A Regression Model Selection Criterion Based on Bootstrap Bumping for Use With Resistant Fitting

by Andrew A. Neath[†] and Joseph E. Cavanaugh[‡]

[†] Department of Mathematics and Statistics
P.O. Box 1653
Southern Illinois University
Edwardsville, IL 62026
E-mail: aneath@siue.edu
Phone: (618) 692-3590
Fax: (618) 692-3147

[‡] Department of Statistics
222 Math Sciences Building
University of Missouri
Columbia, MO 65211
E-mail: cavanaugh@stat.missouri.edu
Phone: (573) 882-4491
Fax: (573) 884-5524

Abstract

We propose a model selection criterion for regression applications where resistant fitting is appropriate. Our criterion gauges the adequacy of a fitted model based on the median squared error of prediction. The criterion is easily computed using the bootstrap "bumping" algorithm of Tibshirani and Knight (1999), which provides a convenient method for obtaining least median of squares model parameter estimates. We present an example to illustrate the merit of the criterion in instances where the underlying data set contains influential values. Additionally, we present and discuss the results of a simulation study which illustrates the effectiveness of the criterion under a wide range of error distributions.

Key words and phrases: Akaike information criterion, bootstrap bumping, Kullback-Leibler information, least median of squares, model selection criterion, regression.

1. Introduction

It is well known that the least squares estimators of regression coefficients are not robust to the influence of individual data points. The study of influential observations has led to variations of the classical normal-error linear model and to the development of alternative “resistant” estimation techniques. Staudte and Sheather (1990) provide an overview.

A companion to the problem of parameter estimation is the problem of model selection, which consists of choosing an appropriate model from a class of candidate models to characterize the data at hand. Such a determination is often facilitated by the use of a model selection criterion. A selection criterion is frequently designed to estimate an expected overall discrepancy, a quantity which reflects the degree of similarity between a fitted approximating model and the operating or “true” model. Various selection criteria have been proposed to target various types of discrepancies. In a modeling problem, a discrepancy should ideally be chosen to evaluate a fitted model based on a principle which is relevant to the application of interest (cf. Linhart and Zucchini, 1986, p. 16). For settings in which resistant fitting is appropriate, we consider a discrepancy which assesses the propriety of a fitted model utilizing the median squared error of prediction. We then propose a model selection criterion which targets this discrepancy. The computation of our criterion employs the bootstrap “bumping” algorithm of Tibshirani and Knight (1999), which provides a convenient procedure for fitting a model based on the least median of squares principle.

Section 2 is devoted to a development of the criterion. In Section 3, we provide an example to illustrate how an improper model may be chosen when the selection procedure does not account for influential observations. We close in Section 4 by presenting and discussing the results of a simulation study which illustrates the effectiveness of the new criterion under a wide range of error distributions.

2. Criterion Development

Assume we observe data $\{(\underset{\sim}{x}_1, y_1), \ldots, (\underset{\sim}{x}_n, y_n)\}$, where the $(\underset{\sim}{x}_i, y_i)$ are independent and identically distributed according to some unknown distribution F. Suppose that each $\underset{\sim}{x}_i$ is a vector of explanatory variables, and each y_i is a scalar response.

Assume that the relationship between y_i and $\underline{x}_i$ is hypothesized to be of the form

$$y_i = h(\underline{x}_i; \underline{\theta}) + \epsilon_i; \qquad i = 1, \ldots, n; \tag{2.1}$$

where $\underline{\theta} \in \Theta$ is a parameter vector, $h(\underline{x}; \underline{\theta})$ is a prediction/response function based on $\underline{x}$ and $\underline{\theta}$, and the ϵ_i are zero mean, independent, identically distributed, and independent of the $\underline{x}_i$.

The number of parameters in $\underline{\theta}$ which must be estimated determines the *dimension* of the model. In the classical linear regression model where $h(\underline{x}; \underline{\theta})$ is linear in $\underline{\theta}$, explanatory variables in $\underline{x}$ are included in the model by estimating the corresponding components of $\underline{\theta}$, and excluded from the model by fixing the corresponding components of $\underline{\theta}$ at zero. The dimension of the model, therefore, corresponds to the number of explanatory variables retained (plus one if an intercept is used). Loosely speaking, our goal is to determine which parameters in $\underline{\theta}$ must be estimated in order for $h(\underline{x}; \underline{\widehat{\theta}})$ to serve as a "good" predictor of y. We wish to make this determination using a method which will not be unduly affected by influential values in the underlying data set. To accomplish our objective, we propose a model selection criterion which gauges the adequacy of the fitted model utilizing the median squared error of prediction.

Before we introduce our selection criterion, we present a brief discussion of model selection criteria based on discrepancies. Our terminology and much of our notation follows that of Linhart and Zucchini (1986).

Let F denote the distribution function which presumably generated the data at hand, and let G denote a distribution function which serves as an approximation to F. Assume that F and G both belong to a class of distribution functions $\mathcal{M}$. We will refer to F as the *operating* or *"true"* model and to G as an *approximating* model. We assume that F is unknown.

A *discrepancy function* is a mapping from $\mathcal{M} \times \mathcal{M}$ to $\Re$ which has the property $\Delta(G, F) \geq \Delta(F, F)$. The purpose of a discrepancy function is to quantify the similarity between two models: thus, $\Delta(G, F)$ should increase as the disparity between G and F increases. It is generally assumed that the approximating model G is parametric, in which case we may write G as $G_{\underline{\theta}}$, $\underline{\theta} \in \Theta$, and use $\Delta(\underline{\theta}) = \Delta(G_{\underline{\theta}}, F)$ to denote the discrepancy.

Although $\Delta(\underline{\theta})$ depends on F and is therefore unknown, a natural estimator of $\Delta(\underline{\theta})$ is provided by the *empirical discrepancy* $\widehat{\Delta}(\underline{\theta}) = \Delta(G_{\underline{\theta}}, \widehat{F})$, where $\widehat{F}$ is the empirical distribution function. A *minimum discrepancy estimator* of $\underline{\theta}$ is defined by $\widehat{\underline{\theta}} = \text{argmin}_{\underline{\theta}}\ \widehat{\Delta}(\underline{\theta})$. The *overall discrepancy*, $\Delta(\widehat{\underline{\theta}}) = \Delta(G_{\underline{\theta}}, F)|_{\underline{\theta}=\widehat{\underline{\theta}}}$, assesses how well the fitted approximating model $G_{\widehat{\underline{\theta}}}$ conforms to the true model F. A model selection criterion is often formulated by constructing a statistic which agrees with $\Delta(\widehat{\underline{\theta}})$ in expectation (at least approximately). Such a criterion can therefore be regarded as an estimator of the expected overall discrepancy $E_F[\Delta(\widehat{\underline{\theta}})]$ (cf. Linhart and Zucchini, 1986, p. 13). As the subsequent discussion will indicate, the natural estimator of $E_F[\Delta(\widehat{\underline{\theta}})]$, namely $\widehat{\Delta}(\widehat{\underline{\theta}}) = \Delta(G_{\underline{\theta}}, \widehat{F})|_{\underline{\theta}=\widehat{\underline{\theta}}}$, is negatively biased. This bias must be corrected if $\widehat{\Delta}(\widehat{\underline{\theta}})$ is to serve as the basis for a model selection criterion targeting $E_F[\Delta(\widehat{\underline{\theta}})]$.

The choice of a discrepancy function for model selection involves a consideration of those properties of a fitted model which are important to the application of interest. Our objective is to develop a criterion for choosing the appropriate form of the prediction/response function h, and to have this determination be resistant to influential values. Thus, the criterion should perform well in settings where the error distribution has thicker tails than the normal distribution. Moreover, to permit the criterion to be broadly applicable, its development should not depend upon a specific error distribution.

For normally distributed errors, a useful discrepancy function is the Gauss discrepancy, based on the mean squared error of prediction:

$$\Delta^{\text{G}}(\underline{\theta}) = E_F[(y - h(\underset{\sim}{x}; \underline{\theta}))^2].^{1} \tag{2.2}$$

A robust alternative to this discrepancy function might be based on the median squared error of prediction:

$$\Delta^{\text{MD}}(\underline{\theta}) = \text{median}_F[(y - h(\underset{\sim}{x}; \underline{\theta}))^2]. \tag{2.3}$$

The empirical discrepancy corresponding to (2.3) is given by

$$\widehat{\Delta}^{\text{MD}}(\underline{\theta}) = \text{median}\{(y_i - h(\underset{\sim}{x}_i; \underline{\theta}))^2; i = 1, \ldots, n\}. \tag{2.4}$$

[1] The approximating model G is defined through the prediction/response function h. A more complete notation for h would therefore be h_G. For simplicity, we omit the subscript G on h.

Thus, the minimum discrepancy estimator of $\underline{\theta}$ associated with Δ^{MD} is the least median of squares estimator (Rousseeuw, 1984).

Estimators of an expected overall discrepancy, $E_F[\Delta(\widehat{\underline{\theta}})]$, may be developed using, among other methods, asymptotics, cross-validation, and bootstrapping. Linhart and Zucchini (1986) give a general outline. For the estimation of $E_F[\Delta^{\text{MD}}(\widehat{\underline{\theta}})]$, we propose a statistic which is based on $\widehat{\Delta}^{\text{MD}}(\widehat{\underline{\theta}})$. This statistic can be conveniently evaluated utilizing the bootstrap "bumping" algorithm of Tibshirani and Knight (1999), and involves a bootstrap-based bias adjustment recommended by Efron (1983, 1986). We will discuss the implementation of the bootstrap-based bias adjustment for estimating an arbitrary expected discrepancy $E_F[\Delta(\widehat{\underline{\theta}})]$. We will then explain how the adjustment can be used along with bootstrap bumping to obtain an estimate of $E_F[\Delta^{\text{MD}}(\widehat{\underline{\theta}})]$.

Consider writing $E_F[\Delta(\widehat{\underline{\theta}})]$ as follows:

$$E_F[\Delta(\widehat{\underline{\theta}})] = E_F[\widehat{\Delta}(\widehat{\underline{\theta}})] + \left\{ E_F[\Delta(\widehat{\underline{\theta}}) - \widehat{\Delta}(\widehat{\underline{\theta}})] \right\}. \tag{2.5}$$

Efron (1983, 1986) refers to the bracketed quantity in (2.5) as the expected *optimism* in judging the fit of a model using the same data as that which was used to construct the fit. It can easily be argued that this quantity is positive, which implies that $\widehat{\Delta}(\widehat{\underline{\theta}})$ serves as a negatively biased estimator of the expected discrepancy $E_F[\Delta(\widehat{\underline{\theta}})]$. In order to correct for this bias, we must estimate the bracketed term. An effective and simple approach involves bootstrapping.

Let $\widehat{\Delta}^*(\underline{\theta})$ be the empirical discrepancy corresponding to a bootstrap sample, and let $\widehat{\underline{\theta}}^*$ be the corresponding bootstrap estimator: $\widehat{\underline{\theta}}^* = \text{argmin}_{\underline{\theta}}\ \widehat{\Delta}^*(\underline{\theta})$. Assume that a sequence of B bootstrap samples are obtained, resulting in a sequence of empirical discrepancies $\widehat{\Delta}_1^*(\underline{\theta}), \ldots, \widehat{\Delta}_B^*(\underline{\theta})$ and corresponding estimates $\widehat{\underline{\theta}}_1^*, \ldots, \widehat{\underline{\theta}}_B^*$. The bootstrap analogue of the bracketed term in (2.5) is

$$E_{\widehat{F}}\left[\widehat{\Delta}(\widehat{\underline{\theta}}^*) - \widehat{\Delta}^*(\widehat{\underline{\theta}}^*)\right],$$

which can be estimated by

$$\widehat{\text{bias}} = \frac{1}{B}\sum_{b=1}^{B}\left[\widehat{\Delta}(\widehat{\underline{\theta}}_b^*) - \widehat{\Delta}_b^*(\widehat{\underline{\theta}}_b^*)\right]. \tag{2.6}$$

Since $\widehat{\Delta}(\widehat{\underline{\theta}})$ serves as an unbiased estimator of $E_F[\widehat{\Delta}(\widehat{\underline{\theta}})]$ in (2.5), an approximately unbiased estimator of the expected discrepancy $E_F[\Delta(\widehat{\underline{\theta}})]$ is given by

$$\widehat{E}_F[\Delta(\widehat{\underline{\theta}})] = \widehat{\Delta}(\widehat{\underline{\theta}}) + \widehat{\text{bias}}. \tag{2.7}$$

The estimator (2.7) can be constructed in any modeling framework conducive to bootstrapping, provided that the empirical discrepancy $\widehat{\Delta}(\underline{\theta})$ and the corresponding minimum discrepancy estimator $\widehat{\underline{\theta}}$ can be conveniently evaluated. In the regression setting where the discrepancy function is Δ^{MD}, the minimum discrepancy estimator is the least median of squares estimator (LMSE). The computation of $\widehat{\underline{\theta}}$, therefore, is not a trivial task. A discussion of some proposed algorithms for computing either the exact LMSE or an approximation can be found in Ryan (1997). A recently proposed method involves treating the minimization as an integer programming problem and applying techniques familiar in operations research (Neath and Sewell, 2000). The approach we use here is based on the bootstrap bumping algorithm (Tibshirani and Knight, 1999). The basic procedure involves taking repeated bootstrap samples, calculating the least squares estimate each time. The least squares estimate from this collection which minimizes the least median of squares condition with respect to the original data is taken as the LMSE. Intuitively, outliers in the original sample will by chance be excluded from some of the bootstrap samples; the least squares estimate for one of these samples should be close to the actual LMSE. Theoretical details are provided in Tibshirani and Knight (1999).

We now outline the computation of our criterion via (2.7) and bootstrap bumping. Let $\{(\underline{x}^*_{1,b}, y^*_{1,b}), \ldots, (\underline{x}^*_{n,b}, y^*_{n,b})\}$ be the b^{th} bootstrap sample selected from $\widehat{F}$, where $b = 1, \ldots, B$. Let $\tilde{\underline{\theta}}^*_b$ be the corresponding *least squares* estimator. Among $\tilde{\underline{\theta}}^*_1, \ldots, \tilde{\underline{\theta}}^*_B$, find

$$\widehat{\underline{\theta}} = \text{argmin}_{\tilde{\underline{\theta}}^*_b} \left[\text{median}\{(y_i - h(\underline{x}_i; \tilde{\underline{\theta}}^*_b))^2; i = 1, \ldots, n\}\right].$$

This $\widehat{\underline{\theta}}$ will function as the effective LMSE of $\underline{\theta}$. Use $\widehat{\underline{\theta}}$ in evaluating the empirical discrepancy $\widehat{\Delta}^{\text{MD}}(\widehat{\underline{\theta}})$ via (2.4).

To compute the bias adjustment (2.6), for each $b = 1, \ldots, B$, find

$$\widehat{\text{bias}}^{\text{MD}}_b = \text{median}\{(y_i - h(\underline{x}_i; \widehat{\underline{\theta}}^*_b))^2; i = 1, \ldots, n\} - \text{median}\{(y^*_{i,b} - h(\underline{x}^*_{i,b}; \widehat{\underline{\theta}}^*_b))^2; i = 1, \ldots, n\}.$$

Here, $\widehat{\underline{\theta}}_b^*$ represents the effective LMSE of $\underline{\theta}$ based on the b^{th} bootstrap sample, which is computed from the bootstrap sample in the same manner that $\widehat{\underline{\theta}}$ is computed from the original sample. Next, evaluate $\widehat{\text{bias}}^{\text{MD}} = (1/B)\sum_{b=1}^{B} \widehat{\text{bias}}_b^{\text{MD}}$. The criterion resulting from (2.7), which we denote as LMS, is given by

$$\text{LMS} = \widehat{\Delta}^{\text{MD}}(\widehat{\underline{\theta}}) + \widehat{\text{bias}}^{\text{MD}}.$$

In practice, we consider a set of candidate models $M_0, M_1, \ldots, M_L$ of the form (2.1), where each M_k is uniquely determined by the parameter vector ${}_k\underline{\theta} \in \Theta(K) \subseteq \Theta$. The fitted model M_k for which LMS is minimized will ideally correspond to the fitted model for which $E_F[\Delta^{\text{MD}}({}_k\widehat{\underline{\theta}})]$ is minimized. Among the fitted candidates, this model is preferred, since it is based on a predictor $h(\underline{\text{x}}; \widehat{\underline{\theta}})$ which tends to minimize the median squared error of prediction for data $(\underline{\text{x}}, y)$ generated under F.

3. An Example

We present an example to illustrate the application of the LMS criterion. We compare the performance of the criterion to an analogous bootstrap-based criterion which is developed assuming normally distributed errors.

We consider candidate models of the form (2.1) where $h(\underline{\text{x}}; \underline{\theta})$ is linear in $\underline{\theta}$. For the explanatory variables in our models, we consider a continuous regressor x, and a binary categorical variable which we will treat with the indicator I. ($I = 1$ if an observation is from group 1; $I = 0$ if an observation is from group 2.) The candidate models are as follows:

$$\begin{array}{lrcl}
M_0: & y & = & \alpha_0 + \epsilon \\
M_1: & y & = & \alpha_0 + \alpha_1 I + \epsilon \\
M_2: & y & = & \alpha_0 + \alpha_1 I + \beta_{11} x + \epsilon \\
M_3: & y & = & \alpha_0 + \alpha_1 I + \beta_{11} x + \beta_{21} x^2 + \epsilon \\
M_4: & y & = & \alpha_0 + \alpha_1 I + \beta_{11} x + \beta_{12} x I + \epsilon \\
M_5: & y & = & \alpha_0 + \alpha_1 I + \beta_{11} x + \beta_{12} x I + \beta_{21} x^2 + \beta_{22} x^2 I + \epsilon
\end{array}$$

The values for the explanatory variables are taken from the second and fourth columns of Table 9.2, Linhart and Zucchini (1986, p. 167). The values for x have been centered by subtracting off the overall mean. The values for the response variable are generated from the operating model

$$y = 97 - 7I - 2x + 1xI + 0.01x^2 + \epsilon,$$

where ϵ is distributed as a mixture of normals, $0.7\ N(0, 5^2) + 0.3\ N(0, 50^2)$; i.e.,

$$\delta\ N(0, 5^2) + (1 - \delta)\ N(0, 50^2) \text{ with } P(\delta = 1) = 0.7 \text{ and } P(\delta = 0) = 0.3.$$

A listing of the data is provided in Table 1, and a plot of the true prediction curves is featured in Figure 1.

The operating model is a special case of the candidate model M_5, the latter which allows groups 1 and 2 to have entirely separate quadratic response functions. We therefore regard M_5 as the most appropriate model, even though it includes an interaction term between x^2 and the indicator I which is absent in the operating model.

We compare the performance of the LMS criterion to a bootstrap-based criterion which targets the Kullback-Leibler discrepancy and assumes normally distributed errors. Define

$$\Delta^{\text{KL}}(\underline{\theta}) = E_F[-2 \ln g_{\underline{\theta}}(\underline{\mathrm{x}}, y)]$$

where

$$g_{\underline{\theta}}(\underline{\mathrm{x}}, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y - h(\underline{\mathrm{x}}; \underline{\theta}))^2}{2\sigma^2} \right\} g_{\underline{\mathrm{x}}}(\underline{\mathrm{x}}).$$

The minimum discrepancy estimator of $\underline{\theta}$ associated with Δ^{KL} is the maximum likelihood estimator, or equivalently, the least squares estimator. Following the procedure in Section 2 for obtaining (2.7), we can develop a bootstrap-based estimator of $E_F[\Delta^{\text{KL}}(\widehat{\underline{\theta}})]$, resulting in a selection criterion appropriate under the normal-error assumption. We will refer to this criterion as KL. (For further consideration of this criterion, see Ishiguro, Morita, and Ishiguro, 1991; Cavanaugh and Shumway, 1997; and Shibata, 1997.)

In computing the criterion scores for each of the fitted candidate models, and in obtaining the least median of squares estimates via bootstrap bumping, 200 bootstrap samples are used. The scores obtained are as follows:

Model	M_0	M_1	M_2	M_3	M_4	M_5
KL	9.595	9.648	8.444	8.354	8.610	8.805
LMS	1649	1880	76	104	135	55

The KL criterion is unable to detect the most adequate model amidst the presence of influential values which result from using the mixture distribution to generate the errors. The criterion picks the underfitted model M_3, while the LMS criterion appropriately chooses the full model M_5.

Scatterplots with the estimated prediction curves are displayed in Figures 2 and 3. The model selected by the KL criterion, fit using least squares, is overly influenced by certain data points and represents a very poor fit. The parallel response curves featured in Figure 2 do not resemble the true curves in Figure 1; predictions for y based on these curves would be highly inaccurate. On the other hand, the model selected by the LMS criterion, fit using least median of squares, is resistant to the influential values and represents a very good fit. The response curves featured in Figure 3 are quite similar to the true curves in Figure 1.

Clearly, one cannot make generalizations regarding the behavior of the LMS criterion based upon a single example. However, our outcome suggests that in outlier-prone settings, the LMS criterion may be more likely than a non-robust competitor to select an adequate candidate model. This notion is further explored in the next section.

4. Simulation Results

We examine the performance of the LMS criterion against other selection criteria in a simulation study. Some of these criteria are bootstrap-based, while some are more computationally simplistic. Some are designed to be resistant, while some are developed assuming normally distributed errors.

Included as a competitor to the LMS criterion will be the KL criterion considered in the previous section. We shall also include a bootstrap-based criterion which targets the Gauss discrepancy (2.2). The minimum discrepancy estimator of $\underset{\sim}{\theta}$ associated with Δ^{G} is the least squares estimator. Again, we develop the bootstrap-based estimator of $E_F[\Delta^{\mathrm{G}}(\widehat{\underset{\sim}{\theta}})]$ following the procedure for obtaining (2.7). We denote this criterion by G.

Asymptotic estimators of expected discrepancies have the advantage of being computationally very simple. We include in our study an asymptotic estimator of the expected Kullback-Leibler discrepancy $E_F[\Delta^{\mathrm{KL}}(\widehat{\underline{\theta}})]$, which is the well-known Akaike (1973, 1974) information criterion (AIC). (A small-sample variant of AIC for use in regression applications is investigated by Hurvich and Tsai, 1989.) We also include an estimator of the expected Gauss discrepancy $E_F[\Delta^{\mathrm{G}}(\widehat{\underline{\theta}})]$, which is often called the Sp criterion (see Linhart and Zucchini, 1986, pp. 112–115). For a candidate model of the form (2.1), let p denote the model dimension, and let $\widehat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(y_i - h(\underline{\mathrm{x}}_i;\widehat{\underline{\theta}}))^2$. AIC and Sp can then be defined as follows:

$$\mathrm{AIC} = \ln(\widehat{\sigma}^2) + \frac{2p}{n}, \tag{4.1}$$

$$\mathrm{Sp} = \frac{n\widehat{\sigma}^2}{(n-p)(n-p-1)}. \tag{4.2}$$

The criteria KL, G, AIC, and Sp are derived in the setting of normally distributed errors, and are therefore not resistant. We consider an additional bootstrap-based criterion which targets a version of the Kullback-Leibler discrepancy which assumes that the errors are distributed as double exponential. Such a criterion should be reasonably robust to the presence of influential values. Let

$$\Delta^{\mathrm{L_1KL}}(\underline{\theta}) = E_F[-2\ln g_{\underline{\theta}}(\underline{\mathrm{x}}, y)]$$

where

$$g_{\underline{\theta}}(\underline{\mathrm{x}}, y) = \frac{1}{2\sigma^2}\exp\left\{-\frac{1}{\sigma^2}|y - h(\underline{\mathrm{x}};\underline{\theta})|\right\} g_{\underline{\mathrm{x}}}(\underline{\mathrm{x}}).$$

The minimum discrepancy estimator of $\underline{\theta}$ associated with $\Delta^{\mathrm{L_1KL}}$ is the least absolute deviations estimator, i.e., the estimator which minimizes $\sum_{i=1}^{n}|y_i - h(\underline{\mathrm{x}}_i;\underline{\theta})|$. Our bootstrap-based estimator of $E_F[\Delta^{\mathrm{L_1KL}}(\widehat{\underline{\theta}})]$ is again developed following the procedure which leads to (2.7). We denote this criterion by L$_1$KL. (Hurvich and Tsai, 1990, propose a similar criterion which also requires Monte Carlo methods.)

An adjusted version of the Gauss discrepancy which is analogous to $\Delta^{\mathrm{L_1KL}}$ is given by

$$\Delta^{\mathrm{L_1G}}(\underline{\theta}) = E_F[\,|y - h(\underline{\mathrm{x}};\underline{\theta})|\,].$$

The minimum discrepancy estimator of $\underline{\theta}$ associated with $\Delta^{\mathrm{L_1G}}$ is the least absolute deviations estimator. Again, we can develop a bootstrap estimator of $E_F[\Delta^{\mathrm{L_1G}}(\widehat{\underline{\theta}})]$ following the method for obtaining (2.7). We call the resulting criterion $\mathrm{L_1G}$.

Finally, we may define robust versions of AIC and Sp, say $\mathrm{L_1AIC}$ and $\mathrm{L_1Sp}$, by using $\widehat{\sigma}^2 = (1/n)\sum_{i=1}^{n} |y_i - h(\underset{\sim}{x}_i; \widehat{\underline{\theta}})|$ in (4.1) and (4.2). The fitted models which are logically associated with these criteria are based on least absolute deviations estimation.

The behavior of the proposed criteria will be compared by simulating a setting where one must decide among seven candidate models $M_0, M_1, \ldots, M_6$. For ease of exposition, we assume our candidate models are nested, with M_k defined by the linear prediction function $h(\underset{\sim}{x};\, {}_k\underline{\theta}) = \theta_0 + \theta_1 x_1 + \ldots + \theta_k x_k$. This corresponds to a practical linear regression setting where the predictor variables can be listed in an order of importance.

In each of our simulation sets, 100 samples of data $\{(\underset{\sim}{x}_1, y_1), \ldots, (\underset{\sim}{x}_n, y_n)\}, n = 20$, are generated from the operating model, $y = 1 + x_1 + x_2 + x_3 + (0x_4 + 0x_5 + 0x_6) + \epsilon$, where $x_1, \ldots, x_6$ are distributed as Uniform(0,6). Thus, the true model is M_3. For every sample, the seven models in the candidate class are fit to the data, the aforementioned criteria are evaluated, and the model selections for each criterion are recorded. Over the 100 samples, the selections are tabulated and summarized. For the computation of the bootstrap-based criteria and the least median of squares estimates, the number of bootstrap samples used is $B = 200$, as in the Section 3 example.

In addition to selecting the correct model, we would like our criteria to favor accurate prediction functions. The values of the estimated prediction function corresponding to the selected models under each criterion are calculated at the points $\underset{\sim}{x}_0 = (0, \ldots, 0)', \underset{\sim}{x}_3 = (3, \ldots, 3)'$, and $\underset{\sim}{x}_6 = (6, \ldots, 6)'$. The values of the true prediction function at these points are 1, 10, 19, respectively. The mean and the standard deviation of the estimates over the 100 data sets are tabulated and summarized.

We present four simulation sets, each of which employs a different type of error distribution. Set 1 considers a normal distribution, Set 2 a double exponential distribution, Set 3 a mixture distribution involving normals, and Set 4 a Cauchy distribution. The relevant densities are pictured in Figure 4. The simulation results are summarized in Tables 2, 3, 4, and 5, respectively.

Set 1 assumes that the errors are normally distributed with variance equal to 1/4. The KL criterion performs very well here, as it should since the assumptions for its development are met. The LMS criterion is very competitive, choosing the correct model for 96 of the samples. The standard error of prediction is higher for the LMS criterion than for the KL criterion. In this setting, least squares estimates are optimal.

The subsequent sets illustrate how the criteria perform under error distributions with increased variance and heavier tails.

Set 2 considers errors which have the double exponential distribution, with the scale parameter $\sigma^2 = 1$ and a variance of $2\sigma^2 = 2 : DE(0, 1)$. The LMS criterion again performs well, selecting the correct model for 91 of the samples. In this set, LMS outperforms the KL criterion. However, the L_1KL criterion outperforms all other criteria. Again, this is not surprising since L_1KL is developed under the assumption of double exponential errors.

Set 3 considers a mixture of normals for the error distribution: $0.8\, N(0, 1/4){+}0.2\, N(0, 16)$. Here, the LMS criterion obtains correct selections on 95 of the trials. The L_1KL criterion also performs well. The increase in the error variance greatly affects KL, and underfitting becomes problematic. This phenomenon is also illustrated in the example in Section 3.

To see how the criteria hold up as the variability of the errors becomes extreme, we simulate errors in Set 4 from a Cauchy distribution. The LMS criterion follows only L_1G in terms of the number of correct selections. With the noise obscuring the true regression relationship, underfitting becomes an issue with all the other criteria. This may explain the L_1G criterion's surprisingly strong performance. In cases where the error variance is not particularly large, such as in Set 1, L_1G has a strong tendency to overfit. As the variance grows larger, this tendency is attenuated.

It should be noted that LMS estimators have an asymptotic efficiency of zero relative to least squares estimators. One should take this into consideration in applying both LMS estimation and the LMS criterion to large-sample modeling problems.

5. Conclusion

The least median of squares estimator has been studied quite extensively. Hawkins (1993) calls the estimator “the standard method of analysis of data when the possibility of severe badly-placed outliers makes an estimate with [resistant properties] desirable.”

Our goal was to create a regression model selection criterion which would perform effectively across a wide range of error distributions, including thicker tailed distributions which tend to introduce influential, outlying values. Targeting a discrepancy based on the median squared error of prediction is natural if least median of squares estimation is favored as a resistant method of model fitting.

The results of our simulation study in Section 4 indicate that the LMS criterion achieves its purpose. While the other criteria perform poorly in at least one of the simulation sets, the LMS criterion consistently ranks near the top in terms of the number of correct model selections and the accuracy of prediction resulting from the chosen models.

The LMS criterion has a characteristic shared by many nonparametric procedures. Although a nonparametric procedure may be outperformed by a parametric competitor in settings consistent with the conditions under which the latter is developed, the nonparametric procedure often performs more effectively over a broader spectrum of settings. Our results indicate that it is possible to find competitors which are preferable to the LMS criterion in specific instances, yet the LMS criterion seems to perform more effectively than these competitors over a diverse collection of applications.

Acknowledgements

The authors wish to thank the referees for valuable suggestions which served to improve the original version of this manuscript.

The work of the second author was supported by a grant from the National Science Foundation, DMS-9704436.

Table 1. Data for Example.

Group 1	$(I = 1)$	Group 2	$(I = 0)$
x	y	x	y
−59.675	125.96	−68.675	277.74
−51.675	167.04	−54.675	271.72
−50.675	164.90	−31.675	150.64
−49.675	210.92	−23.675	144.39
−49.675	163.24	−18.675	137.78
−43.675	149.02	−15.675	133.06
−32.675	140.68	−10.675	125.80
−28.675	130.19	−9.675	117.74
−28.675	130.46	−7.675	114.85
−21.675	101.43	−7.675	101.22
−20.675	115.91	16.325	146.18
−10.675	101.28	27.325	53.17
0.325	87.82	35.325	41.93
0.325	18.04	37.325	37.61
2.325	91.57	45.325	33.33
10.325	148.59	46.325	34.80
16.325	74.44	61.325	13.04
19.325	73.56	80.325	−25.64
29.325	−43.79	97.325	−0.72
48.325	68.72	123.325	5.37

Figure 1. True Response Curves.

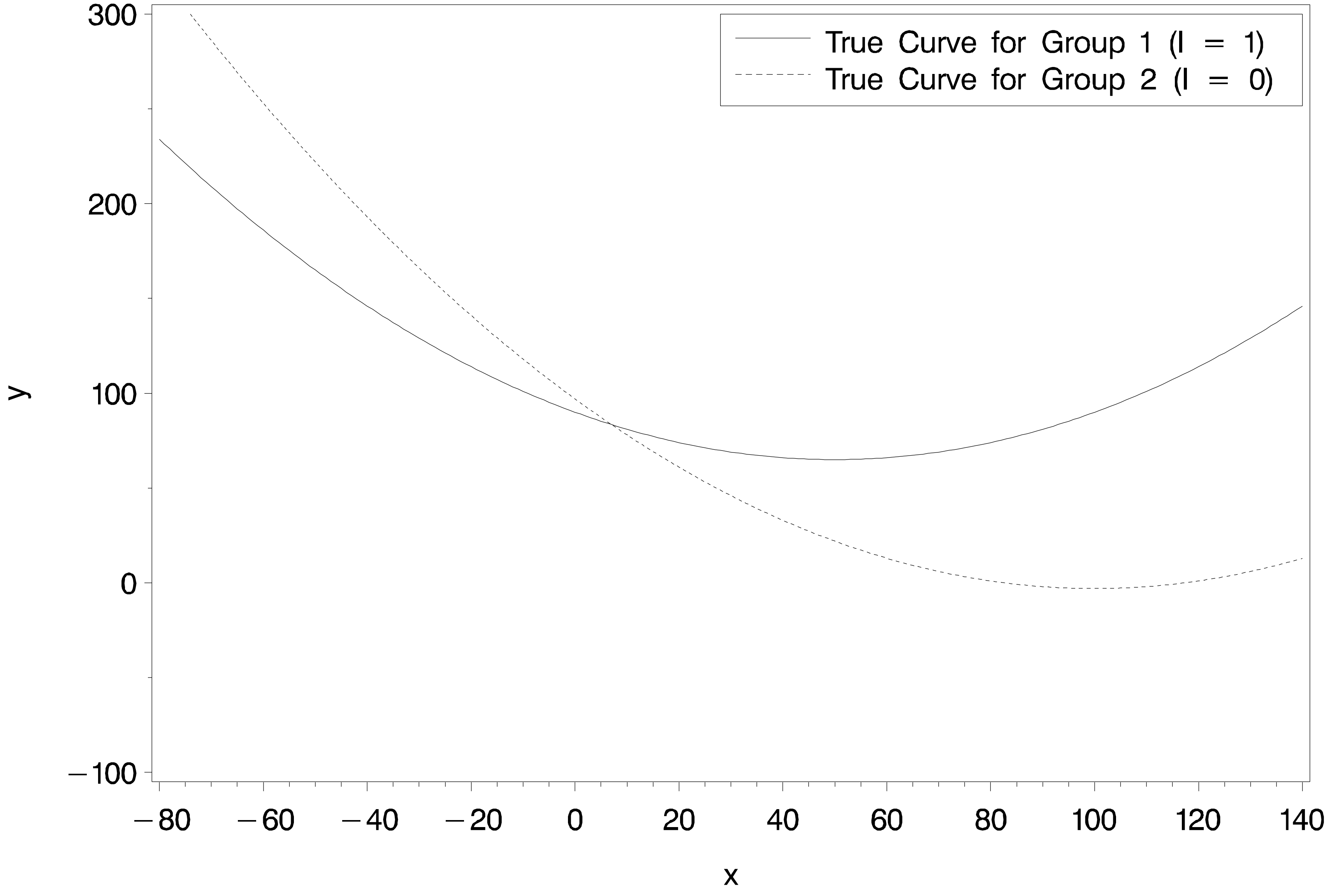

Figure 2. Fitted Model Chosen by KL Criterion.

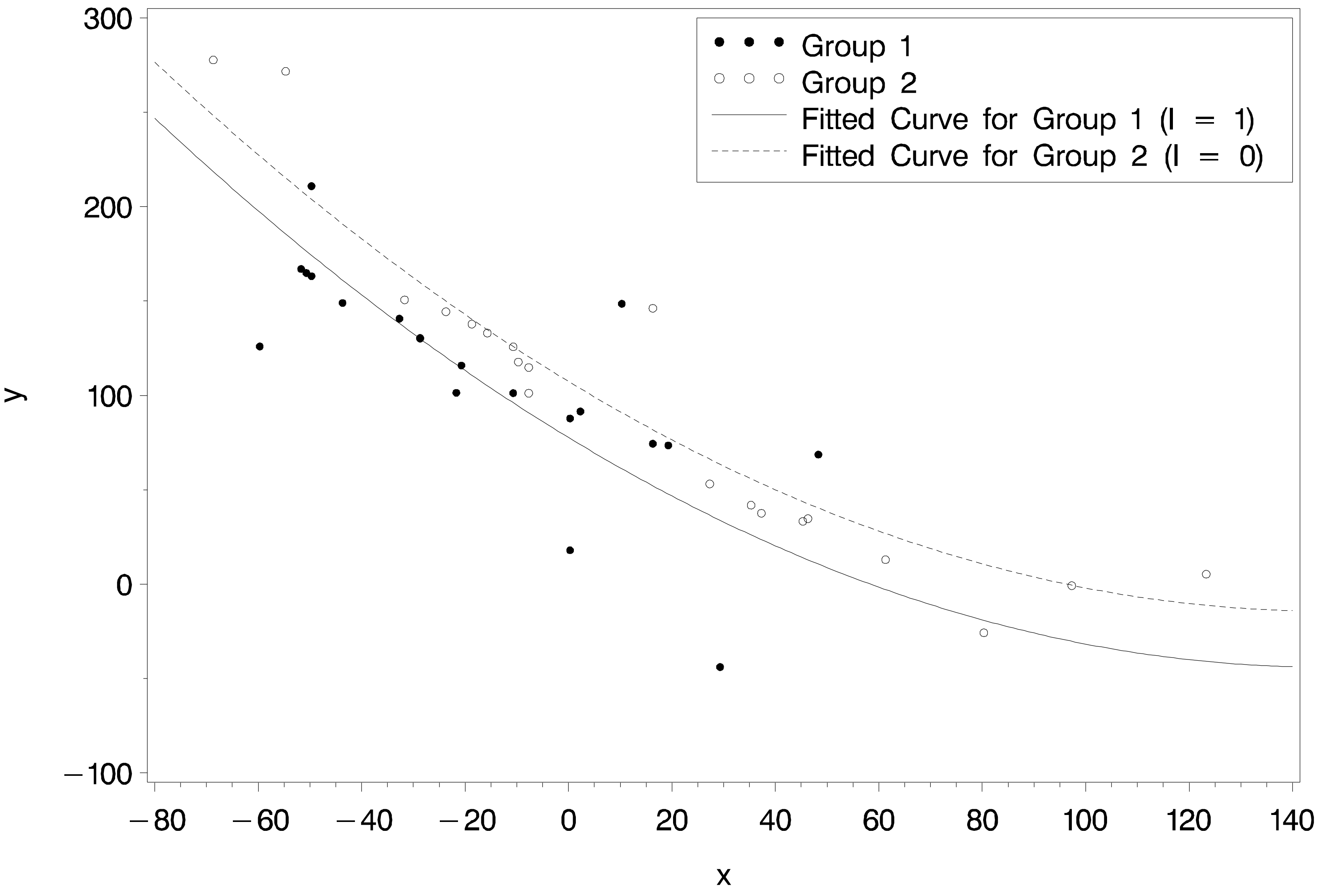

Figure 3. Fitted Model Chosen by LMS Criterion.

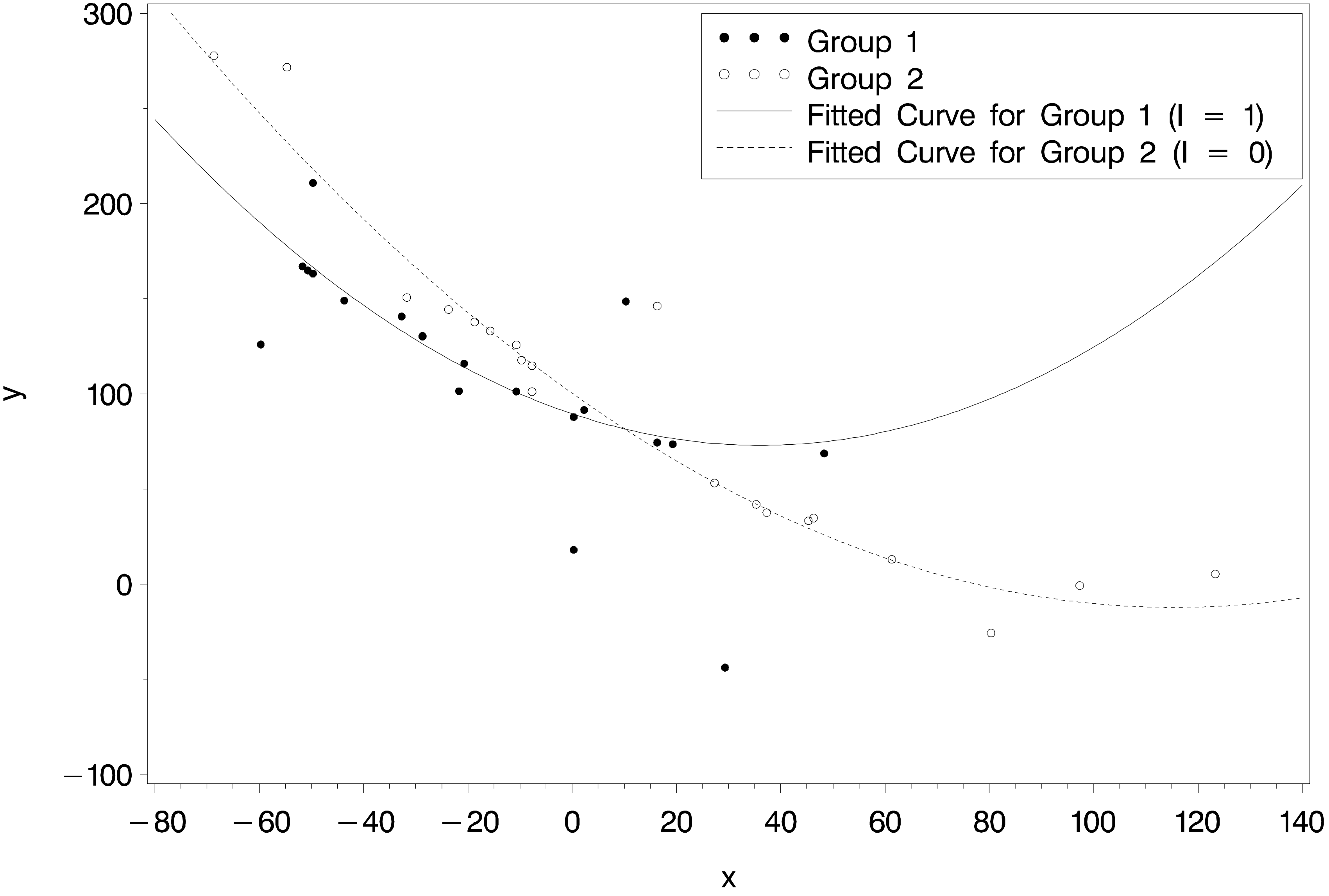

Figure 4. Error Densities for Simulation Sets.

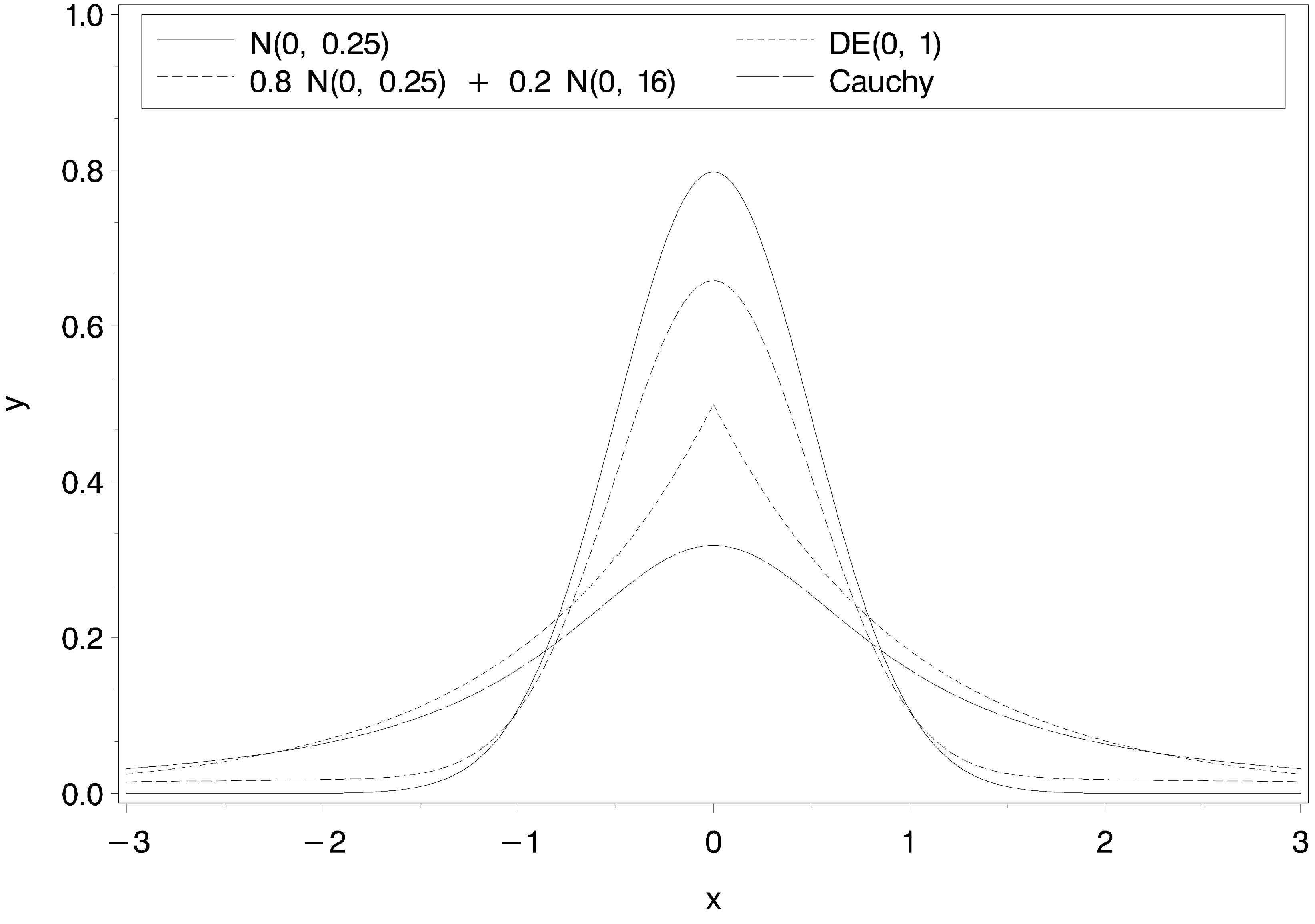

Table 2. Results for Simulation Set 1: $\epsilon \sim N(0,1/4)$

Selections

	AIC	SP	KL	G	L_1AIC	L_1SP	L_1KL	L_1G	LMS
M_0	0	0	0	0	0	0	0	0	0
M_1	0	0	0	0	0	0	0	0	0
M_2	0	0	0	0	0	0	0	0	0
M_3	59	76	98	67	80	91	94	71	96
M_4	15	13	2	18	10	5	4	16	4
M_5	14	4	0	10	3	3	2	4	0
M_6	12	7	0	5	7	1	0	9	0

Predictions

	AIC	SP	KL	G	L_1AIC	L_1SP	L_1KL	L_1G	LMS
$\underset{\sim}{x}_0$	0.998	1.007	1.015	1.026	0.923	0.957	0.969	0.927	1.016
	(0.513)	(0.480)	(0.399)	(0.489)	(0.591)	(0.564)	(0.567)	(0.622)	(0.660)
$\underset{\sim}{x}_3$	9.999	10.001	10.000	10.000	9.992	9.987	9.993	9.994	9.998
	(0.136)	(0.136)	(0.129)	(0.136)	(0.153)	(0.152)	(0.155)	(0.158)	(0.206)
$\underset{\sim}{x}_6$	18.999	18.995	18.985	18.975	19.060	19.017	19.016	19.062	18.980
	(0.499)	(0.473)	(0.403)	(0.490)	(0.585)	(0.533)	(0.526)	(0.610)	(0.647)

Table 3. Results for Simulation Set 2: $\epsilon \sim DE(0,1)$

Selections

	AIC	SP	KL	G	L_1AIC	L_1SP	L_1KL	L_1G	LMS
M_0	0	0	0	0	0	0	0	0	0
M_1	0	0	4	0	0	0	0	0	1
M_2	0	0	7	0	0	0	2	0	4
M_3	66	80	85	71	87	92	97	83	91
M_4	13	9	2	16	6	6	1	9	4
M_5	10	6	0	6	5	1	0	6	0
M_6	11	5	2	7	2	1	0	2	0

Predictions

	AIC	SP	KL	G	L_1AIC	L_1SP	L_1KL	L_1G	LMS
$\underset{\sim}{x}_0$	1.183	1.108	1.439	1.190	0.858	0.896	1.035	0.824	1.430
	(1.173)	(1.075)	(1.570)	(1.147)	(1.319)	(1.230)	(1.207)	(1.344)	(1.756)
$\underset{\sim}{x}_3$	9.986	9.993	9.995	9.992	10.047	10.051	10.060	10.046	10.027
	(0.357)	(0.324)	(0.357)	(0.318)	(0.435)	(0.406)	(0.391)	(0.438)	(0.466)
$\underset{\sim}{x}_6$	18.789	18.878	18.551	18.795	19.237	19.205	19.085	19.269	18.625
	(1.405)	(1.211)	(1.623)	(1.268)	(1.483)	(1.411)	(1.357)	(1.526)	(1.674)

Table 4. Results for Simulation Set 3: $\epsilon \sim 0.8\ N(0,1/4) + 0.2\ N(0,16)$

Selections

	AIC	SP	KL	G	L_1AIC	L_1SP	L_1KL	L_1G	LMS
M_0	0	0	13	0	0	0	0	0	1
M_1	0	0	15	0	0	0	1	0	1
M_2	1	1	38	1	0	1	2	0	1
M_3	62	77	33	75	91	94	96	89	95
M_4	18	14	0	15	5	4	1	8	2
M_5	6	4	0	6	4	1	0	3	0
M_6	13	4	1	3	0	0	0	0	0

Predictions

	AIC	SP	KL	G	L_1AIC	L_1SP	L_1KL	L_1G	LMS
$\underset{\sim}{x}_0$	0.968	0.951	3.938	0.906	1.078	1.124	1.202	1.078	1.042
	(1.892)	(1.845)	(3.327)	(1.835)	(0.191)	(0.187)	(0.491)	(0.175)	(1.181)
$\underset{\sim}{x}_3$	9.989	9.995	10.056	9.998	10.109	10.101	10.072	10.109	9.996
	(0.429)	(0.392)	(0.521)	(0.388)	(0.033)	(0.140)	(0.266)	(0.032)	(0.250)
$\underset{\sim}{x}_6$	19.011	19.038	16.174	19.090	19.141	19.077	18.942	19.139	18.950
	(1.876)	(1.903)	(3.171)	(1.881)	(0.127)	(0.435)	(1.004)	(0.113)	(1.361)

Table 5. Results for Simulation Set 4: $\epsilon \sim$ Cauchy

Selections

	AIC	SP	KL	G	L_1AIC	L_1SP	L_1KL	L_1G	LMS
M_0	24	27	67	28	31	35	42	7	17
M_1	10	12	12	9	9	10	11	11	10
M_2	12	11	14	11	9	9	13	8	12
M_3	30	31	7	35	48	45	34	69	61
M_4	9	9	0	9	2	1	0	3	0
M_5	4	3	0	2	1	0	0	2	0
M_6	11	7	0	6	0	0	0	0	0

Predictions

	AIC	SP	KL	G	L_1AIC	L_1SP	L_1KL	L_1G	LMS
$\underset{\sim}{x}_0$	−1.596	7.616	3.669	−1.659	4.123	4.645	5.463	2.211	3.454
	(27.179)	(58.430)	(35.340)	(35.110)	(4.196)	(4.276)	(4.171)	(2.957)	(4.015)
$\underset{\sim}{x}_3$	6.308	5.983	6.134	6.287	9.820	9.861	9.891	9.855	10.088
	(34.643)	(37.900)	(35.330)	(35.360)	(0.763)	(0.785)	(0.872)	(0.688)	(0.910)
$\underset{\sim}{x}_6$	14.212	4.351	8.600	14.232	15.518	15.077	14.320	17.498	16.721
	(45.490)	(133.380)	(35.850)	(36.940)	(4.475)	(4.383)	(4.403)	(3.329)	(3.831)

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N., Csáki, F. (Editors), *2nd International Symposium on Information Theory*, Akadémia Kiadó, Budapest, pp. 267–281.

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19, 716–723.

Cavanaugh, J. E., Shumway, R. H., 1997. A bootstrap variant of AIC for state-space model selection. *Statistica Sinica* 7, 473–496.

Efron, B., 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* 78, 316–331.

Efron, B., 1986. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* 81, 461–470.

Hawkins, D. M., 1993. The feasible set algorithm for least median of squares regression. *Computational Statistics and Data Analysis* 16, 81–101.

Hurvich, C. M., Tsai, C.–L., 1989. Regression and time series model selection in small samples. *Biometrika* 76, 297–307.

Hurvich, C. M., Tsai, C.–L., 1990. Model selection for least absolute deviations regression in small samples. *Statistics and Probability Letters* 9, 259–265.

Ishiguro, M., Morita, K. I., Ishiguro, M., 1991. Application of an estimator-free information criterion (WIC) to aperture synthesis imaging. In: Cornwell, T. J., Perley, R. A. (Editors), *Radio Interferometry: Theory, Techniques, and Applications*, Astronomical Society of the Pacific, San Francisco, pp. 243–248.

Linhart, H., Zucchini, W., 1986. *Model Selection.* Wiley, New York.

Neath, A.A., Sewell, E.C., 2000. A note on the minimization of the trimmed sum of absolute deviations as an integer program. *Computing Science and Statistics* 31, 227–229.

Rousseeuw, P. J., 1984. Least median of squares regression. *Journal of the American Statistical Association* 84, 871–880.

Ryan, T. P., 1997. *Modern Regression Methods.* Wiley, New York.

Shibata, R., 1997. Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica* 7, 375–394.

Staudte, R. G., Sheather, S. J., 1990. *Robust Estimation and Testing.* Wiley, New York.

Tibshirani, R., Knight, K., 1999. Model search by bootstrap "bumping." *Journal of Computational and Graphical Statistics* 8, 671–686.