A12

A ROBUST DATA SMOOTH BASED ON AN L-ESTIMATOR

Richard T. Carson and Shankar Subramanian, University of California, Berkeley

I. Introduction

Data smooths have become increasingly popular as a means of summarizing the relationship between two variables particularly when a non-linear relationship is suspected or when a graphical display is desired. There are a number of distinct types of data smooths: histograms, nearest-neighbor smooths, kernel smooths, and regression smooths. Of these regression smooths have been the most frequently used because of their speed and ease of interpretation. There are three key features of a regression data smooth. These are its ability to trade off bias and variance, its robustness and resistance properties, and its speed. What we propose in this paper is a regression data smooth based on an L-estimator, trimmed least squares, which represents one attempt to balance these different objectives.

A data smooth by definition estimates the expected value of the variable y_i given the variable x_i. To make the data smooth operational one must place restrictions on the relationship between the x_i and the y_i. The larger the window size used in the smooth the smaller the variance, due to the larger number of points over which the regression is estimated, but the greater the risk of bias if the y_i do not change smoothly with the x_i since a small window size allows the fitted values to change quickly with changes in the x_i. (The ordinary bivariate regression can be seen to simply be a regression data smooth with a window size fixed at the length of the data.) Recent work on regression data smooths has tended to emphasize two approaches to defining window size. The first of these is represented by Friedman and Stuetzle's super smoother (1982) and various spline smooths which allow for a variable number of knots (e.g., Wahba and Wold, 1975). These smooths use either local cross-validation or calculate fitted values for several different window sizes to determine the appropriate window size in different regions of the data space in order to optimize some predefined trade-off between bias and variance (usually mean square error). The second approach uses a fixed window size but assigns decreasing weights to values of x as they get further away from the x_s at which that particular window (or regression) is centered. Cleveland (1979) has combined such a weighting scheme with the robust M-estimator, the bisquare. The data smooth we propose closely follows Cleveland's LOWESS except that it is based on the L-estimator, trimmed least squares. Our data smooth which we will term LOWTESS, for Locally Weighted Trimmed Least Squares, has advantages over LOWESS in some (but certainly not all) situations.

II. Cleveland's LOWESS

Since LOWTESS is basically a modification of LOWESS (Cleveland, 1979), we examine in this section the basic characteristics of LOWESS and define the notation to be used. We will use X and Y to represent the two random variables of interest, and x and y to denote realizations of those variables with subscripts to denote particular values. The fitted value, $\hat{y}_s$, of the y_s at the point x_s depends on all the x_i and y_i in the sample and the properties of the smooth, S.

To define S, one needs to specify the window size - either the fraction of the observations or the range of X values to be included within the window - and how the observations falling within the window are to be used in determining the fitted value y_s. The window size in LOWESS is determined by the fraction, f, of the data points (rounded to the nearest integer) to be included within the window. Cleveland recommends choosing f in the range 0.2 to 0.8, suggesting that a value of 0.5 is a good starting point if the user has little 'feel' for the data. The later implementation of LOWESS in the Bell Laboratories statistical language, S (Becker and Chambers, 1984), uses a default value of 0.667 for f. Obviously, the more data points one has or the greater the suspected nonlinearity, the smaller f should be, while the converse is also true. Cleveland suggest some methods for optimally choosing f but these are generally not practicable for everyday work.

The weight function used by LOWESS is the tricubic. Observations outside the regression window receive zero weight while the weights for observations inside the window decline smoothly (in a symmetric fashion) with the distance from the point x_s at which the window is centered. This weight function is given by

$$W(u) = \begin{cases} (1 - |u|^3)^3 & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

where $u = [(x - x_s)/h_s]$ for any x within the window and h_s is the window width. Cleveland notes that the tricubic weight function provides a "third moment match" of the chi-square distribution with the squared residuals from the smooth in the case where bias is negligible and the underlying distribution is normal.

The next step is to choose the degree, d, of the polynomial in the x_i to be fitted (using a weighted regression) to the y_i within the window. The obvious choice when all the x_i within the window are identical is simply to take the mean of the corresponding y_i as the fitted value, which amounts to taking d=0. Cleveland suggests that a linear fit (d=1) is a good choice because of the computational burden of fitting quadratic or higher-order ($d \geq 2$) polynomials. (We shall see that this is less true of LOWTESS). One interpretation of the degree d of the fitted polynomial is in terms of the order of a Taylor series expansion, where due to the weights and the more local nature of the regression estimated this interpretation is less "inappropriate" than it is in the standard regression case (White, 1980).

The vulnerability of regression smooths to outliers is especially troublesome if the data have a thick-tailed, non-normal distribution or if the relation between the X and Y is not smooth, the more so because the number of data points within the window can be small and because of the large weight given to points near the point x_s where the fitted value is to be found. Cleveland's LOWESS proposal calls for using Beaton and Tukey's (1974) bisquare robust M-estimator.

Let $z_i = (1\ x_i\ x_i^2\ \cdots\ x_i^d)$ and let $W_i = W(x_i)$. The fitted value, $\hat{y}_i$, is given by $z_i\hat{\beta}$, where $\hat{\beta}$ is a solution to

$$\sum_{i=1}^{n} W_i(y_i - z_i\beta)B(W_i^{1/2}(y_i - z_i\beta)/6s) = 0. \qquad (2)$$

Here the biweight weighting function, B, is given by

$$B(z) = \begin{cases}(1-z^2)^2 & \text{if } |z|<1\\ 0 & \text{otherwise}\end{cases} \qquad (3)$$

and s is the median of the $|W_i^{1/2}(y_i - z_i\beta)|$. For a linear fit (d=1), (2) is easily and quickly solved using iteratively reweighted least squares (Coleman et al., 1980), with the least squares estimate of β as the starting value. While several other M-estimators could have been chosen, the biweight has the advantage of having a redescending weight function, assigning zero weight to observations with large residuals.

III. Trimmed Least Squares

An alternative to the family of M-estimators is that of L-estimators which are linear combinations of order statistics. In the simple location case, the best known member of this family of estimator is the α-trimmed mean which includes the ordinary mean (α = 0, no trimming) and the median (α = 0.5) as special cases. A regression analogue of the α -trimmed mean using regression quantiles (Koenker and Basset, 1978; Basset and Koenker, 1982) has been proposed by Koenker and Basset (1978) and Ruppert and Carroll (1980). Extending the concept of quantiles to the linear model, $y_i = z_i\beta + e_i$ (where β is a $p\times 1$ parameter vector), Koenker and Bassett (1978) defined regression quantiles, $\hat{\beta}(\vartheta)$, to be solutions of

$$\min_{\beta} \sum_{i} \rho_{\vartheta}(y_i - z_i\beta),\ 0<\vartheta<1, \qquad (4)$$

where $\rho_{\vartheta}(u) = \vartheta u - uI(u<0)$. Eq. (4) may be recast as a linear programming problem and solved efficiently using a modified version (Fulton, Subramanian and Carson, 1984) of the Barrodale and Roberts algorithm (1974) for the least absolute deviations the estimation problem (i.e., the case *theta* =0.5).

Ruppert and Carroll's (1980) trimmed least squares (TLS) estimator is the least squares estimator computed after dropping the observations corresponding to the $[\alpha n]$th largest and $[\alpha n]$th smallest residuals from a consistent preliminary estimator, β_0. When β_0 is chosen to be the average of $\hat{\beta}(\alpha)$ and $\hat{\beta}(1-\alpha)$, their TLS estimator has an asymptotic distribution (under certain regularity and symmetry assumptions) completely analogous to that of the trimmed mean in the location case.

IV. A Smooth Based on Trimmed Least Squares

Our data smooth simply substitutes TLS for the biweight M-estimator used by Cleveland in LOWESS. LOWTESS and LOWESS produce similar results. It is fairly easy to construct situations where one estimator does quite well in uncovering the underlying "true" relation generating the data while the other does not.

As one is performing a different regression at each data point in the sample space it is easy to see why speed is important in a data smooth. For a linear fit (d=1), LOWTESS, the smooth based on TLS, is slower than LOWESS. Since the LOWESS implementation in S does not make higher-order fits, we have no time comparisons for $d\geq 2$. However, we expect that LOWTESS will be faster than LOWESS for $d\geq 2$ because the time taken to compute regression quantiles increases slower with d than does the time taken for solving the least squares problem. Moreover, we expect that the regression quantile computation can be speeded up substantially by making use of a useful feature of the data smooth. We find that the optimal solution bases (in the LP algorithm) for the regression quantiles computed (in the course of the TLS computations) at two adjacent data points, x_i and x_{i+1}, are very nearly and often the same. As a result, by using the optimal solution bases obtained during the TLS calculations at x_i as the initial bases (in the LP algorithm) at the next data point, x_{i+1}, the number of iterations required to compute the regression quantiles falls very considerably. The fastest TLS-based smooth is when α=0.5, i.e. a smooth based on the LAD estimator.[1]

V. Monte Carlo Results

In the first set of Monte Carlo experiments, the relation between X and Y is linear and the window size of both smooths has been fixed at .667. Three error distributions have been used: a normal, a heavily contaminated normal (30% contamination by a normal distribution with 9 times the variance), and the double exponential. All three error distributions have the same variance. For LOWTESS, results are given for α= 0.1, 0.2 and 0.5 (LAD) and d=1. In the second set of Monte Carlo experiments, the relation between X and Y is quadratic, the window size has been set at 0.4 and the same three error distributions have been used. Here the LOWTESS results are also given for d=2. In the third set of simulations, Y is a cosine function of X, the window size is 0.3, and the same three error distributions have been used. The mean square error from the true curves are given in Tables I, II, and III. In all cases, the mean square error from the actual observations shows a very similar pattern.

The figures provide examples of smooths generated by LOWESS and LOWTESS with different error distributions.

Discussion and Concluding Remarks

Our results at this point are very rough and preliminary. While in almost all cases some variant of LOWTESS performs better than LOWESS, these differences are small and LOWESS usually outdoes most of the variants of LOWTESS. This is particularly true in those cases where there is any curvature in which case LOWTESS must be used with d=2 while LOWESS need not for the same basic level of accuracy. When this is done LOWTESS gives up any advantage in speed to LOWESS. LOWESS appears to be the safer smooth to use.

It has also become apparent in this work how critical the choice of the window size is. Both LOWESS and LOWTESS perform very badly if the window size is poorly chosen. We are currently working on improving the speed of the LAD smooth (which is the fastest of the TLS smooths) to the point where it would be possible to optimize over the window size in the manner of Friedman and Stuetzle's (1982) super-smoother while retaining at the same time some desirable robustness properties.

Table I

MEAN SQUARE ERROR - DEVIATION FROM TRUE LINE

	LOWESS	LOWTESS		
		α=0.1 d=1	α=0.2 d=1	α=0.5 d=1
Normal	81.81	82.32	86.85	108.05
Contaminated Normal	44.66	49.30	44.90	50.31
Double Exponential	50.53	54.66	48.32	47.66

Table II

MEAN SQUARE ERROR - DEVIATION FROM TRUE QUADRATIC

	LOWESS	LOWTESS			
		α=0.1 d=1	α=0.2 d=1	α=0.1 d=2	α=0.2 d=2
Normal	71268	70860	78359	70503	78449
Contaminated Normal	47536	49331	52479	48287	50477
Double Exponential	61952	54055	54941	60028	61637

Table III

MEAN SQUARE ERROR - DEVIATION FROM TRUE COSINE

	LOWESS	LOWTESS			
		α=0.1 d=1	α=0.2 d=1	α=0.1 d=2	α=0.2 d=2
Normal	7.188	7.394	8.716	7.114	8.292
Contaminated Normal	7.120	6.510	7.273	6.449	7.258
Double Exponential	7.259	6.587	7.420	6.705	7.538

Error Distribution: Contaminated Normal

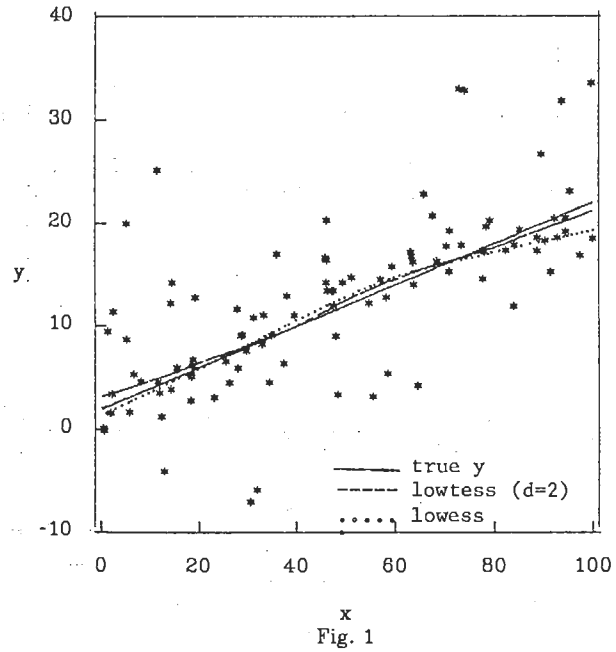

Fig. 1

Error Distribution: Normal

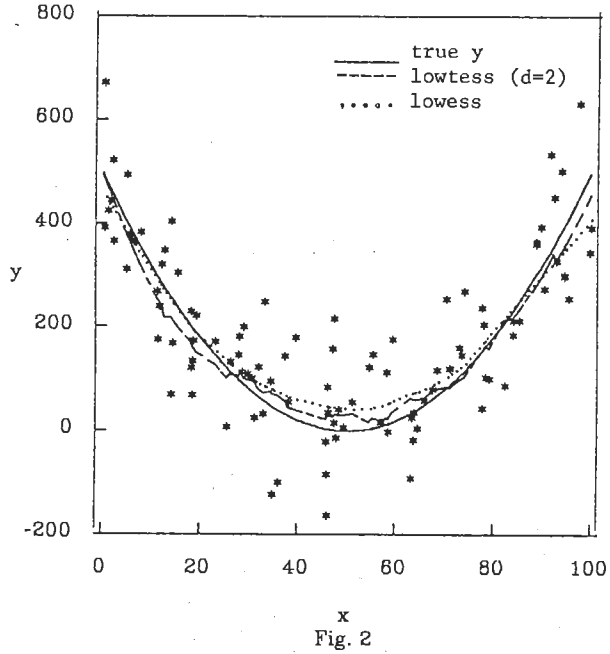

Fig. 2

Error Distribution: Double Exponential

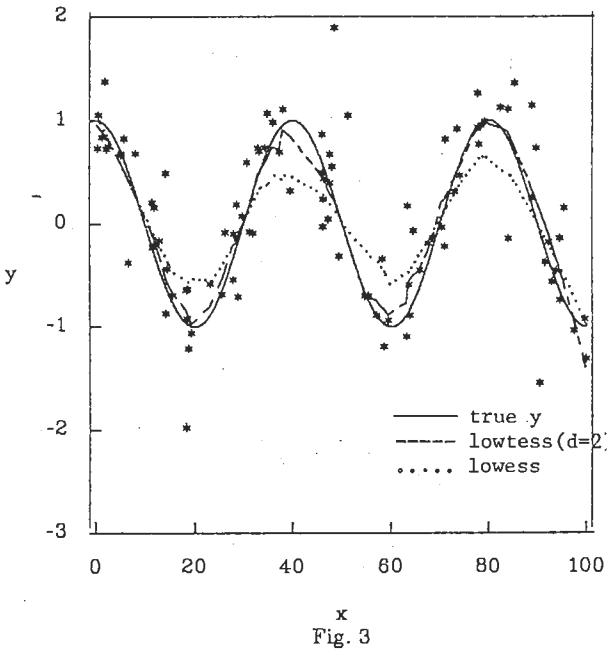

Fig. 3

NOTE

1. The AFK LAD algorithm (Armstrong, Frome & Kung, 1979) allows the user to specify the initial basis. Using this algorithm for the α=0.5 smooth and the trick with the basis elaborated above, the smooth may be speeded up by as much as 30%.

REFERENCES

Armstrong, R.D., Frome, E.L. and Kung, D.S.: "A Revised Simplex Algorithm for the Absolute Deviation Curve Fitting Problem" *Communications in Statistics*, B, (8)1979, 175-190.

Barrodale, I. and F.D.K. Roberts: "Algorithm 478: Solution of an Overdetermined System of Equations in the l_1 Norm", *Communications of the Association for Computing Machinery*, 17(1974), 319-320.

Bassett, G.W. and Koenker, R.W.: "An Empirical Quantile Function for Linear Models with iid Errors," *Journal of the American Statistical Association*, 77(1982), 407-415.

Beaton, A.E. and J.W. Tukey: "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Technometrics*, 16(1974), 147-185.

Becker, R.A. and J.M. Chambers: *S: An Interactive Environment for Data Analysis and Graphics* Belmont, CA: Wadsworth, 1984.

Carroll, R.J. and D. Ruppert: "Robust Estimation in Heteroscedastic Linear Models," *Annals of Statistics*, 10(1982), 429-441.

Chambers, J.M., W.S. Cleveland, B. Kleiner, and P.A. Tukey: *Graphical Methods for Data Analysis* Belmont, CA: Wadsworth, 1983.

Cleveland, W.S.: "Robust Locally Weighted Regression and Smoothing Scatter Plots," *Journal of the American Statistical Association*, 74(1979), 829-836.

Coleman, D., P. Holland, N. Kaden, V. Klema, and S.C. Peters: "A System of Subroutines for Iteratively Re-Weighted Least-Squares Computations," *Association for Computer Machinery. Transactions on Mathematical Software, 6(1980), 327-336.*

Friedman, J. and W. Stuetzle: "Smoothing of Scatter Plots," Technical Report no. ORION006, Department of Statistics, Stanford University, 1982.

Fulton, M., S. Subramanian, and R. T. Carson: "Fast Regression Quantiles Using a Modification of the Barrodale and Roberts l_1 Algorithm," mimeographed, University of California, Berkeley, 1984.

Huber, P.: "Robust Smoothing," in R. Launer and G. Wilkenson, eds., *Robustness in Statistics*, New York: Academic Press, 1979.

Huber, P: *Robust Statistics*, New York: John Wiley, 1981.

Judge, G., W. Griffiths, R.C. Hill, H. Lutkepolhl, T. Lee: *The Theory and Practice of Econometrics*, 2nd. Edition. New York: John Wiley, 1985.

Koenker, R.W. and G.W. Basset: "Regression Quantiles," *Econometrica*, 46(1978), 33-50.

Koenker, R. and G. Bassett: "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, 50(1982), 43-62.

Lehmann, E.L.: *The Theory of Point Estimation*. New York: John Wiley, 1983.

Ruppert, D. and R.J. Carroll: "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, 75(1980), 828-838.

Subramanian, S. and R.T. Carson: "Robust Regression: A Review and Synthesis of Developments Involving Distribution Theory and Non-Spherical Errors," paper presented at the Winter Meeting of the Econometric Society, Dallas, 1984.

Subramanian, S. and R.T. Carson: "Estimation of Trimmed Least Squares in the Presence of Heteroscedasticity," mimeographed, University of California, Berkeley, 1985.

Velleman, P.E.: "Definition and Comparison of Robust Nonlinear Data Smoothing Algorithms," *Journal of the American Statistical Association*, 75(1980), 609-615.

Wahba, G. and S. Wold: "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation," *Communications in Statistics*, 4(1975), 1-17.

White, H.: "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21(1980), 149-170.