Jagannath University Journal of Science
Volume 08, Number I, December 2021, pp. 45–52
https://jnu.ac.bd/journal/portal/archives/science.jsp
ISSN 2224-1698

JnUJSci

Two-Step Automated Robust Model Building

Research Article

Smita Sarker[1] and Md. Siddiqur Rahman[2*]

[12]*Department of Statistics, Jagannath University, Dhaka-1100, Bangladesh*

Received: 21 March 2021
Accepted: 13 November 2021

Abstract: For high-dimensional data sets two-step model building algorithms usually select more important active covariates with fewer noise variates than one-step model building algorithms in liner models. In most of the previous studies, "learning curve", a manual procedure, has been used to determine the size of the reduced set in two-step model building procedure. In this paper, we propose a computer based automatic generation procedure instead of using this manual procedure to determine the size of the "reduced set" in two-step model building procedure. That is, the "reduced set" will be created automatically by computer based automatic generation procedure whichever occurs first from either $R^2 = 0.9$ or number of target variables = 8, and final model is obtained from the reduced set. We also made a comparison between one-step model building procedure and two-step automated robust model building procedure for clean data as well as contaminated data through a simulation study and real data applications. The performance of two-step automated robust model building procedure is much better than the robust one-step model building procedure in contaminated data as well as clean data.

Keywords: *Robust Backward Elimination • MM-estimator • High Breakdown Point • Cross-Validation • Final Prediction Model*

1. Introduction

Linear model is the simplest and widely used model in which the response is estimated using a linear prediction of the entries in the covariates. Two different strategies are considered for linear model selection: (i) One-step model building and (ii) Two-step model building. One-step model building procedure's aim is to build up a final model in one step by using step-by-step algorithm such as Forward Selection (FS) (Khan *et al*. 2007a) or Stagewise (SW) (Khan *et al*. 2007a) or Backward Elimination (BE) (Rahman and Khan 2014) or Least A ngle Regression (LARS) (Efron *et al*. 2004), etc. Though one-step model building algorithms construct better models for high-dimensional data sets, but these algorithms don't guarantee to take all the important covariates in the models. They may also select some noise covariates in the models. Using two-step model building we can almost overcome this problem. Two-step model building procedure consists of two steps. The first step - which we call sequencing - the input variables are sequenced to form a list such that the important covariates are likely to appear at the beginning. The first m covariates of the list, determined by some criteria, form a reduced set from which the final prediction model will be obtained. Different step-by-step algorithms

* Corresponding author: Md. Siddiqur Rahman
Email: rsiddiq11@stat.jnu.ac.bd

are used to sequence all candidate covariates. The second step - which we call segmentation - carefully examines all possible subsets of the covariates in the reduced set in order to select the final prediction model.

Usually large data sets contain a fraction of outliers and other contaminations which are difficult to visualize and clean. The classical model selection algorithms are much affected by these contaminations and fail to select the "correct" linear prediction model. In this situation, in order to reduce this problem satisfactorily, robust algorithms for linear model selection can be used by replacing classical ingredients with their robust counterparts.

In this paper, a robust Backward Elimination (RBE) algorithm is used for sequencing all the covariates (Rahman and Khan, 2014). Most of the previous studies, "learning curve" was used to determine the reduced set. In order to get a reduced set (which is also called short list), we have to plot the learning curve. The size 'm' of the reduced set can be determined at the point where the learning curve does not have a considerable (increasing) slope anymore (Khan *et al*. 2007b). This process is more time consuming and more repellent. In this study, we propose a computer based automatic generation procedure that will sequence the covariates based on their importance by RBE, determine the reduced set, and at last the final prediction model using a robust cross-validation (RCV) procedure.

The rest of the paper is organized as follows. In section 2, we describe our methods and materials**.** In section 3, we present a simulation study that compares the performance of our two-step automated robust model building with robust one-step model building by their predictive powers. Section 4 contains a real data application. We make a conclusion in section 5. Finally, we describe limitations and future scope of this work in section 6.

2. Methods and Materials

In this paper, we focus on two-step model building procedure. One-step model building procedure is briefly discussed in introduction section. Two-step model building procedure consists of two consecutive steps: (i) Sequencing and (ii) Segmentation.

2.1 Sequencing

In this step, we sequence all of the input variables to form a list such that the important covariates are likely to appear at the beginning, and the first m covariates of the list form a reduced set which is considered to obtain the final prediction model. We use RBE (Rahman and Khan 2014) algorithm to sequence all of the covariates.

RBE (Rahman and Khan 2014) algorithm can be summarized as follows:

1. Let, $\boldsymbol{X}$ be the set of all covariates and $\boldsymbol{R}$ be the subset not containing j^{th} covariate. To remove the 1^{st} covariate X_{m1}, calculate the partial correlation $r_{jY.\boldsymbol{R}}$ between X_j and Y after eliminating the linear effect of covariate belonging to $\boldsymbol{R}$ on X_j. Determine, $m_1 = argmin|r_{jY.\boldsymbol{R}}|$.
2. Let, $\boldsymbol{B}$ be a subset containing $(k-1)$ covariates that has been removed from $\boldsymbol{X}$ after $(k-1)$ steps (k = 2,3,…) and $\boldsymbol{R}$ be the subset not containing j^{th} covariate and $\boldsymbol{B}$. To remove the k^{th} covariate X_{mk}, calculate the partial correlation $r_{jY.\boldsymbol{R}}$ between X_j and Y after eliminating the linear effect of X_{m1}, X_{m2},… $X_{m(k-1)}$ on X_j, and then determine $m_k = argmin|r_{jY.\boldsymbol{R}}|$.

Stopping Rule for RBE

At each RBE step, once the most insignificant covariate (among the remaining covariates) is identified, a partial *F*-test can be performed to decide whether to drop this covariate from the model (and continue this process) or to stop. The new most insignificant covariate drops from the model only if the partial *F*-value denoted by $Fpartial$, is smaller than $(0.09,1,n-k-1)$ (say), where k is the current size of the model excluding the new covariate.

Determination of Reduced Set

As the continuation of sequencing step, the first m covariates of the list form a "reduced set" which is considered to obtain the final prediction model. A computer based automatic generation procedure is proposed instead of using manual procedure (i.e. learning curve). That is, the "reduced set" will be created automatically by computer based automatic generation procedure whichever occurs first from either R^2 = 0.9 or number of target variables = 8.

2.2 Segmentation for Determining the Final Prediction Model

When a reduced set of m covariates is obtained, we can go to the second step, called segmentation, to obtain the final prediction model. One reasonable approach is to perform all possible subsets regression on this "reduced set" using appropriate selection criteria (e.g., Akaike Information Criterion (AIC) (Akaike 1973), Mallows' C_p

(Mallows 1973), Final Prediction Error (FPE) (Akaike 1969, 1970), CV(1) (Lachenbruch Mickey 1968), bootstrap estimators (Efron 1983), robust AIC (RAIC) (Ronchetti 1985), robust C_p (RC_p) (Ronchetti Staudte 1994), robust FPE (RFPE) (Yohai 1997), robust cross-validation (RCV) (Ronchetti *et al*. 1997).

The *k*-fold CV estimate of $Err(Z_n, H)$ is given as follows:

$$\hat{Err}^{CV_k} = \frac{1}{R}\sum_{i=1}^{n}\{\frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta\left(x_i, \hat{\beta}\left(Z_n{}^{-k_r(i)}\right)\right)]\} \quad (1)$$

where, for each observation *i*, $Zn^{-kr(i)}$ denotes the data set Z_n without the block containing observation *i* in the r^{th} random run k – fold CV. In this paper, we use k = 5-fold RCV procedure. RCV procedure (which is computationally suitable) is used to evaluate the predictive powers of different subsets of the reduced set of covariates by automatic generation procedure. The subset selected automatically having the largest predictive power makes the final prediction model.

3. Simulation Study

A simulation study is executed for comparing the performances of one-step and two-step automated robust model building procedures. To perform the simulation study, we considered d = 40 candidate covariates. A small number a = 6 of them are nonzero or active covariates (i.e. the covariates that are actually related to the response variable). We describe the data generating process for zero correlation and two non-zero correlation cases (i.e., moderate and high correlation cases) for clean data and contaminated data as follows:

R Programming software has been used to perform this analysis. For zero correlation case, we generated d = 40 covariates from X_j~(0,1);{j = 1,2,…,40}. We also generated the response variable Y using a=6 active covariates with coefficients (7,6), which are repeated three times and ($d - a$) covariates are considered as noise. The variance of the error term is chosen in such a way such that the signal-to-noise ratio equals 2.

For non-zero correlation case, we introduced two latent variables say, L_1 and L_2. Then, the linear statistical model of Y with co-efficient (7,6) can be expressed as:

$$Y = 7L_1 + 6L_2 + \epsilon \quad (2)$$

where, L_i~$N(0,1)$ and the error term ϵ~$N(0, 85/4)$.

The active covariates generated by the latent variables as follows:

$$X_{2(k-1)+i} = L_i + \tau\delta_{2(k-1)+i}; \{i = 1,2; k = 1,2,3\} \text{ and } X_k = u_k; \{k = 7,8,\dots,40\}$$

where, δ_i~$N(0,1)$ and τ is a fixed constant which is chosen to be 1 for moderate correlation of o.5, and 1/3 for high correlation of 0.9.

For zero correlation and two non-zero correlation cases (i.e., moderate and high correlation cases), we generated 1000 data sets each of size 200. We randomly divided each data set into a training sample of size 100 and a test sample of size 100.

Contamination of the Training Data

In order to contaminate the training data, we randomly choose a number of rows and the covariates values of these rows are replaced with 5%, 10% and 20% bad leverage points generated from (50,1) in different simulations. The corresponding response values are also replaced by large numbers generated from (500,1). The probability that any specific row of the training sample will be contaminated with 100α% of bad leverage points is $p = 1 - \exp(\ln(1-\alpha)/q)$, where q is the number of covariates whose values and their corresponding response values we want to contaminate. For all of the correlation cases, we contaminated the training data sets in different ways for measuring the adequacy of the two-step automated robust model building procedure. Different cases of contamination are given as follows:

- Case 1: All active covariates are contaminated.
- Case 2: All candidate covariates are contaminated.
- Case 3: All candidate covariates as well as first 5 noise variables are contaminated.
- Case 4: Most important active covariates as well as first 10 noise variables are contaminated.
- Case 5: All active covariates related to the most important latent variable L_1 are contaminated.
- Case 6: All active covariates related to the least important latent variable L_2 are contaminated.
- Case 7: All active covariates related to each of the two latent variables L_1 and L_2 are contaminated.

Performance of one-step and two-step automated robust model building procedures

At first, we illustrated the performance of one-step and reduced set of two-step automated robust model building procedures by comparing the average, standard deviation (SD), median absolute deviation (mad) of the four quantities (i.e., mean squared prediction error (MSPE), target variables, noise variables and robust R^2). The

standard deviation (SD) and median absolute deviation (mad) are shown in the parentheses. At first, we used training data for fitting our models. Then, we used test data for testing the significance of the fitted models. Robust models are fitted by using a regression MM-estimator (Yohai 1987) because of its high breakdown point (which is 0.5), and high efficiency at the normal distribution.

The performances of one-step and reduced set of two-step automated robust model building procedures in clean data are presented in Table 1 for all correlation cases.

From Table 1, we can see that the reduced set of two-step automated robust model building procedure produces less MSPE and a smaller number of noise variables than the robust one-step model building procedure for all types of correlation cases. At the same time the average robust R^2 values are increased. Also, both methods fit models with almost same number of target variables. For instance, the average of MSPE drops from 39.3 to 33.3, the mean number of noise variables reduces from 5.7 to 1 and the average robust R^2 value increases from 0.87 to 0.89 for zero correlation case.

In sequencing step, we sequence the input variables to form a list such that the good predictors are likely to appear at the beginning, and the first m covariates of the list form a reduced set. Segmentation step carefully examines all possible subsets of the covariates of the reduced set in order to obtain the final prediction model. So, segmentation step produces equal or less MSPE than the reduced set. Since, reduced set of two-step automated robust model building procedure produces less MSPE (i.e. better result) than the robust one-step model building procedure, so two-step automated robust model building procedure will produce better result than the robust one-step model building procedure.

Most important active covariates as well as first 10 noise variables are contaminated

In this case, the covariates which have largest weights are considered as the most important covariates. Each of the covariates (X_1, X_3, X_5) gets the largest weight 7. So, we contaminated 13 covariates (3 actives and 10 noises) as well as their corresponding response values with probabilities 0.0039, 0.0081 and 0.0170 for 5%, 10% and 20% bad leverage points, respectively.

Table 2 represents the performance of one-step and reduced set of two-step automated robust model building procedures for zero correlation case. It shows that the average, SD and mad of MSPE for the reduced set of two-step automated robust model building procedure are smaller than the robust one-step model building procedure for all types of contamination cases.

Table 1. Performance of one-step robust model building and reduced set of two-step automated robust model building in clean data for the zero correlation, moderate correlation and high correlation cases

Cases	Method	a = 6			
		MSPE	Target	Noise	Robust R^2
Zero correlation	One-step	39.3 (11.1) (10.4)	6 (0) (0)	5.7 (4.6) (4.4)	0.87 (0.05) (0.04)
	Reduced set	33.3 (9.1) (8.7)	5.9 (0.3) (0)	1 (1.0) (1.5)	0.89 (0.04) (0.03)
Moderate correlation	One-step	27.3 (7.7) (6.9)	4.6 (0.8) (1.5)	4 (2.8) (3.0)	0.73 (0.09) (0.09)
	Reduced set	26.8 (7.1) (6.7)	4.5 (0.8) (1.5)	3.4 (0.8) (1.5)	0.73 (0.09) (0.09)
High correlation	One-step	15.8 (4.4) (4.1)	3.3 (0.8) (1.5)	5.4 (3.9) (3.0)	0.84 (0.06) (0.05)
	Reduced set	14.7 (4.4) (4.6)	3 (0.8) (1.5)	3.5 (2.1) (1.5)	0.85 (0.06) (0.06)

Table 2. Performance of robust one-step model building and reduced set of two-step automated robust model building for Case 4 with zero correlation

Contaminations	Method	a = 6			
		MSPE	Target	Noise	Robust R^2
5%	One-step	38.5 (10.8) (10.3)	6 (0) (0)	5.3 (4.6) (3.0)	0.87 (0.05) (0.04)
	Reduced set	33.5 (8.6) (8.5)	5.9 (0.3) (0)	0.9 (1.0) (0)	0.89 (0.04) (0.03)
10%	One-step	37.9 (11.1) (9.7)	6.0 (0) (0)	4.7 (4.1) (3.0)	0.87 (0.05) (0.04)
	Reduced set	33.4 (9.3) (9.0)	5.9 (0.3) (0)	1.0 (1.0) (1.5)	0.89 (0.04) (0.03)
20%	One-step	38 (10.7) (9.6)	6.0 (0.1) (0)	3.6 (3.6) (3.0)	0.87 (0.05) (0.04)
	Reduced set	34.2 (9.1) (8.7)	5.9 (0.3) (0)	1.0 (1.0) (1.5)	0.89 (0.04) (0.03)

In addition, the average number, SD and mad of noise variables are also decreased for all types of contamination cases. Also, the average robust R^2 values are increased for all types of contamination cases. The average number of target variables are almost equal for all types of contamination cases. At the same time, although we increased the percentage of contamination, nevertheless the reduced set of two-step automated robust model building procedure remains more stable results than the robust one-step model building procedure.

Table 3. Performance of robust one-step model building and reduced set of two-step automated robust model building for Case 4 with moderate correlation

Contamination	Method	a = 6			
		MSPE	Target	Noise	Robust R^2
5%	One-step	27.3 (7.4) (6.9)	4.6 (0.8) (1.5)	3.9 (2.8) (3.0)	0.73 (0.09) (0.08)
	Reduced set	26.7 (6.9) (6.5)	4.5 (0.8) (1.5)	3.5 (0.8) (1.5)	0.73 (0.09) (0.08)
10%	One-step	27.3 (7.5) (7.2)	4.5 (0.8) (1.5)	3.7 (2.8) (3.0)	0.73 (0.09) (0.08)
	Reduced set	26.8 (7.1) (7.0)	4.5 (0.8) (1.5)	3.5 (0.8) (1.5)	0.73 (0.09) (0.08)
20%	One-step	27.0 (7.4) (6.9)	4.4 (0.8) (1.5)	3.2 (2.6) (3.0)	0.73 (0.09) (0.09)
	Reduced set	27.0 (7.1) (7.2)	4.5 (0.8) (1.5)	3.5 (0.8) (1.5)	0.73 (0.09) (0.08)

Table 3 shows that, for moderate correlation case, the reduced set of two-step automated robust model building procedure produces less MSPE and a smaller number of noise variables than the robust one-step model building procedure for all types of contamination cases. In addition, the average number of target variables are almost equal for all types of contamination cases. Also, the average robust R^2 values are exactly equal for all types of contamination cases.

Table 4 shows that, for highly correlation case, the reduced set of two-step automated robust model building procedure performs better than the robust one-step model building procedure for all types of contamination cases. The average MSPE's are decreased. Also, the average number of target variables remain almost same for all types of contamination cases. At the same time the average R^2 values are increased for all types of contamination cases. In addition, the average number, SD and mad of noise variables are decreased for all types of contamination cases.

Table 4. Performance of robust one-step model building and reduced set of two-step automated robust model building for Case 4 with high correlation

Contamination	Method	a=6			
		MSPE	Target	Noise	Robust R^2
5%	One-step	15.8 (4.5) (4.1)	3.2 (0.8) (1.5)	5.2 (4.1) (3.0)	0.84 (0.06) (0.05)
	Reduced set	14.6 (4.2) (4.3)	2.9 (0.8) (1.5)	3.5 (2.2) (1.5)	0.85 (0.06) (0.07)
10%	One-step	16.3 (4.7) (4.3)	3.2 (0.9) (1.5)	5.3 (4.2) (3.0)	0.84 (0.06) (0.05)
	Reduced set	15.1 (4.6) (4.5)	2.9 (0.8) (1.5)	3.7 (2.1) (1.5)	0.85 (0.06) (0.07)
20%	One-step	16.5 (5.1) (4.2)	3.1 (0.9) (1.5)	5.6 (4.7) (3.0)	0.83 (0.06) (0.06)
	Reduced set	15.2 (4.8) (4.8)	2.8 (0.8) (1.5)	3.6 (2.2) (1.5)	0.85 (0.06) (0.07)

4. Real Data Applications

In this section, a real dataset has been used to check the adequacy and stability of one-step and two-step automated robust model building procedures.

Appliances Energy Prediction Data Set

The measurements of this data set were created by Luis Candanedo, University of Mons (UMONS 2017). Then, this data set was stored up in Uci Machine Learning Repository (Dua & Graff 2017). This data set contains 29 columns of information. We excluded last two non-dimensional columns from the data set. For the same reason, the 1st and 27th columns of data set (i.e. date and dewpoint) were excluded from our study. Then, we considered the 2nd column (i.e. Appliances) as our response variable, and the remaining 24 columns as our candidate covariates. The data set contained 19735 observations and we considered 1^{st} 9860 observations as training data and rest of the observations as test data. We executed one-step and two-step automated robust model building procedures on the data set, and recorded MSPE and Root Trimmed Mean Squared Prediction Error (RTMSPE).

Table 5. Results for original dataset

Methods	Selected covariates	MSPE	RTMSPE
One-step model	1, 19, 4, 3, 12, 9, 13, 8, 7, 22, 21, 16, 11, 15, 5, 23, 18, 20, 10	406.3384	28.38992
Two-step model	1, 19, 4, 3	376.2488	27.08841

Table 5 depicts that robust one-step model building procedure selects 19 covariates in the model, whereas two-step automated robust model building procedure selects only 4 covariates. But, later procedure produces less MSPE than the former one. At the same time, it also produces less amount of RTMSPE. The performances of robust one-step and two-step automated robust model building procedures on three contamination cases are described below:

Case I: The response variable is contaminated

In this case, only a single response variable is contaminated. We contaminated the data by replacing one value of the response variable (say, 12^{th} value 580) by a large number 1000000 and got the following results:

Table 6. Results for Case I dataset

Methods	Selected covariates	MSPE	RTMSPE
One-step model	1, 19, 4, 3, 12, 9, 13, 8, 7, 22, 21, 16, 11,15, 5, 23, 18, 20, 10	401.2596	28.25246
Two-step model	1, 19, 4, 3	372.9660	26.99256

Case II: The candidate covariates are contaminated

In this case, only a single value of a candidate covariate is contaminated. We contaminated the data by replacing one value of 2nd candidate covariate (say, 5th value 19.89) by a large number 1900000 and found the following results:

Table 7: Results for Case II dataset

Methods	Selected covariates	MSPE	RTMSPE
One-step model	1, 19, 4, 3, 12, 9, 13, 8, 7, 22, 21, 16, 11, 15, 5, 23, 18, 20, 10	406.3219	28.38982
Two-step model	1, 19, 4, 3	376.2488	27.08841

Case III: Both the response and candidate covariates are contaminated

In this case, both the response and candidate covariates are contaminated. We contaminated the data by replacing one value of 2nd candidate covariate (say, 16th value 20.56667) by a large number 300000 and the corresponding response value 100 by 5000.

Table 8. Results for Case III dataset

Methods	Selected covariates	MSPE	RTMSPE
One-step model	1, 19, 4, 3, 12, 9, 13, 8, 7, 22, 21, 16, 11, 15, 5, 23, 18, 20, 10	405.9938	28.38490
Two-step model	1, 19, 4, 3	376.3432	27.08706

All the Tables of cases I, II and III show the similar results as Table 5.

5. Conclusion

In this study, we compared robust one-step and two-step automated model building procedures using simulated and real data sets. In simulated data sets, we compared the performance of robust one-step and reduced set of two-step automated robust model building procedures for the zero, moderate and high correlation cases in clean data. We see that the reduced set of two-step automated robust model building procedure performed much better than the robust one-step model building procedure. We also compared the performances of the two models by contaminating the simulated data sets in 7 different ways. As we increased the percentage of bad leverage points, the reduced set of two-step automated robust model building procedure performed much better than the robust one-step model building procedure for all types of contamination cases.

In real data application, the final prediction model selected by the two-step automated robust model building procedure contains very fewer covariates but less test errors than the robust one-step model building procedure for original data set as well as contaminated data sets.

6. Limitations and Further study

There are some limitations in this study. In several occasions, the proposed two-step automated robust model building procedure may omit some correct variables. Because of the limitations of currently used available computers, we used maximum size of the short list = 8 in our proposed computer based automatic generation procedure. If we were used size of the short list larger than 8, it would be more difficult to apply all possible subsets regression on a high-dimensional data sets by using these available computers. In future, when high speed computers will be available, our proposed computer based automatic generation procedure can be extended.

Acknowledgments:

The authors would like to express their sincere gratitude to the Ministry of Science and Technology, Bangladesh Government for providing awarded (National Science and Technology Fellowship) for this research work.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1): 243-247.

Akaike, H. (1970). Statistical prediction identification. *Annals of the institute of Statistical Mathematics*, 22(1): 203-217.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78 (382): 316-331.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., (2004). Least angle regression. *The Annals of statistics*, 32(2): 407-499.

Khan, J. A., Van Aelst, S., & Zamar, R. H. (2007a). Building a robust linear model with forward selection and stepwise procedures. *Computational Statistics & Data Analysis*, 52 (1): 239-248.

Khan, J. A., Van Aelst, S., & Zamar, R. H. (2007b). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, 102 (480): 1289-1299.

Lachenbruch, P. A. & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10 (1): 1-11.

Mallow's, C. L. (1973). Some comments on C_p. *Technometrics*, 15(4): 661-675.

Rahman, M. S. & Khan, J. A. (2014). Building a robust linear model with backward elimination procedure. *Dhaka University Journal of Science*, 62(2): 87-93.

Ronchetti, E. (1985). Robust model selection in regression. *Statistics Probability Letters*, 3(1): 21-23.

Ronchetti, E., Field, C., & Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, 92 (439): 1017-1023.

Ronchetti, E. & Staudte, R. G. (1994). A robust version of Mallows' C_p. *Journal of the American Statistical Association*, 89 (426): 550-559.

Yohai, V. J. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*, 15: 642-656.