = \arg\min_{i<j} MSR(i,j)$$

from which we obtain that the (approximate) LMS estimate:

$$\left(\hat{\alpha}, \hat{\beta}\right) = \left(\alpha_{i_0 j_0}, \beta_{i_0 j_0}\right).$$

Table 1: Forbes's data relating boiling temperature and atmosheric pressure at different altitude

Case	Temperature	Pressure	Case	Temperature	Pressure
1	194.5	20.79	10	201.3	24.01
2	194.3	20.79	11	203.6	25.14
3	197.9	22.40	12	204.6	26.57
4	198.4	22.67	13	209.5	28.49
5	199.4	23.15	14	208.6	27.76
6	199.9	23.35	15	210.7	29.04
7	200.9	23.89	16	211.9	29.88
8	201.1	23.99	17	212.2	30.06
9	201.4	24.02			

Example 1 *In this example we present the results of the LS and LMS fits to Forbes' data. Figure 3 presents the original data, the LS regression fit Pressure* $= -81.06373 + 0.5228924 \times$ *Temperature (red line) and the LMS regression fit Pressure* $= -70.88704 + 0.4715453 \times$ *Temperature (black line). Figure 4 displays the LS residuals (red) and LMS residuals (black). The larger group of 11 points has very small residuals. The Remaining 6 cases have considerably larger residuals and are poorly fit by the LMS line. The LMS analysis suggests that the measurements at "higher altitudes" - and corresponding "lower temperatures" - are quite different from those at "lower altitudes" - and corresponding "higher temperatures" .*

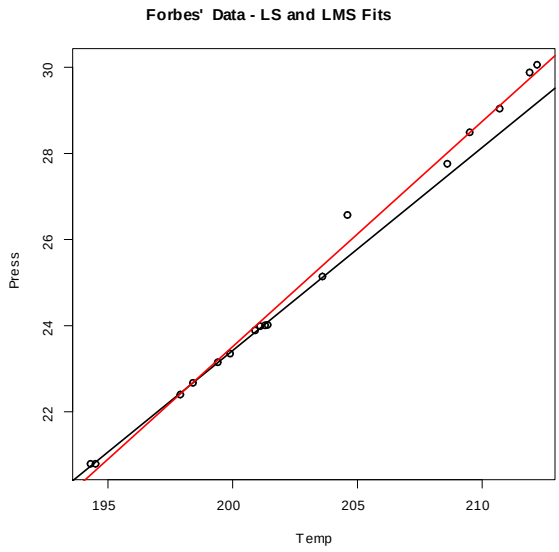

Figure 3: LMS and LS fits for the Forbes' data

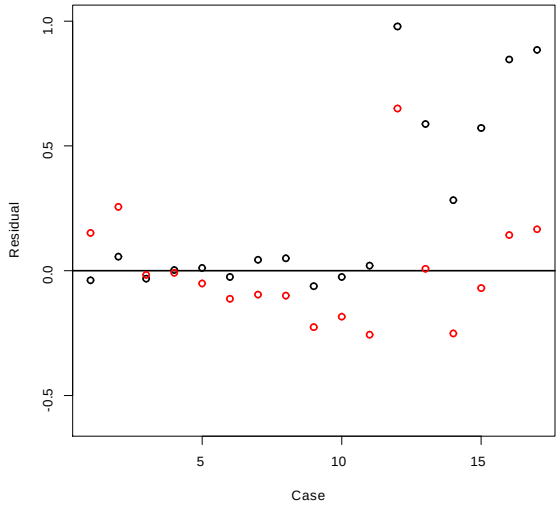

Figure 4: LMS (black) and LS (red) regression residual for the Forbes' data.

Resampling Algorithm for General p

Let's now consider the problem of computing LMS when the number of explanatory variables, p, is larger than one. In this case we may restrict attention the $m = \begin{pmatrix} n \\ \\ p+1 \end{pmatrix}$ subsets of $p+1$ data points. Let $\mathbf{J}$ be the collection of all the subsets $\{i_1, i_2, ..., i_{p+1}\}$ of $\{1, 2, ..., n\}$ such that the $(p+1) \times (p+1)$ matrix

$$\mathbf{X}_{\{i_1,i_2,...,i_{p+1}\}} \quad = \quad \begin{pmatrix} 1 & \mathbf{x}'_{i_1} \\ 1 & \mathbf{x}'_{i_2} \\ \vdots & \vdots \\ 1 & \mathbf{x}'_{i_{p+1}} \end{pmatrix}$$

has full rank. For each subset in $\mathbf{J}$ we can determine the vector of regression coefficients $\boldsymbol{\beta}_{\{i_1,i_2,...,i_{p+1}\}} \in R^{p+1}$ **-** which includes the intercept - such that

$$\begin{pmatrix} y_{i_1} \\ y_{i_2} \\ \vdots \\ y_{i_{p+1}} \end{pmatrix} \quad = \quad \begin{pmatrix} 1 & \mathbf{x}'_{i_1} \\ 1 & \mathbf{x}'_{i_2} \\ \vdots & \vdots \\ 1 & \mathbf{x}'_{i_{p+1}} \end{pmatrix} \boldsymbol{\beta}_{\{i_1,i_2,...,i_{p+1}\}}$$

For each subset in $\mathbf{J}$ we can now compute the corresponding set of n regression residuals

$$r_k^{\{i_1,i_2,...,i_{p+1}\}} \quad = \quad y_k - (1, \mathbf{x}_k)' \, \boldsymbol{\beta}_{\{i_1,i_2,...,i_{p+1}\}}, \qquad k = 1, 2, ..., n.$$

The next step is to calculate the median of the squares of these residuals:

$$MSR\left(\{i_1, i_2, ..., i_{p+1}\}\right) \quad = \quad Med_k \left\{ \left(r_k^{\{i_1,i_2,...,i_{p+1}\}} \right)^2 \right\} \quad \text{for all} \quad \{i_1, i_2, ..., i_{p+1}\} \quad \in \mathbf{J}.$$

The final step is to determine

$$\left\{i_1^0, i_2^0, ..., i_{p+1}^0\right\} \quad = \quad \arg\min_{\mathbf{J}} MSR\left(\{i_1, i_2, ..., i_{p+1}\}\right)$$

from which we obtain that the (approximate) LMS estimate:

$$\hat{\boldsymbol{\beta}} \quad = \quad \boldsymbol{\beta}_{\left\{i_1^0, i_2^0, \ldots, i_{p+1}^0\right\}}.$$

Numerical Complexity of the Resampling Algorithm

The LMS resampling algorithm becomes unfeasible when the number n of data points and the number p of explanatory variables increase. Table 2 gives the number of sub-samples for different values of n and p.

Table 2: Number of sub-samples for different sample sizes (n) and number of explanatory variables (p)

p	n				
	20	40	80	200	500
1	190	780	3160	19900	124750
2	1140	9880	82160	1313400	20708500
3	4845	91390	1581580	64684950	2573031000
5	38760	3838380	300500200	8.240863e+010	2.105769e+013
10	167960	2311801440	1.047768e+013	3.877901e+017	1.094974e+022
15	4845	62852101650	2.695822e+016	1.691526e+023	5.722368e+029

It is clear from Table 2 that evaluating *all* the sub-samples may not be practical (or feasible) for large values of n and p.

A possible approach is to consider only a random subset of sub-samples. In this case we must determine the size of this subset and how the sub-samples are to be chosen. This brings up interesting statistical issues: determining the *sample size* and choosing the *sampling method*.

Rousseeuw suggested to use simple random sampling with replacement (for simplicity) and to choose the sample size big enough to insure a high probability of selecting at least *one sub-sample* that contains only "good data points". Of course, this probability depends on the percentage of outliers in the data. Therefore, to implement Rousseuw's idea we must make some assumption regarding the largest "allowable" percentage of contamination in the data. The most conservative approach is be to take this percentage equal to 50%.

Let ϵ be the maximum allowable fraction of contamination in the data. In this case

$$P\left(\text{choosing a good data point}\right) = \epsilon.$$

Let δ be the probability that a sub-sample of $(p+1)$ data points is free of outliers (a good sub-sample). Under simple random sampling with replacement

$$\delta \quad = \quad P\left(\text{choosing a good sub-sample}\right) = (1-\epsilon)^{p+1}.$$

Suppose that $(1-\alpha)$ is the desired probability of choosing at least one good sub-sample. The required number m of sub-samples needed to achieve this probability satisfies

$$P\left(B\left(m,\delta\right)>0\right) \quad = \quad 1-P\left(B\left(m,\delta\right)=0\right)=\left(1-\alpha\right),$$

where $B\left(m,\delta\right)$ is a Binomial random variable with m trials and probability δ of success. So, m must satisfy the equation

$$\begin{aligned} 1-\left(1-\delta\right)^{m} &= \left(1-\alpha\right) \\ \left(1-\delta\right)^{m} &= \alpha \\ m\log\left(1-\delta\right) &= \log\left(\alpha\right) \\ m &= \frac{\log\left(\alpha\right)}{\log\left(1-\delta\right)} \end{aligned}$$

Therefore,

$$m \quad = \quad \frac{\log\left(\alpha\right)}{\log\left[1-\left(1-\epsilon\right)^{p+1}\right]}$$

where $(1-\alpha)$ is the desired probability of choosing at least one good sub-sample and ϵ the assumed upper bound for the fraction of outliers in the dataset.

Remark 1 *Notice that the number of samples, m, depends on three parameters:*

$\epsilon =$ *assumed maximum fraction of contamination*
$1-\alpha =$ *the desired probability of having a clean sample*
$p =$ *the number of explanatory variables*

Compiquously and fortunately, it doesn't depend on the sample size, n.

The following table gives the values of m for $\alpha = 0.01$, $\epsilon = 0.50$ and several values of p.

p	1	2	3	4	5	6	7	8	9	10	11	12	13
m	17	35	72	146	293	588	1,177	2,356	4,714	9,430	18,861	37724	75449

p	14	15	16	17	18	19	20
m	150,900	301,803	603,607	1,207,216	2,414,434	4,828,869	9,657,740

Remark 2 *Figure 5 shows that when the maximum fraction of contamination is* $\epsilon = 0.5$, *the number* m *of sub-samples needed for a 0.99 probability of a "clean" sub-sample of size* $p + 1$ *follows the exponential formula* $m \approx \exp\{2.2067 + 0.6935p\}$. *The number of sub-samples quickly runs out of control for* p *larger than 10, say. This is a problem that has attracted some current research interest. There are now more efficient algorithms based on the idea of local improvement using "concentration steps". See for example Salibian-Barrera and Yohai (2006) and Salibian-Barrera, Willems, and Zamar, (2008).*

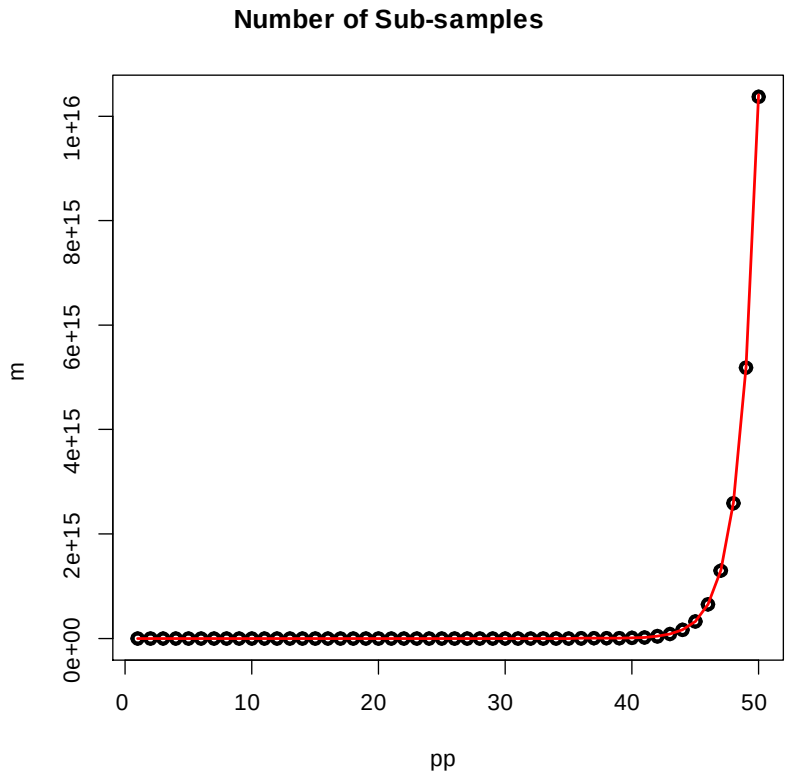

Figure 5: The number m of subsamples when eps = 0.5 and alpha = 0.01.

S-Estimates

The introduction of LMS regression estimates was an important breakthrough in the theory and practice of robustness.

The theoretical appeal of the estimate was greatly strengthened when Martin, Yohai and Zamar (1989) showed that LMS has a strong robustness optimality property - bias-minimaxity - among all M-estimates of regression with general scale. Yohai and Zamar (1993) showed that LMS optimality is even stronger by proving that it extends to larger class of all regression estimates that minimize a function of the regression residuals (which they call "residual admissible regression estimates").

LMS estimates nicely fulfill the first robustness requirement of resistance in the presence of outliers.

Unfortunately, LMS estimates do not fulfill the second robustness requirement of efficiency in the presence of clean data. In fact, LMS estimates are not asymptotically normal and approach their asymptotic non-normal limiting distribution at the slow $\sqrt[3]{n}$-rate instead of the usual $\sqrt{n}$-rate.

Rousseuw and Yohai (1984) introduced the family of regression S-estimates which are robust (have maximal BP $= 1/2$), are affine and regression equivariant and have asymptotic normal distribution with the usual square-root convergence rate.

The basic ideas underlying the definition of regression S-estimates have been already introduced in the first section when we defined robust dispersion M-estimates.

Given a set of $(p+1)$-dimensional points $\{(y_1, \mathbf{x}_1), ..., (y_1, \mathbf{x}_1)\}$ the regression S-estimate is defined as the solution to the minimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{t}} S(\mathbf{t}), \tag{2}$$

where t ranges over R^{p+1} - that is, $\mathbf{t} = (t_0, t_1, ..., t_p)'$, with t_0 representing the intercept parameter. Of course, the key part in the definition of the regression S-estimate is the definition of the function S to be minimized. This function is a robust scale estimate of the regression residuals

$$r_i(\mathbf{t}) \quad = \quad y_i - \mathbf{x}_i'\mathbf{t}, \quad i = 1, 2, ..., n.$$

where the robust scale is defined as in the location-scale model considered the the first section. More precisely, let χ be a (scale) score function satisfying the following properties

(1) χ is continuously differentiable
(2) $\chi(0) = 0$ and $\chi(\infty) = 1$.
(3) $\chi(-y) = \chi(y)$ for all y.
(4) $\chi(y)$ is non-decreasing on $[0, \infty)$.
(5) $0 < b = E_\Phi \chi(y) < 1$.

Commonly used scale score function are Tukey's bi-square loss function

$$\chi(y) = \begin{cases} 3(y/c)^2 - 3(y/c)^4 + (y/c)^6 & \text{if} \quad |y| \leq c \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

and Huber's truncated squares score function

$$\chi(y) = \begin{cases} (y/c)^2 & \text{if} \quad |y| \leq c \\ 1 & \text{otherwise.} \end{cases}$$

The M-scale of the tentative regression residuals $\{r_i(\mathbf{t})\}$ with $\boldsymbol{\beta} = t$ is now implicitly defined by the equation

$$s(\mathbf{t}) = \sup \left\{ s : \frac{1}{n} \sum \chi \left(\frac{r_i(\mathbf{t})}{s(\mathbf{t})} \right) > b \right\}. \tag{4}$$

Notice that, by continuity of χ, $s(t)$ satisfies the equation

$$\frac{1}{n} \sum \chi \left(\frac{r_i(\mathbf{t})}{s(\mathbf{t})} \right) - b = 0 \tag{5}$$

The minimization problem (2) can be approximately solved using the resampling algorithm introduced before for the LMS estimate. Since solving equation (5) is computationally expensive it is crucial to use the "records" step discussed in Section 1 in the case of the location-scale model. That is, if $s^* = s(\mathbf{t}^*)$ is the current minimum and $\mathbf{t}^{**}$ is the vector of coefficients corresponding to next

randomly selected sub-sample, then $s(\mathbf{t}^{**})$ needs to be calculated (and therefore equation (5) needs to be solved) only if

$$\frac{1}{n}\sum \chi\left(\frac{r_i(\mathbf{t}^{**})}{s(\mathbf{t}^{*})}\right) \quad < \quad b.$$

If m represents the number of sub-samples to be evaluated and N represents the number of times equation (5) must be solved then

$$E(N) \quad \approx \quad \log(m) \approx Var(N).$$

(see Problem). By Chevichev's inequality,

$$P\left[|N - \log(m)| < k\sqrt{\log(m)}\right] \quad > \quad 1 - \frac{1}{k^2}.$$

from which it follows that

$$P\left[N < \sqrt{\log(m)}\left(k + \sqrt{\log(m)}\right)\right] > 1 - \frac{1}{k^2} \qquad .$$

The table below gives the values of $m_0 = \sqrt{\log(m)}\left(k + \sqrt{\log(m)}\right)$ for $k = 10$ (and corresponding 99% probability) for several values of m

Number of Sub-samples	Number of Equations
$1,000$	34
$10,000$	40
$100,000$	46
$1,000,000$	51
$10,000,000$	57

Problems

Problem 3 *Show that LMS is an S-estimate with "jump" score function*

$$\chi(y) \quad = \quad \begin{cases} 0 & \text{if} \quad y^2 \leq c \\ \\ 1 & \qquad y^2 > c. \end{cases}$$

What is the value of c?

Problem 4 *Show that a regular S-estimate (with smooth score function satisfying A1-A5) can be viewed as an M-estimate. That is show that the S-estimate satisfies an equation of the form*

$$\sum \psi \left(\frac{y_i - \boldsymbol{\beta}' \mathbf{x}_i}{\hat{\sigma}_n} \right) \mathbf{x}_i \quad = \quad 0$$

for some appropriate ψ and $\hat{\sigma}_n$ (which you must specify).

Problem 5 *Suggest an algorithm to compute S-estimates for simple linear regression. Implement this algorithm using your favorite language and apply it to several datasets including Forbe's and Hooker's datasets. Make sure your program allows the user to choose the value of b.*

Problem 6 *Show that S-estimates are affine and regression equivariant. More precisely, set*

$$\hat{\boldsymbol{\beta}} \quad = \quad \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix}$$

and show that

(1) If $\mathbf{x} \rightarrow A\mathbf{x}$ (with $\mathbf{A}$ invertible) then $\hat{\boldsymbol{\beta}}_1 \rightarrow A^{-1}\hat{\boldsymbol{\beta}}_1$ and $\hat{\beta}_0 \rightarrow \hat{\beta}_0$

(2) If $\mathbf{x} \rightarrow \mathbf{x} + \mathbf{d}$ then $\hat{\boldsymbol{\beta}}_1 \rightarrow \hat{\boldsymbol{\beta}}_1$ and $\hat{\beta}_0 \rightarrow \hat{\beta}_0 - \hat{\boldsymbol{\beta}}_1' \mathbf{d}$

(3) If $y \rightarrow ay$ then $\hat{\boldsymbol{\beta}}_1 \rightarrow a\hat{\boldsymbol{\beta}}_1$ and $\hat{\beta}_0 \rightarrow a\hat{\beta}_0$

(4) If $y \rightarrow y + a + \mathbf{x}'\mathbf{b}$ then $\hat{\boldsymbol{\beta}}_1 \rightarrow \hat{\boldsymbol{\beta}}_1 + \mathbf{b}$ and $\hat{\beta}_0 \rightarrow \hat{\beta}_0 + a$

Tau-Estimates

Regression S-estimates are asymptotically normal at the usual $\sqrt{n}$-rate, robust and regression-affine equivariant. Therefore they are pretty close to what we would like to call good robust regression estimates. However they have a drawback which we can improve upon: they cannot be simultaneously resistant for contaminated data and efficient for clean data. We will see later in this course that the breakdown point of S-estimates is equal to

$$\min\{b, 1-b\}$$

where $b = E\{\chi(Z)\}$, with $Z \sim N(0,1)$ .On the other hand, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is equal to

$$\sigma^2 \frac{E_F\left\{\psi_c^2(Y)\right\}}{\left[E_F\left\{\psi_c'(Y)\right\}\right]^2}\left[E\left(\mathbf{x}\mathbf{x}'\right)\right]^{-1}$$

with $\psi_c(y) = \chi_c'(y)$. It can be shown (the reader can numerically verify this) that to get high efficiency under normal errors c must be chosen rather large and this causes b to become small. In summary, there is a trade-off between Gaussian efficiency and breakdown point in the case of S-estimates. An this trade-off cannot be resolved because there is only one "tuning" constant, c, we can play with. In fact, it has been shown that the maximum efficiency that can be achieved when $b = 1/2$ is very low (below 50%).

There are at least two approaches to get around this difficulty. My favorite one - **warning**: this statement may be suspiciously biased - is to use τ-estimates which were originally introduced by Yohai and Zamar (1988), my first ever published paper!

τ-estimates are based on the idea of minimizing a robust and efficient scale of the regression residuals. The τ-scale is defined as follows:

$$\tau(\mathbf{t}) = \frac{s_n^2(\mathbf{t})}{n}\sum \chi_2\left(\frac{r_i(\mathbf{t})}{s_n(\mathbf{t})}\right) \tag{6}$$

with $s_n(\mathbf{t})$ implicitly defined by the equation

$$\frac{1}{n}\sum \chi_1\left(\frac{r_i(\mathbf{t})}{s_n(\mathbf{t})}\right) = b.$$

Notice that $s_n(\mathbf{t})$ is an M-scale with score function χ_1. The regression τ-estimate is now defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{t}} \tau(\mathbf{t}).$$

It can be shown that the robustness of the τ-estimate is determined by the score function χ_1 and that its efficiency is determined by the score function χ_2. Therefore, the score functions can be chosen to simultaneously achieve any desired level of robustness (breakdown point) and efficiency.

Although τ-estimates have attractive theoretical properties (see Berrendero, Mazzi, Romo and Zamar (199?) and Berrendero and Zamar (2001)), there has not been much progress regarding their computation. This is due in part to the fact that most of the computational effort has been directed toward the Splus implementation of Yohai (198?)'s regression MM-estimates - which can also be simultaneously robust en efficient (see command *lmRob* in Splus 6). Regression MM-estimates will be described and briefly discussed below.

Problems

Problem 7 *Show that τ-estimates satisfy an estimating equation of the form*

$$\frac{1}{n}\sum \Psi_n\left(\frac{r_i(\mathbf{t})}{s_n(\mathbf{t})}\right)\mathbf{x}_i = 0$$

with "adaptive" score function Ψ_n which depends on the data and has the form

$$\Psi_n(y) = w_n\psi_1(y) + (1-w_n)\psi_2(y),$$

where $0 < w_n < 1$ is a weight that depends on the data. When is w_n close to 1? When is w_n close to 0?

Problem 8 *Consider the function*

$$g(y,s) = s^2\chi(y/s). \tag{7}$$

What can you say about the monotonicity (in s) of $g(y,s)$ for given y. Give examples where this monotonicity is (is not) satisfied. In particular, study (7) for the cases of Tukey's and Huber's families of score functions.

Problem 9 *Suppose that χ_2 in (6) is such that $g(y,s)$ given by (7) with $\chi = \chi_2$ is non-decreasing in s for all y. Consider the following "record step" for an algorithm to compute the τ-estimate via re-sampling.*

***Record Step**: Let $\boldsymbol{\beta}^*$, $\boldsymbol{\tau}^*$ and s^* be the current values corresponding to the last observed "record". Let $\boldsymbol{\beta}$ is the next "candidate" to be evaluated for a possible new record. Show that $\boldsymbol{\beta}$ can be discarded if the following two conditions hold:*

$$(1) \quad \frac{1}{n}\sum \chi_1\left(\frac{r_i(\boldsymbol{\beta})}{s^*}\right) > b.$$

$$(2) \quad \sum \chi_2\left(\frac{r_i(\boldsymbol{\beta})}{s^*}\right) > \sum \chi_2\left(\frac{r_i(\boldsymbol{\beta}^*)}{s^*}\right)$$

Determine (numerically if necessary) the expected number of times N condition (1) or condition (2) above are violated. What is the standard deviation of N? Notice that $E(N)$ and $SD(N)$ determine the overall expected computational burden for the proposed resampling algorithm.

Some References

- Andrew F. Siegel (1982). Robust regression using repeated medians. Biometrika, 242-244.

- Peter J. Rousseeuw (1984). Least median of squares regression. Journal of the American Statistical Association, 871-880.

- Rousseeuw, P. and Yohai, V.J. (1984). Robust regression by means of S-estimators. In Robust and nonlinear time series, editors Franke, Hardle and Martin. Lectures Notes in Statistics, vol. 29, 256-272, Springer Vela, New York, 1984.

- Yohai, V.J. and Zamar, R.H. (1988). "High breakdown-point estimates of regression by means of the minimization of an efficient scale". Journal American Statistical Association, 83, 406-413..

- Rousseeuw P.J. and Van Driessen K. (1999) A fast algorithm for the minimum covariance determinant estimator, Technometrics, 41, 212–223.

- Salibian-Barrera, M. and Yohai, V.J. (2006). A fast algorithm for S-regression estimates. Journal of Computational and Graphical Statistics 15, 414-427.

- Salibian-Barrera, M., Willems, G. and Zamar, R.H. (2008). The fast-tau estimator for regression. Journal of Computational and Graphical Statistics, 17, 659-682.