Journal of the Korean
Statistical Society
Vol. 24, No. 2, 1995

Robustness of Minimum Disparity Estimators in Linear Regression Models

Rojin Pak[1]

ABSTRACT

This paper deals with the robustness properties of the minimum disparity estimation in linear regression models. The estimators are defined as statistical quantities which minimize the blended weight Hellinger distance between a weighted kernel density estimator of the residuals and a smoothed model density of the residuals. It is shown that if the weights of the density estimator are appropriately chosen, the estimates of the regression parameters are robust.

KEYWORDS: Linear regression models, Blended weight Hellinger distance, Density estimator, Residual adjustment function, Robustness, Least median squares estimators.

1. INTRODUCTION

We have observed that the least-squares estimator (LS) is very sensitive to outliers because the residuals are squared. As an alternative method to LS, the least absolute values regression estimator (L_1-estimator), which is determined by minimizing

$$\sum_{i=1}^{n} |y_i - x_i\beta|,$$

[1]Department of Statistics, Taejon University, Taejon, 300-716, Korea.

is often used. Later, Huber (1973) proposed the M-estimators which is defined as the statistic minimizing

$$\sum_{i=1}^{n} \rho(y_i - x_i\beta)$$

for some function $\rho(\cdot)$. For example, the least-squares estimator is defined by the function $\rho(r) = r^2/2$ and L_1-estimator by $\rho(r) = |r|$. An appropriately chosen $\rho(\cdot)$ could bound the influence of large residuals, so that some M-estimators turns out quite robust. Rousseeuw and Leroy (1984) developed the least median squares estimator (LMS) given by

$$\underset{\beta}{\text{Minimize}} \underset{i}{\text{med}} \, (y_i - x_i\beta)^2.$$

It turns out that the least median squares estimator is very robust with respect to outliers in y-direction as well as outliers in x-direction. Beran (1977) introduced a robust estimation method, called minimum Hellinger distance estimation, which defines an estimator as a statistic minimizing the Hellinger distance between a parametric model density and a non-parametric density estimator. Later, Lindsay (1994) and Basu and Lindsay (1994) developed minimum disparity estimation, a large subclass of density based minimum distance estimation, of which minimum Hellinger distance estimation is a part. Basu and Lindsay (1994) showed that the minimum disparity estimator (MDE) has attractive efficiency and robustness properties among other robust estimators. In this paper the technique of minimum disparity estimation is applied to the case of linear regression models. While previous studies investigated the robustness of estimators from the mathematical point of view, in this paper the robustness is illustrated through the various graphs. As Beran (1977) and Basu and Lindsay(1994) have shown the efficiency, at the model, of the minimum disparity estimators for location-scale models, the minimum disparity estimators of the regression parameters may also be efficient at the model while some M-estimators and the least median squares estimators are inefficient at the model (Huber (1981), Rousseeuw and Leroy (1987)).

For simplicity of illustration we first consider the simple linear regression

model:

$$y_m = \beta x_m + \epsilon_m,\ 1 \leq m \leq n,$$

where ϵ_m are independently distributed with mean 0 and variance σ^2. In practice, since β is unknown quantity, we assume that the quantities $z_m = (y_m - bx_m)/s$ are defined for any two scalars b and s and that the z_m represent i.i.d. observations with common density, say $g(\cdot)$, when b and s are equal to the true values of β and σ. However for simplicity let us define z_m with the parameters in this paper. One can construct a density estimator for the density of the z_m's :

$$f_n^*(t;\beta) = \int k(t;z,h)d\hat{F}(z) = \frac{1}{n}\sum_{m=1}^{n} k(t;z_m,h), \tag{1.1}$$

where $\hat{F}(z)$ is the empirical distribution function of the z and $k(t;z,h)$ is a smooth family of kernel functions such as normal densities with mean z and standard deviation h. As a counterpart of $f_n^*(t;\beta)$ we construct a kernel smoothed model density, $g^*(t)$, by applying the same smoothing to the model density :

$$g^*(t) = \int k(t;z,h)dG(z), \tag{1.2}$$

where $G(\cdot)$ is the cdf corresponding to $g(\cdot)$. The smoothed model density $g^*(t)$ defined in (1.2) equals $E[f_n^*(t;\beta)]$ when the true parameters are used in defining the z_m's. $f_n^*(t;\beta)$ in (1.1) is not the only type of kernel smoothing that we consider; we will look at the case when the kernel density estimator has the form

$$f_n^*(t;\beta) = \frac{\sum_{m=1}^{n} x_m k(t;z_m,h)}{\sum_{m=1}^{n} x_m} \tag{1.3}$$

and the corresponding smoothed model $g^*(t)$, which equals $E[f_n^*(t;\beta)]$ under the true parameters, is the same as that in (1.2). The $f_n^*(t;\beta)$ in (1.3) is a weighted average of kernels,

$$f_n^*(t;\beta) = \sum_{m}^{n} w_m k(t;z_m,h),$$

where $w_m = x_m / \sum_{m=1}^{n} x_m, \sum_{m=1}^{n} w_m = 1, 1 \leq m \leq n$. These weights should be appropriately chosen on purpose to meet a goal of the individual research as we prefer the weighted least squares method for a particular situation to the usual least squares method. The minimum disparity estimator of β is a statistical quantity which minimizes the disparity $\rho(f_n^*(t;\beta), g^*(t))$, a special type of density based distances between $f_n^*(t;\beta)$ and $g^*(t)$, as a function of β. In general we will treat σ as a nuisance parameter in the problem of estimating β. Detailed discussion of disparities are presented in Lindsay (1994) and Basu and Lindsay (1994). In this paper we will consider a particular family of disparities, called the blended weight Hellinger distances, which can be expressed as :

$$BWHD_\alpha(f_n^*(t;\beta), g^*(t)) = \frac{1}{2}\int \frac{(f_n^*(t;\beta) - g^*(t))^2}{\left(\alpha\sqrt{f_n^*(t;\beta)} + \bar{\alpha}\sqrt{g^*(t)}\right)^2} dt, \quad \alpha \in (0,1],$$

where $\bar{\alpha} = 1 - \alpha$. For $\alpha = 0.5$, the above family generates the twice multiplied Hellinger distance

$$HD(f_n^*(t;\beta), g^*(t)) = \int \left(\sqrt{f_n^*(t;\beta)} - \sqrt{g^*(t)}\right)^2 dt.$$

Under differentiability of $f_n^*(t;\beta)$ with respect to the parameter set of interest, and letting ∇ represent the corresponding gradient, the minimum disparity estimating equations usually have the form

$$\nabla_\beta \rho(f_n^*(t;\beta), g^*(t)) = \int A(\delta_n^*(t)) \nabla_\beta f_n^*(t;\beta) dt = 0,$$

where the function $A(\cdot)$ is called the residual adjustment function (RAF) of the corresponding disparity. Here

$$\delta_n^*(t) = \frac{f_n^*(t;\beta)}{g^*(t)} - 1$$

is the Pearson residual at the point t (Lindsay (1994); Basu and Lindsay (1994)). Since the residuals are functions of the parameter, in the regression case the Pearson residuals involve the parameter through kernel density estimator $f_n^*(t;\beta)$, that is, the parameter is involved in the data part, unlike the

usual parametric estimation under i.i.d.observations, where the parameter is involved in the model part (Basu and Lindsay (1994)). Hence, for the $BWHD$ family of disparities the residual adjustment functions, which are different from those of Basu and Lindsay (1994), have the form

$$A_\alpha(\delta_n^*(t)) = \delta_n^*(t)\left(\alpha\sqrt{\delta_n^*(t)+1}+\bar{\alpha}\right)^{-2} - \frac{\alpha}{2}\frac{\delta_n^{*2}(t)}{\sqrt{\delta_n^*(t)+1}}\left(\alpha\sqrt{\delta_n^*(t)+1}+\bar{\alpha}\right)^{-3},$$

where $\alpha \in (0,1]$.

See Lindsay (1994) for a general discussion of the residual adjustment function and their role in determining the theoretical properties of the estimators.

The rest of the paper is organized as follows: Section 2 provides a discussion of the residual adjustment functions in the case of the $BWHD$ family, and interprets their role in the robustness of the estimators. A numerical example is presented in Section 3.

2. ROBUSTNESS OF THE MINIMUM DISPARITY ESTIMATES

The residual adjustment function $A_\alpha(\delta_n^*(t))$ of the $BWHD$ family is plotted versus $\delta_n^*(t)$ in Figure 1. As the figure shows, the RAF's share the robustness features of RAF in the case of parametric estimation with i.i.d. data from a continuous parametric model (Basu and Lindsay (1994)). The RAF can strongly downweight the estimating component $\nabla_\beta f_n^*(t;\beta)$ with large positive residuals.

In order to understand the role of RAF, it may be helpful to rewrite the estimating equations as below :

$$\int g^*(t)\left[\frac{A(\delta_n^*(t))+1}{\delta_n^*(t)+1}\right]\frac{\nabla_\beta f_n^*(t;\beta)}{f_n^*(t;\beta)}dt = 0. \qquad (2.1)$$

Let us denote $(A(\delta_n^*(t))+1)/(\delta_n^*(t)+1)$ by $W(\delta_n^*(t))$. The $W(\delta_n^*(t))$ in Figure 2 is a weight function depending on the $\delta_n^*(t)$, and $\nabla_\beta f_n^*(t;\beta)/f_n^*(t;\beta)$ acts like

a score function of β. If the $W(\delta_n^*(t))$ is equal to 1, $A(\delta_n^*(t)) = \delta_n^*(t)$, implying that there is no outlier, the estimating equation turns out to be the expectation of $\nabla_\beta f_n^*(t;\beta)/f_n^*(t;\beta)$ with respect to t, that is, averaging $\nabla_\beta f_n^*(t;\beta)/f_n^*(t;\beta)$ over t. If the kernels are standard normal densities, the estimating equation becomes a score function of β which produces the maximum likelihood estimator of β. This can be easily shown by direct calculation of (2.1). $W(\delta_n^*(t))$ will decrease as $\delta_n^*(t)$ increases, so that a smaller weight will be imposed on $\nabla_\beta f_n^*(t;\beta)/f_n^*(t;\beta)$.

The above interpretation of a role of RAF in the robustness of MDE is quite abstract and requires us to keep visualizing shapes of the kernel and the model density. Another way of understanding a role of RAF is to look at RAF as a function of the z_m's, where the z_m's are defined at the true parameters β and σ. We focus on the behavior of one particular residual, and pretend that all of the observations except this particular one are fixed. Then we are able to draw graphs like those in Figure 3. We assume that a residual z follows a normal distribution with mean 0 and variance 1, and that a kernel has a normal distribution with mean z and variance $h = 1/2$. At a particular t, say $t = z$, $A(\cdot)$ is a function of z through

$$\delta^*(z) \propto e^{\frac{1}{2}z^2}.$$

Figure 3 shows that $A(z)$ acts like the $\psi(z)$ function of M-estimation. When a residual is away from 0, $A(z)$ begins to bound the influence of a large residual by some constant, determined by α.

In order to understand the mechanism underlying the MDE in regression, so far, we have restricted our attention to the simple linear regression model. Now, we will turn our attention toward the multiple linear regression model.

Let x_{im} denote m-th observation of the independent variable x_i and let z_m denote m-th residual, $z_m = \beta_0 + \beta_1 x_{1m} + \cdots + \beta_p x_{pm}$, $i = 0, \ldots, p$ and $m = 1, \ldots, n$. Define that for all $i = 0, 1, \ldots, p$

$$f_i^*(t;\boldsymbol{\beta}) = \sum_{m=1}^{n} w_{im} k(t; z_m, h), \quad \sum_{m=1}^{n} w_{im} = 1,$$

$$g_i^*(t) = E[f_i^*(t;\boldsymbol{\beta})],$$
$$\delta_i^*(t) = \frac{f_i^*(t;\boldsymbol{\beta}) - g_i^*(t)}{g_i^*(t)},$$

where $\boldsymbol{\beta}$ denote a vector of the parameters, $\beta_0, \ldots, \beta_p$, and w_{im} are weights corresponding to the independent variable x_i. In fact, $g_i^*(t)$ are identical for all $i = 0, 1, \ldots, p$, and let's call them just $g^*(t)$. Also, let ∇_i, ∇_{ij} and ∇_{ijk} represent the first partial derivative with respect to β_i, the second partial derivative with respect to β_i and β_j and the third partial derivative with respect to β_i, β_j and β_k.

Definition 2.1. Suppose

$$y_m = \boldsymbol{x_m^T}\boldsymbol{\beta} + \epsilon_m, 1 \leq m \leq n,$$

where $\boldsymbol{x_m^T}$, a $(p+1)\times 1$ vector, is the m-th row of the design matrix X, $\boldsymbol{\beta}$ is a $(p+1)\times 1$ vector and ϵ_m are i.i.d. errors with mean 0 and variance σ^2 (known). Then for some appropriate disparity ρ, the minimum disparity estimator $\boldsymbol{\beta_n}$ is defined as the statistic minimizing

$$S(\boldsymbol{\beta}) = \sum_{i=0}^{p} \rho_i(\boldsymbol{\beta}) = \sum_{i=0}^{p} \rho(f_i^*(t;\boldsymbol{\beta}), g_i^*(t))$$

as a function of $\boldsymbol{\beta}$. In particular, for the blended weight Hellinger distance ($BWHD$) family, the minimum disparity estimator of $\boldsymbol{\beta}$ will minimize

$$S(\boldsymbol{\beta}) = \sum_{i=0}^{p} BWHD(f_i^*(t;\boldsymbol{\beta}), g^*(t)) = \sum_{i=0}^{p} \int \left(\frac{f_i^*(t;\boldsymbol{\beta}) - g_i^*(t)}{\alpha\sqrt{f_i^*(t;\boldsymbol{\beta})} + \bar{\alpha}\sqrt{g_i^*(t)}} \right)^2 dt,$$

where $\alpha \in [0, 1]$.

As the definition indicates, the MDE's in the multiple regression are the statistical quantities which minimize the sum of disparities $\rho_i(\cdot)$, for $i = 0, \ldots, p$, corresponding to each regression coefficient, $\beta_0, \ldots, \beta_p$. Hence, the

robustness properties of MDE, that we have discussed in the case of simple linear regression, inherit in the case of multiple linear regression models.

3. EXAMPLE

The Hertzsprung-Russell diagram of the star cluster CTG OB1 (Rousseeuw & Leroy (1987)), which contains 47 stars in the direction of Cygnus, is used in this example. The independent variable is the logarithm of the effective temperature at the surface of the star (T_e), and the dependent variable is the logarithm of its light intensity (L/L_0).

The scatter plot of the data reveals that observations 11, 20, 30 and 34 are outlying from other 43 observations. These four stars are called *giants* in astronomy (Figure 4). Based on this data set, the least-squares method, the least median squares method and the minimum disparity estimation method are used to fit the model and we have the following results:

$$\begin{aligned}
\text{Least-squares} &: Y = 6.793 - 0.4133X, \\
\text{Least median squares} &: Y = -12.76 + 4X \\
\text{Minimum disparity estimation} &: \alpha = 0.5, \quad Y = -14.270 + 4.3603X \\
&\quad\ \alpha = 0.7, \quad Y = -13.385 + 4.1588X \\
&\quad\ \alpha = 0.9, \quad Y = -12.702 + 4.0040X
\end{aligned}$$

The least-squares (LS) fit is influenced by the outliers in great deal, while the least median squares (LMS) fit and the minimum disparity estimation (MDE) fit pass through the main body of observations. LS fit is certainly not robust in the sense that the residuals of the four stars are not significantly large enough to distinguish the four stars from the other stars. As we expected, MDE's are as robust as the estimators by the least median squares method which is so far known as the most robust method in this field.

Robust regression technique try to give mathematical criteria for researchers to ignore or to put less attention on the outliers. For example, it is quite clear from Figure 4 that observation number 11, 20, 30 and 34 are outliers. However, it would be better to say that they are outliers because they are far from the lines fitted by robust regression techniques than just say that they are outliers by sight.

Table 1. Data for the Hertzsprung-Russell Diagram of the Star Cluster CTG OG1. *Source*: Rousseeuw and Leroy (1984).

Index of Star (i)	$\log T_e$ (x_i)	$\log[L/L_0]$ (y_i)	Index of Star (i)	$\log T_e$ (x_i)	$\log[L/L_0]$ (y_i)
1	4.37	5.23	25	4.38	5.02
2	4.56	5.74	26	4.42	4.66
3	4.26	4.93	27	4.29	4.66
4	4.56	5.74	28	4.38	4.90
5	4.30	5.19	29	4.22	4.39
6	4.46	5.46	30	3.48	6.05
7	3.84	4.65	31	4.38	4.42
8	4.57	5.27	32	4.56	5.10
9	4.26	5.57	33	4.45	5.22
10	4.37	5.12	34	3.49	6.29
11	3.49	5.73	35	4.23	4.34
12	4.43	5.45	36	4.62	5.62
13	4.48	5.42	37	4.53	5.10
14	4.01	4.05	38	4.45	5.22
15	4.29	4.26	39	4.53	5.18
16	4.42	4.58	40	4.43	5.57
17	4.23	3.94	41	4.38	4.62
18	4.42	4.18	42	4.45	5.06
19	4.23	4.18	43	4.50	5.34
20	3.49	5.89	44	4.45	5.06
21	4.29	4.38	45	4.55	5.34
22	4.29	4.22	46	4.45	5.34
23	4.42	4.42	47	4.42	4.50
24	4.49	4.85			

Figure 1. Plots of $A(\delta^*(t))$ for MDE.

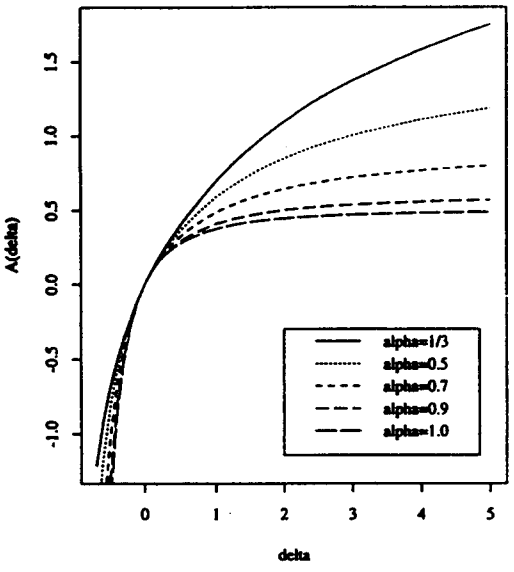


Figure 2. Plots of $W(\delta^*(t))$ for MDE.

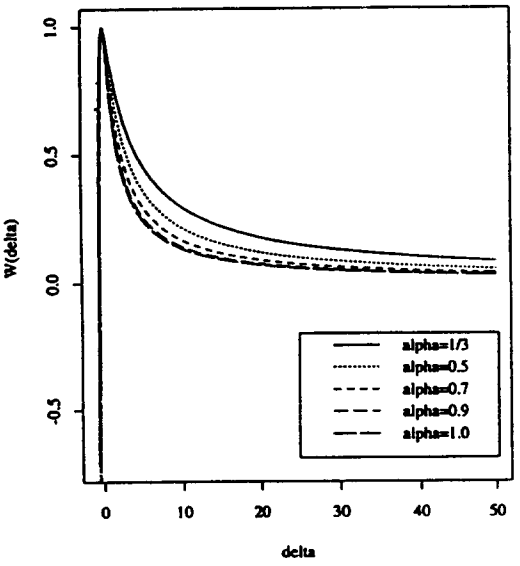


Figure 3. Plots of $A(z)$ for MDE; $z \sim N(0,1)$, $k(t;z,h)$ is $N(z,1)$ and $t = z$.

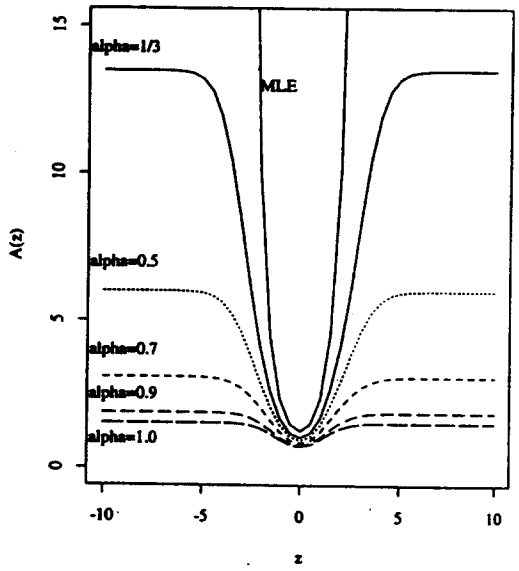
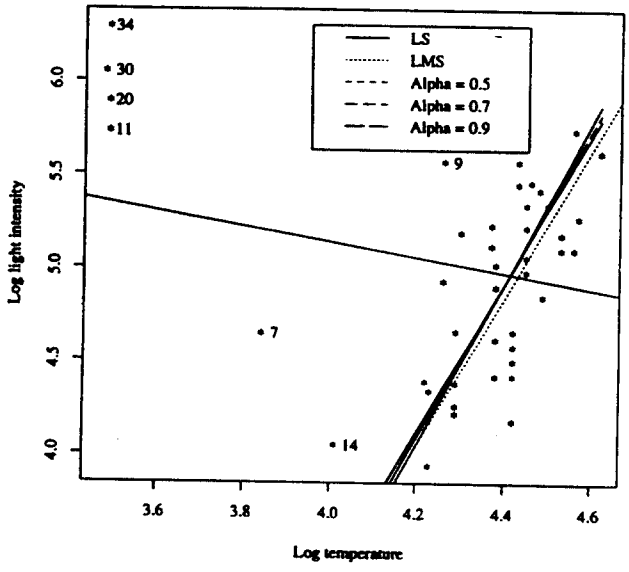


Figure 4. Hertzsprung-Russell Diagram of the Star Cluster CTG OB1; LS-fit, LMS-fit, $MDE(0.5)$-fit, $MDE(0.7)$-fit and $MDE(0.9)$-fit.



REFERENCES

(1) Basu, A. and Lindsay, B.G. (1994). Minimum Disparity Estimation for Continuous Models: Efficiency, Distribution and Robustness. *The Annals of Institute of Statistical Mathematics*, **46**, 683–705.

(2) Beran, R.J. (1977). Minimum Hellinger Distance Estimates for Parametric Models. *The Annals of Statistics*, **5**, 445–463.

(3) Huber, P.J. (1973). Robust Regression: Asymptotic, Conjectures, and Monte Calro. *The Annals of Statistics*, **1**, 799–821.

(4) Huber, P.J. (1981). *Robust Statistics*. John Wiley and Sons, New York.

(5) Lindsay, B.G. (1994). Efficiency Versus Robustness: The Case for Minimum Hellinger Distance and Related Methods. *The Annals of Statistics*, **22**, 1081–1114.

(6) Rousseeuw, P.J. and Leroy, A. (1987). *Robust Regression and Outlier Detection*. John Wiley and Sons, New York.