International Journal of Pure and Applied Mathematics
Volume 37 No. 4 2007, 541-553

AN EFFICIENT ESTIMATOR FOR REGRESSION MEDIAN

A.M. Salem

Department of Mathematics
Faculty of Science
Tanta University
El-Giesh St., Tanta, Gharbia, EGYPT

Abstract: Regression model building is often used in studying the relationship between a dependent variables and a set of independent variables. The effects of outliers on linear regression are examined. To improve this problem, we introduce a new statistical techniques mean median regression have been developed that are not so easily affected by outliers. This is the robust (or resistant) method. The performance of mean median regression estimates is compared with absolute deviation estimates in a simulation study, and we discuss the test for correct specification of the disturbances mainly because of their ability to detect also irregularities in the regressor specification.

AMS Subject Classification: 62J02
Key Words: efficient estimator, regression median, goal programming, simulation study

1. Introduction

Regression analysis is an important statistical tool that is routinely applied in most sciences. There is presently a widespread awareness of the dangers posed by the occurrence of outliers, which may be a result of keypunch errors, misplaced decimal points, recording or transmission errors, exceptional phenomena such as earthquakes or strikes, or members of a different population slipping into the sample. Outliers occur very frequently in real data, and they often go unnoticed because nowadays much data is processed by computers, without careful inspection or screening. Not only the response variable can be outlying,

Received: March 31, 2007

but also the explanatory part, leading to so-called leverage points. Both types at outliers may totally spoil an ordinary least squares (LS) analysis. Often, such influential points remain hidden to the user, because they do not always show up in the usual LS residual plots.

The thinking that robust regression techniques hide the outliers, but the opposite is true because the outliers are far away from the robust fit and hence can be detected by their large residuals from it, whereas the standardized residuals from ordinary LS may not expose outliers at all.

An alternative approach to dealing with outliers in regression analysis is to construct outlier diagnostics. These are quantities computed from the data with the purpose of pinpointing influential observations, which can then be studied and corrected or deleted, followed by an LS analysis on the remaining cases. Diagnostics and robust regression have the same goal, but they proceed in the opposite order: In a diagnostic setting, one first wants to identify the outliers and then fit the good data in the classical way, whereas the robust approach first fits a regression that dose justice to the majority of the data and then discovers the outliers as those points having large residuals from the robust equation. In some applications yield exactly the same result, and then the difference is mostly subjective.

There are many aspects of research devoted to this LAD regression. LAD estimation can be transformed into a goal programming (GP) model so it can be solved by a linear programming algorithm as has been shown in Charnes et al [7].

The regression median (RM) is a regression method designed to produce a regression hyperplane which depends on separate of all observations into two sets with equal numbers. Therefore we can find the RM, this type of estimation reduces the influence of outliers, which proves the robustness of LAD estimation. Bassett et al [5] discovered that LAD estimation differs from regression median (RM) when the data set has a small sample size and for a skewed error distribution. The superiority for RM over LAD for any data set, was discovered by Sueyoshi et al [14], as it has been shown in Salem [13].

In this article we tray to develop a method for robust procedure that can always produce RM estimators for any data set. This robust method starts with dual variables obtained by LAD estimation.

The Durbin-Watson test detect incorrect functional form. The significant test statistic, when checking for normality can also be due to outliers in the data. This is because the null hypothesis for such tests not only maintains that the disturbances are well behaved, but also that they are independent of the regressors, and that the regressors are correctly specified. Failure of any of these conditions can lead to a significant test statistic. Some authers welcome this as an additional merit of their tests, but this lack of robustness is a mixed blessing at best.

2. LAD Estimation

As we known that the least square (LS) procedure is an optimum and also yields the maximum likelihood estimates of the model parameters if the errors are independently and identically distributed (iid) as a normal distribution with mean zero and variance σ^2, i.e. $N(0, \sigma^2)$. Transformations of the data have been used to approximate normality, which is a necessary assumption for testing significance. However, for skewed error distributions, an alternative is that of least absolute deviation (LAD) estimation, which is known to be resistant to outlier effects which can be shown Barrodale [3], Rice et al [12].

Only a brief description of the LAD criterion for the linear regression model may be expressed by

$$\min \sum \left(\varepsilon_i^+ + \varepsilon_i^-\right) .$$

Subject to

$$\beta_0 + \sum_{j=1}^{m} x_{ij}\beta_j + \varepsilon_i^+ - \varepsilon_i^- = y_i, \quad i = 1, 2, \dots, n\,, \qquad \varepsilon_i^+,\ \varepsilon_i^- \geq 0\,, \tag{1}$$

where β_0 and β_j, $j = 1, 2, \dots, m$ are unknown parameters to be estimated, y_i is the dependent variable, x_{ij} is the i-th observation on j-th independent variable ε_i^+ and ε_i^- are positive and negative deviations, respectively as has been shown Charnes et al [9]. The objective of (1) is usually expressed by $\sum_{i=1}^{n} (\nu_i \varepsilon_i^+ + \mu_i \varepsilon_i^-)$ in an ordinary Goal programming (GP) model, where ν_i and μ_i are nonnegative constants representing relative weights to be assigned to positive and negative deviations for each relevant goal y_i. The relative weights are represented by $\nu_i = \mu_i = 1$ in the objective, so (1) may be considered as a GP model with single priority. The dual form of (1) becomes

$$\max \sum_{i=1}^{n} \omega_i y_i\,.$$

Subject to

$$\begin{aligned}&\sum_{i=1}^{n}\omega_i = 0\,, \qquad \sum_{i=1}^{n}\omega_i x_{ij} = 0, \quad j = 1,2,\dots,m\,,\\ &-1 \le \omega_i \le 1, \qquad i = 1,2,\dots,n\,,\end{aligned} \tag{2}$$

where ω_i indicates the dual variable for the i-th observation Toro et al [14].

An interesting transformation is obtained by taking $\omega_i = nP_i - 1$. Then the dual problem becomes

$$\max \sum_{i=1}^{n}(nP_i - 1)y_i\,.$$

Subject to

$$\begin{aligned}&\sum_{i=1}^{n}(nP_i - 1) = 0\,, \qquad \sum_{i=1}^{n}(nP_i - 1)x_{ij} = 0, \qquad j = 1,2,\dots,m\,,\\ &-1 \le nP_i - 1 \le 1, \qquad i = 1,2,\dots,n\,,\end{aligned} \tag{3}$$

which is equivalent to

$$\max \sum_{i=1}^{n} P_i y_i\,.$$

Subject to

$$\begin{aligned}&\sum_{i=1}^{n} P_i = 1\,, \quad \sum_{i=1}^{n} P_i x_{ij} = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \ j = 1,2,\dots,m\,,\\ &0 \le P_i \le \frac{2}{n}, \qquad i = 1,2,\dots,n\,.\end{aligned} \tag{4}$$

As we see from (4) the dual variable P_i, $i = 1,2,\dots,n$, can be considered as a probability because it is between 0 and $2/n$ and the sum for P_i equals to 1. Then by using P_i. We can find the probability location for each observations and then divided the data set into to equal numbers of two data sets (S_A, S_B and S_M) which correspond to $P_i = \frac{2}{n}$, $P_i = 0$ and $0 < P_i < \frac{2}{n}$, respectively. This property follows from the theory of complementary slackness in linear programming as shown in Appa et al [1].

3. Classification of Sample Observations

Now we can divide a data set into two distinct sets which equals numbers of data, by using the probability location P_i, how to determine the regression median (RM) based on the classified data sets, and how to deal with two types of data sets with even and odd observations.

To classificate a data set into two equals sets of observations S_A and S_B, by using the dual variables obtained by LAD estimation the classification can get as follows: $S_A = \{i : P_i > P_M\}$, $S_B = \{i" P_i < P_M\}$ and $S_M = \{i : P_i = P_M\}$ $i = 1, 2, \ldots, n$ where P_M is the middle of the dual variables of all set of observations. S_M contains a single observation whose dual variable is P_M, while S_A and S_B contains two set of equal numbers of observations whose dual variables are larger and smaller than P_M, respectively. The single observation in S_M is defined as the median point, a resulting RM applied to odd observations is required to pass through the median point and the mid point between two means of S_A and S_B, and mid point between the closest two observations in S_A and S_B, respectively.

After the data set classified, the one point in the set S_M is taken as a median point $M_1(x_1, y_1)$, the mean for the observations in the set S_A is $M_A(\bar{x}_A, \overline{y}_A)$, and the mean for the observations in the set S_B is $M_B(\bar{x}_B, \overline{y}_B)$, now we can find the mid-point between the two means for S_A and S_B as: $M_2(x_2, y_2)$

Then the RM must be pass through the two points $M_1(x_1, y_1)$ and $M_2(x_2, y_2)$, and then RM line will be in the form

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{5}$$

where

$$\hat{\beta}_0 = \frac{y_1(x_1 - x_2) - x_1(y_1 - y_2)}{(x_1 - x_2)}$$

is the estimate of β_0, and

$$\hat{\beta}_1 = \frac{y_1 - y_2}{x_1 - x_2}$$

is the estimate of β_1.

The last two relations, are used to force the RM to pass through $M_1(x_1, y_1)$ and $M_2(x_2, y_2)$, and then the RM will be less sensitive with the extreme points.

4. Computational Results

Now we can applied RM into two sets of observations the first one have odd numbers of observations and the second with even numbers of data and making

Observations			Optimal	Method		
	x_i	y_i	P_i	LS $\widehat{y}_i$	LAD $\widehat{y}_i$	RM $\widehat{y}_i$
(1)	9	12	$P_1 = 0.00$	15.537	17.667	16.92
(2)	11	13	$P_2 = 0.00$	16.813	16.333	17.74
(3)	13	20	$P_3 = 0.18$	18.088	19.000	18.56
(4)	12	23	$P_4 = 0.18$	17.540	18.667	18.15
(5)	16	26	$P_5 = 0.18$	20.001	20.000	19.79
(6)	17	18	$P_6 = 0.00$	20.639	20.333	20.20
(7)	19	21	$P_7 = 0.11$	21.915	21.000	21.02
(8)	15	15	$P_8 = 0.00$	19.364	19.667	19.38
(9)	10	18	$P_9 = 0.16$	16.175	18.000	17.33
(10)	21	27	$P_{10} = 0.18$	23.190	21.667	21.84
(11)	22	20	$P_{11} = 0.00$	23.828	22.000	22.25
$\sum$ |residual|				46.116	36.667	36.84
$\sum$ (residual)^2				156.826	175.995	166.996
# of above				5	4	5
# of on				0	2	1
# of below				6	5	5

Table 1:

a comparison between RM and the two estimators, LAD and LS.

Example 1. The first set of observations with odd numbers contains 11 data points in Table 1. We will use three estimators: LS, LAD and RM, and to find the RM linear model, we confirms the theory of complementary slackness so that $\varepsilon_i > 0$, $\varepsilon_i = 0$ and $\varepsilon_i < 0$ corresponding to $P_i = \frac{2}{n}$, $0 < P_i < \frac{2}{n}$ and $P_i = 0$ respectively. The examination of dual variables separates the 11 observations into three sets as: $S_A = \{3, 4, 5, 9, 10\}$, $S_M = \{7\}$ and $S_B = \{1, 2, 6, 8, 11\}$. Here S_M contains middle of the dual variables of all observations is as median $M_1(19, 21)$, the mean of the set S_A is $M_A(14.4, 22.8)$, and the mean of the set S_B is $M_B(14.8, 15.6)$, then the midpoint between M_A and M_B is $M_2\,(16.4, 19.2)$. And then the RM line will pass through the two points M_1, M_2 now the RM line is $\hat{y}_i = 13.2 + 0.41 * x_i$

In Figure 1 there are three regression lines obtained from the three estimators LS, LAD and RM. Here the LS regression line is $\hat{y}_i = 9.8 + 0.64 * x_i$, and the LAD regression line is $\hat{y}_i = 14.7 + 0.33 * x_i$, as we see from Figure 1 the RM

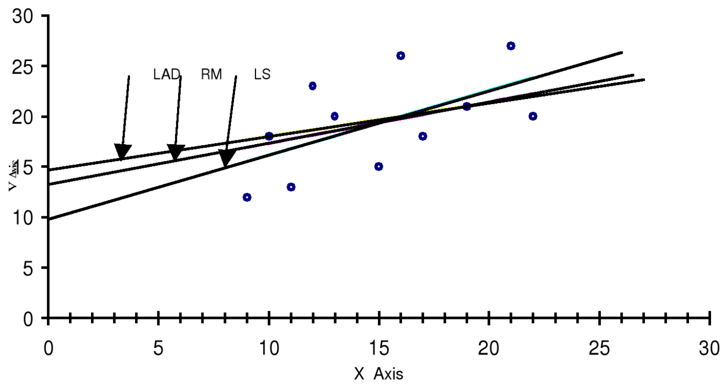

Figure 1: Estimated three regression lines for odd observations

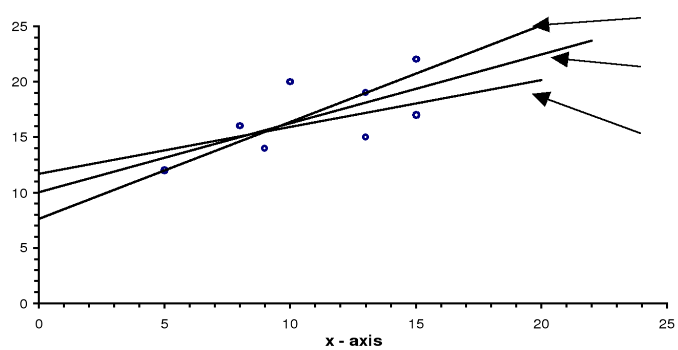

Figure 2: Estimated three regression lines for even observations

passes through the median point, also the LAD regression line passes through the median and the other point, and the LS regression not pass through any point.

Example 2. The second set of observations with even numbers contains 8 data points in Table 2, to find the RM line, where the examination of dual variables separates the 8 observations into two sets as: $S_A = \{2, 3, 5, 6\}$ and $S_B = \{1, 4, 7, 8\}$. Here the midpoint between the two closest point in S_A, S_B is $M_1(9, 15.5)$, and the midpoint between the two mean of S_A, S_B is $M_2(12.25, 16.87)$. Then the RM will pass through the two points M_1 and M_2 the RM line is $\hat{y}_i = 11.692 + 0.423 * x_i$

Figure 2 as we see the three estimators LS, LAD and RM, where the LS regression line is $\hat{y}_i$=10.031+0.622*x_i, and the LAD regression line is $\hat{y}_i$ =

Observations			Optimal	Estimators		
	x_i	y_i	P_i	LS $\widehat{y}_i$	LAD $\widehat{y}_i$	RM $\widehat{y}_i$
(1)	5	12	$P_1 = 0.00$	13.142	12.000	13,808
(2)	8	16	$P_2 = 0.25$	15.008	14.625	15.077
(3)	10	20	$P_3 = 0.25$	16.253	16.375	15.923
(4)	15	17	$P_4 = 0.00$	19.364	20.750	18.039
(5)	13	19	$P_5 = 0.19$	18.120	19.000	17.193
(6)	15	22	$P_6 = 0.25$	19.364	20.750	18.039
(7)	9	14	$P_7 = 0.00$	15.631	15.500	15.500
(8)	13	15	$P_8 = 0.00$	18.120	19.000	17.193
$\sum$ |residual|				16.512	15.500	16.424
$\sum$ (residual)^2				42.034	48.906	41.615
# of above				4	3	4
# of on				0	2	0
# of below				4	3	4

Table 2:

$7.625 + 0.875 * x_i$, and since we have already discussed the RM in detail.

5. Testing Disturbances

In this study we examine whether the disturbances in the regression model (1) are well behaved. As we known, this can also be viewed as a test of whether the higher moments of the dependent variable conform to the assumptions of the model.

Testing for autocorrelation might well be the most intensely researched statistical problems in all of econometrics. Despite heavy competition, the most common procedure is still the Durbin-Watson test (DW). It is based on the assumption that the ε_i in (1) follow a stationary autoregressive process, i.e. that

$$\varepsilon_i = \rho\,\varepsilon_{(i-1)} + \mu_i \quad ,i = 2, \ldots, n \tag{6}$$

where the μ_i's are $NID\,(0,\,\sigma_\mu^2\,)$. Stationarity further implies that all disturbances have common variance σ_ε^2, with

$$\sigma_\varepsilon^2 = (1 - \rho^2)^{-1} \sigma_\mu^2 \tag{7}$$

and that $|\rho| < 1$.

The null hypothesis of no serial correlation is therefore equivalent to H_o : $\rho = 0$. The statistic test in the multivariate case as the DW test is

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}, \tag{8}$$

where $e_i = y_i - \hat{y}_i$ is residuals. A major problem with the DW test used to be that the rejection region depends not only on the significance level $\alpha \circ f$ the test, but also on the regressor vector X. Durbin et al [10] gave the familiar bounds D_L and D_U for D that depends only on α, n such that (when testing against positive serial correlation) $H_o : \rho = 0$ is rejected when $D < D_L$ and accepted when $D > D_U$. No conclusion is possible from their tables when $D_L \leq D \leq D_U$.

Example. In this example we tray to examine whether the disturbances in the regression median model (1) are well behaved. According to DW test, therefore the null hypothesis of no serial correlation ($H_o : \rho = 0$), against ($H_1 : \rho > 0$).

The data set will be used for X and Y are given as

X_i:	3	8	4	2	6	7	5	8	9	6	8	7	9	5	7
Y_i:	8	11	9	6	9	8	5	4	9	8	11	11	12	9	8

Then by using the above steps to estimate the regression coefficient $\hat{\beta}$ and the residuals e to get RM, then from (5) we have

$$\hat{y}_i = 15 - x_i$$

According to equation (8) the statistic test is $D = 1.78$, the upper limit $D_U = 1.36$ and lower limit $D_L = 1.08$ (at significance level $\alpha = 0.05$).

Since $D > D_U$, then we accept $H_o : \rho = 0$, this meaning that there is no serial correlation between the residuals, and then the regression median model is the best fit for data.

6. Simulation Study

It is well known that LAD estimates outperform LS estimates to any error distributions except normal as has been shown Basset et al [4] and Klingman et al [1]. Mangement sciences: Optimization in statistics (vol. 19) Amesterdam: North Holand. Therefore, LS estimates were not compared in this simulation study. The Monte Carlo simulation study was conducted to assess the behavior of RM estimates. The comparing between LAD and RM to investigate how much the latter can improve the statistical efficiency of the former. The study conducted a $3 * 3 * 3$ factorial experiment in which each had 10 replications. Factors used in the simulation were

1. Error distribution (ε)

- Weibull: $F(x) = 1 - e^{-x^c}$, where $c = 0.7, 1.5, 3$
- Normal: $N(0, \sigma^2)$, where $\sigma = 2, 8, 15$
- Laplace:$F(x) = 1 - e^{-\lambda|x|}$, where $\lambda = 0.7, 1.2, 2$

2- sample $size(N) : 5, 10, 15$

3- Regression model $y = 7 + 3x + \varepsilon$, $x \sim U[1, 100]$.

$F(x)$ is a distribution function of a random variable $(x), N(0, \sigma^2)$ denotes a normal distribution ; λ, c are constant that determine the shape of an error distribution. $U[1, 100]$ denotes a uniform distribution on interval $[1, 100]$.

A weibull distribution was used to produce a skewed error. Laplace distribution were employed to generate a long tailed error with outliers. For each error distribution with three different shapes and for each sample size, the estimates of β_0 and β_1 were measured by LAD and RM.

Table 3 presents the ratio of mean square errors (MSE) for β_0 and β_1(MSE(RM)/MSE(LA

where the MSE of each parameter is computed by $MSE = \frac{\sum_{k=1}^{R} (\hat{\beta}_{j\ k} - \beta_j)^2}{R}$, $j = 0, 1$. The term k indicates of replications, $\beta_{j\ k}$ is the estimate of $\beta_j (\beta_0 = 7, \beta_1 = 3)$ for the kth replicates, and R is the number of simulation replications. The study generated 10 replications (i.e., $R = 10$) for a specific combination of its error, shape, and sample size. The notation $\sum_{k=1}^{R} (\hat{\beta}_{j\ k} - \beta_j)^2$ indicates the sum of squared deviations between the jth true parameter (β_j) and its estimate ($\widehat{\beta}_j$). The MSE of the jth parameter is equal to the sum of squared deviations divided by the total number of the simulation replications. The ratio of MSE indicates the MSE of the RM estimate divided by the MSE of the LAD estimate. Since the question of a better estimator arises, this article follows the justification of ratio of MSE criterion proposed by Toro-Vizcarrondo et al [15].

Table 3 reports the ratio of MSE with a decimal number. If the ratio is

Distribution		Sample size 5 $\beta_\circ$	β_1	10 $\beta_\circ$	β_1	15 $\beta_\circ$	β_1	Numb. ≥ 1	Numb. < 1
Normal	$\sigma = 2$	0.9	1.0	1.0	1.0	0.9	0.9	9	9
	$\sigma = 8$	1.0	1.0	0.9	1.0	1.0	1.0		
	$\sigma = 15$	0.9	0.9	1.0	0.9	0.9	0.9		
Lablace	$\lambda = 0.7$	0.9	1.0	0.9	0.9	0.9	0.9	5	13
	$\lambda = 1.2$	0.9	1.0	1.7	1.1	0.9	0.9		
	$\lambda = 2.0$	0.9	1.0	0.9	0.9	0.9	0.9		
Weibull	C=1.7	1.0	1.0	0.9	0.9	0.3	0.1	10	8
	C=1.5	1.0	1.0	0.9	1.0	0.9	0.9		
	C=3.0	1.0	0.9	1.0	1.0	1.0	1.0		
Number (≥ 1)		11		9		4		24	
Number (< 1)		7		9		14			30

Table 3:

less than 1, RM estimates are more efficient than LAD estimates; if the ratio is larger than 1, the opposite is true. The last two columns of the table, when aggregated by distribution, show the frequency count of MSE ratios that were greater than or equal to 1 and less than 1 (respectively) for all simulation runs. The result clearly indicate the efficiency of RM comparing to LAD.

Note that σ, λ and C are constants that determine the shape of each error distribution.

7. Conclusions

This section will discuss advantages of RM and compare it with other regression method (least median of squares (LMS), Atkinson [2], developed a method based on LMS. The purpose of LMS is to identify outliers to be deleted from estimation using an assumption of normal error distribution and an approximate estimate of variance. Therefore, LMS is very useful for confirmatory analysis but it is not designed to produce efficient estimates for a linear regression model. Conversely, the purpose of RM is to yield efficient estimates for any data set without deleting outliers. The LMS obtains robustness by deleting outliers, while RM achieves that property by extending LAD estimation into our computational framework.

This article presented a robust regression method that can yield an RM hyperplane for any data set. The RM method uses LAD dual variables to classify sampled observations into two groups and then produce RM estimates. The simulation study and the two data set confirmed that the RM estimates less sensitive for some error distributions. I hope that this method will be one small step for robust regression using LAD/RM.

According to the Durbin-Watson test we show that there is no autocorrelation between the error for the regression median, this means that our estimator is a suitable estimator in the case of fitting data.

References

[1] G. Appa, C. Smith, On L_i and Chebychev estimation, *Mathematical Programming*, **5** (1973), 73-87.

[2] A.C. Atkinson, Making unmasked, *Biometrika*, **73**, No. 3, (1986), 533-541.

[3] I. Barrodale, L_i approximation and the analysis of data, *Appl. Statist.*, **17** (1968), 51-57.

[4] G. Basset, R. Koenker, Asymptotic theory of least absolute error regression, *Journal of the American Statistical Association*, **73** (1978), 618-622.

[5] G. Basset, R. Koenker, An empirical quantile function for linear model with IID errors, *Journal of the American Statistical Association*, **77** (1982), 407-415.

[6] A. Charnes, W.W. Cooper, Goal programming and constrained regression, *A Comment. Omega*, **3** (1975), 403-409.

[7] A. Charnes, W.W. Cooper, R.O. Ferguson, Optimal estimation of executive compensation by linear programming, *Management Sciencec*, **1** (1955), 138-151.

[8] A. Charnes, W.W. Cooper, T. Sueyoshi, Least square ridge regression and goal programming constrained regression alternatives, *European Journal of Operational Research*, **27** (1986), 147-157.

[9] A. Charnes, W.W. Cooper, T. Sueyoshi, A goal programming constrained regression review of the Bell system beakup, *Management Science*, **34** (1988), 1-26.

[10] J. Durbin, G.S. Watson, Testing for serial correlation in least squares regression II, *Biometrika*, **38** (1951), 159-178.

[11] D. Klingman, J. Mote, Generalized network approaches for solving least absolute value and Chebycheff regression problems, In: *TIMS Studies in*

the Management Sciences: Optimization in statistics (Ed-s: S.H. Zanakis, J.S. Rustagi), Volume 19, Amsterdam, North Holland (1982).

[12] J.R. Rice, J.S. White, Norms for smoothing and estimation, *SIAM Rev.*, **6** (1964), 243-256.

[13] A.M. Salem, A robust method approach for regression median, In: *19-th International Conference for Statistics, Computer Science, Scientific and Social Applications*, Cairo, Egypt (1994), 197-211.

[14] T. Sueyoshi, Y.L. Chang, Goal programming approach for regression median, *Decision Sciences*, **20** (1989), 700-714.

[15] C. Toro-Vizcarrondo, T.D. Wallace, A test of the mean square error criterion for, restrictions in linear regression, *Journal of the American Statistical Association*, **63** (1968), 558-572.