

PURO: A Package for Unmasking Regression Outliers

Enis SINIKSARAN^{1*}, Mehmet Hakan SATMAN²

¹*Istanbul University, Faculty of Political Science, Beyazıt 34116, İstanbul, ,Turkey*

²*Istanbul University, Department of Econometrics, Beyazıt 34116, İstanbul, ,Turkey*

Received:12/03/2010 Accepted:13/09/2010

ABSTRACT

Multiple regression outliers should be identified because of their potential effect on parameter estimates and inferences from the regression model. In recent years, researchers have proposed numerous strategies and procedures to identify the outliers. A *Mathematica* package PURO is introduced which implements seven methods from the latest and most respected outlier detection procedures in the statistics literature.

Keywords: *Regression outlier, Mathematica, Detection procedure, Masking, Swamping, Diagnostic plot.*

1. INTRODUCTION

Identifying regression outliers can be very important task in many fields of science. They should be identified because of their potential effect on parameter estimates and inferences from the regression model. As a result of the need to detect and accommodate the outliers, several outlying measures and influence diagnostics have been designed. See, Belsly et al. (1980), Cook and Weisberg (1982), Rousseeuw and Leroy (1987) for detailed discussions of the identification problem of multiple regression outliers.

If a regression data set contains only a single or a few outliers, standard least squares outlying measures generally work well. However, when a group of observations clustered in an outlying cloud collaborate together, standard diagnostics would fail to identify the outliers (masking) or can identify the inlying observations as outliers (swamping). Although, of these two types of errors, masking can be viewed as a more serious problem than swamping, a good outlier procedure should keep both errors to a minimum. Researchers have proposed numerous strategies and procedures to overcome the insufficiency of standard least squares outlying measures. In Kianifard and Swallow (1990), and Hadi and Simonoff (1993), brief overviews of the procedures can be found.

In the outlier detection tradition, generally there are two ways to evaluate the performance (the rates of masking and swamping) of any proposed outlier identification procedure. The first one is to identify the outliers in some commonly used data sets which have proved intractable to the standard diagnostics. Rousseeuw telephone data, Hertzprung and Russell star data, Hawkins, Bradu and Kass data, Hadi and Simonoff 1993 data, Modified Wood Gravity data, Stackloss data, and body and brain weight data are well known examples of these data sets of which some were created artificially to illustrate the failure of some popular procedures and/or to validate some other methods. Hadi and Simonoff (1993) claim that if an outlier detection procedure fails for first five of these data sets it should not be viewed as a viable method.

The second common way to evaluate the performance of an outlier procedure is to use Monte Carlo simulations in various outlier scenarios depending on the factors such as number of variables in the model, sample size, outlier type, the percentage and the outlying distances of the outliers. Kianifard and Swallow (1990), Wisnowski et al. (2001) and Billor and Kiral (2008) are good examples of Monte Carlo experiments for outlier detection.

Monte Carlo simulations show that many outlier detection procedures perform well. However, the performances of the outlier procedures may vary depending on the factors or model. For instance, some procedures working well for bad leverage points may not

*Corresponding author, e-mail: esiniksaran@istanbul.edu.tr

be successful for vertical outliers, or a good performance in lower dimensions does not guarantee a success in higher dimensions. While some methods are good when the outlying distance is large, some methods are not influenced by the outlying distance. The methods which are good in unmasking rate may not be preferred because of high swamping rates. The methods differ in computation time as well. A very successful procedure may not be in demand because of a high computation time. In short, it seems unreasonable to claim that this or that outlier procedure is the best in all cases. Accordingly, the selection of the appropriate outlier procedure can be subjective based on the researchers' needs and limitations or a priori information about the data at hand such as an estimate of percent or type of outliers.

Almost all of the well known statistical software packages like SPSS, S-Plus or Minitab provide standard regression outlier diagnostics such as hat matrix, Cook's distance, DFBETAS or COVRATIO on the Linear Regression dialog box as options. However, just a few statistical software have implemented multiple regression outliers procedures. Stata or Data Desk implementing Hadi's procedures Hadi (1992) and Hadi (1994) are some exceptions.

The purpose of this paper is to present the *Mathematica* 7.0 package PURO (A Package for Unmasking Regression Outliers) which implements the seven methods from the latest and most respected outlier detection procedures in the statistics literature. *Mathematica* is a very powerful computing environment for all forms of numeric and symbolic computation, and is an exceptionally flexible tool for producing mathematical graphics. The performance of *Mathematica* in precision could be an additional reason to use it. McCullough (1998) proposed a set of tests for assessing statistical software. His method has been applied to the several statistical and econometric software. See, for instance, McCullough (1999) and McCullough and Heiser (2008). In McCullough (2000), he applied his methodology to Mathematica 4.0 and found that the precision of *Mathematica*'s calculations in the areas of linear and nonlinear estimation, random number generation, and statistical distributions are markedly superior to that of other major statistical packages such as SPSS, SAS, S-Plus, and Excel.

In choosing a procedure for the package PURO, we considered the following factors:

- 1) Success in classical data sets and/or in Monte Carlo simulations,
- 2) Ease of implementation,
- 3) Simplicity in interpreting its results,
- 4) Time and space efficiency.

The outlier procedures implemented by our package are HS in Hadi and Simonoff (1993), PY in Pena and Yahoï (1995), SMR in Sebert et al. (1998), BCH in Billor et al (2006), LMS in Rousseeuw (1984), RZ in Rousseeuw and Van Zomeren (1990) and MS in Marchette and Solka (2003).

The motivation for the preparation of the package mainly comes from the shortage of the programs for multiple regression outliers in common statistical software. The other motivating reason is, as explained above, the absence of a unique procedure which is best in all cases. We believe that, for this reason, a user of an outlier detection software should see the results of several outlier procedures which may enable her to combine them into one result set, and hence can take initiative in making the final decision about the outliers and outlier clusters within her data. Hopefully, PURO providing the results of seven reliable outlier procedures can satisfy such a user's need. The program file PURO.m and the demo file demoPURO.nb of the package can be freely downloaded from <http://www...>

In the following section, we briefly describe the multiple outlier detection methods used in the package. In Section 3, the package is introduced and its features are presented. This section provides an application for the package to show how the results of the procedures should be interpreted correctly.

2. Brief descriptions of outlier detection procedures in the package

Before giving the descriptions of the procedures, we need to fix the terminology and notation in the paper. First, let us illustrate the common terms (leverage points, outliers and regular points) used in outlier detection literature in an example of simple linear regression, which can be found in several books and articles, see, for instance, Rousseeuw and Van Zomeren (1990).

Regression outliers are observations that lie outside the regression plane defined by the bulk of the data. If the response value of an outlier is distant from responses of the clean cases, it is called a vertical outlier. If an observation has an outlying value in the space of explanatory variables (x-space), it is called a leverage point. If a leverage point is outlier, it is called a bad leverage point, otherwise it is called a good leverage point.

Figure 1 shows a scatterplot of a data set of 30 observations. The bulk of the data have a linear pattern. These 27 data points can be called clean, regular, well-fitted or inlying observations. Based on the definitions above, we can say that the points *A* and *C* are outliers; *A* is a vertical outlier and *C* is a bad leverage point. *B* is a leverage point but not an outlier and hence a good leverage point since it fits to the linear pattern of the clean cases. In multiple regression, however, especially when the number of independent variables is greater than two, it is no longer possible to identify the outliers and leverage points visually in a scatterplot as the data can not be viewed in total. Nevertheless, the definitions remain the same as in the simple linear regression case.

The aim of the standard regression diagnostics and the outlier detection procedures is to find the vertical outliers and bad leverages. We will now give the brief descriptions of the procedures implemented in our package. In these descriptions, we consider the standard linear regression model

$$y = X\beta + \varepsilon$$

where y is an $n \times 1$ vector of responses, X is an $n \times p$ matrix of predictor variables, β is an $n \times 1$ vector of parameters and ε an $n \times 1$ vector of random errors. The projection matrix $H = X(X'X)^{-1}X'$ will be denoted as hat matrix and h_{ij} is the ij th element of that matrix.

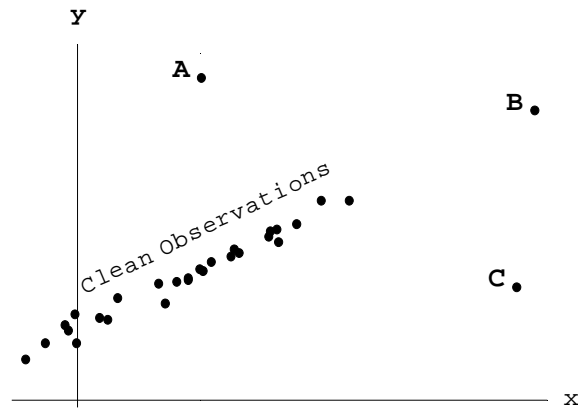


Figure 1. 27 clean and 3 suspected points in a simple regression example.

2.1 HS procedure

HS proposed by Hadi and Simonoff (1993) is a forward search algorithm. It begins with a least squares fit of the regression model to the data. Next, observations are ordered by the adjusted residuals, $a_i = e_i / \sqrt{1 - h_{ii}}$

where e_i is the OLS residual. The first $p+1$ observations constitute a basic subset. This subset is iteratively increased to a clean subset M of size $h = \text{integer part of } (n+p-1)/2$ by using the scaled residuals. Then, the absolute values of d_i are calculated where d_i is the internally studentized residual if the observation is in M or the prediction error if the observation is not in M . The $h+1$ arranged observations in ascending order according to $|d_i|$ constitute the new subset M . The procedure continues to add observations until the value of $(s+1)$ th order statistic of $|d_i|$ exceeds $t_{(\alpha/2(s+1), s-p)}$ where s is the size of current subset M . Observations $s+1$ to n are declared as outliers. Our package provides the outliers found by the procedure as a set.

2.2 PY procedure

PY proposed by Pena and Yahoï (1995) searches for breakpoints in the ordered components within the eigenvectors of the influence matrix $M = EDHDE / ps^2$ where H is the hat (projection) matrix, E is the diagonal matrix that has the residuals on the main diagonal, D is a diagonal matrix with the elements $(1 - h_{ii})^{-1}$, and s^2 is the mean square error estimate of the variance. When the ratio of components exceeds 2.5, then all ordered observations after this value are considered as the prospective regression outliers.

The package presents the final set of outliers found by the procedure.

2.3 SMR procedure

SMR which is a clustering based approach is suggested by Sebert et al (1998). In this procedure, the first job is to calculate the Euclidean distances between pairs of standardized predicted values and residuals from an OLS fit of the data. Then, observations are clustered by using single linkage clustering algorithm. Based on Mojena's stopping rule, the cluster tree (dendrogram) is cut at a height of $\bar{h} + 1.25s_h$ and hence the final clusters are formed, where \bar{h} is the average of the tree cluster heights for all $N-1$ clusters and s_h is the unbiased standard deviation of the heights of the $N-1$ clusters. The purpose is to identify the bulk of the data as the inliers. All other observations are considered as outliers. See, Kim and Krzanowski (2007) for a modified version of SMR procedure.

The package provides the prospective clusters and outliers detected by the procedure. The cluster tree which is useful to detect the outlier clusters visually is also presented.

2.4 BCH procedure

BCH proposed by Billor et al. (2006) has two main stages. In the first stage, the high leverage observations in the x -space which is spanned by the independent variables of regression are identified by the algorithms in Hadi (1992) and Hadi (1994) to constitute a clean subset. Then, the procedure uses an iteratively re-weighted least squares routine where the weights are based on the robust distances to derive the final solution. The method is complemented by a simple diagnostic display in which the squared normalized distances are plotted versus the squared normalized residuals. This plot would be very useful in exploring the nature of all the observations, distinguishing among outliers, good leverages, and regular observations. The data points can be interpreted depending on where they locate on this plot. Billor et al. (2006) suggest the following

- The points in the upper right region are both outliers and high leverage points. In other words, they are bad leverages.
- The points in the lower right region are high leverage points but not outliers. In other words, they are good leverages.
- The points in the upper left region are outliers but not high leverage points. In other words, they are vertical outliers.
- The points in the lower left region are good or regular observations.
- The outliers suggested by the BCH procedure are provided by the package as a set. The diagnostic plot of the procedure is also presented.

2.5 LMS method and RZ diagnostic plot

Least median of squares (LMS) which was first introduced by Rousseeuw (1984) as a high breakdown

method for regression minimizes h th largest squared residual and can resist the effect of nearly 50% of contamination in the data where h is integer part of $(n+p+1)/2$. The LMS procedure is first implemented in a computer software package entitled Program for ROBust reGRESSION (PROGRESS) developed by Rouseeuw and Leroy (1987). Rouseeuw and Hubert (1997) prepared a new version of PROGRESS to facilitate its inclusion in S-PLUS and SAS/IML software. The algorithm in PURO is the same as in PROGRESS. This algorithm is to draw sub-sample of p different observations repeatedly. We consider the number of sub-samples as $p \times 500$, as suggested in Rouseeuw and Leroy (1987). If the standardized LMS residual of a data point is outside the interval $[-2.5, 2.5]$ then it is declared as a regression outlier. We need to point out that there is no guarantee that the results of the LMS method will be the same on each run since the random procedure is used. The reader should also be aware of that even if the random algorithm used here takes less time than the brute force algorithms; it may be still time consuming when the sample size and/or the number of the independent variables are high.

The minimum volume ellipsoid (MVE) method introduced by Rouseeuw (1985) seeks the ellipsoid of minimum volume containing at least h cases. After obtaining the MVE estimates, robust distances (RD) to the center of the ellipsoid can be calculated. If the robust distance of a data point exceeds the cutoff value $\sqrt{\chi^2_{p,0.975}}$, then it is labeled as a multivariate outlier.

Rouseeuw and Van Zomeren (1990) calculated robust distances and compared them with Mahalanobis distances for some classical data sets to show the robustness of MVE. Although MVE is generally used to detect multivariate outliers in data sets without a regression structure, when MVE is applied to x -space of regression and its results are combined with those of LMS, the nature of regression outliers can be explored. Rouseeuw and Zomeren's diagnostic plot in which LMS residuals are plotted versus robust distances of MVE does that job well by classifying the data into regular observations, vertical outliers, good leverage points, and bad leverage points. In Figure 2, we see an illustration of the plot we will call RZ diagnostic plot in the rest of the paper. MVE suggests that points to the right of the vertical border line through $\sqrt{\chi^2_{p,0.975}}$ (The points in Region IV, V and VI)

are leverage points, whereas LMS suggests that points outside the tolerance band $[-2.5, 2.5]$ (The points in Region I, III, IV and VI) are regression outliers. Thus, we can expect to see the clean data points in Region II, bad leverages in Region IV and VI, the vertical outliers in Region I and III and finally the good leverages in Region V.

The package provides the outliers explored by LMS method, and the RZ diagnostic plot to display the nature of the regression outliers.

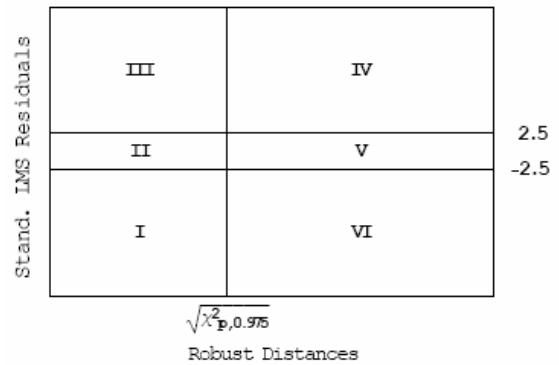


Figure 2. RZ diagnostic plot.

2.6. MS data image

The basic idea of MS procedure proposed by Marchette and Solka (2003) is to map the data into an image which provides a way to detect outliers and clusters within the data visually. In the method, after choosing a suitable distance metric, an $n \times n$ interpoint distance matrix (a_{ij}) is created where the ij th element is the distance between observations i and j , and n is the number of observations. Then, the matrix is transformed to a grayscale picture which is composed of shades of gray, varying from white at zero values to black at the maximum value. To investigate the potential outliers and clusters, one looks for some shades in the image. Generally, a "v" or "+" of dark is indicative of potential outliers. Figure 3 shows the examples of MS displays for two different data sets. On the left data image, the dark "v" indicates the observation 1 is outlier, on the right one, the dark "+" indicates the observations 5, 6 and 7 are outliers.

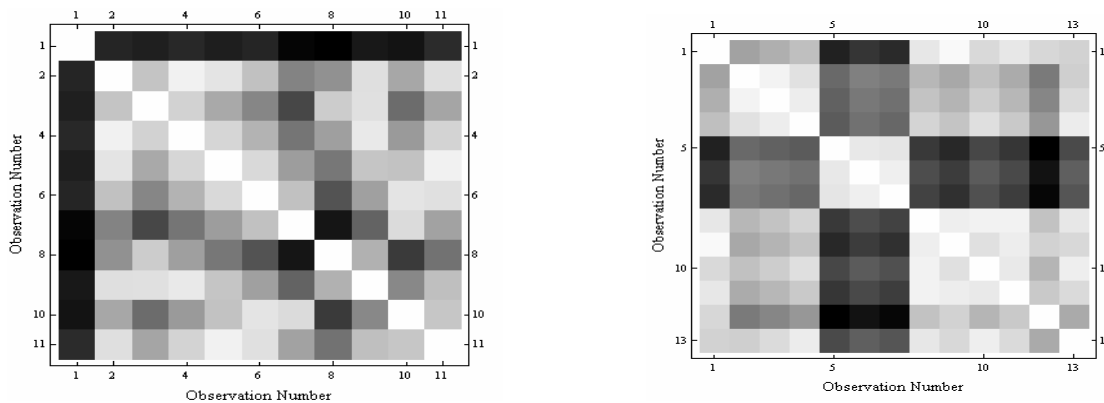


Figure 3: Two examples of MS data images

In the package, MS data image of a data set is given for Mahalanobis distances.

3. THE PACKAGE AND AN APPLICATION

Using the package is very easy, and no serious familiarity with Mathematica is required.

The Mathematica package file PURO.m must be installed in some location so that it can be used by Mathematica Kernel. A simple way of doing this (there are many other ways which can be found in Mathematica web resources) is to copy the file into the

Mathematica \triangleq 7.0 \triangleq AddOns \triangleq Applications

folder on your computer's hard drive. Then, open a Mathematica notebook and load the PURO package by typing the following Mathematica command:

```
<<PURO.m
```

There is only one function to use:

```
DetectOutliers[data,option1,option2].
```

The user must specify two options in the function. The options must be set to **True** or **False**. First option deals with the intercept of the model. When the user selects **True** for this option, PURO performs a regression with a constant term. Otherwise, the program yields a regression through the origin.

As indicated before, two of the seven outlier procedures employed in PURO may be time consuming when the sample size and/or the number of the independent variables are high. These are LMS method and RZ diagnostic plot. In Table 1 and 2, we see the CPU times of the program with and without these two methods for different values of n (the sample size) and p (the number of the independent variables). All times are averaged over 25 runs on a Windows XP Professional system, P4, 2.7 GHz, 512 MB RAM. Naturally, CPU times of PURO increase when n and p are increased. We also see that including LMS procedure and RZ diagnostic plot makes significant differences for large values of n and p . That is not so surprising since LMS and MVE inherently require excessive computations. The second option allows the user to exclude these two methods from the entire set. When the user selects **True** for this option, PURO performs all the procedures. Otherwise, the program does not apply LMS and RZ procedures, and gives the results of the remaining procedures.

In brief, there are four possible usages of the function

1) `DetectOutliers[data, True, True]`: A regression with intercept and all procedures performed
 The data set of 32 observations given in Table 3 is an artificial one created to illustrate the features of the package. As seen from the scatter plot in Figure 5, the majority of the data has a linear pattern. The observation

- 2) `DetectOutliers[data, False, True]`: A regression through the origin and all procedures performed
 3) `DetectOutliers[data, True, False]`: A regression with intercept and LMS, and RZ procedures excluded
 4) `DetectOutliers[data, False, False]`: A regression through the origin and LMS, and RZ procedures excluded

The regression data in *Mathematica* must be a list of vectors. Each vector consists of the observed values of the independent variables and the associated response variable. Write your data as follows:

```
myData={{x11,x12,...,y1},{x21,x22,...,y2},...}
```

where x_{ik} is the value of the i th case of the k th independent variable and y_i is the value of the i th case of the response variable.

When you enter your regression data and type any of the four functions above Mathematica kernel makes all the calculations and gives the prospective outlier sets and graphics.

Table 1. CPU times of PURO (in seconds) without LMS and RZ.

n	p				
	1	2	3	5	10
20	0.54	0.56	0.59	0.61	0.67
50	1.41	1.46	1.54	1.57	1.68
100	3.90	4.05	4.34	5.17	5.48
200	14.57	15.23	16.66	19.21	23.39
500	107.14	118.96	134.44	159.96	228.68

Table 2. CPU times of PURO (in seconds) with LMS and RZ.

n	p				
	1	2	3	5	10
20	1.21	1.74	2.49	4.19	10.81
50	2.27	3.17	4.37	7.22	17.49
100	5.23	6.72	9.30	14.09	32.34
200	16.89	21.35	25.23	35.33	69.82
500	114.59	131.50	145.89	178.99	261.26

28 is a good leverage point, the observation 29 is a vertical outlier, and the observations 30, 31 and 32 are bad leverage points. Let us examine the performances of the standard diagnostics and the outlier detection procedures for this data set.

Table 3. Artificial Data

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x	6.680	13.70	10.66	11.98	9.590	16.32	16.40	12.15	18.85	4.800	6.230	7.230	1.990	13.91	10.02	15.97
y	9.960	17.81	14.67	16.27	14.34	21.07	19.58	16.06	24.63	10.46	10.08	12.52	4.920	18.71	11.92	21.00
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
x	3.150	14.21	11.32	12.60	20.44	4.490	4.310	15.83	4.910	11.25	17.51	37.55	11.37	38.53	39.12	37.64
y	6.910	18.29	15.09	17.05	24.82	8.520	9.170	20.19	7.060	14.85	21.87	43.45	41.22	15.03	13.71	12.63

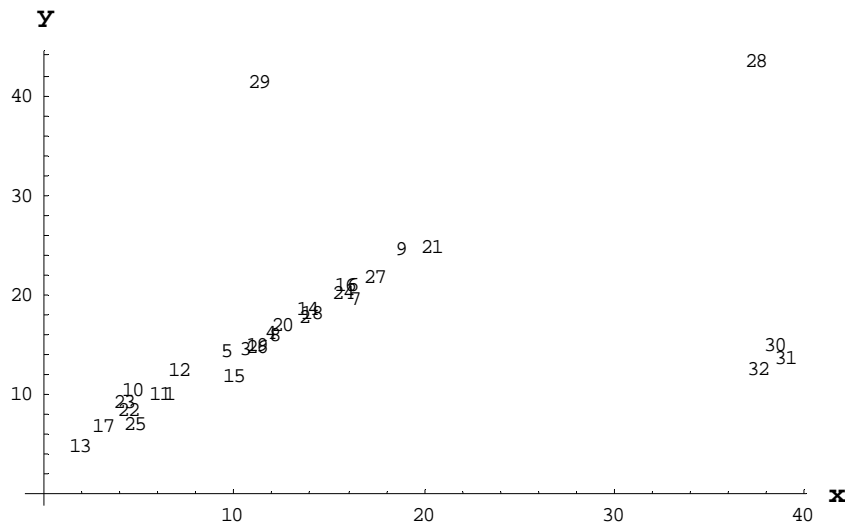


Figure 5. Scatter plot for the artificial data.

There are three submenus of PURO: Procedures, Standard Diagnostics and Visual Diagnostics. Procedures menu has five submenus which show the results of five procedures described in Sections 2.1 through 2.5. In Fig. 6, we see these submenus and the outliers identified by HS procedure. The other procedures suggest the following sets:

- SMR : {{28},{29},{30,31,32}}
- LMS : {10,15,29,30,31,32}
- PY : {28,29,30,31,32}

BCH : {29,30,31,32}

SMR suggests the observations 28, 29, 30, 31, 32 are outliers. It also suggests that these suspected observations form three different groups. In LMS method, all outliers are found and two swappings occur (observations 10 and 15). HS procedure finds all the outliers. The procedure is also successful in excluding the good leverage point 28 from the bad leverage points. PY identifies all the outliers with one swamping. BCH procedure finds all the outliers without a swamping

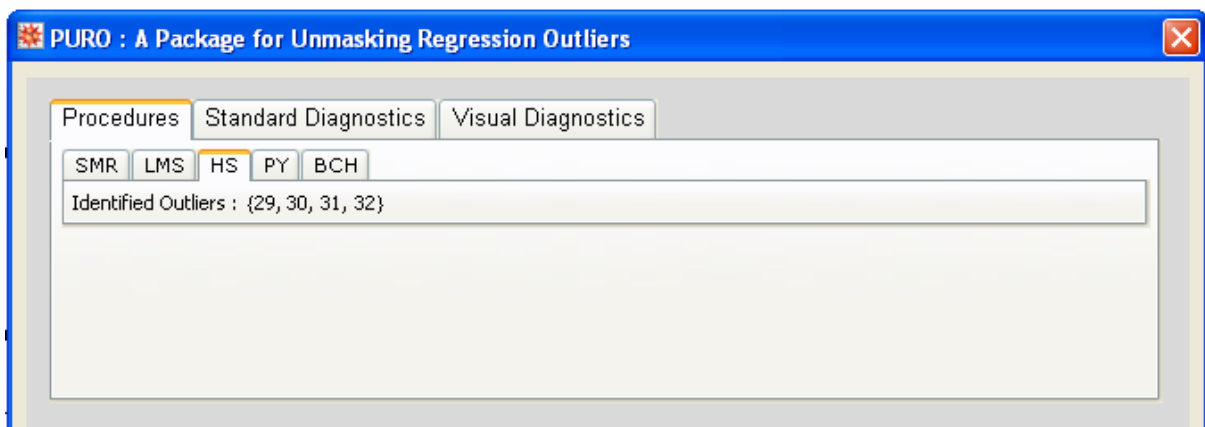


Figure 6. Procedures submenu of PURO and the identified outliers by HS procedure for the artificial data.

In Figure 7, we see the Standard Diagnostics submenu and the outliers identified by COVRATIO diagnostic. The other standard diagnostics suggest the following sets:

Hat Diagonals : {28, 30, 31, 32}

- Cook Distances : {}
- DFBETAS : {28, 29, 30, 31, 32}
- Stud. Res. : {28, 29}
- DFFITS : {28, 29}

Among the six standard diagnostics, Hat Diagonals and DFBETAS seem more successful than the other four

diagnostics.

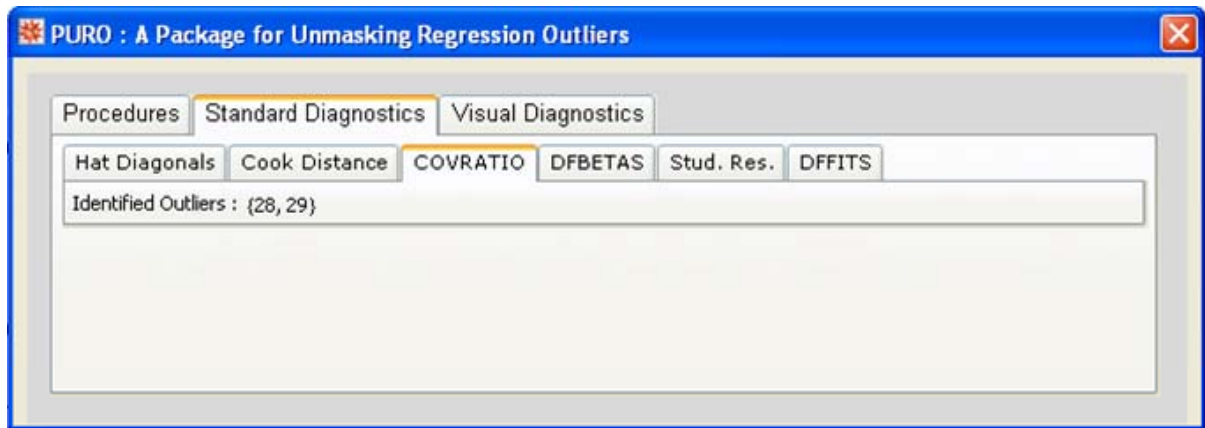


Figure 7. Standard Diagnostics submenu of PURO and the identified outliers by COVRATIO diagnostic for the artificial data

Visual Diagnostics submenu presents four displays: SMR cluster tree, RZ diagnostic plot, BCH diagnostic plot and MS data images. In Figure 8-11, we see these four displays.

Referring to SMR cluster tree, it can be seen that below the red line there are four groups formed. Going across the tree from left to right, Group 1 consists of observations 1-27, Group 2 has only one member which is the data point 28, Group 3 consists of the observations 30, 31 and 32, and finally Group 4 is a one-member cluster (observation 29) like the Group 2. Since the Group 1 contains the majority of the observations, this subset is the regular (inlying) observations. Thus the observations 28, 29, 30, 31, 32 forming the other groups are identified as the outliers. RZ diagnostic plot in Figure

6 finds all the outliers. Recalling Figure 2, it can be seen that it is also successful in displaying the nature of the data, distinguishing among bad leverage points, vertical outliers, good leverage points, and the regular points. As mentioned in the above passages, BCH identifies all the outliers. The diagnostic plot of the BCH procedure in Figure 10 displays clearly the nature of all the data points. As one should expect, the regular observations 1-27 are in the lower left region, vertical outlier 29 is in the upper left region, good leverage point 28 is in the lower right region and finally all the bad leverage points 30, 31 and 32 are in the upper right region. MS data images in Figure 11 seems successful in identifying the outliers since there is a dark "v" including the observations 28-32 in the upper right corner of the plots. There is one swamping in the procedure.

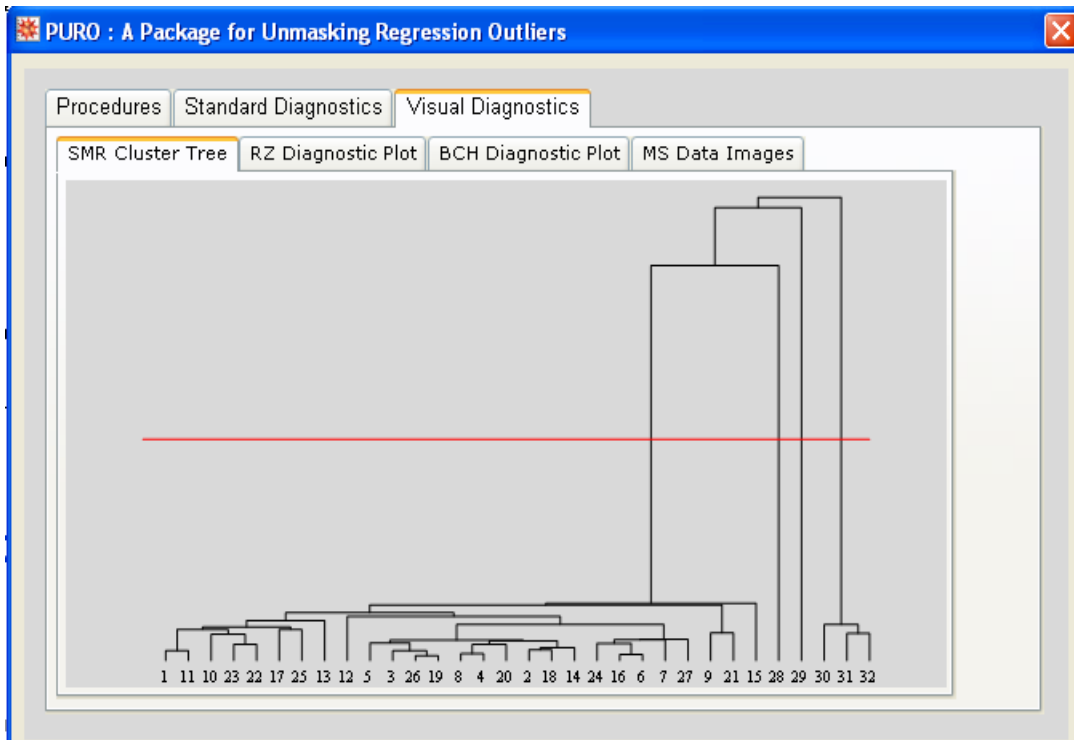


Figure 8. Visual Diagnostics submenu of PURO and the SMR cluster tree for the artificial data.

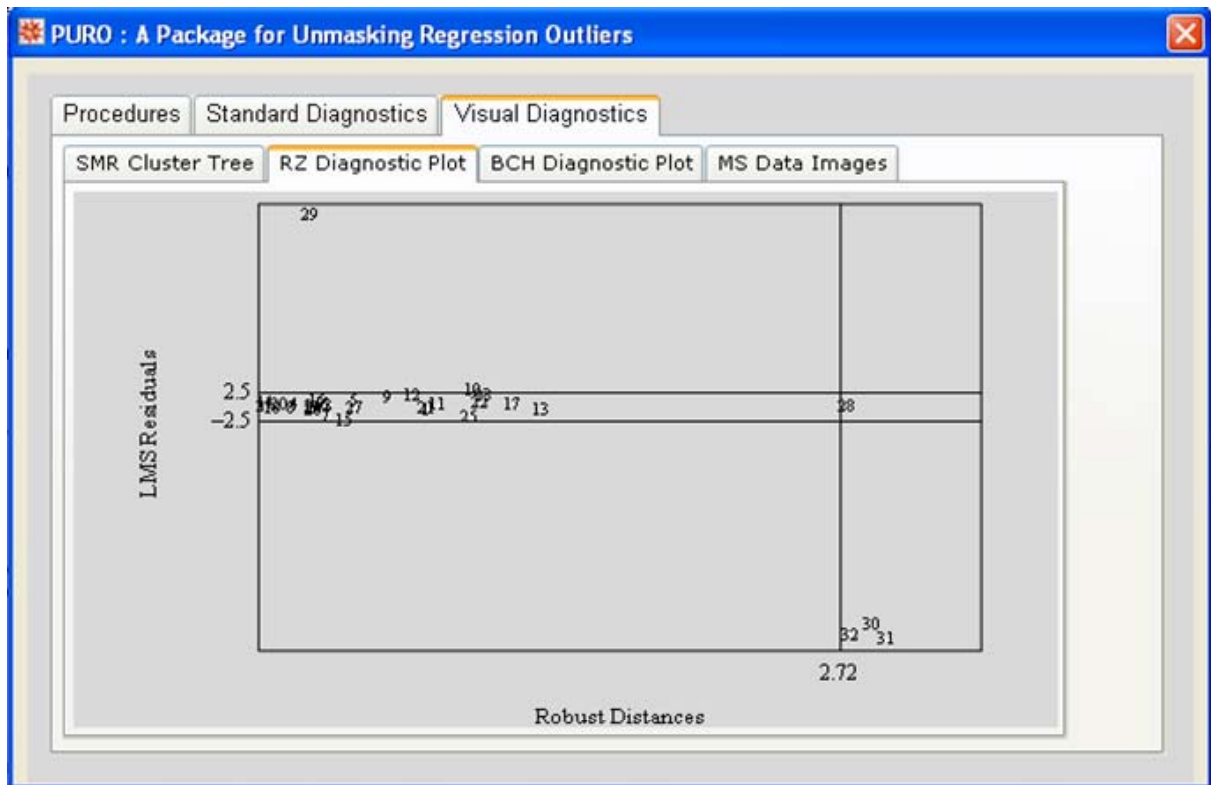


Figure 9. RZ diagnostic plot for the artificial data.

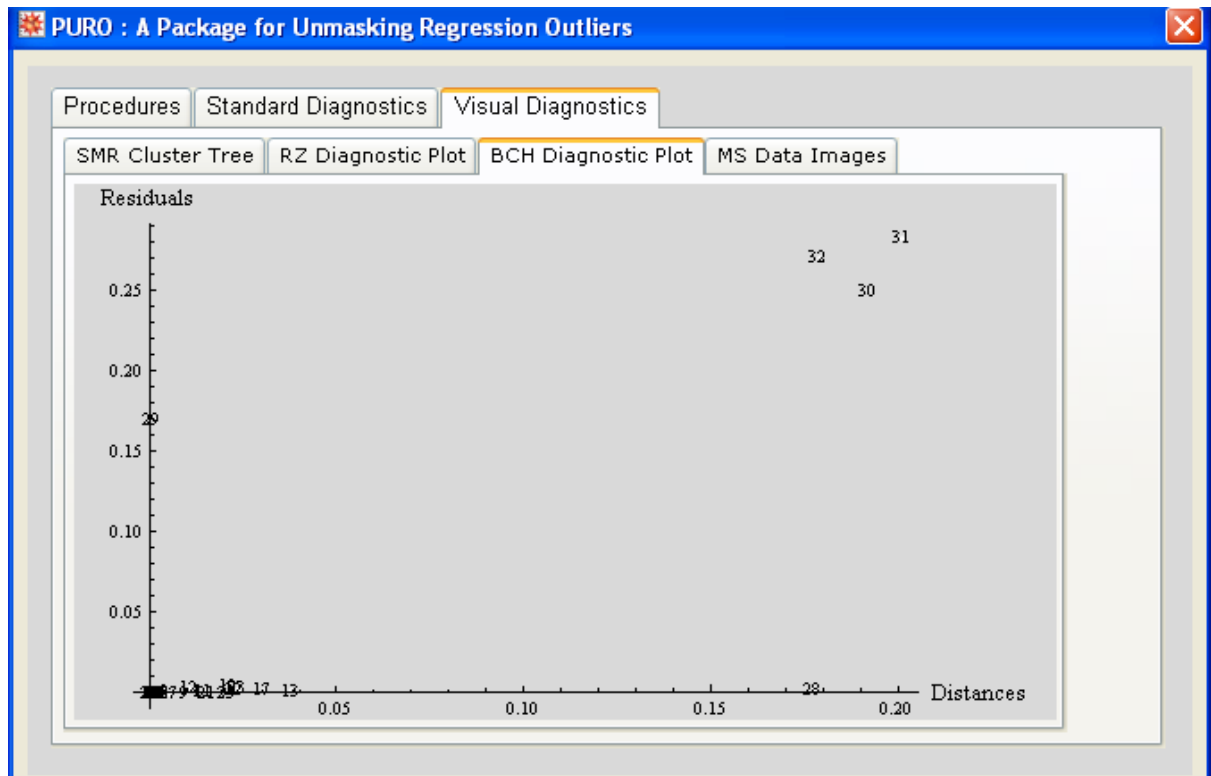


Figure 10. BCH diagnostic plot for the artificial data.

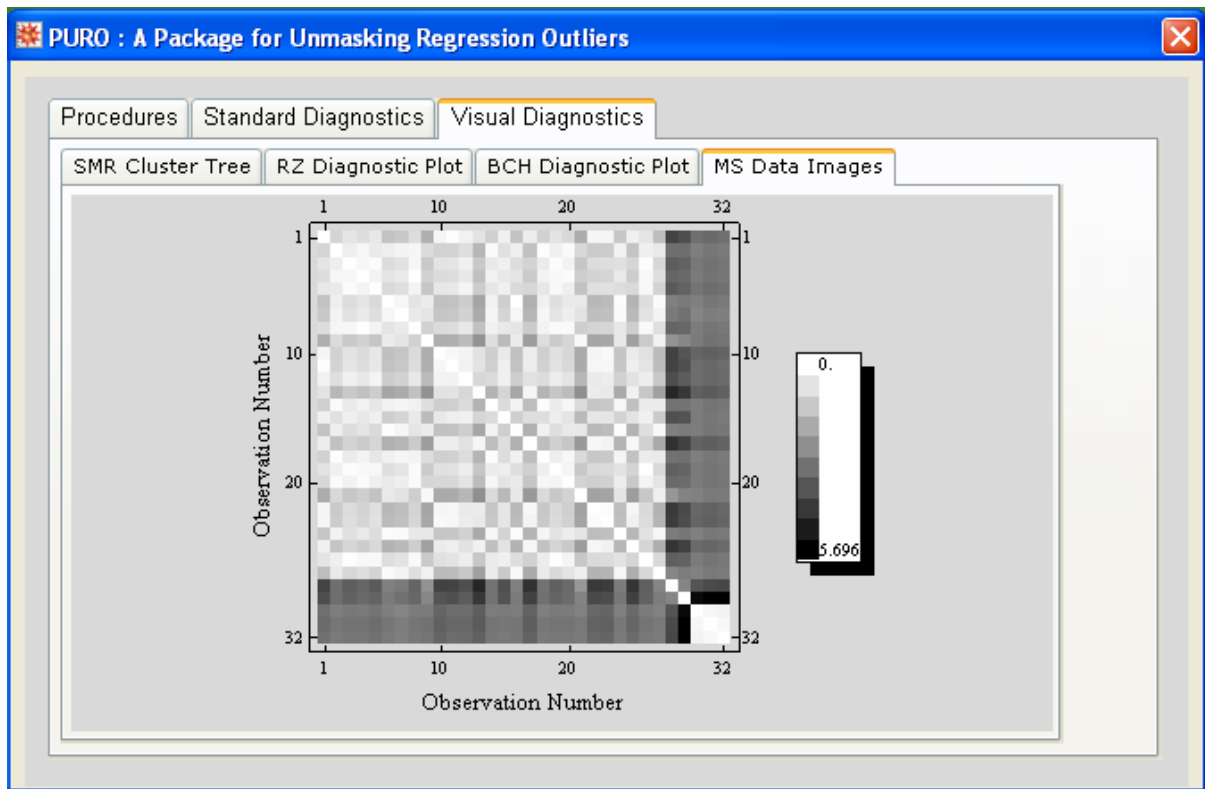


Figure 11. MS Data images of interpoint distances matrix for the artificial data.

4. CONCLUDING REMARKS

In this paper, we presented our package PURO which implements some reliable outlier detection procedures in statistics literature. These procedures include both direct methods from algorithms and indirect methods from robust regression estimators. They also include purely graphical methods.

The procedures HS, PY, LMS and MS identify the outliers without distinguishing among vertical outliers, good leverages, and bad leverages, whereas both RZ and BCH are proposed to classify the data into clean observations, vertical outliers, bad leverage points, and good leverage points. Although SMR procedure is not proposed to display the nature of outliers, each outlier cluster found by the procedure possibly include different type outliers.

Although the outliers identified from each procedure may not be necessarily the same, the user would expect a general concordance between the procedures. For instance, the observations identified by both RZ and BCH diagnostic plots as the vertical outliers, good leverages and bad leverages will possibly lie in the different groups of SMR procedure, as in the application for the artificial data in the previous section.

In the final analysis, the user can take initiative in making decisions about the outliers in her data by combining all the results to one resulting set.

REFERENCES

- [1] Belsley, D.A., Kuh, E. Welsch, R., "Regression Diagnostics". *Wiley and Sons*, N.Y. and Toronto. (1980).
- [2] Billor, N., Chatterjee, S., Hadi, A.S., "A Re-Weighted Least Squares Method for Robust Regression Estimation", *American Journal of Mathematical and Management Sciences*, 26, Nos. 3&4, 229-252 (2006).
- [3] Billor, N., Kiral, G., "A comparison of multiple outlier detection methods for regression data", *Comm. in Stat.*, 37, 521-545 (2008).
- [4] Cook, R.D., Weisberg, S., "Residuals and Influence in Regression. London: *Chapman and Hall* (1982).
- [5] Hadi, A.S., "Identifying multiple outliers in multivariate data", *J. Roy. Statist. Soc. Ser. B* 54, 761-777 (1992).
- [6] Hadi, A.S., Simonoff, J.S., Procedures for the identification of multiple outliers in linear models, *J. Amer. Statist. Assoc.*, 88, 1264-1272 (1993).
- [7] Hadi, A.S., "A modification of a method for the detection of outliers in multivariate samples", *J. Roy. Statist. Soc. Ser. B* 56, 393-396 (1994).
- [8] Kianifard, F., Swallow, W., A Monte Carlo comparison of some procedures for identifying outliers in linear regression, *Commun. Statist. Part A Theory Methods*, 19, 1913-1938 (1990).
- [9] Kim, S., Krzanowski, W.J., , Detecting multiple outliers in linear regression using a cluster method combined with graphical visualization, *Comp. Stat.*, 22, 109-119 2007
- [10] Marchette, D.J., Solka, J.L., 2003, Using Data Images for Outlier Detection, *Comput. Statist. Data Anal.*, 43, 541-552.

- [11] McCullough, B. D., 1998 Assessing the reliability of statistical software: Part I. *The American Statistician* 52, 358–366.
- [12] McCullough, B. D., 1999, “Assessing the reliability of statistical software: Part II”. *The American Statistician* 53, 149-159.
- [13] McCullough, B. D., 2000, The Accuracy of *Mathematica 4* as a Statistical Package., *Computational Statistics* 15 , 279-299.
- [14] McCullough, B. D. , Heiser, D. A., 2008, On the accuracy of statistical procedures in Microsoft Excel 2007, *Computational Statistics and Data Analysis* 52, 4570-4578.
- [15] Pena, D., Yahoï, V.J., 1995, The detection of Influential Subsets in Linear Regression by Using an Influence Matrix, *J. Roy. Statist. Soc. Ser. B* 57(1), 145-156.
- [16] Rousseeuw, P.J, 1984, Least median of squares regression, *J. Amer. Statist. Assoc.* 79, 871-881.
- [17] Rousseeuw, P.J, 1985, Multivariate Estimation with High Breakdown Point, *Mathematical Statistics and Applications*, ed. by W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Dordrecht: Reidel Publishing Company, 283-297.
- [18] Rousseeuw, P.J., Leroy, A.M., 1987, “Robust Regression and Outlier Detection”, *Wiley*, New York.
- [19] Rousseeuw, P.J, Van Zomeren, B.C., 1990, Unmasking multivariate outliers and leverage points, *J. Amer. Statist. Assoc.*, 85, 633-639.
- [20] Rousseeuw, P.J, Huber, M., 1997, URL. <http://www.agoras.ua.ac.be/abstract/Recdev97.htm>
- [21] Sebert, D.M., Montgomery, D.C., Rollier, D., 1998, A clustering algorithm for identifying multiple outliers in linear regression, *Comput. Statist. Data Anal.* 27, 461-484.
- [22] Wisnowski, J.W. , Montgomery, D.C., Simpson, J.R., 2001, “A comparative analysis of multiple outlier detection procedures in the linear regression model”, *Comput. Statist. Data Anal.*, 36, 351–382.