Chapter 6

Proposed Methodology

Introduction

The proposed methodology (which will be termed Outlier Resistant Model Robust Regression, or ORMRR) is motivated by the desire to formulate a tactic that handles varying degrees of model misspecification in the presence of outliers in an automatic fashion. It is also meant to be an estimator that performs well in small samples (for example, a sample as small as n = 10 to 20) as well as large. We make use of a paper by Mays and Birch (1996), which develops an estimator that improves upon model robust procedures proposed by Einsporn and Birch (1988) and (1993). These methods were formulated and developed for the case of normal errors with constant variance. The proposed methodology utilizes the method of Mays and Birch in an outlier robust manner.

The combined use of robust parametric and nonparametric procedures, as opposed to the classical procedures, in this model robust context complicates the estimator in several ways, the most obvious of which are that extra parameters are involved (in the estimator) and thus require estimation. In addition, the robust techniques require iterative methods for solution. The latter difficulty not only complicates the calculation of the estimator, but also the theoretical properties of the estimator.

6.A Model Robust Regression

Model Robust Regression 1 (MRR1), as it will be referred to here, was developed by Einsporn and Birch (1988) and (1993) as a means of combining parametric and nonparametric fits via a convex combination

$$\hat{y}_i^{MRR1} = (1-\lambda)\hat{y}_i^{OLS} + \lambda\hat{y}_i^{NP},$$

where $\hat{y}_i^{OLS}$ is the fitted value using ordinary least squares, $\hat{y}_i^{NP}$ is the fitted value using a nonparametric procedure, and $\lambda \in [0,1]$ is the mixing parameter.

Mays and Birch (1996) developed another model robust procedure and termed it Model Robust Regression 2 (MRR2) which appears to be an improvement over the MRR1 procedure. The idea is to supplement the parametric fit with a nonparametric fit by utilizing the fitted model

$$\hat{y}_i^{MRR2} = \hat{y}_i^{OLS} + \lambda\hat{r}_i^{NP},$$

where $\hat{r}_i^{NP}$ is the fitted value to the ith residual from the OLS fit using a nonparametric procedure. As in MRR1, $\lambda \in [0,1]$. MRR2 was developed to avoid some problems that occur with MRR1 (See Mays and Birch (1996)). Theoretical properties of MRR1 and MRR2 were developed based on the assumption of a true underlying model of the form

$$y_i = m(x_i) + \varepsilon_i.$$

The user's specified model is a linear model, and we rewrite the true function m in terms of the user's model as

$$\begin{aligned} y_i &= m(x_i) + \boldsymbol{\varepsilon}_i \\ &= \mathbf{x}_i^{P'}\boldsymbol{\beta} + f(x_i) + \boldsymbol{\varepsilon}_i \end{aligned}$$

where f is an unknown nonlinear function. Written in matrix notation, we have that $\mathbf{y} = \mathbf{X}^P\beta + \mathbf{f} + \varepsilon$. Our research is an attempt to take advantage of the flexibility of MRR2, but develop a procedure that is robust to the presence of outliers.

6.B Development of Methodology

The ORMRR estimator is comprised of two separate stages of *robust* estimation. The first will be a robust parametric fit to the raw data using a linear model (via M-estimation). The second will be a robust nonparametric smoothing of the residuals that result from the robust parametric fit (via RLLR). These estimates will be combined through the mixing parameter. These stages of estimation lead to a fitted model of the form

$$\hat{y}_i^{ORMRR} = \mathbf{x}_i^{P'}\hat{\boldsymbol{\beta}}^M + \lambda\,\hat{r}_i^{RLLR},$$

where $\hat{r}_i^{RLLR}$ is the smoothed value of the ith residual from the robust parametric fit and $\lambda \in [0,1]$ is the mixing parameter. The notation for the ith residual from the robust parametric fit will be r_i, and these values constitute the response variable in the robust nonparametric fit.

Expanding the estimator to more fully understand its components, we have that

$$\hat{\mathbf{y}}^{ORMRR} = \mathbf{H}^M\mathbf{y} + \lambda\mathbf{H}^{RLLR}\mathbf{r} \qquad (6.B.1)$$

using the following notation

- $\mathbf{H}^M$ is the M-estimation Hat matrix such that

  $$\mathbf{H}^M = [\mathbf{h}_i^{M'}] = [\mathbf{x}_i^{P'}(\mathbf{X}_p{}'\mathbf{W}^M\mathbf{X}_p)^{-1}\mathbf{X}_p{}'\mathbf{W}^M] \qquad (6.B.2)$$

  Note that the columns for the $\mathbf{X}^P$ matrix in (6.B.2) are the elements chosen for the parametric model.

- $\mathbf{W}^M$ is the weight matrix for M-estimation with diagonal elements of the form $w_i^M = \psi^M(r_i^*)/r_i^*$, where ψ^M is the ψ function chosen for the robust *parametric* fitting portion of the procedure, $r_i^* = \dfrac{y_i - \mathbf{x}_i^{P'}\hat{\beta}^M}{\hat{\sigma}}$, and the M-estimate is given by $\hat{\boldsymbol{\beta}}^M = (\mathbf{X}^{P'}\mathbf{W}^M\mathbf{X}^P)^{-1}\mathbf{X}^{P'}\mathbf{W}^M\mathbf{y}$, and $\hat{\mathbf{y}}^M = \mathbf{H}^M\mathbf{y}$.

- $\mathbf{H}^{RLLR}$ is the Hat matrix from the Robust Local Linear Regression (RLLR) fit *to the residuals,* where the ith row of $\mathbf{H}^{RLLR}$ is given by

  $$\mathbf{h}_i^{RLLR'} = \mathbf{x}_i^{NP'}(\mathbf{X}^{NP'}\mathbf{W}_i^{RLLR}\mathbf{X}^{NP})^{-1}\mathbf{X}^{NP'}\mathbf{W}_i^{RLLR}\mathbf{y}$$

  where $\mathbf{X}^{NP}$ is the matrix for the nonparametric local, *linear* fit.

- $\mathbf{W}_i^{RLLR}$ is the weight matrix for the RLLR fit to the residuals, with diagonal elements of the form $w_{ij}^{RLLR} = h_{ij}\cdot\delta_j$, where h_{ij} is the usual kernel weight applied to the residuals for observation r_j when predicting at $x = x_i$, $\delta_j = \psi^{RLLR}(r_j^{**})/r_j^{**}$, ψ^{RLLR} is the ψ function chosen for the robust nonparametric portion of the fit, and r_j^{**} is the rescaled residual from the robust nonparametric fit *to the residuals*.

In order to obtain the Hat matrix for ORMRR, we rewrite the fitted model, using the notation above, as

$$\begin{aligned}\hat{\mathbf{y}}^{\mathrm{ORMRR}} &= \mathbf{H}^{\mathrm{M}}\mathbf{y} + \lambda\mathbf{H}^{\mathrm{RLLR}}[\mathbf{y} - \mathbf{X}^{\mathrm{P}}\hat{\boldsymbol{\beta}}^{\mathrm{M}}] \\ &= \mathbf{H}^{\mathrm{M}}\mathbf{y} + \lambda\mathbf{H}^{\mathrm{RLLR}}\mathbf{y} - \lambda\mathbf{H}^{\mathrm{RLLR}}\mathbf{H}^{\mathrm{M}}\mathbf{y} \\ &= \left[\mathbf{H}^{\mathrm{M}} + \lambda\mathbf{H}^{\mathrm{RLLR}} - \lambda\mathbf{H}^{\mathrm{RLLR}}\mathbf{H}^{\mathrm{M}}\right]\mathbf{y} \\ &= \left[\mathbf{H}^{\mathrm{M}} + \lambda\mathbf{H}^{\mathrm{RLLR}}\left[\mathbf{I} - \mathbf{H}^{\mathrm{M}}\right]\right]\mathbf{y} \\ &= \mathbf{H}^{\mathrm{ORMRR}}\mathbf{y}\end{aligned}$$

The mixing parameter λ is usually chosen via a cross-validation (cv) type of criterion that is frequently selected to be an unbiased estimate of the mean squared error of the fit. The interpretation of λ is in the context of lack-of-fit of the specified model. A value of λ near zero has the implication that the procedure has determined that the model is correct. A value of λ near 1 indicates a model that is misspecified. One should be careful, however, in these interpretations, in that λ only gives an indication whether or not the model is correctly specified, and not the *degree* to which it may be misspecified.

The intricacy of the estimator can be seen in the explanation of the notation of (6.B.1). The elements w_i^{RLLR} and w_i^{M} in the weight matrices are random variables since they depend on residuals. If we assume that the bandwidth b_n and the mixing parameter λ are both data driven (a common assumption), then both should be considered as random variables. In addition, different fitted values will be obtained through the different choices of ψ^{M}, ψ^{RLLR}, and K (the kernel function), and the multitude of possibilities for parameter choices in these functions.

6.C Example of Outlier Resistant Model Robust Regression

To illustrate the advantages of the ORMRR method, consider the model

$$E(y|x) = -.4(x-.5)^2 \ +.5 + (\cosh(x) + \tanh(x)),$$

which contains a linear component and a nonlinear component. Consider the following fit of a quadratic model to simulated data from this underlying model via M-estimation. The true underlying mean function is represented in Figure 6.1 by the solid curve, and the M-estimate is given by the dashed curve. Note the outlying observations that occur at x = 0.0 and at x = 1.8, and their lack of an effect on the fit of the quadratic model. The robust fit of the quadratic model produces a reasonable fit, but the fit is poor in several areas due to model misspecification. This is especially evident in the region x = -2.5 to x = -0.8, where the user specified quadratic model overestimates the mean function, and in the region x = -0.6 to x = 1.6 where the estimated function dips well below the mean function.

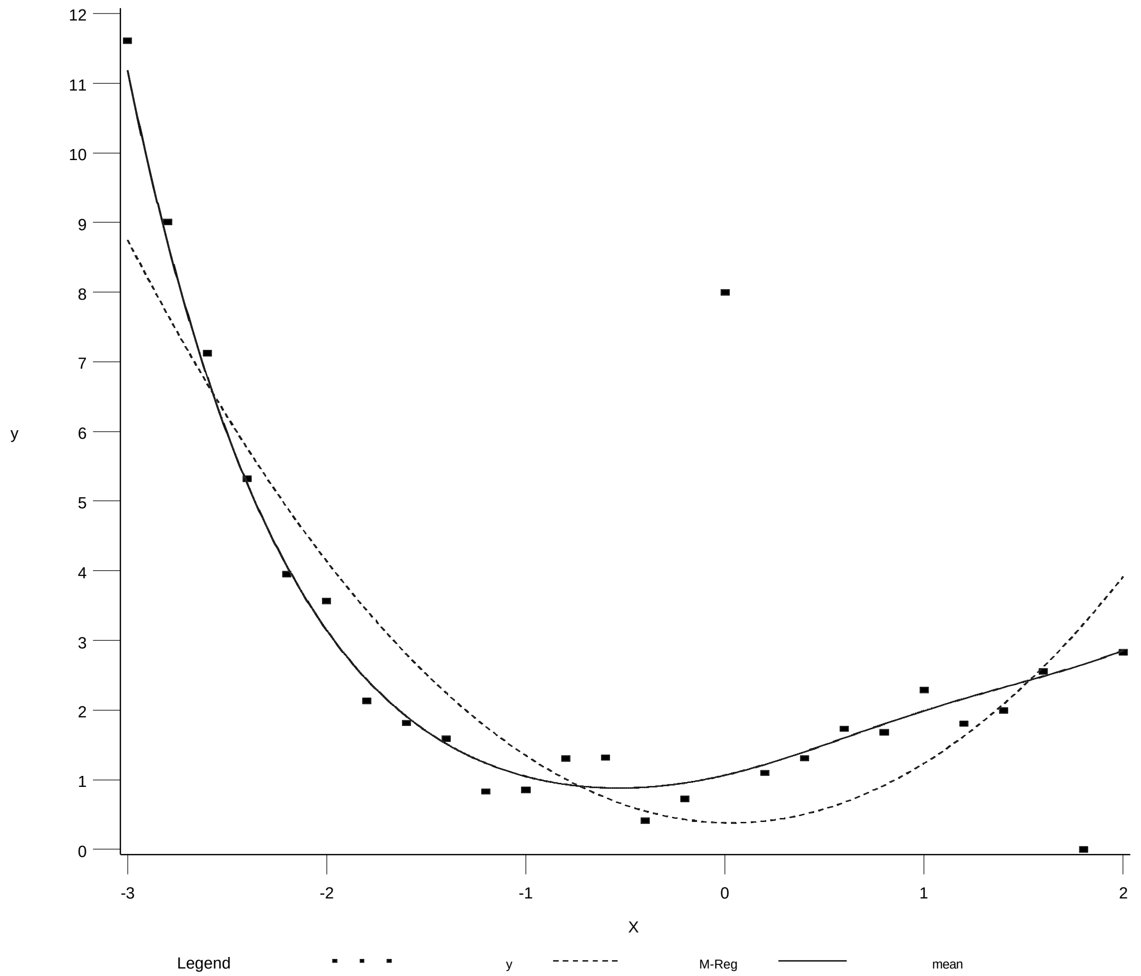

Figure 6.1 Plot of a robust quadratic fit (using M-regression) to a data set generated from the model $E(y|x) = -.4(x-.5)^2 +.5 + (\cosh(x) + \tanh(x))$ ($\sigma = 0.25$) with outliers at the points x = 0.0 and x = 1.8. The model is obviously insufficient in the region x = -2.5 to x = -0.8, and in the region x = -0.6 to x = 1.6.

Consider now the RLLR fit to the residuals that resulted from the initial parametric fit using M-estimation in Figure 6.2. In regression, a common analytical technique is to look for the residuals to be randomly distributed about zero. If they are not, then the general conclusion is that either the model is doing a poor job of fitting the data (that is, significant lack of fit), or the variance is heterogeneous. The plot of the residuals in Figure 6.2 demonstrates an obvious trend. The dashed curve is the RLLR fit to these residuals.

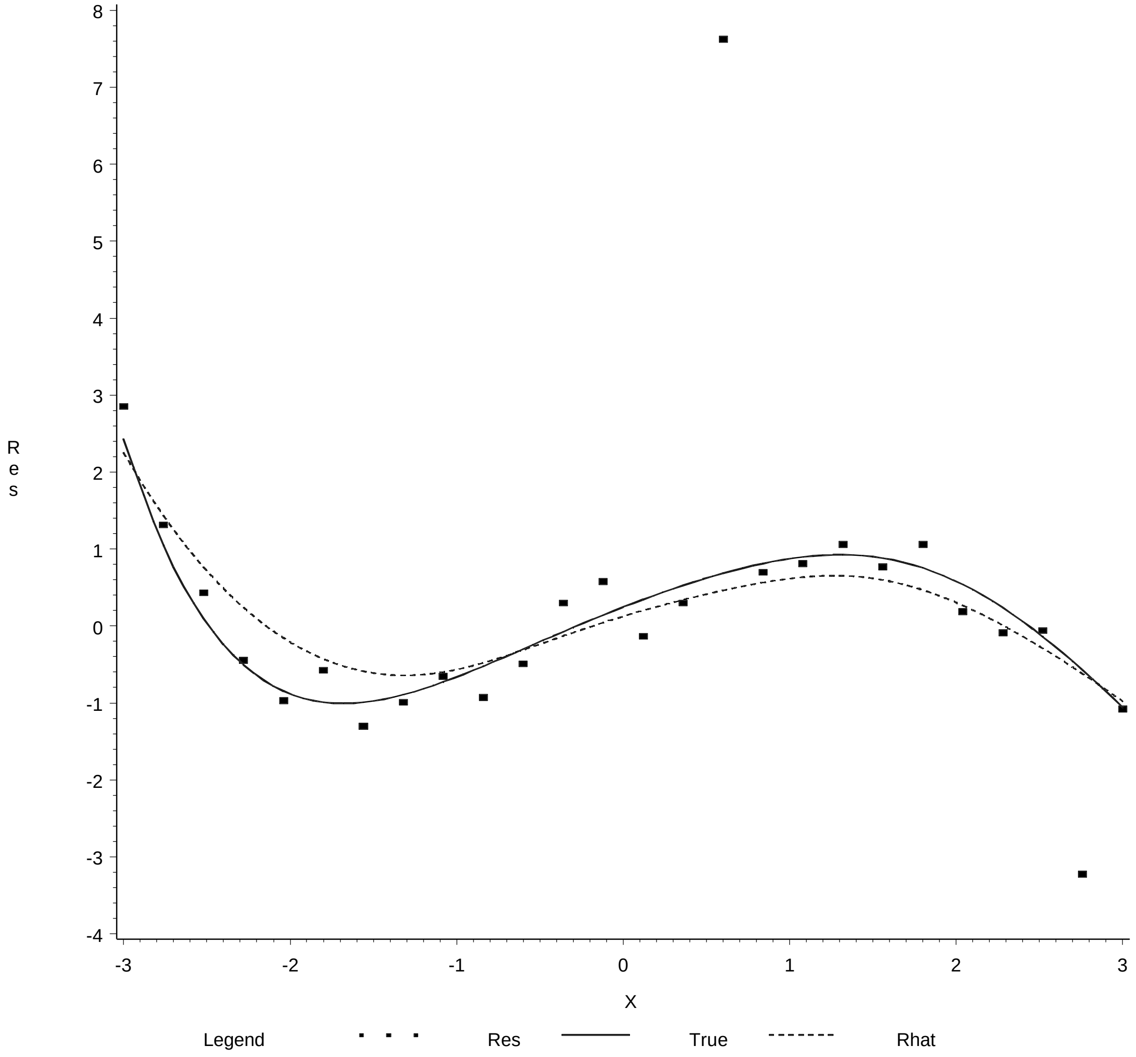

Figure 6.2 Plot of a RLLR fit to residuals from the fit to the raw data using M-regression in Figure 6.1. Note outliers at x = 0.0 and at x = 1.8. The bandwidth used here is $b_n = 0.45$.

Adding back a portion of these residuals to the initial parametric fit results in the final ORMRR fit demonstrated in Figure 6.3.

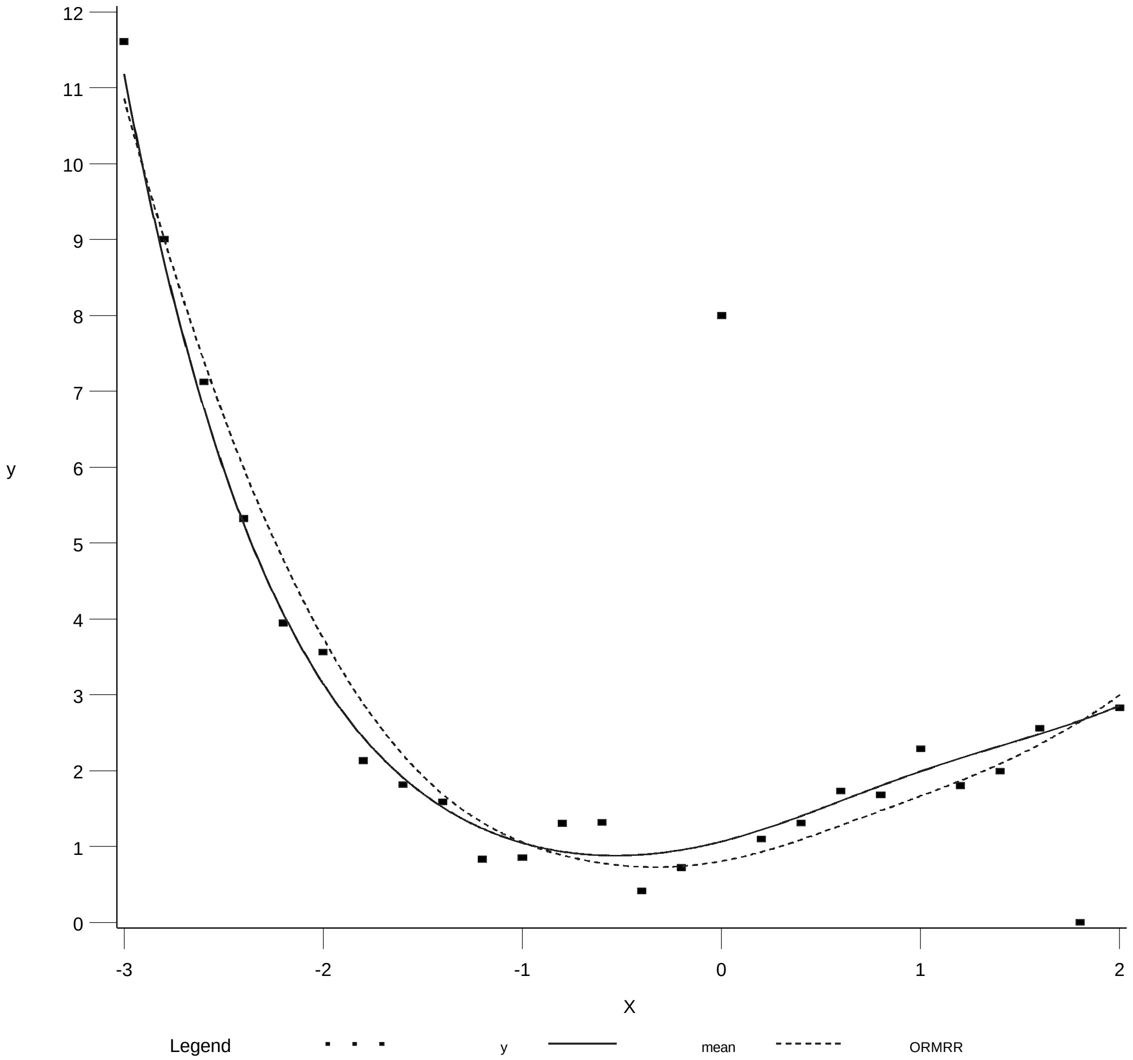

Figure 6.3 Plot of ORMRR fit to a data set generated from the model $E(y|x) = -.4(x-.5)^2 + .5 + (\cosh(x) + \tanh(x))$ with outliers at the points x = 0.0 and at x = 1.8.

This demonstrates the effectiveness of the ORMRR procedure in a specific situation that includes a significant amount of model misspecification and a few small outlying observations.

Figure 6.4 is a comparison of the ORMRR fit to the Loess fit. Notice that the Loess fit appears to be slightly more variable and it does not do a good job of fitting the data at the left boundary. These observations were downweighted as possible outliers by the method that Loess utilizes for identifying outliers.

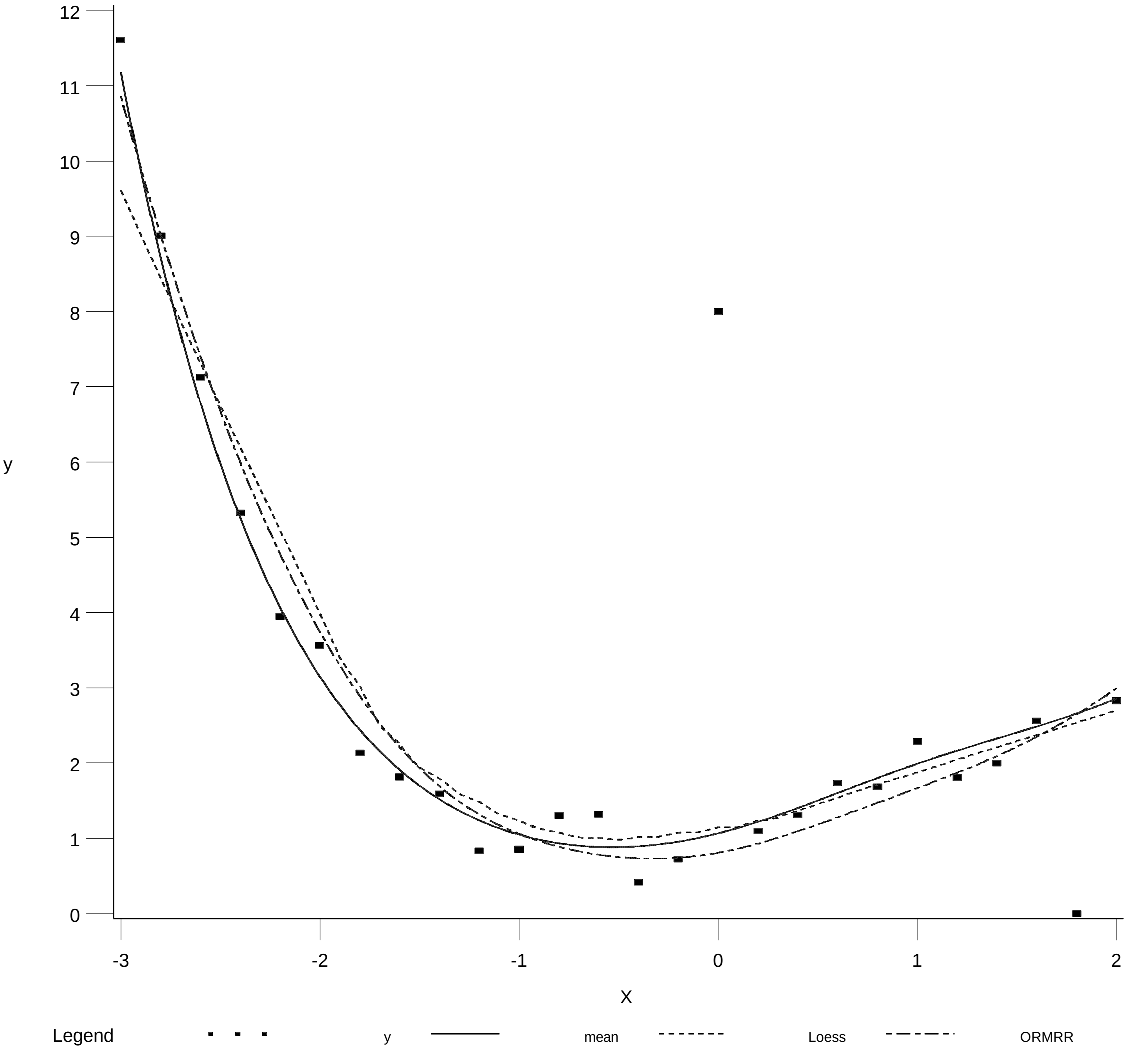

Figure 6.4 Plot of a robust nonparametric fit (using Loess) as compared with a fit using ORMRR to a data set generated from the model E(y|x) = -.4(x-.5)^2 +.5 + (cosh(x) + tanh(x)) with outliers injected at the points x = 0.0 and at x = 1.8. The fit using Loess appears to be somewhat more variable than that of ORMRR and performs poorly at the left boundary.