Chapter 7

Theoretical Comparisons

7.A Introduction

The need for the development of methodology that can be used to compare the previously discussed techniques motivates the theoretical derivations that are to follow in this chapter. These developments will be utilized to determine if Outlier Resistant Model Robust Regression (ORMRR) is a viable technique, and if so, in which situations it performs the best. Based on the evaluation of ORMRR in several select situations, we hope to be able to demonstrate the wide applicability of the procedure to a variety of data sets to which an analyst may be exposed.

A common evaluation mechanism of fitted regression models in particular, and estimators in general, is mean squared error (mse), the expected squared distance between the estimator and the quantity being estimated, which is equivalent to squared bias plus variance. For a given situation, actual numeric values of the mse for the competing fitting procedures will be calculated for comparison purposes. The mse is an obvious choice as there is, as a general rule in the regression setting, a trade-off between the amount of bias and the amount of variance associated with the estimates.

A good example of this trade-off, as will be detailed in Section 7.B, is the nonparametric fit produced by local linear regression. As the bandwidth $b_n \to 0$, the fit resembles a "connect-the-dots" picture (highly variable), while the bias tends toward zero. Conversely, as $b_n \to \infty$, the fit tends toward a straight line (small variability), but suffers from bias problems. A fit that has both a moderate amount of bias and variance is considered to be a good, stable fit.

Another application of the mse is its usefulness in acting as a criterion for parameter selection. Since the mse of a fitting procedure depends, in part, on parameters such as the bandwidth and mixing parameter (as will be demonstrated in the next section), the optimal (in terms of mse) parameters b_{opt} and λ_{opt} should be chosen in some way to *minimize* the mse of the

fitted values associated with that procedure. The actual computations involved in using the mse formulas (that will be presented in Section 7.B) to arrive at the optimal parameters will be discussed in Section 7.C, and will also be investigated further in Chapter 8.

7.B MSE Criteria

This section introduces the derivations of mse formulas for M-Regression, ORMRR, Loess, and RLLR (those of OLS are omitted since they are well-documented in the literature (see Myers (1990), e.g.)). The final form of the bias and variance of the fits for the vector of predicted values ($\hat{\mathbf{y}}$) are given, and the reader will be referred to the Appendix B for details of the derivations. Note that once the bias and variance formulas are derived, the mse follows directly since it is a function of only bias and variance.

Formulations of the bias and variance are performed using the technique employed by both Speckman (1988) and Mays and Birch (1996) in which the fitting parameters, namely, the bandwidth and mixing parameter (where appropriate), are assumed to be fixed quantities. This assumption simplifies the development of, while also resulting in reliable forms of, the mse quantities. In addition, the formulas are asymptotic (following the results in Huber (1981) for M-Regression; note also that the nonparametric techniques are evaluated from the viewpoint that they are little more than local applications of M-Regression).

The mse of the fits associated with any particular method depend on a variety of factors, one of which is the underlying model. As described in Chapter 6, the general form of the model will be expressed as

$$\begin{aligned} \mathbf{y} &= \mathbf{m} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}^{P}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon}\,. \end{aligned} \qquad (7.B.1)$$

This specification is necessary in order to complete the bias and variance derivations that, together, define the mean squared error of the estimator. Recall that $\boldsymbol{\beta}$ is the vector of coefficients that provide the "best" linear fit (the $\boldsymbol{\beta}$ that would result if fitting to actual means

(rather than raw data) using OLS to fit a model that is linear in the parameters), and **f** is the portion of the mean function **m** that cannot be explained by the linear model.

Another factor that influences the mse derivations is the distribution of the random errors. In the current research, the noise associated with the data is assumed to come from a contaminated normal distribution ($\varepsilon \sim CN(\pi, \sigma_1, \sigma_2) = (1-\Delta) \cdot N(0, \sigma_1) + \Delta \cdot N(0, \sigma_2)$, where Δ is a Bernoulli random variable with probability of success π). This allows us the flexibility of assuming that any number of outliers can occur at any location in the regressor space. This is a non-restrictive assumption that will make the results easily generalized to a variety of situations.

Based on this assumption of a contaminated normal distribution, the errors have mean zero and constant variance

$$\sigma_\varepsilon^2 = (1-\pi)\cdot\sigma_1^2 + \pi\cdot\sigma_2^2 .$$

The common variance σ_ε^2 will be referred to as σ^2 in the subsequent mse formulas.

M-Regression

The development of the bias and variance formulas for M-Regression is a logical starting point since subsequent methods that will be discussed all have a similar theoretical basis as M-Regression. Recall that the fitted values for M-Regression are of the form

$$\hat{\mathbf{y}}^M = \mathbf{X}^P \cdot \hat{\boldsymbol{\beta}}^M = \mathbf{X}^P (\mathbf{X}^{P\prime}\mathbf{W}\mathbf{X}^P)^{-1}\mathbf{X}^{P\prime}\mathbf{W}\mathbf{y} = \mathbf{H}^M\mathbf{y} .$$

The fully iterated M-estimator $\hat{\boldsymbol{\beta}}^M$ is asymptotically normal with mean $\boldsymbol{\beta}$ and variance

$$V^2(\mathbf{X}^{P\prime}\mathbf{X}^P)^{-1} \text{ , where } V^2 = \sigma^2 \frac{E\left\{\psi^2\left(\frac{\varepsilon}{\sigma}\right)\right\}}{E\left\{\psi'\left(\frac{\varepsilon}{\sigma}\right)\right\}}$$

(Huber, 1981). Below are the derivations for the asymptotic bias and asymptotic variance for $\hat{\mathbf{y}}^M$. All other derivations are in appendix A. The asymptotic bias of the vector of predicted values that result from M-Regression is given by

$$
\begin{aligned}
\text{Bias}\left(\hat{\mathbf{y}}^{M}\right) &= \mathrm{E}(\hat{\mathbf{y}}^{M}) - \mathrm{E}(\mathbf{y}) \\
&= \mathrm{E}(\mathbf{H}^{M}\mathbf{y}) - \mathrm{E}(\mathbf{y}) \\
&\rightarrow \mathbf{H}^{OLS}\mathrm{E}(\mathbf{y}) - \mathbf{m} \qquad \text{(since weights are data based, and these calculations are asymptotic)} \\
&= \mathbf{H}^{OLS}(\mathbf{X}^{P}\boldsymbol{\beta} + \mathbf{f}) - (\mathbf{X}^{P}\boldsymbol{\beta} + \mathbf{f}) \\
&= \mathbf{H}^{OLS}\mathbf{X}^{P}\boldsymbol{\beta} + \mathbf{H}^{OLS}\mathbf{f} - \mathbf{X}^{P}\boldsymbol{\beta} - \mathbf{f} \\
&= \mathbf{X}^{P}\boldsymbol{\beta} + \mathbf{H}^{OLS}\mathbf{f} - \mathbf{X}^{P}\boldsymbol{\beta} - \mathbf{f} \qquad (\text{since } \mathbf{H}^{OLS}\mathbf{X}^{P} = \mathbf{X}^{P}) \\
&= -(\mathbf{I} - \mathbf{H}^{OLS})\mathbf{f} \,. \qquad (7.B.2)
\end{aligned}
$$

The asymptotic variance of this vector of predicted values is given by

$$
\begin{aligned}
\text{Var}\left(\hat{\mathbf{y}}^{M}\right) &= \text{Var}(\mathbf{H}^{M}\mathbf{y}) \\
&\rightarrow \mathbf{H}^{OLS}\text{Var}(\mathbf{y})\mathbf{H}^{OLS\prime} \\
&= \mathbf{H}^{OLS} \cdot \mathrm{V}^{2}\mathbf{I} \cdot \mathbf{H}^{OLS\prime} \\
&= \mathrm{V}^{2} \cdot \mathbf{H}^{OLS}. \qquad (7.B.3)
\end{aligned}
$$

Note that the magnitude of the bias of M-Regression is directly related to the degree of departure of the *true* underlying model from the *specified* linear model (this difference is determined by the vector **f**). If the true underlying model *is* linear, then **f** = **0**, and the fits based on the M-estimator are asymptotically unbiased. In addition, the asymptotic bias of the M-estimator is identical to that of the OLS estimator, and the asymptotic covariance matrices differ only by a constant.

The expressions used to calculate the asymptotic bias and variance at non-data points may be obtained in a similar manner. If x_0 is any value of x, and $\mathbf{x}_0^{P\prime} = (1\ \mathrm{x}_0\ \mathrm{x}_0^2\ \ldots)$ is the vector that corresponds to x_0 in the parametric model space , then

$$\text{Bias}\left\{\hat{\mathrm{y}}^{M}(\mathrm{x}_0)\right\} = \mathbf{x}_0^{P\prime}(\mathbf{X}^{P\prime}\mathbf{X}^{P})^{-1}\mathbf{X}^{P\prime}\mathbf{f} - \mathrm{f}(\mathrm{x}_0) \qquad (7.B.4)$$

$$\text{Var}\left\{\hat{\mathrm{y}}^{M}(\mathrm{x}_0)\right\} = \mathrm{V}^{2}\mathbf{x}_0^{P\prime}(\mathbf{X}^{P\prime}\mathbf{X}^{P})^{-1}\mathbf{x}_0^{P} \qquad (7.B.5)$$

Robust Local Linear Regression / Loess

The bias and variance derivations for these two estimators are practically identical, which results from the fact that they differ only by the neighborhood weighting functions. For notation purposes, let $\mathbf{H}^{LLR} = \left[h_{ij}^{LLR}\right]$, i = 1,…,n, j = 1,…,n be the "smoother matrix" of *neighborhood* weights (non-robust weights) for the corresponding nonparametric technique; that is, weights calculated using the tricube function for Loess, and the simplified normal for RLLR. Using this notation, recall that the non-robust LLR fitted value at the i^{th} location is given by

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

where $\mathbf{x}_i^{NP'} = (1\ x_i)$ is the vector that corresponds to x_i in the nonparametric model space, and the diagonal weight matrix $\mathbf{W}_i^{LLR} = < h_{ij}^{LLR} >, j = 1,\ldots,n$. Based on the above notation, the vector of predicted values at the n data locations for local linear regression can be expressed as

$$\hat{\mathbf{y}}^{LLR} = \mathbf{H}^{LLR}\mathbf{y} = \begin{bmatrix} \mathbf{h}_1^{LLR'} \\ \vdots \\ \mathbf{h}_n^{LLR'} \end{bmatrix}.$$

Asymptotically, both $\mathbf{H}^{RLLR} \to \mathbf{H}^{LLR}$ and $\mathbf{H}^{LOESS} \to \mathbf{H}^{LLR}$. That is,

$$\mathbf{H}_{Asym}^{RLLR} = \mathbf{H}_{Asym}^{LOESS} = \mathbf{H}^{LLR}$$

(see Appendix B.2). Then the asymptotic bias and variance of the vector of predicted values that result from the *robust* procedures RLLR and Loess (see Appendix B.2) are given by

$$\text{Bias}(\hat{\mathbf{y}}^{NP}) = -(\mathbf{I} - \mathbf{H}^{LLR})\mathbf{m} \qquad (7.B.6)$$

$$\text{Var}(\hat{\mathbf{y}}^{NP}) = V^2\mathbf{H}^{LLR}\mathbf{H}^{LLR'}. \qquad (7.B.7)$$

Note that these formulas are identical to those for the non-robust local linear regression, except for the V^2 value (just as in the parametric case). This is due to the asymptotics of the results and the reliance of theory similar to that of M-Regression.

These formulas demonstrate the well-known fact about nonparametric regression, that a trade-off exists between bias and variance, which is controlled by the size of the bandwidth b_n. As the bandwidth $b_n \to 0$, note that $\mathbf{H}^{LLR} \to \mathbf{I}$, and the bias of the nonparametric estimator tends to zero. However, the variances of the individual estimates tend toward their maximum as the covariance matrix tends to the identity matrix. Conversely, as $b_n \to \infty$, the diagonal elements of the matrix $\mathbf{H}^{LLR}\mathbf{H}^{LLR'}$ tend toward a minimum, while the difference between $\mathbf{I}$ and $\mathbf{H}^{LLR}$ is maximized, resulting in maximum bias.

The bias and variance of the nonparametric estimators at a given location x_0 are given by

$$\text{Bias}\left\{\hat{y}^{NP}(x_0)\right\} = \mathbf{h}_0^{LLR'}\mathbf{f} - f(x_0) \qquad (7.B.8)$$

$$\text{Var}\left\{\hat{y}^{NP}(x_0)\right\} = V^2\mathbf{h}_0^{LLR'}\mathbf{h}_0^{LLR} \qquad (7.B.9)$$

where $\mathbf{h}_0^{LLR'} = \mathbf{x}_0^{NP'}(\mathbf{X}^{NP'}\mathbf{W}_0^{LLR}\mathbf{X}^{NP})^{-1}\mathbf{X}^{NP'}\mathbf{W}_0^{LLR}$ and $\mathbf{x}_0^{NP'} = (1 \;\; x_0)$ is a point in the LLR model space.

ORMRR

Recall that the ORMRR vector of estimates is given by

$$\begin{aligned}\hat{\mathbf{y}}^{ORMRR} &= \mathbf{H}^{ORMRR}\mathbf{y} \\ &= \mathbf{H}^{M}\mathbf{y} + \lambda\mathbf{H}^{RLLR}\mathbf{r} \\ &= [\mathbf{H}^{M} + \lambda\mathbf{H}^{RLLR}(\mathbf{I} - \mathbf{H}^{M})]\mathbf{y}.\end{aligned}$$

Note that as $n \to \infty$, $\mathbf{H}^{ORMRR} \to [\mathbf{H}^{OLS} + \lambda\mathbf{H}^{LLR}(\mathbf{I} - \mathbf{H}^{OLS})] = \mathbf{H}_{Asym}^{ORMRR}$ (see Appendix B.3 for details). Based on this result, the asymptotic bias and variance of $\hat{\mathbf{y}}^{ORMRR}$ are given by

$$\text{Bias}(\hat{\mathbf{y}}^{ORMRR}) = -(\mathbf{I} - \mathbf{H}_{Asym}^{ORMRR})\mathbf{f} \qquad (7.B.10)$$

$$\text{Var}(\hat{\mathbf{y}}^{ORMRR}) = V^2\mathbf{H}_{Asym}^{ORMRR}\mathbf{H}_{Asym}^{ORMRR'}. \qquad (7.B.11)$$

Note that for the mixing parameter $\lambda = 0$, $\mathbf{H}_{Asym}^{ORMRR} = \mathbf{H}^{OLS}$, and the asymptotic bias and variance of $\hat{\mathbf{y}}^{ORMRR}$ are identical to that of M-Regression.

As $b_n \to 0$ (for a nonzero value of λ), the estimator behaves like the nonparametric methods in that the bias tends to zero and the variance tends toward the maximum (both due to overfitting/undersmoothing the data). On the other hand, as $b_n \to \infty$, the opposite is true, resulting in an increase in bias and a decrease in variance (due to underfitting/oversmoothing).

The above expressions are the bias and variance formulas for a fixed value of the bandwidth, the "optimal" value, denoted b_{opt}. This optimal bandwidth value is chosen to be the value that minimizes the mse of the nonparametric fit to the residuals. The bias and variance that underlie this *conditional* mse are in terms of the estimated residuals $\hat{\mathbf{r}} = \mathbf{H}^{RLLR} \cdot \mathbf{r}$. Also from Appendix B.3, these conditional bias and variance formulas are given by

$$\text{Bias}(\hat{\mathbf{r}}) = -(\mathbf{I} - \mathbf{H}^{LLR})(\mathbf{I} - \mathbf{H}^{OLS})\mathbf{f} \qquad (7.B.12)$$

$$\text{Var}(\hat{\mathbf{r}}) = V^2 \mathbf{H}^{LLR}(\mathbf{I} - \mathbf{H}^{OLS})\mathbf{H}^{LLR'} \qquad (7.B.13)$$

Once the optimal bandwidth is chosen to minimize the mse that results from (7.B.12) and (7.B.13), the optimal mixing parameter is chosen to minimize the mse that results from (7.B.10) and (7.B.11). The details of the calculation of these optimal parameters are given in Section 7.F.

Below are the bias and variance at any location x_0, using $\mathbf{x}_0^{P'} = (1\ x_0\ x_0^2\ \ldots)$, where the order is determined by the user's specified parametric model, and $\mathbf{x}_0^{NP'} = (1\ x_0)$, which reflects the use of RLLR in the nonparametric portion of ORMRR (as noted previously, $\mathbf{x}_0^{P'}$ and $\mathbf{x}_0^{NP'}$ are the vectors denoting x_0 in the parametric and nonparametric model spaces, respectively). The bias and variance formulas are given by

$$\text{Bias}\left\{\hat{y}^{ORMRR}(x_0)\right\} = \mathbf{h}_0^{Asym'}\mathbf{f} - f(x_0) \qquad (7.B.14)$$

$$\text{Var}\left\{\hat{y}^{ORMRR}(x_0)\right\} = V^2 \cdot \mathbf{h}_0^{Asym'}\mathbf{h}_0^{Asym} \qquad (7.B.15)$$

where $\mathbf{h}_0^{Asym'} = \mathbf{x}_0^{P'}(\mathbf{X}^{P'}\mathbf{X}^{P})^{-1}\mathbf{X}^{P'} + \lambda\,\mathbf{h}_0^{LLR'}(\mathbf{I} - \mathbf{H}^{OLS})$.

7.C Applications of the MSE Formulas

The previous section developed the bias vector and covariance matrix for the vector of fitted values, along with the bias and variance formulas at isolated points, associated with the fitting techniques previously discussed. These derivations will be used to summarize the quality of the fits of these techniques, and this section discusses a selection of mse quantities that result from the formulas in Section 7.B. It will be these *mse* quantities that will be utilized for comparison purposes in subsequent sections.

The *average mean squared error* (AMSE) will be employed to determine the optimal parameters, which were alluded to in the previous section. If **b** is the vector of biases that result from equation (7.B.2), and **v** is a vector that contains the diagonal elements (the variances) from the covariance matrix that results from equation (7.B.3), for example, then using the (nx1) vector of ones denoted by **1**, AMSE is defined as

$$\text{AMSE} = \frac{\mathbf{b}'\mathbf{b} + \mathbf{v}'\mathbf{1}}{n}. \qquad (7.C.1)$$

$$= \frac{\sum_{i=1}^{n} \left\{ \text{Bias}^2(\hat{y}_i) + \text{Var}(\hat{y}_i) \right\}}{n} \qquad (7.C.2)$$

$$= \frac{\sum_{i=1}^{n} \text{MSE}(\hat{y}_i)}{n} \qquad (7.C.3)$$

The optimal bandwidth and mixing parameter, where appropriate, will be chosen to minimize this quantity for each procedure. Note that the calculation of AMSE does not depend on the raw data, but is an asymptotic quantity that depends only on the values of the regressor variable, the parameters of the fitting procedure, and the error distribution.

Finding the optimal parameters serves two critical purposes. The first is that, in so doing, the minimum values of the AMSE are calculated, and the procedures can be compared based upon the best they can be expected to perform (in terms of AMSE). In other words, the theoretical minimum AMSE of ORMRR is a constant for a given situation that can be calculated

and compared with the minimum AMSE of the other methods, in order to determine whether or not it is even necessary to proceed and study the performance of ORMRR when applied to data.

Another benefit of calculating the optimal parameters is its usefulness in the search for a criterion that can be used for data-driven selection of the fitting parameters (such as cross-validation or plug-in methods). The parameters chosen by a criterion such as PRESS can be compared to the optimal values calculated using the theoretical AMSE in order to quantify how well the data-driven technique is performing.

Another form of the mse is integrated mean squared error (IMSE), which is defined as

$$\text{IMSE} = \int_{\mathbf{x}} \left[m(x) - \hat{m}(x) \right]^2 dx . \qquad (7.C.4)$$

This mse formula evaluates the fit across the entire range of the data. The complexity of this integral leads to an approximation of IMSE (termed AIMSE, *approximate integrated mean squared error*). AIMSE is identical to AMSE in the respect that it is the average of mse's at multiple locations; however, the mse's are calculated at 1000 equally-spaced locations from 0 to 1, the range of the transformed regressor variable, and not at just the n data points. The AIMSE is calculated for M-Regression, for example, using equations (7.B.4) and (7.B.5) for 1000 x_0 values.

In subsequent sections, the optimal parameters for each method will be calculated using AMSE (where appropriate), and then the competing methods will be compared using AIMSE. An example will be presented to clarify the use of the different mse values, and to provide further demonstrations of the fits of some of the competing methods.

7.D Example

Consider the underlying model

$$E(y|x) = 12\left(\frac{x^2}{x^2 + 10.322} + 1 \right), \quad -3 \le x \le 3. \qquad (7.D.1)$$

from which a set of data will be generated. A set of 19 observations (equally spaced in the regressor space) will be generated from this model, based on the Contaminated Normal error

distribution CN(.1, $\sigma_1 = 0.75, \sigma_2 = 3.0$). Based upon these assumptions, the values of AIMSE presented in Table 7.1 can be calculated for the competing methods.

Table 7.1 Optimal AIMSE values for procedures for model in (7.D.1) and error distribution CN(.1, $\sigma_1 = 0.75, \sigma_2 = 3.0$).

Procedure	**Loess**	**ORMRR**	**RLLR**	**MREG**	**OLS**
AIMSE	0.26184	0.26359	0.28275	0.38001	0.42651

Note that these are computed using the optimal parameters (where appropriate), the selection of which will be discussed in Section 7.F. In calculating the AIMSE for ORMRR, M-Regression, and OLS, the quadratic model in one regressor was the assumed parametric model.

Consider Figure 7.1, which illustrates the true mean function for this example, and the corresponding optimal quadratic fit. Note that the points along the true mean function represent the (nx1) vector of mean values **m**. Notice also that the quadratic function somewhat resembles the true mean function across the range of the data, but misses the dip in the middle of the data, and the flat portions of the curve to either side of the center. In addition, it does a relatively poor job at the endpoints, which are frequently of interest to the researcher.

The structure in the true mean function that is not captured by the quadratic fit is illustrated in Figure 7.2. In addition, the points along this structure represent the vector **f** in the model representation of (7.B.1). This structure is what we hope to capture with the nonparametric portion of the ORMRR fit to the residuals.

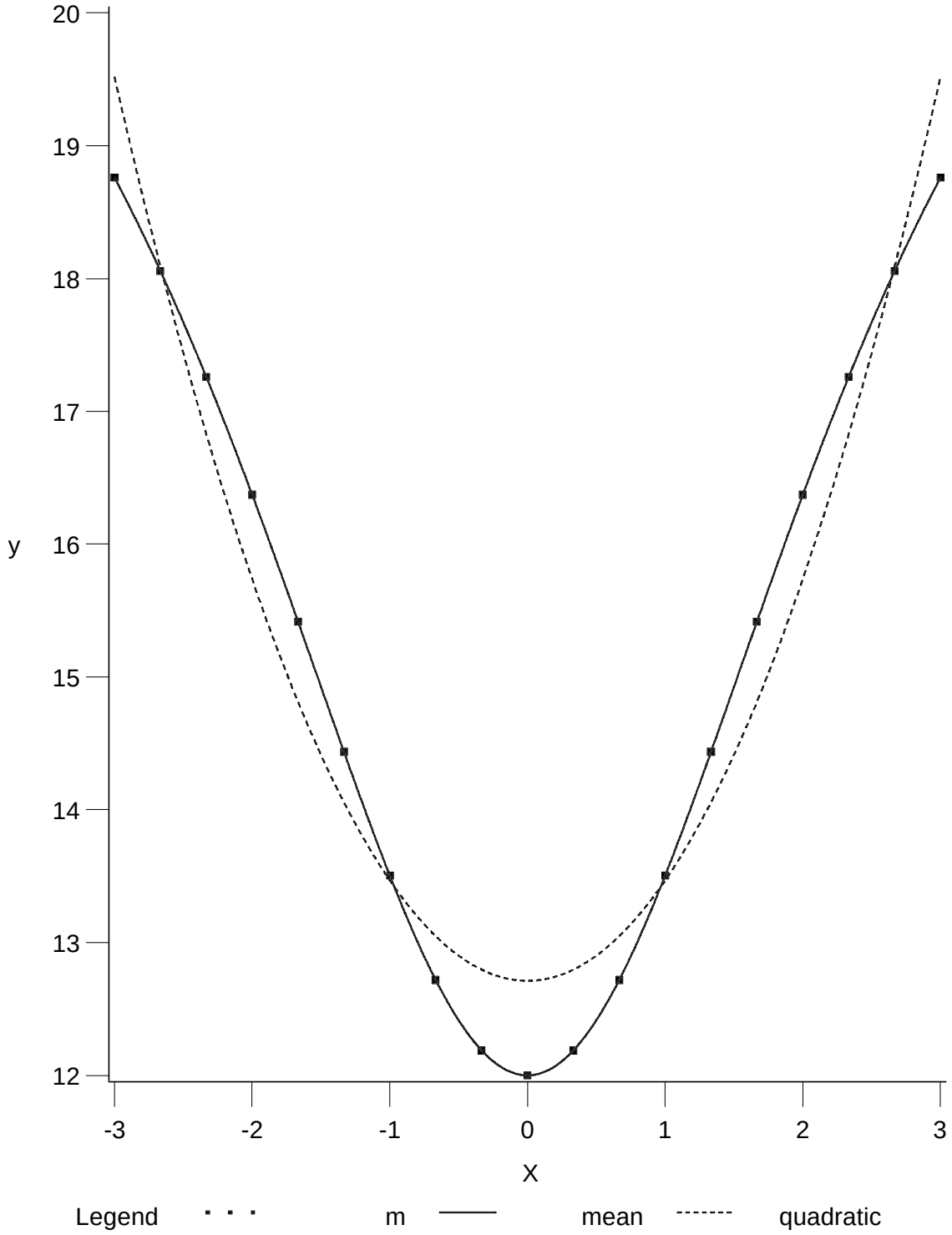

Figure 7.1 Plot of true mean function for model in (7.D.1), along with portion explained by a quadratic model

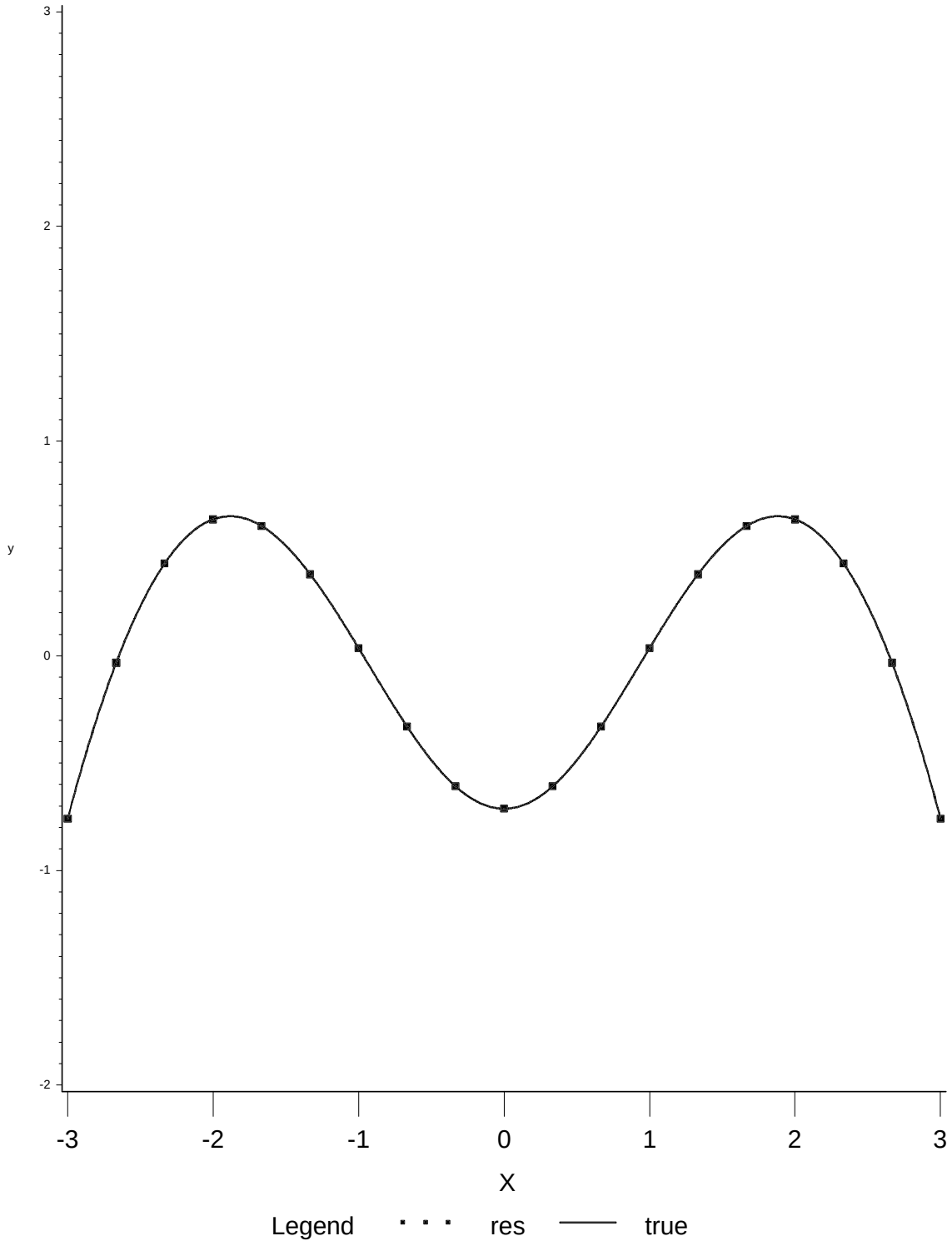

Figure 7.2 Plot of nonlinear portion of model in (7.D.1)

Figure 7.3 is a plot of the raw data generated from the model in (7.D.1) using the CN(.1, $\sigma_1 = 0.75, \sigma_2 = 3.0$), and the fit that results using OLS. This example serves only to, once again, highlight the shortcomings of techniques that ignore any possible outliers.

Figure 7.4 is a plot of the fit generated by using M-Regression. This fit resembles the quadratic model overlaid in Figure 7.1. M-Regression does an excellent job in identifying outliers, but it misses the dip in the center of the data and the flat parts of the curve to either side of the data. Since the M-Regression fit is the fit obtained using a quadratic model, the highest expectation for the resulting curve is for it to resemble the quadratic function in Figure 7.1 (which it achieves very nicely).

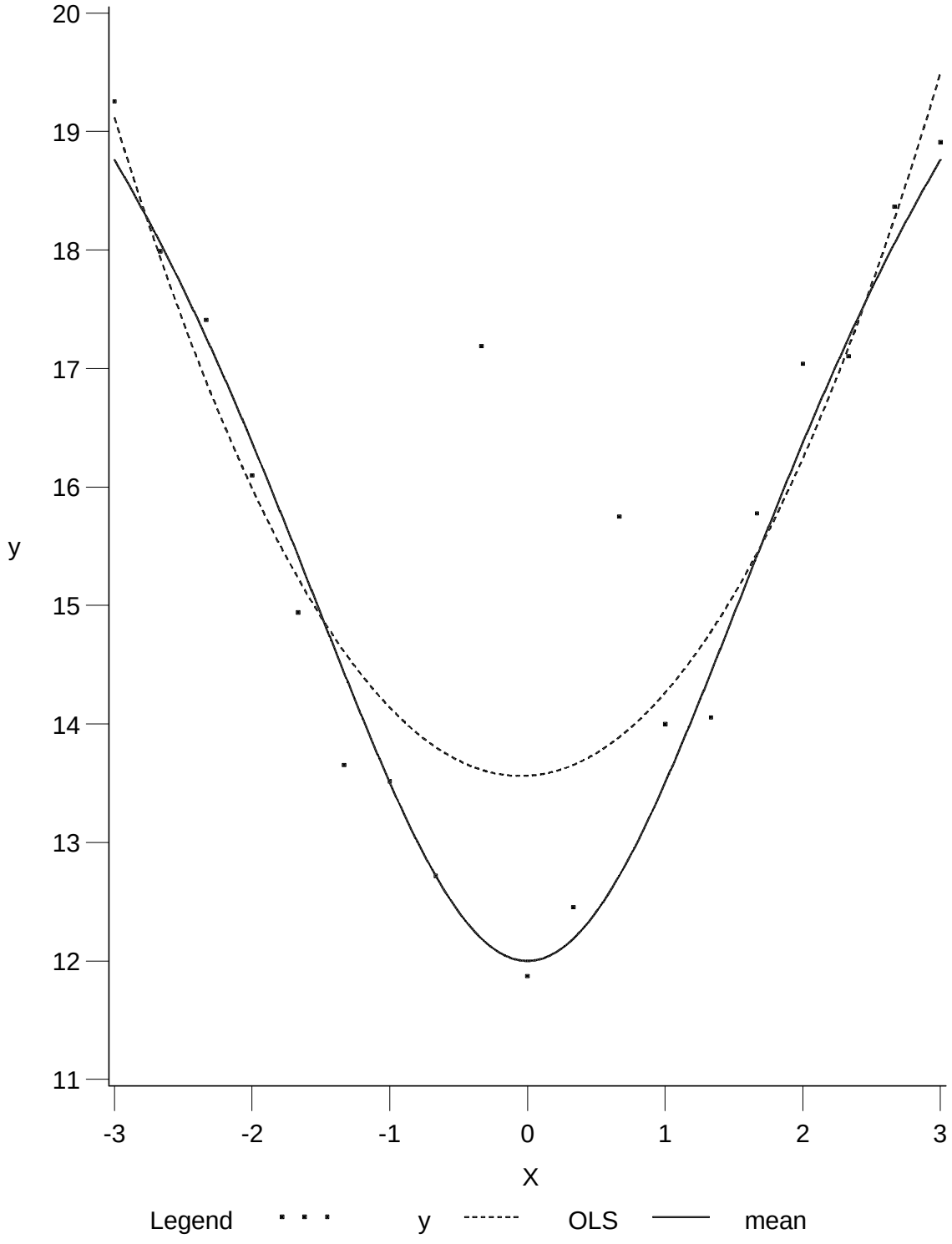

Figure 7.3 Plot of OLS fit (and mean function) for model in (7.D.1)

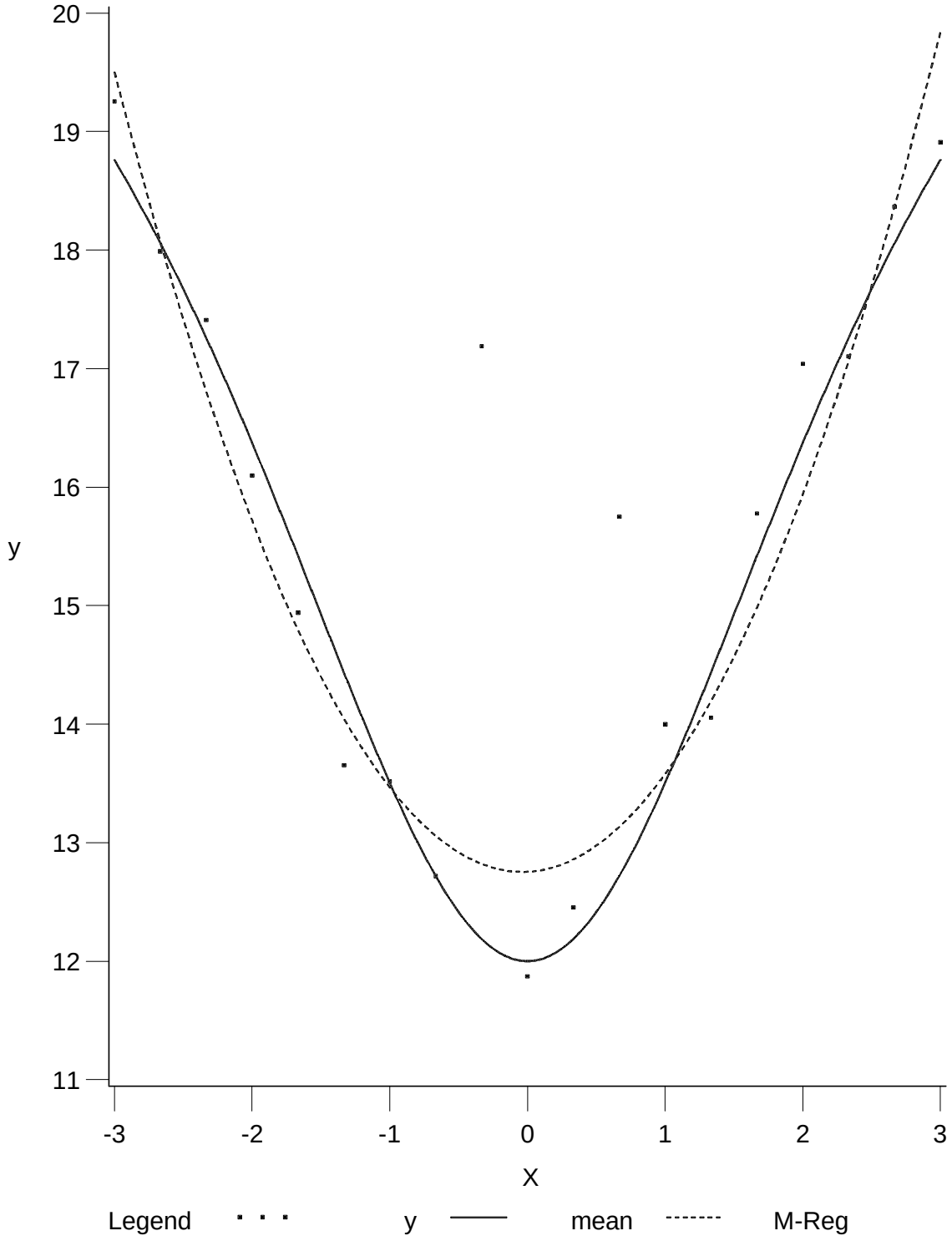

Figure 7.4 Plot of M-Regression fit (and mean function) for model in (7.D.1)

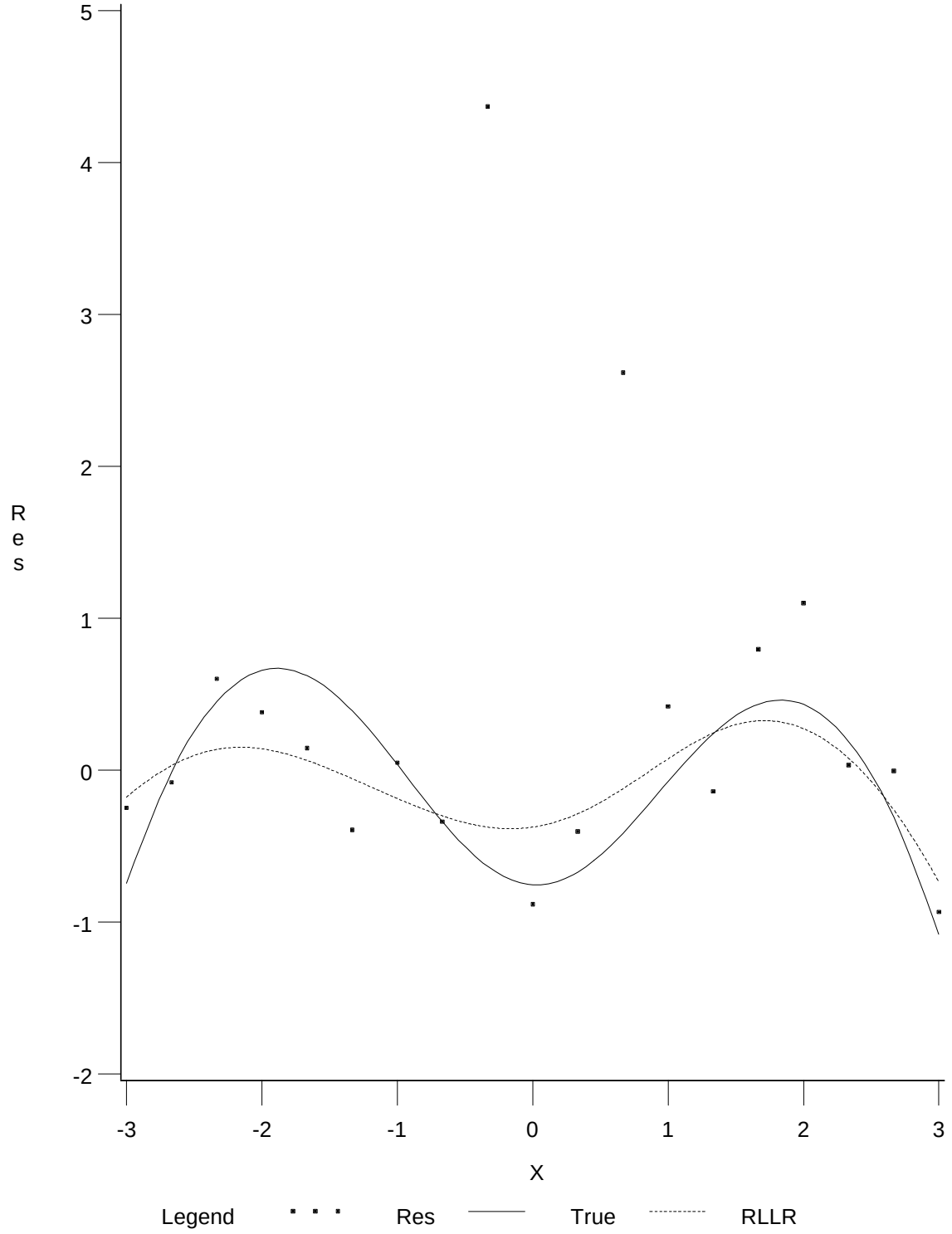

Figure 7.5 Plot of RLLR fit to residuals from M-Regression fit to original data, as compared to true amount of remaining structure left in the residuals.

Figure 7.5 displays the resulting residuals from the M-Regression fit, and the corresponding smoothing of these residuals using RLLR. The expectation for the fit to the residuals is that it resemble the structure of the nonlinear portion of the underlying model that was presented in Figure 7.2. Though not as pronounced, the trend is similar to the true structure and also does a good job of obtaining a fit that is robust to the outliers.

The resulting ORMRR estimate is displayed in Figure 7.6 (the optimal parameters b_{opt} = 0.16 and λ_{opt} = 0.93 are used). Notice that the predicted values in the center of the data are not drastically improved through the residual fit (the smoothed residuals are almost zero at that location), however the fitted values demonstrate a significant improvement over the M-Regression fit in the flat area of the curve on the right, and especially toward the endpoints.

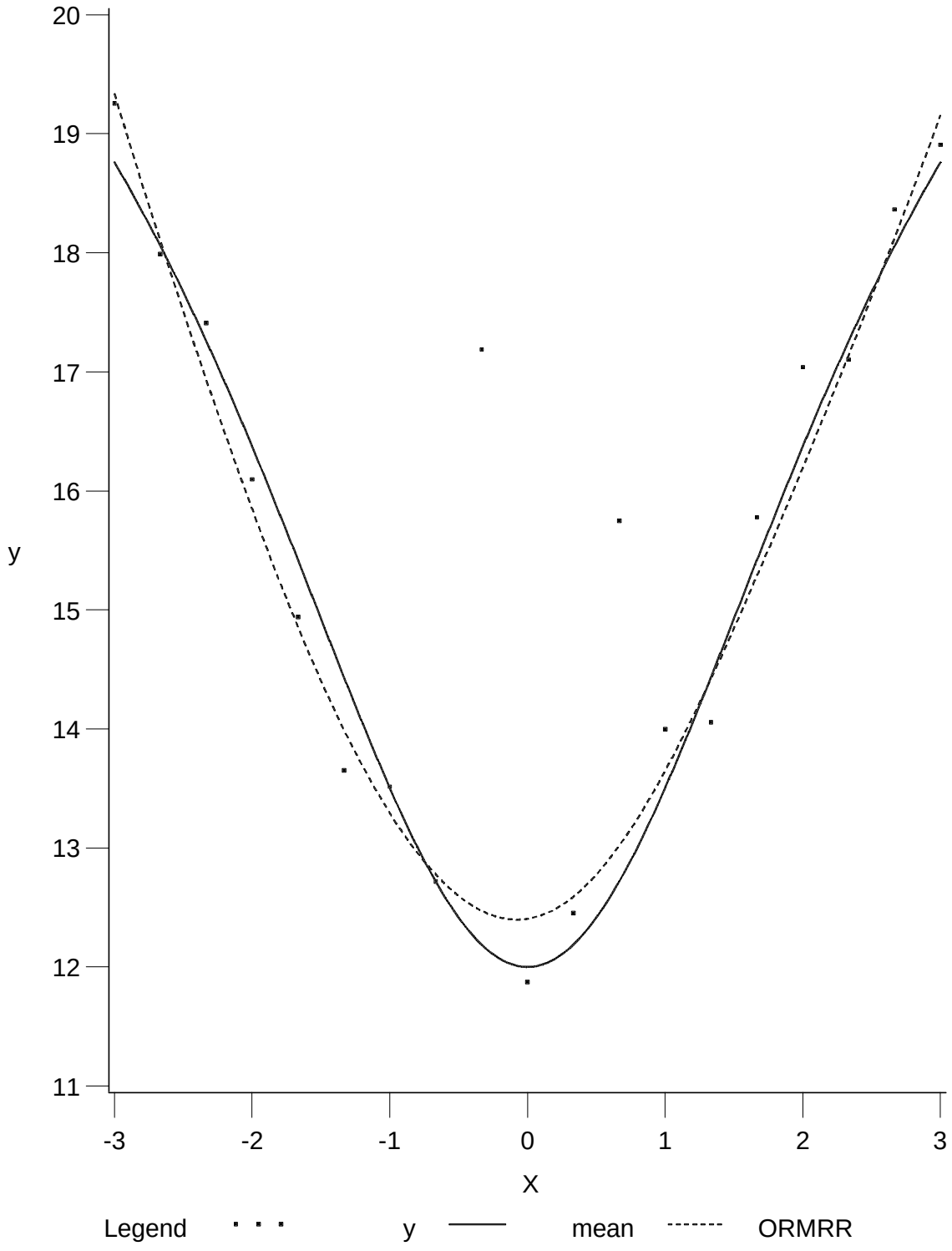

Figure 7.6 Plot of ORMRR fit (and mean function) for model in (7.D.1)

The last two plots (Figures 7.7 and 7.8) for this example illustrate the nonparametric fits obtained from Loess and RLLR. Both are good illustrations of the overall advantages and disadvantages of nonparametric techniques in a regresssion setting.

In general (as with all nonparametric fitting procedures), the bias of the fit is relatively small (note that the optimal window percentage for Loess ($f_{opt} = 0.45$) and the optimal bandwidth ($b_{opt} = 0.1298$) for RLLR were used to obtain the fits). Loess does an adequate job in all areas except the center of the data. However, the lack of a global underlying fitted model manifests itself in the overall variability of the fit. Without the benefit of some underlying structure, Loess has a difficult time distinguishing between "good" data and "bad" data in the center of the regressor space.

The fit provided by RLLR appears to have less of a variability problem (at least with this data set), however, like Loess, it seems to have a difficult time distinguishing between good and bad data. As mentioned before, the weighting scheme of RLLR results in a smoother set of predicted values.

These observations also demonstrate the motivation behind the proposed fitting method. What ORMRR gives up in terms of bias by not utilizing a purely local fitting procedure, it more than makes up for in terms of variability by relying on a global model that contributes to the overall fit.

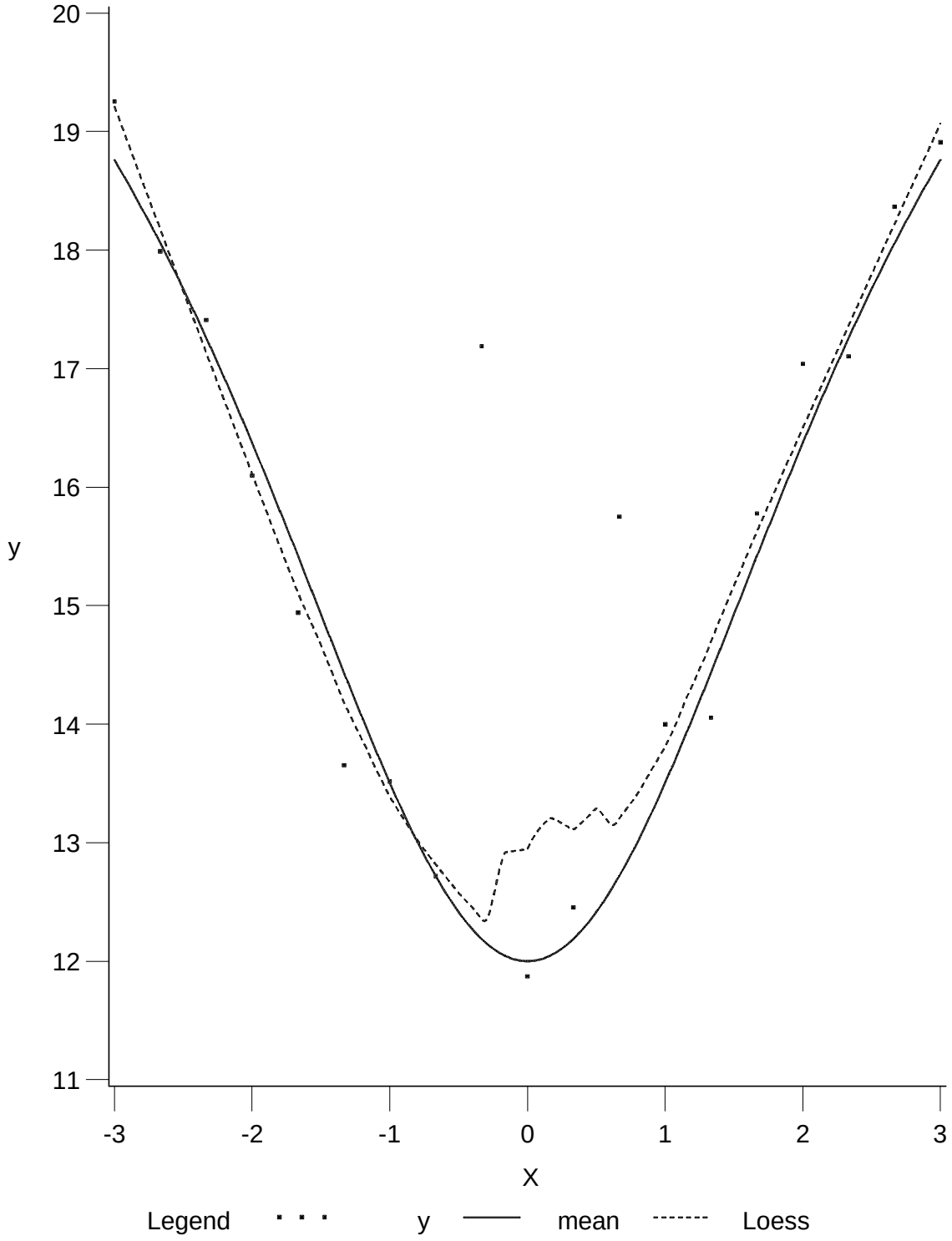

Figure 7.7 Plot of Loess fit (and mean function) for model in (7.D.1)

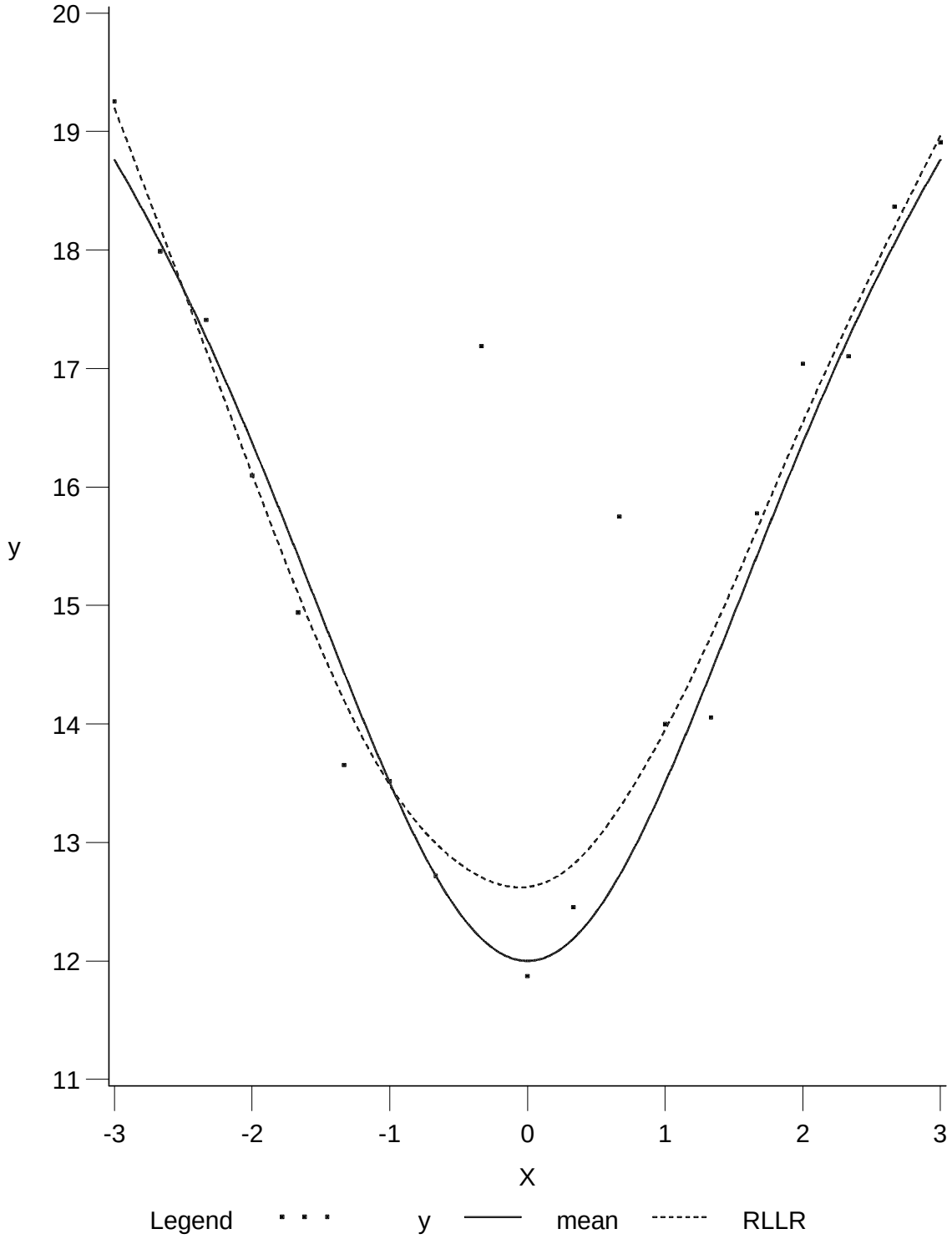

Figure 7.8 Plot of RLLR fit (and mean function) for model in (7.D.1)

Consider an empirical comparison of these methods using the above example. For each fitting procedure, calculate the predicted value (based on the nineteen data points in the example) at 1000 locations in the regressor space. For each location, calculate the squared distance between the predicted value and the conditional mean of y give the particular value of x at that location. Then average these across the 1000 locations to get an empirical mean squared error value. The observed values of this empirical mse (emse) for the estimators considered above are displayed in Table 7.2.

Table 7.2 Empirical mse values for model in (7.D.1) with error distribution CN(.1, $\sigma_1 = 0.75$, $\sigma_2 = 3.0$).

Procedure	**Loess**	**ORMRR**	**RLLR**	**MREG**	**OLS**
emse	0.29278	0.10619	0.18431	0.22987	0.32127

Practical applications of this mse value will be discussed in Section 7.F. The optimal bandwidth of 0.1298 was used for RLLR.

7.E Description of Model

As demonstrated in the above example, and upon careful inspection of the bias and variance formulas for each method, it is evident that the evaluation of the theoretical mse values requires the functions $\mathbf{X\beta}$ and **f**, as well as the sample size and the values of the regressors. Also needed is the distribution of the error terms. In order to compare the competing methods, we must specify these "design" parameters. Once these values are specified, the AIMSE values can be calculated for each of the competing methods and compared.

The requirement for having $\mathbf{X\beta}$ and **f** (and thus **m**) indicates that a true underlying model must be stipulated in order to complete the AIMSE calculations. Consider the model

$$E(y|x) = 12\left(\frac{x^2}{x^2 - 25\cdot\sqrt[4]{\gamma} + 28} + 1\right), \quad -3 \le x \le 3, \quad \gamma \in [0,1] \qquad (7.E.1)$$

as the function **m** (in 7.B.1). This model is one that, based on the parameter γ, ranges from a form that is almost identically quadratic in the regressor ($E(y|x) = \mathbf{X\beta} = \beta_0 + \beta_1 x + \beta_2 x^2$) to

forms with much more curvature than and very different from a quadratic model. The parameter γ is referred to as a misspecification parameter, as it controls the degree to which the true underlying model deviates from an approximately quadratic model (note that the example in the previous section used model (7.E.1) with $\gamma = 0.5$). As before, the error distribution is assumed to be the contaminated normal CN(π, σ_1, σ_2). Figure 7.9 is a plot of the model for the various values of γ. Note that, as with the graph of the example in Figure 7.1, these true mean curves deviate in varying degrees, as a function of γ, from the quadratic function, both in the center of the data (at the dip), and in the flat parts to either side of the center.

The separation between this model and a quadratic is due to the fact that the quadratic function has no points of inflection, while the model used for the example has two. As the value of γ increases, these points of inflection occur at locations closer to the value $x = 0$ (the center of the data), causing increasingly larger differences between the example models and the quadratic.

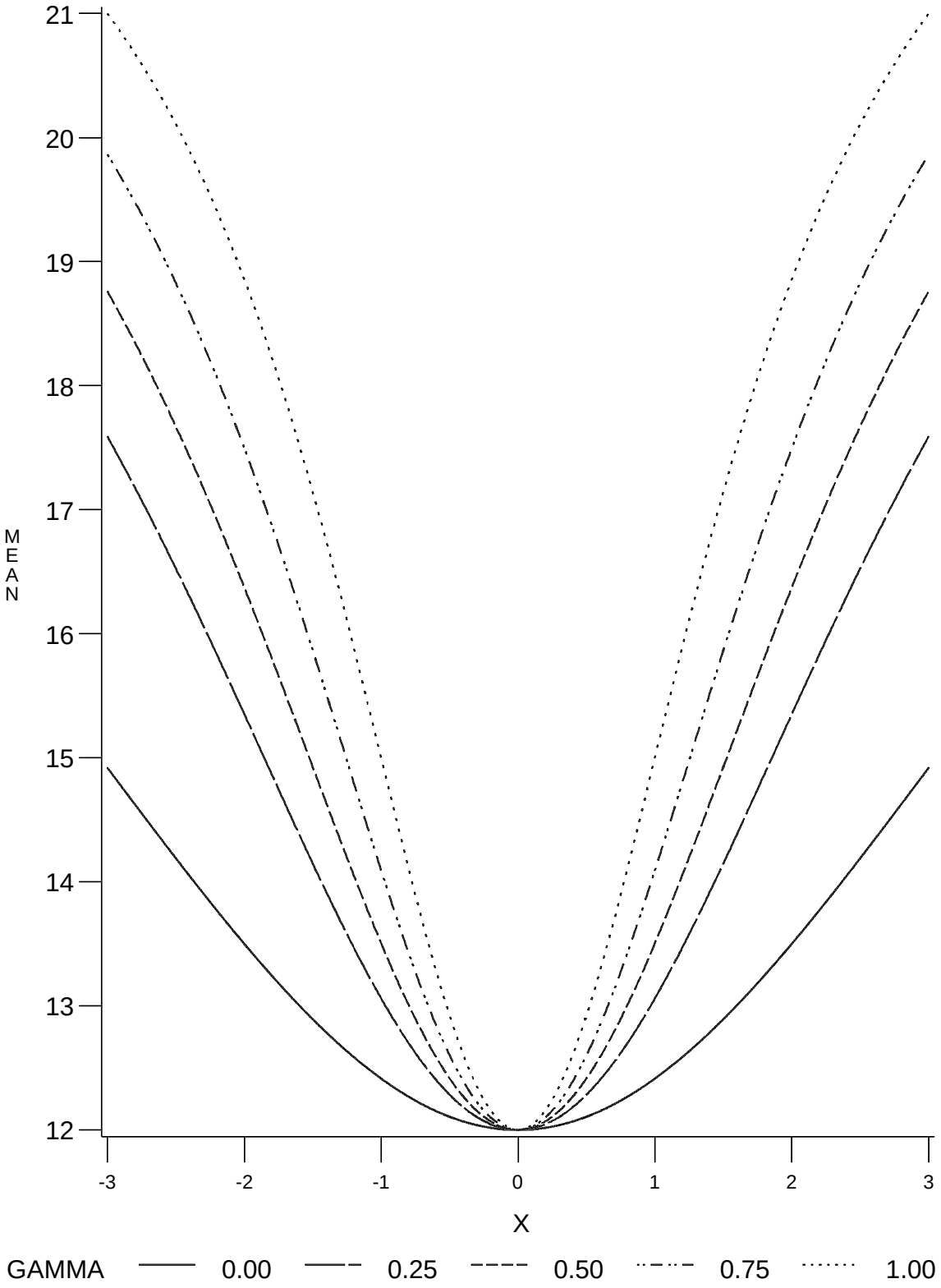

Figure 7.9 Plot of base model to be used in simulations for varying degrees of model misspecification γ

7.F Verification of Bias and Variance Calculations

The formulas presented in Section 7.B are asymptotic derivations that require empirical verification. Consider the empirical form of the mse described in Section 7.D. The verification will involve this mse value (calculated at 250 locations rather than 1000 locations for computing considerations), calculated for multiple data sets.

This mse value will actually be calculated for 500 data sets (for a given situation such as n = 19, $\gamma = 0.25$, $\pi = 0.1$, $\sigma_1 = 0.75$, $\sigma_2 = 3.0$); each of the data sets is generated using Monte Carlo simulation techniques. Once these 500 mse values are calculated, they are then averaged together. This value will be termed INTMSE and will be computed using the optimal parameters (bandwidth and mixing parameter, where appropriate) for each of the data sets. This value is then compared to the unique value of AIMSE that results in that scenario (also using the optimal parameters). We hope to observe that the INTMSE values are very close to the AIMSE values, which will validate the theoretical formulas presented in Section 7.B. We expect that the two will become more and more similar with an increase in sample size, as the theoretical formulas are asymptotic.

The last topic that needs to be covered before the discussion of the actual calculated values of both AIMSE and INTMSE is the calculation of the optimal parameters. These are chosen as the values of the parameters that minimize the AMSE of the method.

There are several reasons that AMSE, as opposed to AIMSE, is chosen as the minimization criterion. The first is that it *is* a mse type of statistic, hence it provides a quantification of the bias and the variance of the fitted values at the n data points (which is the type of quantification that we desire). In addition, it provides discrimination between the techniques that is equivalent to that of AIMSE. Finally, and more importantly, it provides us with optimal values that we can hope to achieve via data-driven techniques. Since the resources of the data-driven methods are restricted to the n data points only, it is reasonable to use a criterion to choose the optimal parameters that also depends only on the n data locations (see Mays and Birch (1996)).

The method of calculation of the optimal parameters begins with three starting values for the parameter of interest, which results in three distinct AMSE values. Based on these three values, a binary search algorithm is utilized to determine the parameter value that minimizes the AMSE value. The search is restricted by the range of possible and/or feasible values for the parameter of interest. For example, $0 \leq f \leq 1$ when f is the window width parameter for Loess.

The procedure for calculating the optimal parameters for ORMRR is unique in that two parameters are actually needed (all other methods considered have zero or one). The parametric portion of the fit is free of bandwidth and mixing parameter considerations. The resulting residuals from this parametric fit are smoothed using RLLR, which utilizes a bandwidth value. The optimal bandwidth for the residual fit is found, according to the above techniques, using the conditional (conditioned upon the parametric fit) AMSE derived from equations (7.B.12) and (7.B.13).

Next, the optimal mixing parameter is determined from the AMSE of the final ORMRR fit derived from equations (7.B.10) and (7.B.11). The bandwidth is fixed at b_{opt} for these calculations (thus the optimal bandwidth and mixing parameter are determined separately, not as a pair).

Table 7.3 presents the AIMSE and INTMSE values produced by the ORMRR method for four different sample sizes and across a range of γ values between 0 and 1 for each sample size. Each row of the table represents a specific group of the values that must be specified in order that these two mse quantities be computable (that is, sample size, values of the regressor, error distribution, value of γ).

As mentioned above, these results are offered as a means of quantifying how accurate the theoretical formulas are for the proposed methodology. The last column is the relative error between INTMSE and the theoretically optimal mse value AIMSE.

In this example, the parameters of the Contaminated Normal distribution that were selected yield the following error structure,

$$\varepsilon_i \sim CN(.1,\ \sigma_1 = 0.75, \sigma_2 = 3.0).$$

This results in the value for the parameter V^2, which is part of the variance calculations for all of the procedures outlined above,

$$V^2 = \sigma^2 \cdot \frac{E(\psi^2)}{[E(\psi')]^2} = (1.40625) * (.77018) = 1.08.$$

Note that $\frac{E(\psi^2)}{[E(\psi')]^2}$ was found by using the Integrate function in Mathematica, and that the ψ function used is Huber's ψ with $c_H = 1.345$. All of the robust procedures (M-Regression, RLLR, Loess, and ORMRR) were programmed to use this particular ψ function as the weighting mechanism for generating the residual based weights in order to make fair comparisons between the procedures.

Table 7.3 Theoretically optimal (AIMSE) and simulated (INTMSE) mse values using the optimal smoothing parameters for the ORMRR estimator for $\pi = 0.10$, $\sigma_1 = 0.75$, $\sigma_1 = 3.0$. Comparison is used to verify validity of theoretical mse formulas for ORMRR presented in Section 7.D.

n	**γ**	**AIMSE**	**INTMSE**	**Rel. Error**
	0	0.28226	0.27989	0.008397
	0.25	0.36272	0.34578	0.046703
10	*0.5*	0.43487	0.49056	0.128061
(inc=2/3)	*0.75*	0.50491	0.62460	0.237052
	1	0.57835	0.74392	0.286280
	0	0.16042	0.13705	0.14568
	0.25	0.2197	0.19487	0.113018
19	*0.5*	0.26359	0.28755	0.090899
(inc=1/3)	*0.75*	0.30451	0.31193	0.024367
	1	0.34959	0.38567	0.103207
	0	0.10626	0.08391	0.210333
	0.25	0.15458	0.14622	0.054082
30	*0.5*	0.18513	0.1744	0.057959
(inc=.207)	*0.75*	0.21362	0.20773	0.027572
	1	0.24512	0.24619	0.004365
	0	0.06683	0.04576	0.315278
	0.25	0.10407	0.09492	0.087922
50	*0.5*	0.12418	0.11246	0.094379
(inc=.122)	*0.75*	0.14316	0.13705	0.042680

	1	0.16422	0.15961	0.028072

Recall that we expected, as a result of the asymptotic nature of the results, to observe improved performance of the formulas as the sample size increased. As illustrated in Table 7.3, the relative error *does* generally decrease after a sample size of 10, and is somewhat stable thereafter. These relatively small relative errors provide evidence of accuracy of the formulas presented in Section 7.B for the proposed estimator, even in relatively small finite sample sizes.

Table 7.4 offers the same type of information as that given in Table 7.3 for all competing procedures. Not only can this be used to compare the simulated values with the optimal values for any given procedure, but also for comparison of the performance of the procedures when using the optimal parameters for calculation of INTMSE (this can be interpreted as the best the procedures can be expected to perform on average when applied to actual data). The calculations for OLS are offered merely for comparison purposes (and to verify the fact that the simulations were performed/programmed correctly), as the mse formulas for these procedures are well-defined in the literature.

Note that OLS, as expected in the presence of outliers, performs poorly and results in the worst AIMSE and INTMSE in every scenario. In addition, this procedure, which is heavily reliant on the specified model, suffers from dramatic increases of the INTMSE values as the degree of model misspecification increases.

Also as expected, M-Regression has the minimum INTMSE for $\gamma = 0$, with a significant improvement over the values for OLS (for similar comparisons between OLS and M-Regression using Monte Carlo techniques, see Agard and Birch (1993)). In addition, as anticipated when more information about the true curve is available, both mse values for M-Regression decrease as the number of observations in the data set increase. However, the procedure performs in an increasingly poor manner as the degree of model misspecification increases.

The nonparametric procedures, on the other hand, are inferior for little or no model misspecification, but perform extremely well as both the degree of misspecification and the sample size increase. They are particularly poor for small sample sizes, indicating their complete reliance on the data (which results from the fact that no particular model form is specified). Note the

similarities between RLLR and Loess; the two procedures perform in a very similar manner, demonstrating that, at this point, neither is clearly better than the other.

As with the parametric procedures, the *formulas* for the nonparametric procedures perform relatively poorly for the smaller sample sizes, but improve significantly as n increases. This is, again, due to the asymptotic nature of the derived bias and variance formulas.

ORMRR is very competitive with M-Regression with little or no model misspecification, and very competitive with the nonparametric techniques with moderate to large degrees of misspecification and larger sample sizes. These results indicate that, at the very least, the technique should theoretically be optimal or near optimal among these or similar procedures across a wide range of model misspecification and sample sizes for a relatively moderate amount of contamination ($\pi = .10$).

Table 7.4 Comparison of AIMSE (bold) and INTMSE (non-bold) for all procedures using the optimal smoothing parameters for $\pi = 0.10$, $\sigma_1 = 0.75$, $\sigma_1 = 3.0$. Table is for verification of mse formulas (Section 7.B) for all procedures and for comparison of performance of these procedures.

n	γ	Loess	ORMRR	RLLR	MREG	OLS
	0	0.34890	0.27989	0.43347	0.28002	0.36144
		0.32626	**0.28226**	**0.35072**	**0.28226**	**0.36483**
	0.25	0.57248	0.34578	0.45843	0.40197	0.44745
		0.43172	**0.36272**	**0.42734**	**0.37913**	**0.46169**
10	*0.5*	0.54661	0.49056	0.56018	0.56702	0.59863
		0.44024	**0.43487**	**0.46510**	**0.53478**	**0.61735**
	0.75	0.66725	0.62460	0.63719	0.94242	0.94112
		0.49155	**0.50491**	**0.50983**	**0.83952**	**0.92208**
	1.0	0.76083	0.74392	0.76521	1.46407	1.57970
		0.59735	**0.57835**	**0.57134**	**1.46456**	**1.54713**
	0	0.18906	0.13705	0.22929	0.12805	0.21246
		0.19495	**0.16042**	**0.21133**	**0.16042**	**0.20692**
	0.25	0.23431	0.19487	0.26573	0.22416	0.28299
		0.24276	**0.21970**	**0.25954**	**0.24323**	**0.28973**
19	*0.5*	0.29832	0.28755	0.31795	0.37316	0.42918
		0.26184	**0.26359**	**0.28275**	**0.38001**	**0.42651**
	0.75	0.30551	0.31193	0.31897	0.65808	0.69848
		0.30337	**0.30451**	**0.30995**	**0.65375**	**0.70025**
	1.0	0.38274	0.38567	0.37946	1.24451	1.26725
		0.32898	**0.34959**	**0.34684**	**1.22886**	**1.27536**
	0	0.12119	0.08391	0.14595	0.08391	0.13533
		0.13386	**0.10626**	**0.14641**	**0.10626**	**0.13668**
	0.25	0.15158	0.14622	0.18230	0.16162	0.21534
		0.16423	**0.15458**	**0.18095**	**0.18529**	**0.21571**
30	*0.5*	0.16882	0.17440	0.18704	0.30349	0.34295
		0.18154	**0.18513**	**0.19767**	**0.31695**	**0.34737**
	0.75	0.18428	0.20773	0.21192	0.57110	0.61845
		0.20190	**0.21362**	**0.21711**	**0.58217**	**0.61259**
	1.0	0.22727	0.24619	0.24429	1.14048	1.17857
		0.23533	**0.24512**	**0.24332**	**1.14338**	**1.17380**
	0	0.07298	0.04576	0.08063	0.04934	0.09220
		0.08702	**0.06683**	**0.09653**	**0.06683**	**0.08553**
	0.25	0.09844	0.09492	0.11120	0.12866	0.17044
		0.10885	**0.10407**	**0.12014**	**0.14402**	**0.16271**
50	*0.5*	0.10684	0.11246	0.11896	0.26060	0.30310
		0.12006	**0.12418**	**0.13166**	**0.27314**	**0.29184**
	0.75	0.12369	0.13705	0.13700	0.53177	0.56411
		0.13416	**0.14316**	**0.14503**	**0.53413**	**0.55283**
	1.0	0.15847	0.15961	0.15735	1.09357	1.12104
		0.15170	**0.16422**	**0.16292**	**1.08842**	**1.10711**

7.G Fit Diagnostics

Here we would like to present some formulas that might prove useful in the evaluation of the fit obtained using ORMRR. These formulas have not been studied via Monte Carlo simulations (as are those that are presented in Chapter 7), and thus are presented only as suggestions to be employed at the user's discretion.

The first is a confidence interval on the conditional mean of y given a particular value of x = x_0. Following the development of such a confidence interval for the linear model fitted using ordinary least squares, a proposed (1-α)x100% confidence interval on E(y|x_0) is

$$\hat{y}_i^{ORMRR} \pm t_{(n-tr(H^{ORMRR})),\alpha/2}\sqrt{\hat{V}^2 \mathbf{h}_0^{ORMRR}\mathbf{h}_0^{ORMRR\prime}}, \qquad (7.G.1)$$

where $\mathbf{h}_0^{ORMRR\prime} = \mathbf{x}_0^{P\prime}(\mathbf{X}^{P\prime}\mathbf{W}^{M}\mathbf{X}^{P})^{-1}\mathbf{X}^{P\prime} + \lambda\mathbf{x}_0^{NP\prime}(\mathbf{X}^{NP\prime}\mathbf{W}_0^{RLLR}\mathbf{X}^{NP})^{-1}\mathbf{X}^{NP\prime}(\mathbf{I}-\mathbf{H}^{M})$ and the quantity V^2 is estimated by

$$\hat{V}^2 = \frac{(mad)^2\left(\frac{n^2}{n-R}\right)\sum\psi(r_i^*)^2}{\left[\sum\psi'(r_i^*)\right]^2}. \qquad (7.G.2)$$

Another quantity that may prove useful is an outlier diagnostic based on the residual from the ORMRR fit. Since, theoretically, the ORMRR fit extracts as much information from the data about the model as possible, the information in the data that is not used (the residuals) should contain only information on the outlier nature of the observations. Most outlier diagnostics in the literature *are* based on residuals from the fitted model. If one is fitting a parametric model and it is misspecified, then obviously the residual will include not only a variance component, but also a lack of fit component due to the wrong model being used. An R-student type of statistic can be developed using the variance calculation (assuming that $\mathbf{H}^{ORMRR}$ is constant)

$$\begin{aligned} Var(\mathbf{y}-\hat{\mathbf{y}}^{ORMRR}) &= Var[(\mathbf{I}-\mathbf{H}^{ORMRR})\mathbf{y}] \\ &= (\mathbf{I}-\mathbf{H}^{ORMRR})\cdot V^2\mathbf{I}\cdot(\mathbf{I}-\mathbf{H}^{ORMRR})' \\ &= V^2(\mathbf{I}-\mathbf{H}^{ORMRR})\cdot(\mathbf{I}-\mathbf{H}^{ORMRR})' \qquad (7.G.3) \end{aligned}$$

Thus, (7.G.3) is the estimate of the variance covariance matrix of the residuals from the ORMRR fit. Consider now the standardized residual from the ORMRR fit

$$a_i = \frac{y_i - \hat{y}_i^{ORMRR}}{\hat{V}\sqrt{cv_{ii}}}, \qquad (7.G.4)$$

where cv_{ii} is the i^{th} diagonal element of the variance-covariance matrix $(\mathbf{I} - \mathbf{H}^{ORMRR}) \cdot (\mathbf{I} - \mathbf{H}^{ORMRR})'$ developed in (7.G.3), and $\hat{V}^2$ is defined in (7.G.2).

This statistic is a straightforward calculation from the ORMRR model using similar motivations as those upon which the R-student statistic was based. We hope to eventually study these diagnostics in depth, not only in the model robust setting, but also in the context of purely nonparametric fitting.