M.Sc. in Applied Statistics MT2004

Robust Statistics

The classical books on this subject are Hampel *et al*. (1986); Huber (1981), with somewhat simpler (but partial) introductions by Rousseeuw & Leroy (1987); Staudte & Sheather (1990). The dates reflect the development of the subject: it had tremendous growth for about two decades from 1964, but failed to win over the mainstream. I think it is an important area that is used a lot less than it ought to be.

Univariate statistics

Outliers are sample values that cause surprise in relation to the majority of the sample. This is not a pejorative term; outliers may be correct, but they should always be checked for transcription errors. They can play havoc with standard statistical methods, and many *robust* and *resistant* methods have been developed since 1960 to be less sensitive to outliers.

The sample mean $\overline{y}$ can be upset completely by a single outlier; if any data value $y_i \to \pm\infty$, then $\overline{y} \to \pm\infty$. This contrasts with the sample median, which is little affected by moving any single value to $\pm\infty$. We say that the median is *resistant* to *gross errors* whereas the mean is not. In fact the median will tolerate up to 50% gross errors before it can be made arbitrarily large; we say its *breakdown point* is 50% whereas that for the mean is 0%. Although the mean is the optimal estimator of the location of the normal distribution, it can be substantially sub-optimal for distributions close to the normal. Robust methods aim to have high efficiency in a neighbourhood of the assumed statistical model.

Why will it not suffice to screen data and remove outliers? There are several aspects to consider.

1. Users, even expert statisticians, do not always screen the data.
2. The sharp decision to keep or reject an observation is wasteful. We can do better by down-weighting dubious observations than by rejecting them, although we may wish to reject completely wrong observations.
3. It can be difficult or even impossible to spot outliers in multivariate or highly structured data.
4. Rejecting outliers affects the distribution theory, which ought to be adjusted. In particular, variances will be underestimated from the 'cleaned' data.

For a fixed underlying distribution, we define the *relative efficiency* of an estimator $\tilde{\theta}$ relative to another estimator $\widehat{\theta}$ by

$$RE(\tilde{\theta};\widehat{\theta}) = \frac{\text{variance of } \widehat{\theta}}{\text{variance of } \tilde{\theta}}$$

since $\widehat{\theta}$ needs only RE times as many observations as $\tilde{\theta}$ for the same precision, approximately. The asymptotic relative efficiency (ARE) is the limit of the RE as the sample size $n \to \infty$. (It may be defined more widely via asymptotic variances.) If $\widehat{\theta}$ is not mentioned, it is assumed to be the optimal estimator. There is a difficulty with biased estimators whose variance can be small or zero. One solution is to use the mean-square error, another to rescale by $\theta/E(\widehat{\theta})$. Iglewicz (1983) suggests using var $(\log \widehat{\theta})$ (which is scale-free) for estimators of scale.

We can apply the concept of ARE to the mean and median. At the normal distribution $ARE(\text{median};\text{mean}) = 2/\pi \approx 64\%$. For longer-tailed distributions the median does better; for the t distribution with five degrees of freedom (which is often a better model of error distributions than the normal) $ARE(\text{median};\text{mean}) \approx 96\%$.

The following example from Tukey (1960) is more dramatic. Suppose we have n observations $Y_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$ and we want to estimate σ^2. Consider $\widehat{\sigma}^2 = s^2$ and $\tilde{\sigma}^2 = d^2\pi/2$ where

$$d = \frac{1}{n}\sum_i |Y_i - \overline{Y}|$$

and the constant is chosen since for the normal $d \to \sqrt{2/\pi}\,\sigma$. The $ARE(\tilde{\sigma}^2; s^2) = 0.876$. Now suppose that each Y_i is from $N(\mu, \sigma^2)$ with probability $1 - \epsilon$ and from $N(\mu, 9\sigma^2)$ with probability ϵ. (Note that both the overall variance and the variance of the uncontaminated observations are proportional to σ^2.) We have

ϵ (%)	$ARE(\tilde{\sigma}^2; s^2)$
0	0.876
0.1	0.948
0.2	1.016
1	1.44
5	2.04

Since the mixture distribution with $\epsilon = 1\%$ is indistinguishable from normality for all practical purposes, the optimality of s^2 is very fragile. We say it lacks *robustness of efficiency*.

There are better estimators of σ than $d\sqrt{\pi/2}$ (which has breakdown point 0%). Two alternatives are proportional to

$$\begin{aligned} \textit{IQR} &= X_{(3n/4)} - X_{(n/4)} \\ \textit{MAD} &= \underset{i}{\text{median}}\,\{|Y_i - \underset{j}{\text{median}}\,(Y_j)|\} \end{aligned}$$

(Order statistics are linearly interpolated where necessary.) At the normal,

$$\begin{aligned} \textit{MAD} &\to \text{median}\,\{|Y - \mu|\} \approx 0.6745\sigma \\ \textit{IQR} &\to \sigma[\Phi^{-1}(0.75) - \Phi^{-1}(0.25)] \approx 1.35\sigma \end{aligned}$$

(We refer to $MAD/0.6745$ as the MAD estimator, calculated by function `mad` in S-PLUS.) Both are not very efficient but are very resistant to outliers in the data. The MAD estimator has ARE 37% at the normal (Staudte & Sheather, 1990, p. 123).

Consider n independent observations Y_i from a location family with pdf $f(y-\mu)$ for a function f symmetric about zero, so it is clear that μ is the centre (median, mean if it exists) of the distribution of Y_i. We also think of the distribution as being not too far from the normal. There are a number of obvious estimators of μ, including the sample mean, the sample median, and the MLE.

The *trimmed mean* is the mean of the central $1-2\alpha$ part of the distribution, so αn observations are removed from each end. This is implemented by the function `mean` with the argument `trim` specifying α. Obviously, `trim=0` gives the mean and `trim=0.5` gives the median (although it is easier to use the function `median`). (If αn is not an integer, the integer part is used.)

Most of the location estimators we consider are *M-estimators*. The name derives from 'MLE-like' estimators. If we have density f, we can define $\rho = -\log f$. Then the MLE would solve

$$\min_\mu \sum_i -\log f(y_i - \mu) = \min_\mu \sum_i \rho(y_i - \mu)$$

and this makes sense for functions ρ not corresponding to pdfs. Let $\psi = \rho'$ if this exists. Then we will have $\sum_i \psi(y_i - \widehat{\mu}) = 0$ or $\sum_i w_i\,(y_i - \widehat{\mu}) = 0$ where $w_i = \psi(y_i - \widehat{\mu})/(y_i - \widehat{\mu})$. This suggests an iterative method of solution, updating the weights at each iteration.

Examples of M-estimators

The mean corresponds to $\rho(x) = x^2$, and the median to $\rho(x) = |x|$. (For even n any median will solve the problem.) The function

$$\psi(x) = \begin{cases} x & |x| < c \\ 0 & \text{otherwise} \end{cases}$$

corresponds to *metric trimming* and large outliers have no influence at all. The function

$$\psi(x) = \begin{cases} -c & x < -c \\ x & |x| < c \\ c & x > c \end{cases}$$

is known as *metric Winsorizing*[2] and brings in extreme observations to $\mu \pm c$. The corresponding $-\log f$ is

$$\rho(x) = \begin{cases} x^2 & \text{if } |x| < c \\ c(2|x| - c) & \text{otherwise} \end{cases}$$

and corresponds to a density with a Gaussian centre and double-exponential tails. This estimator is due to Huber. Note that its limit as $c \to 0$ is the median, and as $c \to \infty$ the limit is the mean. The value $c = 1.345$ gives 95% efficiency at the normal.

Tukey's *biweight* has

$$\psi(t) = t\Big[1 - \left(\frac{t}{R}\right)^2\Big]_+^2$$

[2] A term attributed by Dixon (1960) to Charles P. Winsor.

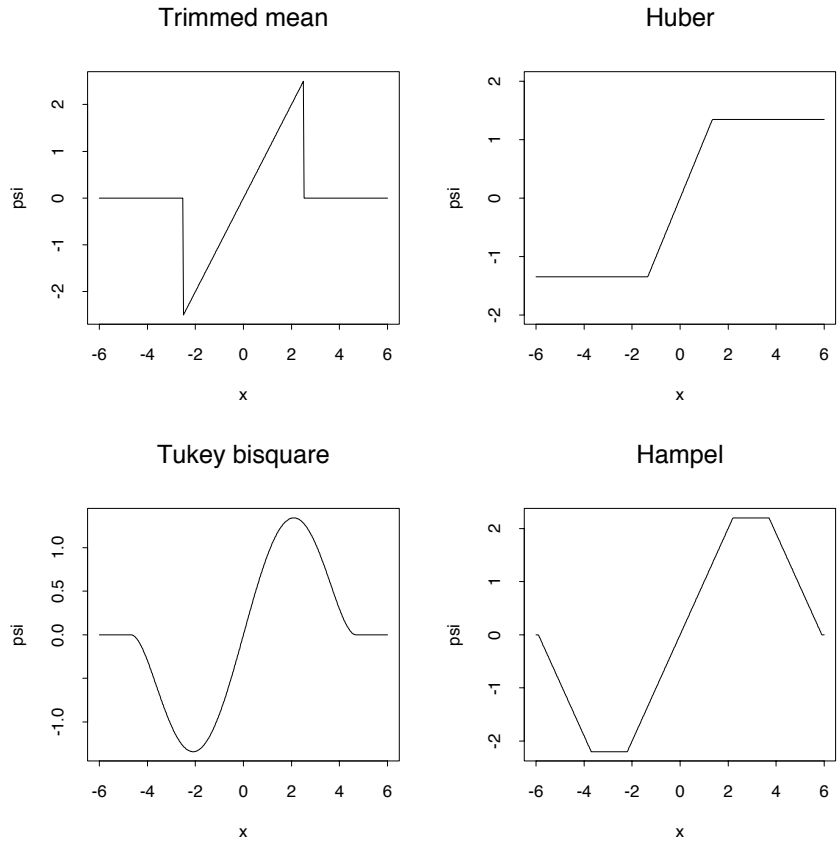

Figure 1: The ψ-functions for four common M-estimators.

where $[\,]_+$ denotes the positive part of. This implements 'soft' trimming. The value $R = 4.685$ gives 95% efficiency at the normal.

Hampel's ψ has several linear pieces,

$$\psi(x) = \text{sgn}(x) \begin{cases} |x| & 0 < |x| < a \\ a & a < |x| < b \\ a(c - |x|)/(c - b) & b < |x| < c \\ 0 & c < |x| \end{cases}$$

for example, with $a = 2.2s, b = 3.7s, c = 5.9s$. Figure 1 illustrates these functions.

There is a scaling problem with the last four choices, since they depend on a scale factor (c, R or s). We can apply the estimator to rescaled results, that is,

$$\min_{\mu} \sum_i \rho\left(\frac{y_i - \mu}{s}\right)$$

for a scale factor s, for example the MAD estimator. Alternatively, we can estimate s in a similar way. The MLE for density $s^{-1}f((x - \mu)/s)$ gives rise to the equation

$$\sum_i \psi\left(\frac{y_i - \mu}{s}\right)\left(\frac{y_i - \mu}{s}\right) = n$$

which is not resistant (and is biased at the normal). We modify this to

$$\sum_i \chi\left(\frac{y_i - \mu}{s}\right) = (n-1)\gamma$$

for bounded χ, where γ is chosen for consistency at the normal distribution, so $\gamma = E\,\chi(N)$. The main example is "Huber's proposal 2" with

$$\chi(x) = \psi(x)^2 = \min(|x|, c)^2 \tag{1}$$

In very small samples we need to take account of the variability of $\widehat{\mu}$ in performing the Winsorizing.

If the location μ is known we can apply these estimators with $n-1$ replaced by n to estimate the scale s alone.

Examples

We give two datasets taken from analytical chemistry (Abbey, 1988; Analytical Methods Committee, 1989a,b). The dataset `abbey` contains 31 determinations of nickel content ($\mu g\, g^{-1}$) in SY-3, a Canadian syenite rock, and `chem` contains 24 determinations of copper ($\mu g\, g^{-1}$) in wholemeal flour. These data are part of a larger study that suggests $\mu = 3.68$.

```
> sort(chem)
 [1]  2.20  2.20  2.40  2.40  2.50  2.70  2.80  2.90  3.03
[10]  3.03  3.10  3.37  3.40  3.40  3.40  3.50  3.60  3.70
[19]  3.70  3.70  3.70  3.77  5.28 28.95
> mean(chem)
[1] 4.2804
> median(chem)
[1] 3.385
> location.m(chem)
[1] 3.1452
    ....
> location.m(chem, psi.fun="huber")
[1] 3.2132
    ....
> mad(chem)
[1] 0.52632
> scale.tau(chem)
[1] 0.639
> scale.tau(chem, center=3.68)
[1] 0.91578
> unlist(huber(chem))
     mu       s
 3.2067 0.52632
> unlist(hubers(chem))
     mu       s
 3.2055 0.67365
```

The sample is clearly highly asymmetric with one value that appears to be out by a factor of 10. It was checked and reported as correct by the laboratory. With such a distribution the various estimators are estimating different aspects of the distribution and so are not comparable. Only for symmetric distributions do all the location estimators estimate the same quantity, and although the true distribution is unknown here, it is unlikely to be symmetric.

```
> sort(abbey)
 [1]   5.2   6.5   6.9   7.0   7.0   7.0   7.4   8.0   8.0
[10]   8.0   8.0   8.5   9.0   9.0  10.0  11.0  11.0  12.0
[19]  12.0  13.7  14.0  14.0  14.0  16.0  17.0  17.0  18.0
[28]  24.0  28.0  34.0 125.0
> mean(abbey)
[1] 16.006
> median(abbey)
[1] 11
> location.m(abbey)
[1] 10.804
> location.m(abbey, psi.fun="huber")
[1] 11.517
> unlist(hubers(abbey))
     mu      s
 11.732 5.2585
> unlist(hubers(abbey, k=2))
     mu      s
 12.351 6.1052
> unlist(hubers(abbey, k=1))
     mu      s
 11.365 5.5673
```

Note how reducing the constant `k` (representing c) reduces the estimate of location, as this sample (like many in analytical chemistry) has a long right tail.

Robust *vs* resistant regression

We will see in the section on *Linear Models* the concept of *resistant regression*, which is about non-disastrous behaviour in the presence of incorrect data points. In the terminology introduced here, resistant regression has a high breakdown point, approaching 50%. We considered replacing least-squares by one of

LMS Least median of squares: minimize the median of the squared residuals. More generally, for LQS, minimize some quantile (say 80%) of the squared residuals.

LTS Least trimmed squares: minimize the sum of squares for the smallest q of the residuals. Originally q included just over 50%, but S-PLUS has switched to 90%.

However, either involves very much more computing than least squares, as the minimands are not differentiable. Both do show up the effect of multiple outliers, as they concentrate on fitting just over 50% of the data well. In doing so they are less efficient when there are no outliers (LMS more so than LTS).

To illustrate some of the problems, consider an example. Rousseeuw & Leroy (1987) give data on annual numbers of Belgian telephone calls. Figure 2 shows the least squares line, an

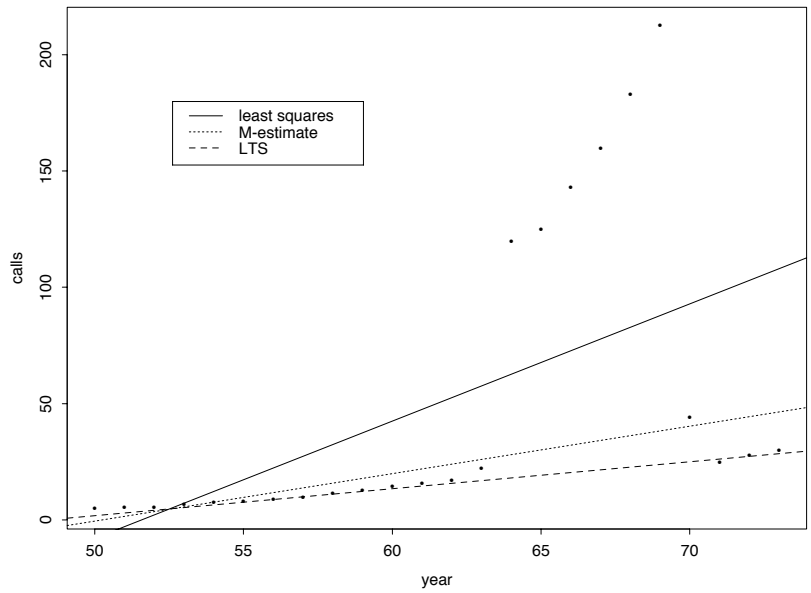

Figure 2: Millions of phone calls in Belgium, 1950–73, from Rousseeuw & Leroy (1987), with three fitted lines.

M-estimated regression and the least trimmed squares regression. The `lqs` line is $-56.16 + 1.16\,$`year`. Rousseeuw & Leroy's investigations showed that for 1964–9 the total length of calls (in minutes) had been recorded rather than the number, with each system being used during parts of 1963 and 1970.

Resistant regression

A succession of more resistant regression estimators was defined in the 1980s. The first to become popular was

$$\min_b \operatorname*{median}_i |y_i - \boldsymbol{x}_i b|^2$$

called the *least median of squares* (LMS) estimator. The square is necessary if n is even, when the central median is taken. This fit is very resistant, and needs no scale estimate. It is however very inefficient, converging at rate $1/\sqrt[3]{n}$. Furthermore, it displays marked sensitivity to central data values: see Hettmansperger & Sheather (1992) and Davies (1993, §2.3).

Rousseeuw suggested least trimmed squares (LTS) regression:

$$\min_b \sum_{i=1}^{q} |y_i - \boldsymbol{x}_i b|^2_{(i)}$$

as this is more efficient, but shares the same extreme resistance. The recommended sum is over the smallest $q = \lfloor (n + p + 1)/2 \rfloor$ squared residuals. (Earlier accounts differed.)

This was followed by *S-estimation*, in which the coefficients are chosen to find the solution to

$$\sum_{i=1}^{n} \chi\Big(\frac{y_i - \boldsymbol{x}_i b}{c_0\, s}\Big) = (n - p)\beta$$

with smallest scale s. Here χ is usually chosen to be the integral of Tukey's bisquare function

$$\chi(u) = u^6 - 3u^4 + 3u^2, \ |u| \le 1, \qquad 1, \ |u| \ge 1$$

$c_0 = 1.548$ and $\beta = 0.5$ is chosen for consistency at the normal distribution of errors. This gives efficiency 28.7% at the normal, which is low but better than LMS and LTS.

In only a few special cases (such as LMS for univariate regression with intercept) can these optimization problems be solved exactly, and approximate search methods are used. (Marazzi, 1993 is a good source. Most of these methods work by looking at least-squares fits to around q of the data points, and more or less randomly try a large sample of such fits.)

Theory for robust regression

In a regression problem there are two possible sources of errors, the observations y_i and the corresponding row vector of p regressors $\boldsymbol{x}_i$. Most robust methods in regression only consider the first, and in some cases (designed experiments?) errors in the regressors can be ignored. This is the case for M-estimators, the only ones we consider in this section.

Consider a regression problem with n cases $(y_i, \boldsymbol{x}_i)$ from the model

$$y = \boldsymbol{x}\beta + \epsilon$$

for a p-variate row vector $\boldsymbol{x}$.

M-estimators

If we assume a scaled pdf $f(e/s)/s$ for ϵ and set $\rho = -\log f$, the maximum likelihood estimator minimizes

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{x}_i b}{s}\right) + n \log s \tag{2}$$

Suppose for now that s is known. Let $\psi = \rho'$. Then the MLE b of β solves the non-linear equations

$$\sum_{i=1}^{n} \boldsymbol{x}_i \psi\left(\frac{y_i - \boldsymbol{x}_i b}{s}\right) = 0 \tag{3}$$

Let $r_i = y_i - \boldsymbol{x}_i b$ denote the residuals.

The solution to equation (3) or to minimizing over (2) defines an M-estimator of β.

A common way to solve (3) is by iterated re-weighted least squares, with weights

$$w_i = \psi\left(\frac{y_i - \boldsymbol{x}_i b}{s}\right) \Big/ \left(\frac{y_i - \boldsymbol{x}_i b}{s}\right) \tag{4}$$

The iteration is only guaranteed to converge for *convex* ρ functions, and for redescending functions (such as those of Tukey and Hampel; page 3), equation (3) may have multiple roots. In such cases it is usual to choose a good starting point and iterate carefully.

Of course, in practice the scale s is not known. A simple and very resistant scale estimator is the MAD about some centre. This is applied to the residuals about zero, either to the current residuals within the loop or to the residuals from a very resistant fit (see the next subsection).

Alternatively, we can estimate s in an MLE-like way. Finding a stationary point of (2) with respect to s gives

$$\sum_i \psi\left(\frac{y_i - \boldsymbol{x}_i b}{s}\right)\left(\frac{y_i - \boldsymbol{x}_i b}{s}\right) = n$$

which is not resistant (and is biased at the normal). As in the univariate case we modify this to

$$\sum_i \chi\left(\frac{y_i - \boldsymbol{x}_i b}{s}\right) = (n-p)\gamma \tag{5}$$

An example

Library `MASS` includes a model-fitting function `rlm`. By default Huber's M-estimator is used with tuning parameter $c = 1.345$. By default the scale s is estimated by iterated MAD, but Huber's proposal 2 can also be used.

```
> summary(lm(calls ~ year, data=phones), cor=F)
              Value Std. Error  t value Pr(>|t|)
(Intercept) -260.059  102.607     -2.535    0.019
       year    5.041    1.658      3.041    0.006
Residual standard error: 56.2 on 22 degrees of freedom
> summary(rlm(calls ~ year, maxit=50, data=phones), cor=F)
              Value Std. Error  t value
(Intercept) -102.622   26.608     -3.857
       year    2.041    0.430      4.748
Residual standard error: 9.03 on 22 degrees of freedom
> summary(rlm(calls ~ year, scale.est="proposal 2", data=phones), cor=F)
Coefficients:
              Value Std. Error  t value
(Intercept) -227.925  101.874     -2.237
       year    4.453    1.646      2.705
Residual standard error: 57.3 on 22 degrees of freedom
```

As Figure 2 shows, in this example there is a batch of outliers from a different population in the late 1960s, and these should be rejected completely, which the Huber M-estimators do not. Let us try a re-descending estimator.

```
> summary(rlm(calls ~ year, data=phones, psi=psi.bisquare), cor=F)
Coefficients:
             Value Std. Error t value
(Intercept) -52.302   2.753    -18.999
       year   1.098   0.044     24.685
Residual standard error: 1.65 on 22 degrees of freedom
```

This happened to work well for the default least-squares start, but we might want to consider a better starting point, such as that given by `init="lts"`.

MM-estimation

It is possible to combine the resistance of these methods with the efficiency of M-estimation. The MM-estimator proposed by Yohai, Stahel & Zamar (1991) (see also Marazzi, 1993, §9.1.3) is an M-estimator starting at the coefficients given by the S-estimator and with fixed scale given by the S-estimator. This retains (for $c > c_0$) the high-breakdown point of the S-estimator and the high efficiency at the normal. At considerable computational expense, this gives the best of both worlds.

Function `rlm` has an option to implement MM-estimation.

```
> summary(rlm(calls ~ year, data=phones, method="MM"), cor=F)
Coefficients:
             Value Std. Error t value
(Intercept) -52.423   2.916    -17.977
       year   1.101   0.047     23.366

Residual standard error: 2.13
```

S-PLUS has a function `lmRob` in library section `robust` which implements a slightly different MM-estimator with similar properties, and comes with a full set of method functions, so can be used routinely as a replacement for `lm`. Let us try it on the `phones` data.

```
> library(robust, first = T)
> phones.lmr <- lmRob(calls ~ year, data = phones)
> summary(phones.lmr, cor = F)
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) -52.541   3.625    -14.493   0.000
       year   1.104   0.061     18.148   0.000

Residual scale estimate: 2.03 on 22 degrees of freedom
Proportion of variation in response explained by model: 0.494

Test for Bias:
            Statistics P-value
 M-estimate      1.401   0.496
LS-estimate      0.243   0.886
```

This works well, rejecting all the spurious observations. The 'test for bias' is of the M-estimator against the initial S-estimator; if the M-estimator appears biased the initial S-estimator is returned.

References

Abbey, S. (1988) Robust measures and the estimator limit. *Geostandards Newsletter* **12**, 241–248.

Analytical Methods Committee (1989a) Robust statistics — how not to reject outliers. Part 1. Basic concepts. *The Analyst* **114**, 1693–1697.

Analytical Methods Committee (1989b) Robust statistics — how not to reject outliers. Part 2. Inter-laboratory trials. *The Analyst* **114**, 1699–1702.

Davies, P. L. (1993) Aspects of robust linear regression. *Annals of Statistics* **21**, 1843–1899.

Dixon, W. J. (1960) Simplified estimation for censored normal samples. *Annals of Mathematical Statistics* **31**, 385–391.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) *Robust Statistics. The Approach Based on Influence Functions*. New York: John Wiley and Sons.

Hettmansperger, T. P. and Sheather, S. J. (1992) A cautionary note on the method of least median squares. *American Statistician* **46**, 79–83.

Huber, P. J. (1981) *Robust Statistics*. New York: John Wiley and Sons.

Iglewicz, B. (1983) Robust scale estimators and confidence intervals for location. In *Understanding Robust and Exploratory Data Analysis*, eds D. C. Hoaglin, F. Mosteller and J. W. Tukey, pp. 405–431. New York: John Wiley and Sons.

Marazzi, A. (1993) *Algorithms, Routines and* S *Functions for Robust Statistics*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. New York: John Wiley and Sons.

Staudte, R. G. and Sheather, S. J. (1990) *Robust Estimation and Testing*. New York: John Wiley and Sons.

Tukey, J. W. (1960) A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*, eds I. Olkin, S. Ghurye, W. Hoeffding, W. Madow and H. Mann, pp. 448–485. Stanford: Stanford University Press.

Yohai, V., Stahel, W. A. and Zamar, R. H. (1991) A procedure for robust estimation and inference in linear regression. In *Directions in Robust Statistics and Diagnostics, Part II*, eds W. A. Stahel and S. W. Weisberg. New York: Springer-Verlag.