Robust Statistical Methods

Anthony C. Atkinson

London School of Economics, UK

Abstract

Robust statistical methods are intended to behave well in the presence of departures from the model that explains the greater part of the data. A contamination model for the data is that the observations y have density

$$f(y) = (1-\epsilon)f_1(y,\theta_1) + \epsilon f_2(y,\theta_2). \tag{1}$$

The simplest example is when $f_1(y,\theta_1) = \phi(\mu,\sigma^2)$, the normal distribution. When there is no contamination ($\epsilon = 0$) the minimum variance unbiased estimator of μ is the sample mean $\bar{y}$. Now suppose that there is some contamination. In the finite sample case, even with $\epsilon = 1/n$, the sample mean has unbounded bias as the observation from $f_2(.)$ becomes increasingly extreme. The estimate breaks down as the observation goes to $\pm\infty$. Asymptotically (as $n \to \infty$) the sample mean has zero breakdown.

The sample median, on the other hand is not so affected. Asymptotically up to half the observations can be moved arbitrarily far away from μ with the median providing an unbiased estimator. However, the variance of the median is asymptotically $\pi/2$, so that the efficiency of the median as an estimator of location is 0.637, although the breakdown point is 50%. An aim of robust statistics is to find estimators that are unbiased in the presence of contamination whilst achieving the Cramer-Rao lower bound. Of course, such estimators do not exist, but breakdown can be traded against variance inflation.

The trade-off is achieved through the use of M-estimators and their extensions. Given an estimator of μ, say $\tilde{\mu}$, the residuals are defined as

$$r_i(\tilde{\mu}) = y_i - \tilde{\mu}. \tag{2}$$

As is well known, the least squares estimate of μ, which is also the maximum likelihood estimate, minimizes the sum of squares

$$\sum_{i=1}^{n} \{r_i(\mu)/\sigma\}^2. \tag{3}$$

Of course the value of σ is irrelevant.

Traditional robust estimators attempt to limit the influence of outliers by replacing the square of the residuals in (3) by a function ρ of the residuals which is bounded. The M (Maximum likelihood like) estimate of μ is the value that minimizes the objective function

$$\sum_{i=1}^{n} \rho\{r_i(\mu)/\sigma\}. \tag{4}$$

Of the numerous form that have been suggested for $\rho(.)$ (Adrews et al., 1972, Hampel et al., 1986, Huber and Ronchetti, 2009) perhaps the most popular choice is Tukey's Biweight function

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if } |x| \leq c \\ \frac{c^2}{6} & \text{if } |x| > c, \end{cases} \tag{5}$$

where c is a crucial tuning constant. For small $x, \rho(x)$ behaves like (3). For large $|x|$ the residuals are constant; the effect of extreme observations is mitigated.

In equation (4) it is assumed that σ is known, yielding the estimate $\tilde{\mu}_M(\sigma)$. Otherwise, an M-estimator of scale $\tilde{\sigma}_M$ is defined as the solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\{r_i(\mu)/\sigma\} = K_c, \tag{6}$$

where both μ and σ are iteratively jointly estimated. K_c and c are related constants which are linked to the breakdown point of the estimator of μ.

Regression, which will be the subject of two of the talks, is more difficult. If the contamination is only in the y direction, M-estimation is appropriate. However, if the x values may also be outlying, leverage points may be present. Then, not only is ordinary least squares exceptionally susceptible to the presence of outliers, but so are M-estimates. Instead, very robust methods, with an asymptotic breakdown point of 50% of outliers are to be preferred.

Very robust regression was introduced by Rousseeuw (1984) who developed suggestions of Hampel (1975) that led to the Least Median of Squares (LMS) and Least Trimmed Squares (LTS) algorithms.

In the regression model $y_i = x_i^T\beta + \epsilon_i$, the residuals in (2) become $r_i(\tilde{\beta}) = y_i - x_i^T\tilde{\beta}$. The LMS estimator minimizes the hth ordered squared residual $r^2_{[h]}(\beta)$ with respect to β, where $h = \lfloor (n + p + 1)/2 \rfloor$ and $\lfloor . \rfloor$ denotes integer part.

The convergence rate of $\tilde{\beta}_{\text{LMS}}$ is $n^{-1/3}$. Rousseeuw (1984, p. 876) also suggested Least Trimmed Squares (LTS) which has a convergence rate of $n^{-1/2}$ and so better properties than LMS for large samples. As opposed to minimising the median squared residual, $\tilde{\beta}_{\text{LTS}}$ is found to

$$\text{minimize } SS_{\text{T}}\{\hat{\beta}(h)\} = \sum_{i=1}^{h} e_i^2\{\hat{\beta}(h)\}, \tag{7}$$

where, for any subset $\mathcal{H}$ of size h, the parameter estimates $\hat{\beta}(h)$ are straightforwardly obtained by least squares.

Unlike M-estimation, these procedures do not require an estimate of σ^2. However, the estimate is required for outlier detection. Let the minimum value of (7) be $SS_{\text{T}}(\tilde{\beta}_{\text{LTS}})$. The estimator of σ^2 is based on this residual sum of squares. However, since the sum of squares contains only the central h observations from a normal sample, the estimate needs scaling. The factors come from the general results of Tallis (1963) on elliptical truncation.

The LMS and LTS estimators are least squares estimates from carefully selected subsets of the data, asymptotically one half for LTS. If there are no, or only a few, outliers, such estimates will be inefficient. To increase efficiency, reweighted versions of the LMS and LTS estimators can be computed, using larger subsets of the data. These estimators are found by giving weight 0 to observations which are determined to be outliers when using the parameter estimates from LMS or LTS. Least squares is then applied to the remaining observations.

An alternative to these forms of very robust estimation is deletion of outliers, starting from a fit to all the data (Cook and Weisberg, 1982, Atkinson, 1985). If there are few outliers, the resulting estimators will be based on most of the data and so will be more efficient than those based on smaller subsets. However, in the presence of many outliers these backwards methods can fail. Atkinson and Riani (2000) suggest a Forward Search (FS) in which least squares is used to fit the model to subsets of the data of increasing size. The process stops when all observations not used in fitting are determined to be outliers. See Atkinson et al. (2010) for a recent discussion of the FS.

In LMS and LTS inference is made from models fitted to subsets of the data of one or two sizes, with perhaps subsets of three different sizes for the reweighted versions. Instead, in the FS the model is progressively fitted to subsets of increasing size. The procedure needs both to reject all outliers, in order to provide unbiased estimates of the parameters, and to use as many observations as possible in the fit in order to enhance efficiency. One thread in the session will be the improved properties of the estimates that result from using this flexible, data-dependent subset size for parameter estimation.

A second thread in the session has to do with efficient computation. The LMS and LTS estimates used are approximations found by least squares fitting to many subsets of observations. As a consequence LMS and LTS estimation (and, in general, all algorithms of robust statistics) spend a large part of the computational time in sampling subsets of observations and then computing parameter estimates from the subsets. In addition, each new subset has to be checked as to whether it is in general position (that is, it has a positive determinant). For these reasons, when the total number of possible subsets is much larger than the number of distinct subsets used for estimation, an efficient method is needed to generate a new random subset without checking explicitly if it contains repeated elements. We also need to ensure that the current subset has not been previously extracted. A lexicographic approach can be found that fulfills these requirements.

In addition to data analysis, robust techniques can be employed in the design of experiments. The model is (1) with $f_1(.)$ typically a regression model and $f_2(.)$ a departure, specified to some extent. In Box and Draper (1963) interest is in protecting second-order response surface models from biases from omitted third-order terms. Only the second-order model will be fitted to the data. The methods of optimum experimental design (Fedorov, 1972, Atkinson et al., 2007) require that a model, or models, be specified. In a series of papers Wiens and co-workers (Wiens and Zhou, 1997, Wiens, 1998, Fang and Wiens, 2000, Wiens, 2009) extend optimum design to partially specified situations. For example, Fang and Wiens (2000) bound the departure between the fitted and true models. They also allow for the possibility of heteroscedastic errors, bounding the magnitude of departure from homoscedasticity. With loss function the average mean squared error of prediction, I-optimal (Atkinson et al., 2007, §10.6) designs are obtained when the data are homoscedastic and the polynomial model is correct ($f_2(.) = 0$). When these conditions do not hold, the robust design replaces the support points of the optimum design with clusters of observations at nearby but distinct sites.

References

[1] Andrews, D.F., P.J. Bickel, F.R. Hampel, W.J. Tukey, and P.J. Huber (1972). *Robust Estimates of Location: Survey and Advances.* Princeton, NJ: Princeton University Press.

[2] Atkinson, A.C. (1985). *Plots, Transformations, and Regression.* Oxford: Oxford University Press.

[3] Atkinson, A.C., A.N. Donev, and R.D. Tobias (2007). *Optimum Experimental Designs, with SAS.* Oxford: Oxford University Press.

[4] Atkinson, A.C. and M. Riani (2000). *Robust Diagnostic Regression Analysis.* New York: SpringerVerlag.

[5] Atkinson, A.C., M. Riani, and A. Cerioli (2010). The forward search: theory and data analysis (with discussion). *J. Korean Statist. Soc. 39*, 117–134.

[6] Box, G.E.P. and N.R. Draper (1963). The choice of a second order rotatable design. *Biometrika 50*, 335–352.

[7] Cook, R.D. and S. Weisberg (1982). *Residuals and Influence in Regression.* London: Chapman and Hall.

[8] Fang, Z. and D.P. Wiens (2000). Integer-valued, minimax robust designs for estimation and extrapolation in heteroscedastic, approximately linear models. *J. Amer. Statist. Assoc. 95*, 807–818.

[9] Fedorov, V.V. (1972). *Theory of Optimal Experiments.* New York: Academic Press.

[10] Hampel, F., E.M. Ronchetti, P. Rousseeuw, and W.A. Stahel (1986). *Robust Statistics*. New York: Wiley.

[11] Hampel, F.R. (1975). Beyond location parameters: robust concepts and methods. *Bull. Int. Statist. Inst. 46*, 375–382.

[12] Huber, P.J. and E.M. Ronchetti (2009). *Robust Statistics*, 2nd Ed. New York: Wiley.

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

[14] Tallis, G.M. (1963). Elliptical and radial truncation in normal samples. *Ann. Math. Statist. 34*, 940–944.

[15] Wiens, D.P. (1998). Minimax robust designs and weights for approximately specified regression models with heteroscedastic errors. *J. Amer. Statist. Assoc. 93*, 1440–1450.

[16] Wiens, D.P. (2009). Robust discrimination designs. *J. R. Stat. Soc. Ser. B Stat. Methodol. 71*, 805–829.

[17] Wiens, D.P. and J. Zhou (1997). Robust designs based on the infinitesimal approach. *J. Amer. Statist. Assoc. 92*, 1503–1511.