Dhaka Univ. J. Sci. 58(2): 291-295, 2010 (July)

Robust Stepwise Algorithms for Linear Regression: A Comparative Study

Md Siddiqur Rahman[1] and Jafar A. Khan[2]

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

[2]Department of Statistics, Biostatistics and Informatics, Dhaka University, Dhaka-1000, Bangladesh

Received on 03.11.2009. Accepted for Publication on 24.02.2010.

Abstract

For building a linear prediction model, Forward selection (FS) (Weisberg 1985) and Least Angle Regression (LARS) (Efron *et al.* 2004) are two efficient stepwise procedures for sequencing the candidate predictors. Both the methods yield poor results when data contain outliers and other contaminations. Khan *et al.* (2007a) and Khan *et al.* (2007b) proposed robust versions of LARS (RLARS) and FS (RFS), respectively, which are computationally very suitable and scalable to large high-dimensional datasets. However, no comparison has been made between RFS and RLARS. In this study, we compare these two stepwise algorithms. We conduct an extensive simulation study to compare the number of correct covariates identified by these two algorithms in linear regression. We also apply these algorithms to empirical data. Based on our simulation study and real-data application, the efficiency of RFS appears to be better than that of RLARS.

Key words: Computational complexity; Forward Selection; Least angle regression; Linear regression; Robust prediction; Stepwise procedure; Winsorization.

I. Introduction

Robust model selection has not received much attention in the robustness literature. Most of the work related to robust model selection in regression has focused on the development of robust selection criteria that can be used to compare models. Seminal works that address this issue include those of Ronchetti (1985) and Ronchetti and Staudte (1994), which introduced robust versions of two selection criteria, the Akaike information criterion and Mallows' C_p. Maronna, Martin and Yohai (2006) proposed a robust final prediction error (FPE) criterion (see also the S-PLUS documentation), whereas Muller and Welsh (2005) proposed a robust selection criterion based on a stratified bootstrap procedure. Robust selection criteria for more general models have been developed by Cantoni and Ronchetti (2001) for generalized linear models and Ronchetti and Trojani (2001) for generalized method of moments. In the later context, model selection can make use of indirect inference (see Genton and Ronchetti 2003; Jiang and Turnbull 2004). Atkinson and Riani (2002) proposed an added-variable t test for variable selection in the context of regression based on the forward search procedure. Morgenthaler, Welsch, and Zenide (2003) constructed a selection technique to simultaneously identify the correct model structure as well as any unusual observations. Ronchetti, Field and Blanchard (1997) proposed robust model selection by cross-validation. A major drawback of most robust model selection methods is that they are very time-consuming, because they require the robust fitting of a large number of submodels. One exception is a model selection procedure based on the Wald test (Sommer and Huggins 1996), which requires the computation of estimates only from full models.

Classical forward selection (FS) (see, e.g., Weisberg 1985, chap. 8) and classical least angle regression (LARS) proposed by Efron *et al.* (2004) are two computationally efficient techniques, but yield poor results when the data contain outliers and other contaminations. Khan *et al.* (2007a) and Khan *et al.* (2007b) proposed robust versions of LARS (RLARS) and FS (RFS) respectively, which are computationally very suitable and scalable to large high-dimensional datasets.

Both the RFS and RLARS model selection strategies sequence the input variables to form a list such that the good predictors appear in the beginning. Though RFS and RLARS are computationally suitable stepwise procedures, a comparison of these two methods has not been done yet. In this paper, we compare the RFS procedure with the RLARS through a simulation study. We also compare these two procedures through empirical data analysis.

The rest of the article is organized as follows. In section 2, we review the classical FS and LARS. In section 3, we review RFS and RLARS. In section 4, we present the result of simulation study and compare the sequences produced by RFS and RLARS procedures. Section 5 contains a real-data application. We conclude in section 6.

II. FS and LARS Algorithms

In this section we review two important step by step algorithms: FS and LARS.

(a) Forward Selection (FS)

Let $X_1, X_2, \cdots, X_d$ be *n* dimensional vectors representing the covariates, and Y the *n*-dimensional vector representing the response. By location and scale transformations we can always assume that the variables have been standardized to have mean zero and unit length. The FS procedure selects the covariate (X_1, say) that has the largest absolute correlation $|r_{1y}|$ with Y and calculates the residual vector $Y - r_{1y}X_1$. All other covariates are then 'adjusted for X_1' and entered into competition. That is, each X_j $(j \neq 1)$ is regressed on X_1 and the corresponding residual vector $Z_{j.1}$ (which is orthogonal to X_1) is obtained. The correlations of these $Z_{j.1}$ with the residual vector $Y - r_{1y}X_1$, called partial correlations between X_j and Y adjusted for X_1, decide the next variable (X_2, say) to enter the regression model. All other covariates are then 'adjusted for X_1 and X_2' and entered into further competition and so on. We continue adding one covariate at each step, until a stopping criterion is met. We need $(d-1)$ steps to get the ordering of all d predictors.

(b) Least Angle Regression (LARS)

Efron *et al.* (2004) proposed LARS, which is closely related to the SFS and LASSO (Tibshirani 1996) procedures. LARS is a stylized version of the Stagewise procedure and provides an ordering in which the covariates enter a regression model. This sequence is usually the same as LASSO or SFS but is obtained by greatly reducing the computational burden.

The SFS procedure enters variables in small steps in the regression model to prevent exclusion of correlated predictors from top of the sequence. But this method often becomes time-consuming, because thousands of tiny steps are often taken in the direction of the same variable. LARS solves this problem by analytically determining the optimal step size for each variable.

Let $Y, X_1, X_2, \cdots, X_d$ be the variables, standardized using their corresponding mean and standard deviation. Let r_j denote the correlation between X_j and Y, and let R_X be the correlation matrix of the covariates $X_1, X_2, \cdots, X_d$. Suppose that X_m has the maximum absolute correlation r with Y, and denote $s_m = \text{sign}(r_m)$. Then X_m becomes the first active variable, and the initial fit $\hat{Y}=0$ should be modified by moving along the direction of $s_m X_m$ a certain distance, γ, which can be expressed in terms of correlations between the variables. By determining γ, LARS simultaneously identifies the new covariate that enters the model, that is, the second active variable.

As soon as we have more than one active variable, LARS modifies the current fit $\hat{Y}$ along the *equiangular direction*, the direction that has equal angle (correlation) with all active covariates. Moving along this direction ensures that the correlation of each active covariate with the residual decreases equally. Let A be the set of the subscripts corresponding to the active variables. The standardized equiangular vector B_A is derived. Note that we do not need the direction B_A itself to determine which covariate enters the model next; we need only the correlation of all variables (active and inactive) with B_A. These correlations can be expressed in terms of the correlation matrix of the variables. LARS modifies the current fit by moving along B_A up to a certain distance γ_A, which can be determined from the correlations of the variables.

III. Robustification of FS and LARS Algorithms

In the last section, we review FS and LARS algorithms which can be expressed in terms of sample means, variances, and correlations. Because of these non-robust building blocks, these algorithms are sensitive to contamination in the data. A simple robustification of these algorithms can be achieved by replacing the non-robust ingredients of the algorithms by their robust counterparts. The choices of rapidly computable robust center and scale measures are straightforward: median (med) and median absolute deviation (mad), which are used to robustly standardize the data. Unfortunately, good available robust correlation matrix estimators are computed from the d-dimensional data and thus are very time consuming (see, e.g., Rousseeuw and Leroy 1987). On the other hand, robust univariate approaches (Huber 1981) are very sensitive to correlation outliers (outliers that are not detected by univariate analyses but affect the classical correlation).

One solution is to derive correlations among pairs of variables from an affine-equivariant bivariate covariance estimator. A computationally efficient choice is the bivariate M-estimator defined by Maronna (1976). Alternatively, the robust correlation estimator of Gnanadesikan and Kettnring (1972) or the related orthogonalized Gnanadesikan-Kettnring estimator (Maronna and Zammar 2002) can be used. For very large, high-dimensional data sets, we need an even faster robust correlation estimator. Huber (1981) introduced the idea of univariate winsorization of the data and suggested that classical correlation coefficients be calculated from the winsorized data. Alqallaf, Konis, Martin and Zamar (2002) reexamined this approach for the estimation of individual elements of a high-dimensional correlation matrix. For n univariate observations $x_1, x_2, \cdots, x_n$, the transformation is given by

$$u_i = \psi_c((x_i - med(x_i))/mad(x_i)),\ i = 1,2,\cdots,n,$$

where the Huber score function $\psi_c(x)$ is defined by $\min(\max(-c, x), c)$, with c a tuning constant chosen by the user (e.g., $c = 2 \text{ or } 2.5$). Note that in our case, $med(x_i) = 0$ and $mad(x_i) = 1$, because we use med and mad to robustly standardize the data. The univariate winsorization approach can be computed very rapidly, but unfortunately it does not take into account the orientation of the bivariate data.

To remedy this problem, Khan *et al.* (2007a) proposed a *bivariate winsorization* of the data based on an initial robust bivariate correlation matrix R_0 and a corresponding tolerance ellipse. Outliers are shrunken to the border of this ellipse by using the bivariate transformation

$$u = \min(\sqrt{c/D(x)}, 1)x, \quad \text{with } x = (x_1, x_2)^t.$$

Here, $D(x)$ is the Mahalanobis distance based on some initial bivariate correlation matrix R_0. For the tuning constant c, we use $c = 5.99$, the 95% quantile of the χ_2^2 distribution. The choice of R_0 is discussed later.

The Initial Correlation Estimate. Choosing an appropriate correlation matrix R_0 is an essential part of bivariate winsorization. In principle, we could use any robust bivariate scatter estimate, but for computational convenience, Khan *et al.* (2007a) proposed a new method called adjusted winsorization. This method considers quadrants relative to the coordinatewise medians (which in this case are 0 due to the robust standardization of the data) and uses two tuning constants to perform univariate winsorization of the data. A larger tuning constant, c_1, is used to winsorize the points lying in the two diagonally opposed quadrants that contain most of the standardized data (called the “major quadrants”). A smaller tuning constant, c_2, is used to winsorize the remaining data. In this article we use $c_1 = 2$ and $c_2 = \sqrt{h}c_1$, where $h = n_2 / n_1$, n_1 is the number of observations in the major quadrants and $n_2 = n - n_1$. The initial correlation matrix R_0 is obtained by computing the classical correlation matrix of the adjusted winsorized data. The adjusted winsorization handles correlation outliers much better than univariate winsorization. By using bivariate winsorization, the outliers are shrunken to the boundary of the larger ellipsoid and thus appropriately downweighted so that a robust correlation estimate is obtained. Although the initial adjusted winsorization and the resulting bivariate winsorization are not affine-equivariant, they can be computed very rapidly and can appropriately handle correlation outliers.

IV. Simulations

To compare the behavior of RLARS with RFS, we consider a simulation setting similar to that used by Frank and Friedman (1993). We first create a linear model,

$$Y = L_1 + L_2 + \cdots + L_k + \sigma\varepsilon_i, \qquad \text{(i)}$$

with k latent variables, where $L_1, L_2, \cdots, L_k$ and ε are independent standard normal variables. The value of σ is chosen so that the single-to-noise ratio is 3. A set of d candidate predictors is created as follows. Let $e_1, e_2, \cdots, e_d$ be independent standard normal variables and let

$$X_i = L_i + \tau\, e_i, \qquad i = 1, 2, \cdots, k,$$

$$X_{k+1} = L_1 + \delta\, e_{k+1},$$

$$X_{k+2} = L_1 + \delta\, e_{k+2},$$

$$X_{k+3} = L_2 + \delta\, e_{k+3},$$

$$\vdots$$

$$X_{3k-1} = L_k + \delta\, e_{3k-1},$$

$$X_{3k} = L_k + \delta\, e_{3k},$$

and

$$X_i = e_i,$$
$$i = 3k+1, 3k+2, \cdots, d.$$

The constants $\delta = \sqrt{5}$ and $\tau = \sqrt{0.5}$ are chosen so that $\operatorname{corr}(X_1, X_{k+1}) = \operatorname{corr}(X_1, X_{k+2}) = \operatorname{corr}(X_2, X_{k+3}) = \cdots = \operatorname{corr}(X_k, X_{3k}) = \frac{1}{3}$. Note that covariates $X_1, X_2, \cdots, X_k$ are low noise perturbations of the latent variables and constitute our target covariates. Variables $X_{3k+1}, X_{3k+2}, \cdots, X_d$ are independent noise covariates and variables $X_{k+1}, X_{k+2}, \cdots, X_{3k}$ are noise covariates that are correlated with the target covariates.

To allow for a fraction $\in$ of outliers, we consider the following sampling distributions, listed in increasing order of difficulty:

(1) $\varepsilon \sim N(0,1)$, no contamination

(2) $\varepsilon \sim (1-\in)N(0,1) + \in N(0,1)/\mathit{uniform}(0,\ 1)$, symmetric, slash contamination

(3) $\varepsilon \sim (1-\in)N(0,1) + \in N(20,1)$, asymmetric, shifted normal contamination

(4) Same as (2), except that contaminated cases come along with high leverage X-values (normal random variables with mean 50 and variance 1 in our simulation)

(5) Same as (3), but with high leverage outliers, as described in (4).

To compute RFS and RLARS, we generate 1000 independent samples of size $n = 150$ from the five simulation designs just described, with $k = 10$ latent variables and $d = 50$ candidate covariates. For each data set, we sequence the variables using RFS and RLARS.

To summarize the simulation results, for each sequence we determine the number t_m of target variables included in the first m sequenced variables, with m ranging between 1 and 20 for each of the methods. Fig. 1 shows the average (over 1000 data sets) of t_m for RFS and RLARS and sampling situations. We display here the results for the case where $\in = .10$.

From figure 1(a), we see that RFS performs better than RLARS procedure in the uncontaminated case. From Fig. 1(b)-1(e), we also see that RFS performs better than RLARS procedure under contamination. In the design without leverage but asymmetric, shifted normal contamination, RFS shows slightly better performance than RLARS [Fig. 1(c)]. All the figures show that RFS procedure is much less affected than RLARS.

V. Example

In this section, we use a real data set to further compare the performance between RFS and RLARS algorithms. We use a data set given in the CAED Report 17, Iowa State University, presented in 1963 (Draper and Smith 1998).

The response variable is the corn yield (bu/acre). There are 9 covariates, which we number from 1 to 9. Though this analysis may not be of particular scientific interest, it will demonstrate the ability of RFS and RLARS algorithms to identify the correct covariates in linear regression.

RFS sequences the covariates in the following order: (1, 6, 2, 5, 3, 8, 7, 9, 4). When RLARS is applied to the data, we obtain the following different sequence of covariates: (1, 6, 7, 8, 2, 5, 3, 9, 4). Table 1 presents the results of fitted least squares linear regressions (including the intercept) of the selected covariates taken from each of the above sequences.

Comparing the values of robust R^2 (robust coefficient of multiple determination) measure, such as $R^2 = 1 - median(e^2)/(mad(Y))^2$, where e is the vector of residuals from the robust fit (see also Rousseeuw and Leroy 1987), it may generally be concluded that the efficiency of RFS is more than that of RLARS in selecting the target covariates in the linear regression model.

VI. Conclusions

FS and LARS are popular and computationally suitable algorithms for building linear prediction models, but they are sensitive to outliers. Though RFS is more aggressive fitting technique than RLARS, RFS has much better performance compared to the RLARS under both the uncontaminated and contaminated cases.

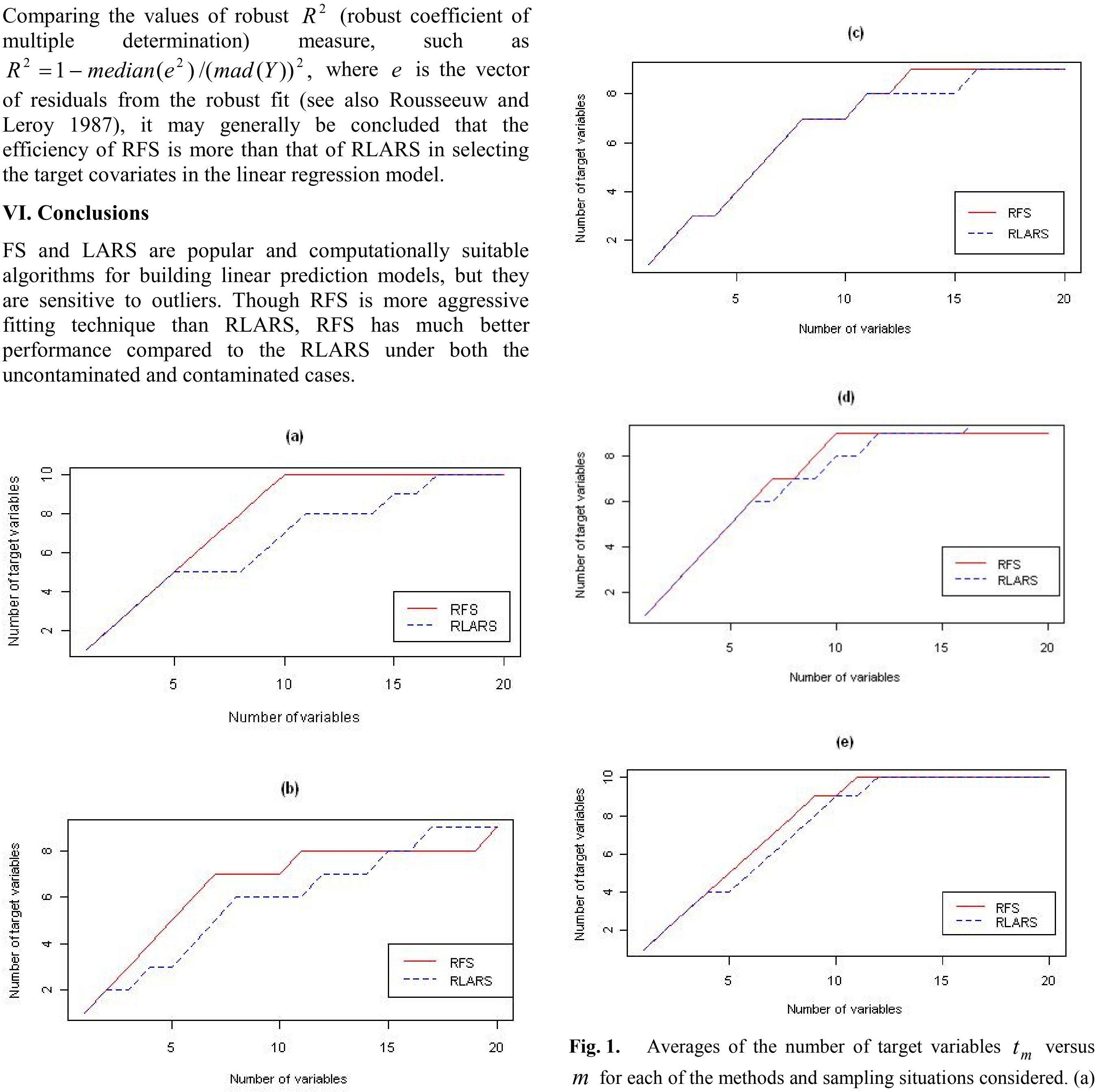

Fig. 1. Averages of the number of target variables t_m versus m for each of the methods and sampling situations considered. (a) No contamination; (b) slash contamination; (c) normal contamination; (d) slash contamination/high leverage; (e) normal contamination/high leverage. We generated data sets of $d = 50$ predictors, $k = 10$ latent variables and 10% of contamination $(\in = 0.1)$ (— RFS; --- RLARS).

Table. 1. Performance of RFS and RLARS

Number of covariates	Sequence of covariates in RFS	Sequence of covariates in RLARS	robust R^2 for RFS	robust R^2 for RLARS
3	1, 6, 2	1, 6, 7	0.852	0.812
4	1, 6, 2, 5	1, 6, 7, 8	0.902	0.848
5	1, 6, 2, 5, 3	1, 6, 7, 8, 2	0.906	0.854

1. Alqallaf, F. A., K. P. Konis, R. D. Martin, and R. H. Zamar, 2002. Scalable Robust Covariance and Correlation Estimates for Data Mining. *Proceedings of the Seventh ACM SIGKD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, 14-23.

2. Atkinson, A. C., and M. Riani, 2002. Forward Search Added-Variable t Tests and the Effect of Masked outliers on Model Selection, *Biometrika*, 89, 939-946.

3. Contoni, E, and E. Ronchetti, 2001. Robust Inference for Generalized Linear Models. Journal of the American Statistical Association, 96, 1022-1030.

4. Draper, N., and H. Smith, 1998. Applied Regression Analysis. 3rd Ed.. New York: John Wiley and Sons. 362.

5. Efron, B. E., T. Hastie, I. Johnstone, R. Tibshirani. 2004. Least Angle Regression, *The Annals of Statistics*, 32, 407-451.

6. Frank, I, and J. H. Friedman, 1993. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35, 109-148, New York: Springer-Verlag.

7. Genton, M. G., and E. Roncchetti, 2003. Robust Indirect Inference. *Journal of the American Statistical Association*, 98, 67-76.

8. Gnanadesikan, R., and J. R. Kettenring, 1992. Robust Estimates, Residuals, and Ouliers Detection With Multiresponse Data. *Biometrica*, 28, 81-124.

9. Huber, P. J., 1981. *Robust Statistics.* Wiley, New York.

10. Jiang, W., and B. Turnbull, 2004. The Indirect Method: Inference Based on Intermediate Statistics: A Synthesis and Examples. *Statistical Science*, 19, 239-263.

11. Khan, J. A., S. Van Aelst, R. H. Zamar, 2007a. Robust Linear Model Selection Based on Least Angle Regression. *Journal of the American Statistical Association*, 102, 1289-1299.

12. Khan, J. A., S. Van Aelst, R. H. Zamar, 2007b. Building a Robust Linear Model with Forward Selection and Stepwise Procedures. *Computational Statistics and Data Analysis* (*CSDA*), **52(1)**, 239-248.

13. Maronna, R. A., 1976. Robust M-estimators of Location and Scatter. *The Annals of Statistics*, 4, 51-67.

14. Maronna, R. A., and R. H. Zamar, 2002. Robust Estimates of Location and Dispersion for High-Dimensional Datasets. *Technometrics*, 44, 307-317.

15. Maronna, R. A., R. D. Martin, and V. J. Yohai, 2006. *Robust Statistics: Theory and Methods*, New York: Wiely.

16. Morgenthaler, S., R. E. Welsch, and A. Zenide, 2003. Algorithms for Robust Model Selection in Linear Regression, in *Theory and Applications of Recent Robust Methods*, eds. M Hubert, G. Pison, A. Struyf, and S. Van Aelst, Basel, Switzerland: Birkhauser-Verlag, 195-206.

17. Muller, S., and A. H. Welsch, 2005. Outlier Robust Model Selection in Linear Regression. *Journal of the American Statistical Association*, 100, 1297-1310.

18. Ronchetti, E., 1985. Robust Model Selection in Regression. *Statistics and Probability Letters*, 3, 21-23.

20. Ronhetti, E., and R. G. Staudte, 1994. A Robust Version of Mallows's C_p. *Journal of the American Statistical Association*, 89, 550-559.

21. Ronhetti, E., and F. Trojani, 2001. Robust Inference With GMM Estimators. *Journal of Econometrics,* 101, 37-69.

22. Ronhetti, E., C. Field, and W. Blanchard, 1997. Robust Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, 92, 1017-1023..

23. Rousseeuw, P. J., and A. M. Leroy, 1987. *Robust Regression and Outlier Detection*, New York: Wiley-Interscience.

24. Sommer, S., and R. M. Huggins, 1996. Variable Selection Using the Wald Test and Robust C_p. *Journal of the Royal Statistical Society*, Ser. B, 45, 15-29.

25. Tibshirani, R., 1996. Regression Shrinkage and Selection Using via the Lesso. *Journal of the Royal Statistical Society*, Ser. B. 58, 267-288.

26. Weisberg, S., 1985. Applied Linear Regression. 2nd Ed.. Wiley, New York.