Chapter 8

A Simulation Study

8.A Introduction

The previous chapter developed theoretical mean squared error formulas for the competing procedures considered in the current research. In addition, simulations (for 10% contamination in the error distribution) were utilized to verify that the theoretical formulas were accurate. This chapter considers two additional proportions of contamination, and compares the procedures (in terms of INTMSE) for these situations. Each INTMSE value is based upon 500 simulated data sets, each of which is fit using the optimal parameter(s), where appropriate.

The simulations thus far have been done using the optimal parameters. Since these depend on the true underlying model, obviously they will be unknown quantities for anything but contrived data. Also in this chapter, a criterion is introduced that can be used to determine what values should be used for the parameters when fitting a given data set. This criterion is a cross-validation quantity (similar to PRESS) that is formulated in such a way as to provide protection against outliers. The performance of this criterion will be investigated by comparing the parameter values it selects with the theoretically optimal parameters. In addition, a comparison will be offered between the INTMSE observed when using parameters chosen by the criterion and INTMSE observed when the theoretically optimal parameters are used.

8.B Simulation Comparison of Procedures Using Optimal Parameters

Consider now a comparison of the competing procedures for the model described in (7.E.1),

$$E(y|x) = 12\left(\frac{x^2}{x^2 - 25\cdot\sqrt[4]{\gamma} + 28} + 1\right), \quad -3 \le x \le 3, \quad \gamma \in [0,1], \qquad (8.B.1)$$

across a range of sample sizes, degrees of model misspecification, *and* degrees of contamination. The procedures that are compared in this chapter, as in Chapter 7, are ORMRR, Loess, M-Regression, OLS and Robust Local Linear Regression (RLLR). Note that in this section the simulations were performed using the *optimal* parameter values that were chosen according to the procedure described in Chapter 7.

The motivation for this study is to determine if our procedure is competitive when using the optimal parameters. If ORMRR performs well under these circumstances, then we have reason to believe that, using a data-driven parameter selection criterion, our procedure can provide a superior choice among applied fitting methods. The optimal parameters used by the competing procedures are as follows: for Loess, the parameter is a *proportion* of observations that receive positive weight when predicting at a given location; for RLLR, it is the *bandwidth* for the nonparametric fit to the raw data; and for ORMRR, the two parameters are the bandwidth for the nonparametric (RLLR) fit to the residuals, and the mixing parameter that determines what portion of the fitted residuals is added back to the initial robust parametric fit.

Tables 8.1, 8.2, and 8.3 offer comparisons, in terms of INTMSE (the simulated value of AIMSE), of these five estimators for $\pi = 0.05$, 0.10, and 0.25 (5%, 10%, and 25% contamination in the error distribution), respectively. The results are similar to those discussed for $\pi = 0.10$ in Chapter 7 (Table 8.2 is identical to Table 7.4, except for the absence of the AIMSE values). The INTMSE is highlighted for those procedures that have the minimum or close to the minimum INTMSE among all the procedures considered here.

As observed for $\pi = 0.10$ in Chapter 7, M-Regression performs the best of all procedures when there is no model misspecification. However, it does a very poor job even with moderate amounts of model misspecification. In addition, INTMSE for M-Regression (and for all procedures) is directly related to both the proportion of contamination and the degree of model misspecification, and is inversely related to sample size. These are intuitive properties of the INTMSE values and provide evidence that the simulations were performed correctly.

As expected, ORMRR outperforms the nonparametric procedures in cases of little or no model misspecification, while conceding little in terms of MSE to M-Regression. This is

especially true when considering the smaller sample sizes, as this is apparently when the parametric portion of ORMRR contributes the most in terms of stability of the fitted values.

In cases of moderate to severe model misspecification, ORMRR is consistently better than either Loess or RLLR in smaller sample sizes, and still very competitive in the larger sample sizes. We expect the nonparametric procedures to perform well in the case where there is moderate to severe model misspecification, since the specified quadratic model is of increasingly less benefit to the ORMRR fitting procedure the more it deviates from the true model. Note that this trend is also evident in the simulations of Mays and Birch (1996) in comparisons of Model Robust Regression 2 and local linear regression.

Table 8.1 Simulated mse values for Loess, ORMRR, M-Regression, and Robust Local Linear Regression estimators with π **= 0.05**. Fits are based on theoretically *optimal* parameter values.

n	g	***Loess***	***ORMRR***	***M-Reg***	***RLLR***	***OLS***
	0.00	0.29061	**0.21799**	**0.21715**	0.32607	0.26118
	0.25	0.38424	**0.27878**	0.29047	0.34011	0.32214
10	*0.50*	0.40345	**0.37874**	0.48391	0.41590	0.49695
	0.75	0.52727	**0.48634**	0.82327	**0.48247**	0.83945
	1.00	**0.59612**	**0.59521**	1.48867	**0.60301**	1.49228
	0.00	0.16224	**0.11647**	**0.11585**	0.18934	0.17181
	0.25	0.19819	**0.16953**	0.18976	0.21693	0.23039
19	*0.50*	0.22967	**0.21972**	0.33356	0.23632	0.36347
	0.75	**0.24424**	**0.25549**	0.61722	**0.25883**	0.64361
	1.00	**0.29765**	**0.30799**	1.20570	**0.30100**	1.21419
	0.00	0.10367	**0.06895**	**0.06866**	0.11214	0.09986
	0.25	**0.12901**	**0.12239**	0.15332	0.14524	0.18075
30	*0.50*	**0.14261**	**0.14354**	0.28558	0.15311	0.30616
	0.75	**0.16696**	**0.17326**	0.55825	**0.17348**	0.58478
	1.00	**0.18672**	0.20062	1.12348	**0.19700**	1.13324
	0.00	0.06734	**0.04095**	**0.04071**	0.06722	0.05795
	0.25	**0.08247**	**0.08115**	0.12217	0.09251	0.14127
50	*0.50*	**0.09885**	**0.09655**	0.25481	**0.09829**	0.26651
	0.75	**0.10168**	**0.11285**	0.52122	**0.11237**	0.53437
	1.00	**0.11632**	0.12982	1.07900	**0.12714**	1.09358

Table 8.2 Simulated mean squared error values for Loess, ORMRR, M-Regression, and Robust Local Linear Regression estimators with π = **0.10**.

n	g	*Loess*	*ORMRR*	*MREG*	*RLLR*	*OLS*
	0	0.34890	**0.27989**	**0.28002**	0.43347	0.36144
	0.25	0.57248	**0.34578**	0.40197	0.45843	0.44745
10	*0.5*	0.54661	**0.49056**	0.56702	0.56018	0.59863
	0.75	0.66725	**0.62460**	0.94242	**0.63719**	0.94112
	1.0	0.76083	**0.74392**	1.46407	0.76521	1.57970
	0	0.18906	**0.13705**	**0.12805**	0.22929	0.21246
	0.25	0.23431	**0.19487**	0.22416	0.26573	0.28299
19	*0.5*	**0.29832**	**0.28755**	0.37316	0.31795	0.42918
	0.75	**0.30551**	**0.31193**	0.65808	**0.31897**	0.69848
	1.0	**0.38274**	**0.38567**	1.24451	**0.37946**	1.26725
	0	0.12119	**0.08391**	**0.08391**	0.14595	0.13533
	0.25	**0.15158**	**0.14622**	0.16162	0.18230	0.21534
30	*0.5*	**0.16882**	**0.1744**	0.30349	0.18704	0.34295
	0.75	**0.18428**	**0.20773**	0.57110	0.21192	0.61845
	1.0	**0.22727**	0.24619	1.14048	0.24429	1.17857
	0	0.07298	**0.04576**	**0.04934**	0.08063	0.09220
	0.25	**0.09844**	**0.09492**	0.12866	0.11120	0.17044
50	*0.5*	**0.10684**	**0.11246**	0.26060	**0.11896**	0.30310
	0.75	**0.12369**	**0.13705**	0.53177	**0.13700**	0.56411
	1.0	**0.15847**	**0.15961**	1.09357	**0.15735**	1.12104

The results indicate that ORMRR is a reliable procedure that performs well across a range of sample sizes, degrees of model misspecification, and proportions of contamination. In particular, it is superior in most situations and is reliable in all situations. As mentioned above, the benefit of ORMRR over the nonparametric methods is most evident in smaller sample sizes and smaller amounts of model misspecification. It is also preferable to M-Regression for any amount of misspecification of the model specified by the user (for all samples sizes and proportions of contamination of the error distribution).

Table 8.3 Simulated mean squared error values for Loess, ORMRR, M-Regression, and Robust Local Linear Regression estimators with π = **0.25**.

n	g	*Loess*	*ORMRR*	*MREG*	*RLLR*	*OLS*
	0.00	**0.53490**	**0.55210**	**0.54991**	0.74184	0.64430
	0.25	0.79370	**0.69593**	**0.69394**	0.86490	0.78002
10	*0.50*	**0.71904**	0.79240	0.81827	1.03189	0.93205
	0.75	**1.03990**	1.10646	1.24709	1.13569	1.23883
	1.00	1.45335	**1.16409**	1.74638	1.38314	1.90334
	0.00	0.31454	**0.21870**	**0.21766**	0.36963	0.38824
	0.25	0.36050	**0.33491**	**0.33650**	0.44412	0.46138
19	*0.50*	0.52772	**0.45161**	0.48757	0.53503	0.59355
	0.75	0.72244	**0.57204**	0.78473	**0.58494**	0.86566
	1.00	0.78467	**0.64054**	1.36799	0.67294	1.47199
	0.00	0.19054	**0.13618**	**0.13559**	0.23653	0.24000
	0.25	0.29621	**0.21053**	**0.21728**	0.30703	0.34078
30	*0.50*	0.32714	**0.30297**	0.37659	**0.31366**	0.45974
	0.75	**0.36347**	**0.35917**	0.63922	**0.37839**	0.72406
	1.00	0.42639	**0.39521**	1.21678	0.42463	1.28606
	0.00	0.13077	**0.07788**	**0.07753**	0.14661	0.15822
	0.25	0.17633	**0.14782**	0.16181	0.19126	0.23825
50	*0.50*	0.21449	**0.19138**	0.29523	**0.20077**	0.37241
	0.75	**0.23110**	**0.22667**	0.57176	**0.23505**	0.63468
	1.00	**0.24884**	0.26546	1.13626	0.26458	1.19208

8.C Data-Driven Parameter Selection

Data driven techniques for determining parameter values were discussed in Chapter 4, where both of the basic forms of selection criteria were covered in some detail. Plug-in methods use mse formulas which are estimated using candidate values of the parameters(s) in question. The parameter value(s) that result in the minimum mse value is (are) selected by the criterion as those to be used in the final fit. Cross-validation techniques are based on the idea of building a model with n-1 observations (that is, with the i^{th} data point left out) and then evaluating the fit *at*

the i^{th} data point, i = 1, ..., n. The most familiar cross-validation criterion is the PRediction Error Sum of Squares (PRESS), given by

$$\text{PRESS} = \sum_{i=1}^{n} (e_{i,-i})^2 , \qquad (8.C.1)$$

(see Stone, (1974)) where $e_{i,-i} = y_i - \hat{y}_{i,-i}$ is the PRESS residual, and $\hat{y}_{i,-i}$ is the predicted value at the i^{th} data location when the i^{th} value of the response is not used to calculate the fit.

For nonparametric regression, PRESS has been studied as a bandwidth selection criterion and found to select bandwidths that are smaller than the optimal bandwidth (because it penalizes more for bias that it does for variance, resulting in an undersmoothed fit). Einsporn introduced PRESS*, which penalizes PRESS for small bandwidths through an adjustment in the denominator, and is given by

$$\text{PRESS}* = \frac{\sum_{i=1}^{n} (e_{i,-i})^2}{n - \text{trace}(\mathbf{H}^{(\bullet)})} , \qquad (8.C.2)$$

where $\mathbf{H}^{(\bullet)}$ is the Hat matrix of the procedure under consideration. Since trace($\mathbf{H}^{OLS}$) = R, the number of parameters in the fit, this value can be interpreted for any procedure as the equivalent number of parameters being utilized in the overall fit to the data. If trace($\mathbf{H}^{(\bullet)}$) is "large" (which is the result of undersmoothing, a consequence of a small bandwidth), then the fit is somewhat equivalent to fitting a higher order polynomial, resulting in a variable fit. This results in a small denominator for PRESS*, which in effect penalizes small bandwidths.

Mays and Birch (1996) studied the properties of PRESS* and found that the adjustment overcorrects in such a way that PRESS* chooses too large a bandwidth. He introduces PRESS**, which adds an additional penalty for too large a bandwidth, and is given by

$$\text{PRESS}** = \frac{\sum_{i=1}^{n} (e_{i,-i})^2}{n - \text{trace}(\mathbf{H}^{(\bullet)}) + (n-1)\dfrac{\text{SSE}_{max} - \text{SSE}(\theta)}{\text{SSE}_{max}}} , \qquad (8.C.3)$$

where $SSE_{max} = \sum (y_i - \bar{y})^2$ (like the value of SS_{total} in ANOVA for regression), $SSE(\theta) = \sum (y_i - \hat{y}_{i,-i}(\theta))^2$, and θ is the parameter in question. The additional quantity in the denominator approaches zero as $b_n \to \infty$, and it approaches n-1 as $b_n \to 0$. Thus, smaller bandwidths result in a larger denominator due to this adjustment, which results in a smaller value of PRESS**. Note that this criterion can be utilized for selection of both the bandwidth and the mixing parameter that play a role in MRR2 and ORMRR. When selecting the bandwidth, however, the fit is to the residuals, and thus y_i is replaced by r_i in 8.C.3.

The problem with using this selection criterion in the presence of outliers is precisely the same as the problem with using OLS as a fitting procedure when outliers are present, in that the quadratic loss function allows individual observations to exert an inordinate amount of influence over the value of PRESS**. Consider the robust version of PRESS** proposed in (8.C.4) that will be studied via Monte Carlo simulation in the following section. The notation used is $d_\rho(\theta)$, and the criterion is given by

$$d_\rho(\theta) = \frac{\sum_{i=1}^{n} \rho(e_{i,-i})}{n - \text{trace}(\mathbf{H}^{(\bullet)}) + (n-1)\dfrac{\text{Sum}(\rho_{max}) - \text{Sum}(\rho(\theta))}{\text{Sum}(\rho_{max})}}, \qquad (8.C.4)$$

where $\rho(\bullet)$ is Huber's rho function, $\text{Sum}(\rho_{max}) = \sum \rho(y_i - \hat{y}^M_{loc})$, $\text{Sum}(\rho(\theta)) = \sum \rho(y_i - \hat{y}_i(\theta))$, and $\hat{y}^M_{loc}$ is the M-estimate of location for the values of the response variable y. Recall that Huber's rho is the integral of Huber's ψ function and is defined as

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & |u| \le c_H \\ c_H \cdot |u| - \frac{1}{2}c_H^2 & |u| > c_H \end{cases}.$$

The motivation for this alteration of PRESS** is the same as that for using M-Regression over OLS. That is, the use of a bounding function limits the amount of influence that any given observation can have on the value of the criterion. This adjustment allows $d_\rho(\theta)$ to behave in a similar fashion to PRESS**, even in the presence of outliers.

Several criteria were evaluated via simulations, including robust and non-robust versions of PRESS and PRESS*, in addition to other variations of PRESS**. Of those studied, $d_\rho(\theta)$ performed the best by far, and only the results of the simulations involving $d_\rho(\theta)$ are presented here.

The simulation experiment to study the behavior of $d_\rho(\theta)$ in selecting the parameter θ is the same as for the previous simulations considered. That is, the same underlying model (7.E.1) is used, across the same sample size values and degrees of model misspecification. Obtaining the results is highly computer-intensive, hence only one value for the amount of contamination of the error distribution is considered (10%).

The table values that will be presented are the result of several steps. Initially, for a given situation (a particular value of n, γ, and π), the parameter values for ORMRR are selected for a given simulated data set using $d_\rho(\theta)$ as the parameter selection criterion. This is repeated 500 times, and the reported bandwidth and mixing parameter values are actually the *average* of these 500 parameters selected for the 500 simulated data sets. Table 8.4 presents the parameters selected by $d_\rho(\theta)$, along with the optimal parameters for comparison purposes, for ORMRR. The results for sample size n = 10 are omitted. In practice we found that the criterion performed poorly for such a small sample size.

Note that the theoretically optimal bandwidths decrease and the optimal mixing parameters increase as γ moves from 0 to 1. The reason is obvious, since an increase in γ results in an increase in the amount of curvature in the data for the model in (7.E.1). A smaller bandwidth must be used to capture this additional curvature, while an increasingly larger mixing parameter is necessary because of the lack of fit of the user's specified model. By inspection of the values in Table 8.4, it is apparent that the parameter values selected by $d_\rho(\theta)$ demonstrate this same trend, indicating that $d_\rho(\theta)$ can at least be considered a reasonable choice as a parameter selection criterion.

Even though the data driven values appear to be somewhat off in some cases (especially for $\gamma = 0$), they must be considered in the context of the corresponding value of γ. If $\gamma = 0$, the bandwidth value selected is practically irrelevant since there should be little or no structure in the

residuals. This is also true for the mixing parameter. If no model misspecification exists, then even for a mixing parameter of 1, little should be added back to the parametric fit since there should be relatively little structure captured by the nonparametric fit to the residuals (as the residuals should be randomly scattered about zero).

Table 8.4 Average bandwidths and mixing parameters selected across simulated data sets via $d_\rho(\theta)$ as compared with theoretically optimal values.

n	γ	Bandwidth	Bandwidth (by $d_\rho(\theta)$)	Mixing Parameter	Mixing Parm (by $d_\rho(\theta)$)
	0.00	1.00000	0.177406	0.02500	0.399031
	0.25	0.21953	0.177188	0.79688	0.455090
19	0.50	0.15586	0.113688	0.93438	0.686288
	0.75	0.12227	0.097219	1.00000	0.811020
	1.00	0.10000	0.084906	1.00000	0.916422
	0.00	1.00000	0.156109	0.02500	0.338926
	0.25	0.18750	0.137422	0.86875	0.528058
30	0.50	0.13633	0.109719	0.97656	0.743315
	0.75	0.10859	0.090781	1.00000	0.891571
	1.00	0.08945	0.079188	1.00000	0.961699
	0.00	1.00000	0.130234	0.02500	0.297698
	0.25	0.15820	0.108719	0.93125	0.515071
50	0.50	0.11797	0.087453	1.00000	0.850084
	0.75	0.09551	0.075500	1.00000	0.967983
	1.00	0.07930	0.067938	1.00000	0.990898

A consideration that is more important than the parameter values chosen is the value of INTMSE observed when fitting using $d_\rho(\theta)$ as the parameter selection criterion. The goal of selecting parameter values that are near optimal is secondary to obtaining a satisfactory overall fit (that is, obtaining INTMSE values when using $d_\rho(\theta)$ as the selection criterion that are near the INTMSE values observed when using the optimal parameters).

Table 8.5 provides information on several quantities of interest. First, it quantifies how well ORMRR performs using $d_\rho(\theta)$ as the parameter selection criterion by displaying the resulting INTMSE values (column 2), along with the INTMSE values observed from using the optimal parameters (column 1, which is directly from Table 8.2).

In addition, values for Loess are included for comparison purposes. Three INTMSE values are provided. This includes using the optimal parameters (column 3, which is directly from Table 8.2), using $d_\rho(\theta)$ as the selection criterion, and using the fraction f = 0.5 for all situations. The latter column of values is included because of the reality in applied statistics that quite often practitioners are less willing to use a complex technique because of the additional time it may require over some other technique that may be less complicated. Nonparametric regression is not an exception. Both Minitab and S-Plus use f = 0.5 as the default smoothing parameter value for Loess, and in communication with Dr. Stephen Marron, he expressed the opinion that many users will accept the default because of either a lack of understanding of other options, or because they do not feel that further effort in selecting the parameter via data driven techniques is worthwhile.

Table 8.5 Comparison of INTMSE values: theoretically optimal, simulated using optimal parameters, and simulated using parameters chosen via $d_\rho(\theta)$ ($\pi = 0.1$).

n	ORMRR INTMSE (opt parms)	ORMRR INTMSE (using $d_\rho(\theta)$)	Loess INTMSE (opt parms)	Loess INTMSE (using $d_\rho(\theta)$)	Loess INTMSE f = 0.5
	0.13705	0.201477	0.18906	0.254993	0.1980164
	0.19487	0.228690	0.23431	0.32958	0.2555738
19	0.28755	0.278486	0.29832	0.32539	0.3112824
	0.31193	0.343906	0.30551	0.36595	0.4249751
	0.38567	0.414216	0.38274	0.46435	0.654435
	0.08391	0.109047	0.12119	0.15764	0.120414
	0.14622	0.158896	0.14246	0.18274	0.1572721
30	0.17440	0.200817	0.16605	0.21740	0.1935419
	0.20773	0.244765	0.18395	0.25232	0.2768725
	0.24619	0.260614	0.22471	0.229718	0.4546181
	0.04576	0.068798	0.07298	0.09486	0.0744161
	0.09492	0.114705	0.09502	0.12093	0.115751
50	0.11246	0.124849	0.10199	0.13432	0.1585382
	0.13705	0.138101	0.12137	0.13278	0.2292314
	0.15961	0.157251	0.15847	0.14560	0.3936657

The reason for including these values is to exhibit the superiority of ORMRR over the method that many users employ on a regular basis. It is obvious from the table that this approach

is only competitive when there is little curvature in the data ($\gamma \approx 0$), and in such cases a large smoothing parameter (such as $f = 0.5$) is similar to fitting a global parametric model.

The INTMSE values for Loess when using $d_\rho(\theta)$ as the selection criterion performs relatively well, but is still beaten on almost every occasion by ORMRR. In some cases, ORMRR even performs better with the data driven parameter selection than the values for Loess using the optimal parameters, indicating that ORMRR truly is superior to Loess at least for situations similar to the one being considered here. The results are very encouraging for the single regressor case, and we hope to continue to show improvement over existing methods by extending ORMRR to the multiple regressor case.