Chapter 5

Robust Nonparametric Regression

Introduction

The nonparametric techniques described in Chapter 4 were developed for processes that produced random errors that followed a normal distribution with constant variance. Most of these techniques are asymptotically equivalent to kernel regression, which is not robust to outliers. They are based on a local least squares type of algorithm, which is optimal for processes with normal errors, but performs poorly for errors from heavy-tailed distributions. This motivates the development of the robust nonparametric procedures.

In a groundbreaking article on robust nonparametric regression, Cleveland (1979) introduced a robust local polynomial regression algorithm. The focus of research in the area of nonparametric regression had, up until that point, focused on non-robust versions of smoothing splines and kernel regression.

The idea behind the majority of the forms of robust nonparametric regression is analogous to that of robust parametric regression. An initial local fit is obtained via OLS or some high breakdown method. Large rescaled residuals from this fit are then downweighted (usually via some appropriately chosen ψ - type function, as in parametric M-regression) and a new fit is obtained via IRLS. Iteration follows until the final fit is obtained.

5.A Local M-Estimation

Härdle and Gasser (1984) proposed an estimator $\hat{m}_i$ which minimizes

$$\sum_{i=1}^{n} h_{ij}^{KER} \cdot \rho\left(\frac{y_i - m_i}{\hat{\sigma}}\right). \qquad (5.A.1)$$

The objective function in (5.A.1) is identical to that of the parametric M-estimator of location except for the kernel weights h_{ij}^{KER}. If $\rho(u) = u^2$ in (5.A.1), then $\hat{m}_i$ is the kernel regression estimator.

Hall and Jones (1990) developed the local M-estimator of location further and extend Gasser and Müller's estimator to the case where the values of the regressor are random. As in M-estimation for parametric regression, the choice of a tuning parameter remains as a problem that must be addressed. An additional complication in the nonparametric setting is the freedom to also choose a bandwidth. Hall and Jones emphasized the urgency of the choice of ψ (and its corresponding parameters), since it is demonstrated that the asymptotic MSE of the estimator is a function of the moments of the distribution of $\psi(\varepsilon)$. They also proposed the use of a cross-validation criterion for the simultaneous selection of the bandwidth and the tuning parameter.

The local M-estimator of location is a direct application of M-estimation to the idea of a local fitting procedure. It is a robust local location model fit at each prediction location. Thus, the ψ function downweights observations that have large rescaled residuals, while the kernel weights downweight observations that are remote to x. The local location M-estimator of m at x_i is of the form, using the IRLS algorithm,

$$\hat{m}(x_i) = 1 \cdot (\mathbf{1}'\mathbf{W}_i^{NP}\mathbf{1})^{-1}\mathbf{1}'\mathbf{W}_i^{NP}\mathbf{y},$$

where **1** is an (nx1) vector of ones, $\mathbf{W}_i^{NP}$ = diagonal (w_{ij}), $w_{ij} = h_{ij}^{KER} \cdot \delta_{ij}$, $\delta_{ij} = \psi(r_{ij}^*)/r_{ij}^*$, and r_{ij}^* is the rescaled residual for observation j when predicting at the i^{th} location.

The subscript "i", for δ_{ij} above, is necessary since the residual for each observation is dependent upon the current location of prediction. As before, iteration is necessary since the weights in wls are a function of the unknown parameters. The vector **1** reflects the location model portion of this estimator, indicating that the estimate at x_i is a weighted average of local observations.

The local M-estimator of location has bias problems similar to that of kernel regression in finite samples as a result of the utilization of observations that are not identically distributed (since they have different means as a function of x). However, as in kernel regression, Härdle and

Gasser assume that as $n \to \infty$, the bandwidth $b_n \to 0$, and $n \cdot b_n \to \infty$, which results in the estimator having the property of weak consistency ($\hat{m}(x) \xrightarrow{P} m(x)$). Note that, in order to obtain the unique solution to (5.A.1), a monotone ψ function must be used. This is due to the local minimum of the objective function shown to exist by Birch (1980) if a redescending ψ function is used.

Fan, Hu, and Truong (1994) recently filled a conspicuous gap in the literature caused by the absence of methodology for fitting a local M-type polynomial regression curve. This estimator is given by $\hat{m}_i(x_i) = \mathbf{x}_i^{NP\prime}\hat{\boldsymbol{\beta}}_i$, where $\mathbf{x}_i^{NP\prime}$ is the model vector for location i that reflects the degree of the local polynomial fit, which is defined by the minimization, with respect to $\boldsymbol{\beta}_i$, of

$$\sum_{j=1}^{n} h_{ij}^{KER} \cdot \rho\left(\frac{y_j - \mathbf{x}_j'\beta_i}{\hat{\sigma}}\right).$$

This leads to the IRLS solution for $\boldsymbol{\beta}_i$ given by

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^{NP\prime}\mathbf{W}_i^{NP}\mathbf{X}^{NP})^{-1}\mathbf{X}^{NP\prime}\mathbf{W}_i^{NP}\mathbf{y},$$

where $\mathbf{W}_i^{NP} = \text{diagonal}\ (h_{ij}^{KER} \cdot \delta_{ij})$ and $\delta_{ij} = \psi(r_{ij}^*) / r_{ij}^*$.

The motivation for use of this type of estimator will be developed by highlighting some of the shortcomings of the location M-estimator of location (using NW weights):

- it has bias of order $O(b_n)$ (as do most procedures that estimate the function locally by a constant), while most estimators in the literature (such as local polynomial fitting procedures and spline methods) have bias of order $O(b_n^2)$
- the bias term varies directly with both $m'(x)$ and $f_X'(x)$, which indicates that the estimator cannot adapt well to data with varying density of the x valuess, or with significant curvature of the regression function
- as with kernel regression, there is substantial bias that requires modification at the endpoints.

Fan *et al* (1994) prove that $\hat{m}(x_i)$ is asymptotically normally distributed, and the bias does not depend on either $m'(x)$ or $f_X'(x)$, which reduces the amount of estimation necessary for plug-in type bandwidth selection procedures.

5.B Locally Weighted Regression and Smoothing Scatterplots (Loess/Lowess)

Loess, a robust method of local polynomial regression, was proposed by Cleveland (1979). The procedure is similar to that of local M-estimation, but with a few distinct differences. An initial LPR fit is obtained using r-NN weights based on a kernel function (that is, kernel weights such that exactly r observations in a neighborhood of x_i receive positive weight). These kernel weights, when calculated according to the r-NN scheme of Loess will be denoted h_{ij}^{KER-L}, j=1,…,n . Note that the kernel function that will be used is the tricube weight function, as suggested by Cleveland. The parameter f, which is specified by the user in order to obtain the Loess fit, determines the value of r, the number of observations that receive positive weight when predicting at any particular location in the regressor space. For more discussion on r-NN weights, see Chapter 4.

The initial non-robust LPR fit results in a set of n initial residuals, which give rise to a scale estimate $\hat{\sigma}$. The residuals are rescaled and then transformed into weights by a weight function such as the bisquare function (see Chapter 3). The function used in this research, which is selected for comparison purposes, is Huber's psi function. This set of weights δ_j, j=1,…,n are multiplied by the kernel weights and then used in another iteration of IRLS. This results in the updated weight matrix for prediction at x_i taking the form $\mathbf{W}_i^{Loess} = \text{diagonal}\,(h_{ij}^{KER-L} \cdot \delta_j)$. That is, the vector of estimated coefficients that determine the local polynomial fit at the i^{th} location is given by

$$\hat{\boldsymbol{\beta}}_i^{Loess} = (\mathbf{X}^{NP\prime}\mathbf{W}_i^{Loess}\mathbf{X}^{NP})^{-1}\mathbf{X}^{NP\prime}\mathbf{W}_i^{Loess}\mathbf{y},$$

where $\mathbf{W}_i^{LOESS}$ is defined above, and $\mathbf{X}^{NP}$ is the model matrix for the nonparametric procedures.

Note that the residual weights are not a function of i but remain *constant* across all points of prediction. This process is iterated a total of t times, or until convergence of the predicted values. As mentioned above, the weighting function for the residuals that is used by Loess will be the same as that for the proposed methodology, namely Huber's psi function with $c_H = 1.345$.

The use of residual weights based on a global fit makes Loess unique from local M-estimation. This intuitively seems to be a more reasonable approach, since a residual from a fit that is centered about x_i has little meaning if it corresponds to a location remote from x_i. Examples that we studied for both Loess and M-type local linear regression confirmed our intuitions about which residual weighting scheme was most appropriate. Several data sets were fitted at a grid of locations in the regressor space.

The variable residual weighting scheme (across points of prediction) of the M-type LPR resulted in highly noisy fits. Recall that, in the M-type LPR, there were n sets of n weights given by $\delta_i' = (\delta_{i1}\ \delta_{i2}\ \ldots\ \delta_{in})$, i = 1, …,n. This scheme is the result of fitting a line to all the data points and downweighting based on residuals and location. The downweighting of observations based on residuals from a fitted line at a remote location does not have an intuitive appeal and does not appear to work well in practice. The constant residual weighting scheme of Cleveland (one vector of residual based weights $\delta' = (\delta_1\ \delta_2\ \ldots\ \delta_n)$) resulted in smooth fits that appeared to follow the data trends well.

Cleveland suggests that a linear function for the polynomial fit, and that t = 2 iterations, prove adequate for most applications. He also suggests a smooth kernel function (such as the tricube) in order to obtain a smooth fit.

5.C M-Type Smoothing Splines

The smoothing spline estimate $\hat{m}(x)$ of the true underlying function m(x) minimizes $n^{-1}\sum(y_i - m(x))^2$ subject to $\int (m^{(c)}(x))^2 dx < M$, which turns out to be a polynomial of degree

2c-1 between consecutive x-values. Robustification of this objective function is given by Cox (1983), where the robust estimate of the regression function $\hat{m}(x)$ minimizes

$$n^{-1}\sum \rho(y_i - m(x)) + \lambda \int (m^{(c)}(x))^2 dx$$

where ρ is a convex function that is symmetric about 0, such as the one proposed by Huber (see Huber (1981) for a general discussion of robust regression).

Cox's proposal is a direct extension of the robust methods of M-estimation to the spline regression estimator. The influence of an outlier is decreased via the ρ function, which replaces the non-robust quadratic loss function.

5.D L_1 Nonparametric Regression

Wang and Scott (1994) outlined the L_1 nonparametric estimator and a few advantages it maintains over estimates in the previous sections. The authors proposed several universally desirable properties of a robust nonparametric estimator, including

- a non-iterative method
- an outlier-resistant method
- a method that can easily be extended to higher dimensions.

The proposed estimator $\hat{m}_i$ minimizes the function

$$\sum_{j=1}^{n} h_{ij} |y_j - \hat{m}_i(x_j)|$$

where $\hat{m}_i(x_j)$ is a polynomial in x_j. Notice that this is the local M-estimator using the special ρ function $\rho(u) = |u|$. This ρ function is special because of the constant slope on either side of 0, and the lack of any parameters in its definition.

The L_1 estimator also does not have some of the complexities/undesirable properties of the aforementioned estimators, such as the iteration required for kernel, spline, and Loess smoothers. An additional advantage is that there are no tuning parameters that must be estimated or specified in addition to the bandwidth. This also relieves the user of the need for a scale estimate, since

there is no need to rescale the residuals, as the L_1 norm is of constant form for all values of the residuals.

The L_1 estimator has a closed form solution via linear programming techniques, which avoids the problems of iteration. See Wang and Scott (1994) for an outline of the solution algorithm for the L_1 estimator.

5.E Robust Local Linear Regression (RLLR)

The method that will be used in this research is termed Robust Local Linear Regression (RLLR) and is based on local M-estimation while utilizing an advantageous component of Loess. It appears to fit as well as any of the robust nonparametric procedures studied. The constant weighting scheme described previously will be employed because of its intuitive appeal and its observed feasibility in applications. The difference in this estimator from Loess is that r-NN weights will not be used, and the neighborhood weights will be determined using the simplified normal kernel function $K(u) = e^{-u^2}$. Instead of the parameter f utilized by Loess, a data driven global bandwidth is used to determine the window width of influential observations in predicting at a certain location.

The estimated coefficients for the local polynomial fit using RLLR when predicting at location i are given by

$$\hat{\boldsymbol{\beta}}_i^{RLLR} = (\mathbf{X}^{NP\prime}\mathbf{W}_i^{RLLR}\mathbf{X}^{NP})^{-1}\mathbf{X}^{NP\prime}\mathbf{W}_i^{RLLR}\mathbf{y},$$

where $\mathbf{W}_i^{RLLR} = \text{diagonal}\ (h_{ij}^{KER-R} \cdot \delta_j)$. As mentioned above, the neighborhood weights for prediction at the i^{th} location h_{ij}^{KER-R} j=1,…,n are determined using the simplified normal kernel function, and the residual-based weights δ_j, j=1,…,n are determined using Huber's psi function.

This estimator is intuitively appealing in that it allows the data to dictate the optimal support of the kernel function (which is equivalent to determining the optimal value of the bandwidth). It will play a key role in the final form of the proposed estimator in Chapter 6.

5.F Example of Robust Local Linear Regression

Consider the last example in Chapter 4 in which an outlier had a local effect on the nonparametric fit. A desirable property of a robust nonparametric fit is that it mimic what the nonparametric fit would be if the outlier were not present, which is actually an interpretation of a robust procedure - one that results in fitted values that would have been obtained were the outlier(s) not present. Figure 5.1 is a plot of a RLLR fit to the same data set used as an example in Chapter 4.

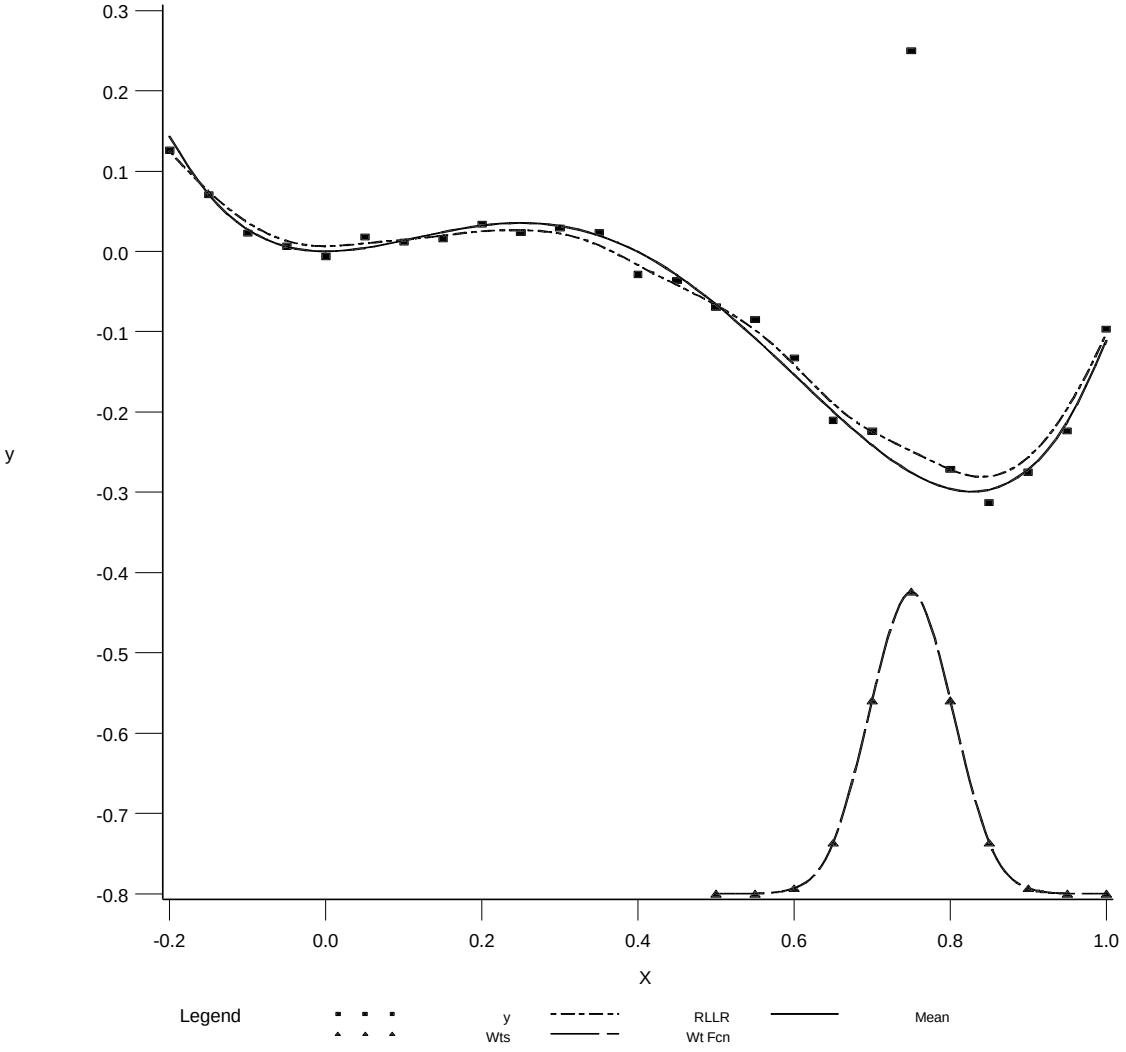

Figure 5.1 Plot of a robust local linear regression fit to a simulated data set with an outlier. The estimator remains virtually unaffected by the outlier and mimics the local linear fit for the data without the outlier (see Chapter 4).

The outlier in this data set has essentially no influence on the fitted values from RLLR. As mentioned before, however, the nonparametric fit is more variable than the parametric one, and that variability increases with the value of σ. An incorporation of the parametric and nonparametric fits provides a means of reducing variability *and* bias. This concept will be discussed in the next chapter.