Communications for Statistical Applications and Methods
2019, Vol. 26, No. 6, 575–582

https://doi.org/10.29220/CSAM.2019.26.6.575
Print ISSN 2287-7843 / Online ISSN 2383-4757

Unified methods for variable selection and outlier detection in a linear regression

Han Son Seo[1,a]

[a]Department of Applied Statistics, Konkuk University, Korea

Abstract

The problem of selecting variables in the presence of outliers is considered. Variable selection and outlier detection are not separable problems because each observation affects the fitted regression equation differently and has a different influence on each variable. We suggest a simultaneous method for variable selection and outlier detection in a linear regression model. The suggested procedure uses a sequential method to detect outliers and uses all possible subset regressions for model selections. A simplified version of the procedure is also proposed to reduce the computational burden. The procedures are compared to other variable selection methods using real data sets known to contain outliers. Examples show that the proposed procedures are effective and superior to robust algorithms in selecting the best model.

Keywords: outliers, regression diagnostics, robustness, variable selections

1. Introduction

Selecting variables and detecting outliers are important issues in the field of regression diagnostics. Many methods are suggested focusing on these issues separately. However it is well known that the process of model selection depends on the order in which the variable selection and outlier identification are performed (Adams, 1991; Blettner and Sauerbrei, 1993; Hoeting *et al.*, 1996). There are a number of approaches in performing variable selection and outlier detection simultaneously. Chatterjee and Hadi (1988) studied the joint impact of simultaneous omission of a variable and an observation on a regression equation. They gave a testing statistic for the significance of a variable when an observation is omitted. Menjoge and Welsch (2010) presented a backward selection method performing on the data matrix augmented by appending dummy variables to the original data. Hoeting *et al.* (1996) offered a simultaneous approach based on Bayesian posterior model probabilities. McCann and Welsch (2007) applied least angle regression (LARS) to the augmented matrix. Kim *et al.* (2008) suggested to identify potential outliers and to perform all possible subset regressions for the mean-shift outlier model. Recently Kong *et al.* (2018) provided the theoretical results in terms of high breakdown point, full efficiency and outlier detection consistency for mean-shift outlier model. Many variable selection methods using robust estimators or robust measures have also been proposed. Ronchetti and Staudte (1994) suggested a robust version of Mallow's C_p. Ronchetti *et al.* (1997) provided a robust algorithm for model selection using M-estimator and cross-validation. Wisnowski *et al.* (2003) showed that the cross-validation cannot be applied to contaminated datasets using a robust estimation scheme and suggested resampling methods for cross validation and Bootstrap estimators

[1] Department of Applied Statistics, Konkuk University, Korea, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea.
E-mail: hsseo@konkuk.ac.kr

Published 30 November 2019 / journal homepage: http://csam.or.kr

of prediction error measures in robust regression. Dupuis and Victoria-Feser (2011, 2013) used robust regression algorithms for model selection in large data sets. In this article we develop variable selection methods using outlier detection methods. Section 2 starts by discussing the measures for variable selection and proposes unified procedures for variable selection and outlier detection. Several applications to real data sets are given to show the performance of the procedure. Section 3 contains a few concluding comments.

2. Variable selection with outlier detection

Model comparisons are complicated in the presence of outliers. If outlier detection methods are applied and some observations are excluded from the data in estimating models, two models usually become non-nested models. The likelihood ratio test for non-nested models is not available because the exact distribution of the test statistic depends on unknown parameters. Many approaches have been suggested for testing non-nested models (Efron, 1984; Vuong, 1989; Royston and Thompson, 1995; Godfrey, 1998). Most of them use asymptotic or empirical distribution of testing statistics under the general conditions. As an alternative to tests, the measures quantifying the relative quality of models for a given set of data can be compared. Akaike information criterion (AIC) (Akaike, 1973) uses log likelihood, adds a penalty for the number of parameters and permits comparison between non-nested models. Bayesian information criterion (BIC) (Schwarz, 1978) is suggested to overcome the inconsistency of AIC. Both AIC and BIC are not appropriate for comparing models for different sizes of data. They are biased and have a tendency to increase with a large size of data. C_p also depends on the sample size. In this study we use an adjusted-R^2 as a measure of model comparisons. It is biased and not recommended in some aspects. But it does not have a severe tendency by the size of data and the dimension of variables. The dimension of final model can be determined by the direct minimization of a measure or as the number of variables associated with a significant change in a measure. Wisnowski *et al.* (2003) tried two approaches and showed that the latter prediction selects the true model more successfully. We use the latter procedure for the determination of the final model.

2.1. All possible regressions approach

There have been many methods proposed for variable selection and outlier identification. Some methods often consider identified outliers as potential outliers. (Hoeting *et al.*, 1996; Kim *et al.*, 2008). Kim *et al.* (2008) suggested a two-step method using the mean-shift outlier model. The first step identifies potential outliers and the second step is to execute all possible subset regressions of the mean-shift outlier model containing potential outliers. This approach enables simultaneous variable selection and identification of outliers because the mean-shift outlier model produces the same residual sums of squares as the model fitted excluding the relevant observations. However, it requires to fix the same observations as potential outliers for all possible subset regressions.

We suggest a method which does not need to set potential outliers in advance by detecting outliers in each model of all possible subset regressions. The sequential method of Hadi and Simonoff (1993) is used to identify outliers, which is not computationally intensive and resistant to masking and swamping effects. It starts with an initial clean subset of observations and increases the size of the clean subset until the remaining observations are declared as outliers. Outlyingness on a subset corresponding to the observations of the minimum residual sum of squared is determined by t-test using the order statistic of internally studentized residuals. A unified procedure for variable selection and outlier detection is outlined as follows.

1. Set up a one-predictor model and perform Hadi and Simonoff's procedure to identify outliers.

Table 1: All possible subset regressions for the Stackloss data

Variables	Outliers	Adjusted-R^2
X_1	4, 21	0.9556*
X_2		0.7542
X_3	1, 2, 3, 4	0.2640
X_1, X_2	1, 3, 4, 21	0.9688*
X_1, X_3	4, 21	0.9553
X_2, X_3		0.7449
X_1, X_2, X_2	1, 3, 4, 21	0.9692**

"$*$" = the best model among k-models; "$**$" = the best model overall.

Table 2: All possible subset regressions for the Scottish Hill racing data

Variables	Outliers	Adjusted-R^2
X_1	7, 18, 33	0.9593*
X_2	7, 11, 17, 18, 33, 35	0.6825
X_1, X_2	7, 18, 33	0.9855**

"$*$" = the best model among k-models; "$**$" = the best model overall.

2. Fit the model again after deleting the identified outliers and calculate adjusted-R^2 of the model.

3. Repeat the step 1 and 2 for all one-predictor models.

4. Determine the best one-predictor model based on adjusted-R^2s.

5. Find the best k-predictors model (for $k = 2, \ldots, p$) by the same procedure.

6. Finally determine the best model among the p best models considering the rule of parsimony.

We provide several examples with the same data sets as those used by Kim *et al.* (2008) and Hoeting *et al.* (1996) to compare the results.

Example 1. Stackloss data

The Stackloss data (Brownlee, 1965) have 21 observations from a plant for the oxidation of ammonia as a stage in the production of nitric acid. They consist of three independent variables, X_1 (air flow), X_2 (inlet temperature), X_3 (concentration of acid) and the response variable Y (stack loss). The Stackloss data have studied by many authors (Rousseeuw and van Zomeren, 1990; Hoeting *et al.*, 1996; Kim *et al.*, 2008; Sebert *et al.*, 1998). It is generally accepted that the explanatory variable X_3 should be dropped from the first order linear model and that observations (1, 3, 4, 21) are outliers. Kim *et al.* (2008) and Hoeting *et al.* (1996) also showed the same results. Applying Hadi-Simonoff's method to each model of all possible regressions, sets of outliers are identified and adjusted-R^2s are calculated. Table 1 shows that the two best models based on adjusted-R^2s are (X_1, X_2) and (X_1, X_2, X_3). Both of two models detected the same observations as outliers. Removing the outliers and performing the partial F-tests, the model (X_1, X_2) is selected as the best model.

Example 2. Scottish Hill racing data

Scottish Hill racing data (Atkinson, 1986) includes two independent variables X_1 (distance), X_2 (climb) and the response variable Y (time) with 35 observations. The model (X_1, X_2) with outliers (7, 18, 33) is selected as the best subset model. This result agrees with Kim *et al.* (2008) and Hoeting *et al.* (1996) (Table 2).

Table 3: All possible subset regressions for the Wood Gravity data

Variables	Outliers	Adjusted-R^2	Variables	Outliers	Adjusted-R^2
X_1		0.3621	X_4		0.4027
X_2		0.3781	X_5		0.2210
X_3	4, 6, 8, 19	0.5386*			
X_1, X_2		0.6699*	X_2, X_4		0.4715
X_1, X_3	4, 6, 8, 19	0.6577	X_2, X_5		0.3883
X_1, X_4		0.5137	X_3, X_4	4, 6, 8, 19	0.5044
X_1, X_5		0.3562	X_3, X_5	4, 6, 8, 19	0.5711
X_2, X_3	4, 6, 8, 19	0.6213	X_4, X_5		0.5159
X_1, X_2, X_3		0.7499	X_1, X_4, X_5		0.5275
X_1, X_2, X_4		0.6586	X_2, X_3, X_4	4, 6, 8, 19	0.5906
X_1, X_2, X_5		0.6836	X_2, X_3, X_5	4, 6, 8, 19	0.6716
X_1, X_3, X_4		0.5489	X_2, X_4, X_5		0.5101
X_1, X_3, X_5	4, 6, 8, 19	0.7517*	X_3, X_4, X_5	4, 6, 8, 19	0.6699
X_1, X_2, X_3, X_4		0.7451	X_1, X_3, X_4, X_5	4, 6, 8, 19	0.9421**
X_1, X_2, X_3, X_5		0.7570	X_2, X_3, X_4, X_5	4, 6, 8, 19	0.7918
X_1, X_2, X_4, X_5		0.6630			
X_1, X_2, X_3, X_4, X_5	4, 6, 8, 19	0.9375*			

Observations (4, 6, 7, 8, 11, 19) were used as an initial set of outliers in the Hadi-Simonoff's procedure.
"$*$" = the best model among k-models; "$**$" = the best model overall.

Example 3. Modified Wood Gravity data

The Modified Wood Gravity data are a five-predictor data set and are contaminated by replacing cases 4, 6, 8 and 19 with outliers (Rousseeuw, 1984). Observations (4, 6, 7, 8, 11, 19) are identified as potential outliers by many methods. We use a set of observations (4, 6, 7, 8, 11, 19) as an initial outliers in the Hadi-Simonoff's procedure. Table 3 shows the results of all possible models. Based on adjusted-R^2s, the model (X_1, X_3, X_4, X_5) with outliers (4, 6, 8, 19) is selected as the best model.

2.2. A sequential procedure

Performing all possible regressions needs a lot of computations. We suggest a sequential procedure for simultaneous variables selection and outliers detection to reduce computational work. The sequential procedure finds the best model of the next stage based on the current best model. Each variable not involved in the current best model is added to the current best model and then adjusted-R^2 is computed after deleting outliers identified by an outlier detection method. One variable is selected by continuing this process for all variables in the current best model and is added to the current best model. We call this a provisional best model. The final model of the next stage is determined by exchanging a variable involved in the provisional best model with a variable not involved. The details of the procedure are described as follows.

Let S be a set of explanatory variables and $\rho(S)$ be a measure of the process computed from the model of S after deleting identified outliers. Let M_k be the best set of k variables. M_1 is determined as $M_1 = \arg_{X_i} \max[\rho(X_i) : X_i \in E]$, where $E = \{X_1, X_2, \ldots, X_p\}$ is the set of p variables.

Given M_k, M_{k+1} is determined as follows.

- Step 1. Each variable not involved in M_k is added to the current best model and then adjusted-R^2 is computed after deleting outliers identified by Hadi-Simonoff's procedure.

  The provisional best set of $k+1$ variables, denote it as T_{k+1}, is the best model among them.

  Let M_k^c be the complement of M_k and define $Q_k(X_i) = M_k \bigcup \{X_i\}$ where $X_i \in M_k^c$.

$$T_{k+1} = \arg_S \max[\rho(S) : S \in A_k],$$

Table 4: Results of the sequential procedure applied to the Wood Gravity data

Variables	Outliers	Adjusted-R^2
X_3	4, 6, 8, 19	0.5386*
X_1, X_2		0.6699*
X_1, X_3, X_5	4, 6, 8, 19	0.7517*
X_1, X_3, X_4, X_5	4, 6, 8, 19	0.9421**
X_1, X_2, X_3, X_4, X_5	4, 6, 8, 19	0.9375*

"*" = the best model among k-models; "**" = the best model overall.

where $A_k = \{Q_k(X_i)|X_i \in M_k^c\}$, the collection of $Q_k(X_i)$'s for all $X_i \in M_k^c$.

- Step 2. Exchange a variable in the provisional best model with a variable not in the provisional best model and then adjusted-R^2 is computed after deleting outliers. The final best set of $k+1$ variables is the best model among them.

  Let T_{k+1}^c be the complement of T_{k+1} and $R_{k+1}(X_i, X_j) = (T_{k+1} - \{X_i\}) \bigcup \{X_j\}$ where $X_i \in M_k$ and $X_j \in T_{k+1}^c$.

  The final best set of $k+1$ variables, M_{k+1}, is

$$M_{k+1} = \arg_S \max[\rho(S) : S \in B_{k+1}],$$

  where $B_{k+1} = \{R_{k+1}(X_i, X_j)|X_i \in M_k, \ X_j \in T_{k+1}^c\}$, the collection of $R_{k+1}(X_i, X_j)$'s for all $X_i \in M_k$ and $X_j \in T_{k+1}^c$.

The sequential approach requires $p(p^2+5)/6$ regressions, whereas all possible regressions does $2^p - 1$. Table 4 compares the computational cost between a sequential procedure and all possible regression procedures for given p.

Several examples are provided using the same data sets as those used by Kim *et al.* (2008) and Hoeting *et al.* (1996) for comparisons. Both the sequential procedure and the all possible regressions approach obtain the same result for the Stackloss data, the Scottish Hill Racing data and the Modified Wood Gravity data because the number of variables, p, is small. Table 4 shows the best models for k-predictors model selected by applying the sequential procedure to the Modified Wood Gravity data. Model (X_1, X_3, X_4, X_5) with outliers (4, 6, 8, 19) is the best model based on adjusted-R^2.

The sequential procedure is applied to the Mortality data which is available in the Data and Story Library of StatLib (StatLib, 1996). The response variable is age-adjusted mortality, and the potential predictors are 14 variables measuring demographic characteristics of the cities, climate characteristics or recording the pollution potential of air pollutants. The Mortality data are also used by McCann and Welch (2007). McCann and Welch (2007) suggested to append dummy variable identity matrix to the design matrix for robustness and to use LARS for variable selection. They proposed three algorithms. The first algorithm, called "LARSD-T", fits a LS regression on the variables provided by LARS and selects the final model using t-statistics for each variable. McCann and Welch (2007)'s other algorithm starts by taking a sample from the data and selecting variables from LARS. The nested models of the selected variables are constructed by the LARS-ordering and the MAD of the residuals from the LARS are calculated for each model. These steps are repeated a large number of times and the best 1% models based on MAD are selected. A variables is finally selected if it appears in 50% or more of the top 1% best models (LARS-CV). The same algorithm as LARS-CV using the data appended by dummy variables is "LARSD-CV", which was designed to get samples free of leverage outliers and to detect additive outliers.

Table 5: Models selected by various algorithms applied to the Mortality data

Algorithm	Variables selected	MAD
LARS-CV	$X_7, X_4, X_6, X_{14}, X_5$	24.40
LARSD-CV	$X_7, X_4, X_6, X_{14}, X_5, X_3, X_9$	19.15
LARSD-T	X_7, X_4, X_{14}	26.88
Sequential Proc	$X_7, X_{14}, X_4, X_8, X_6, X_{12}, X_2, X_{11}$	18.85

Part of results adopted from McCann and Welch (2007).

Table 6: Best subsets scores on cleaned data of the Mortality data

Algorithm	Parameters	C_p	AIC	BIC
LARS-CV	6	7.82	313.24	316.32
LARSD-CV	8	8.74	313.77	318.43
LARSD-T	4	22.15	326.07	326.21
Sequential Proc	6	5.84	298.10	311.50

Part of results adopted from McCann and Welch (2007). AIC = Akaike information criterion; BIC = Bayesian information criterion.

Tables 5 contains the results of McCann and Welch (2007)'s algorithms and the sequential procedure applied to the Mortality data. Three robust algorithms, LARS-CV, LARSD-CV and LARSD-T selected variables, $(X_7, X_4, X_6, X_{14}, X_5)$, $(X_7, X_4, X_6, X_{14}, X_5, X_3, X_9)$, and (X_7, X_4, X_{14}) respectively. LARSD-CV showed the best performance among them based on MAD of LTS residuals and had a MAD value, 19.15. The sequential procedure selected variables $(X_7, X_{14}, X_4, X_8, X_6, X_{12}, X_2, X_{11})$ and its MAD is 18.85.

Another meaningful comparison is to check how the best models would score on a clean data. The clean version of the Mortality data is obtained by removing 9 points that had a standardized residual with a magnitude of 2.5 or larger after any of the LTS. McCann and Welch (2007) reported the performance of the robust algorithms on this cleaned set. For the comparison with McCann and Welch's methods, the size of significance is adjusted in the sequential procedure to get the best model with 9 outliers. Table 6 contains the performance of methods on a clean version of Mortality data using C_p, AIC, and BIC. Among McCann and Welch's methods, LARS-CV performed better than others based on AIC (313.24) and BIC (316.32). LARSD-CV has a C_p value, 8.74 which is close to the size of selected variables. The sequential procedure selected variables $(X_7, X_{14}, X_4, X_6, X_{11}, X_9)$ and detected 9 observations (2, 14, 27, 31, 36, 52, 55, 57, 58) as outliers. The scores of this model, BIC = 311.5, AIC = 298.1, and $C_p = 5.84$, show that the sequential procedure performs better than robust algorithms.

3. Concluding remarks

A unified procedure for variable selection and outlier detection has been provided using all possible subset regressions. The suggested procedure does not depend on the order in which variable selection and outlier identification are performed. It does not require presetting potential outliers prior to selecting variables. Examples with several "benchmark" data sets showed that the proposed method is effective in selecting variables when outliers exist.

A sequential procedure is also suggested that does not investigate all possible models if p is large and all possible subset regressions consider too many models. Examples demonstrate that a sequential procedure contains less computations and is superior to robust algorithms in selecting the best model. It is not easy to find an appropriate measure for variable selection when outliers exist. Instead of adjusted-R^2, a generalized criterion for model comparison (Busemeyer and Wang, 2000) can be used

in the suggested procedures.

Acknowledgement

This paper was supported by Konkuk University in 2018.

References

Adams JL (1991). A computer experiment to evaluate regression strategies. In *Proceeding of American Statistical Association Section on Statistical Computing*, 55–62.

Atkinson AC (1986). [Influential observations, high leverage points, and outliers in linear regression]: Comment: aspects of diagnostic regression analysis, *Statistical Science*, **1**, 397–402.

Akaike H (1973). *Information theory and an extension of the maximum likelihood principle*, In B. N. Petrov & F. Csaki (Eds), Second International Symposium on Information theory, Budapest, Akademiai Kiado.

Blettner M and Sauerbrei W (1993). Influence of model-building strategies on the results of a case-control study, *Statistics in Medicine*, **12**, 1325–1338.

Brownlee KA (1965). *Statistical Theory and Methodology in Science and Engineering* (2nd ed.), Wiley, New York.

Busemeyer JR and Wang Y (2000). Model comparisons and model selections based on generalization criterion methodology, *Journal of Mathematical Psychology*, **44**, 171–189.

Chatterjee S and Hadi AS (1988). Impact of simultaneous omission of a variable and an observation on a linear regression equation, *Computational Statistics & Data Analysis*, **6**, 129–144.

Dupuis DJ and Victoria-Feser MP (2011). Fast robust model selection in large datasets, *Journal of the American Statistical Association*, **106**, 203–212.

Dupuis DJ and Victoria-Feser MP (2013), Robust VIF regression with application to variable selection in large data sets, *Annals of Applied Statistics*, **7**, 319–341.

Efron B (1984). Comparing non-nested linear models, *Journal of the American Statistical Association*, **79**, 791–803.

Godfrey LG (1998) Tests of non-nested regression models: Some results on small sample behaviour and the bootstrap, *Journal of Econometrics*, **84**, 59–74.

Hadi AS and Simonoff JS (1993). Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, **88**, 1264–1272.

Hoeting J, Raftery AE, and Madigan D (1996). A method for simultaneous variable selection and outlier identification in linear regression, *Computational Statistics & Data Analysis*, **22**, 251–270.

Kim S, Park SH, and Krzanowski WJ (2008). Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model, *Journal of Applied Statistics*, **35**, 283–291.

Kong D, Bondell HD, and Wu Y (2018). Fully efficient robust estimation, outlier detection and variable selection via penalized regression, *Statistica Sinica*, **28**, 1031–1052.

McCann L and Welsch RE (2007). Robust variable selection using least angle regression and elemental set sampling, *Computational Statistics & Data Analysis*, **52**, 249–257.

Menjoge RS and Welsch RE (2010). A diagnostic method for simultaneous feature selection and outlier identification in linear regression, *Computational Statistics & Data Analysis*, **54**, 3181–3193.

Ronchetti E, Christopher Field C, and Blanchard W (1997). Robust linear model selection by Cross-

Validation, *Journal of the American Statistical Association*, **92**, 439.
Ronchetti E and Staudte RG (1994). A robust version of Mallows's C_p, *Journal of the American Statistical Association*, **89**, 550–559.
Rousseeuw PJ (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.
Rousseeuw PJ and van Zomeren BC (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, **85**, 633–639.
Royston P and van Thompson SG (1995). Comparing non-nested regression models, *Journal of the American Statistical Association*, **51**, 114–127.
Schwarz JH (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461–464.
Sebert DM, Montgomery DC, and Rollier DA (1998). A clustering algorithm for identifying multiple outliers, *Computational Statistics & Data Analysis*, **27**, 461–484.
StatLib (1996). Department of Statistics, Carnegie Mellon University, Data and Story Library, Datafile Name:SMSA, Reference: U.S. Department of Labor Statistics.
Vuong QH (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica*, **57**, 307–333.
Wisnowski JW, Simpson JR, Montgomery DC, and Runger GC (2003). Resampling methods for variable selection in robust regression, *Computational Statistics & Data Analysis*, **43**, 341–355.

Received July 13, 2019; Revised September 5, 2019; Accepted September 10, 2019