{cikhaj.tex July 31, 1998
prepared for XIII Summer School of Biometry in Cikhaj (Brno), Czech Republic, 21–25. 09. 1998 }

FINDING REGRESSIONAL OUTLIERS BY DYNAMIC PROJECTIONS

Anna Bartkowiak
Institute of Computer Science, University of Wrocław
Przesmyckiego 20, Wrocław 51-151, Poland,
e-mail: aba@ii.uni.wroc.pl

Abstract

Atypical observations hidden in the data may play quite an disastrous role in a fitted regression, especially when commonly used outlier detection techniques like computing leverages, Mahalanobis distances, ordinary and studentized residuals, DFFits, cross-validations – do not detect them.

However (multivariate) outliers can be detected quite easily by graphical techniques, e.g. scatterplot matrices, spin plots or dynamic projections using the grand tour method. We find here specially useful the grand tour.

When we know about the outliers and their location in the multivariate space, then we may account for them by special regression models.

We illustrate our considerations using the Modified Wood Gravity data from Rousseeuw and Leroy (1987).

Key words and phrases: multivariate outlier, regression model, rotation, projection, dynamic graphics

1 Introduction, short review of methods for finding outliers

The role and influence of atypical observations, called also outliers, has been considered since long by many statisticians and data analysts. The principal questions are: How to detect such atypical observations, how to define outliers, how to detect them in the given data, how to asses their influence [1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 13, 14, 15, 16].

Graphical methods have been always considered as a great help when identifying the outliers and their impact in the established model.

Classically the graphical methods are used for constructing various kinds of static scatterplots displaying various kinds of residuals or so called regression diagnostics (leverages, Cook distances, Mahalanobis distances, DFFits, DFBetas, etc.). Eventually these diagnostics are visualized in a dynamic spin plot (spinner) displaying 3 variables.

Quite recently a possibility of viewing truly multivariate data using sequentially fast rotations and projections has been offered [17, 18, 6, 8]. Generally the method is called the grand tour.

In all these works a non contaminated subset of the data was sought, i.e. a subset that does not contain any outlier.

The Least Median method (LM) described in Rousseeuw and Leroy [15] was supposed to yield one such subset. Another interesting proposals were given a.o. by Atkinson [1, 2, 3] – by introducing so called stalactite plots, Hadi and coauthors [11, 12, 13, 14] – aiming at a 'clean' subset, Rocke and Woodruff [16] – proposing a hybrid method from the two mentioned, and Bartkowiak & Szustalewicz [6, 8] using the grand tour method.

In the following we will concentrate on the grand tour method.

2 Data used for illustrations

We will use for our illustrations the modified data on Wood Specific Gravity. These data may be found in [15], Table 8. They contain data for a regression of one variable (y, wood specific gravity) on 6 predictors. The data come from a larger data set, however the set in the book [15] contains only $n = 20$ data vectors, from which four (no. 4, 6, 8, 19, – called in the following FOUR) were deliberately contaminated.

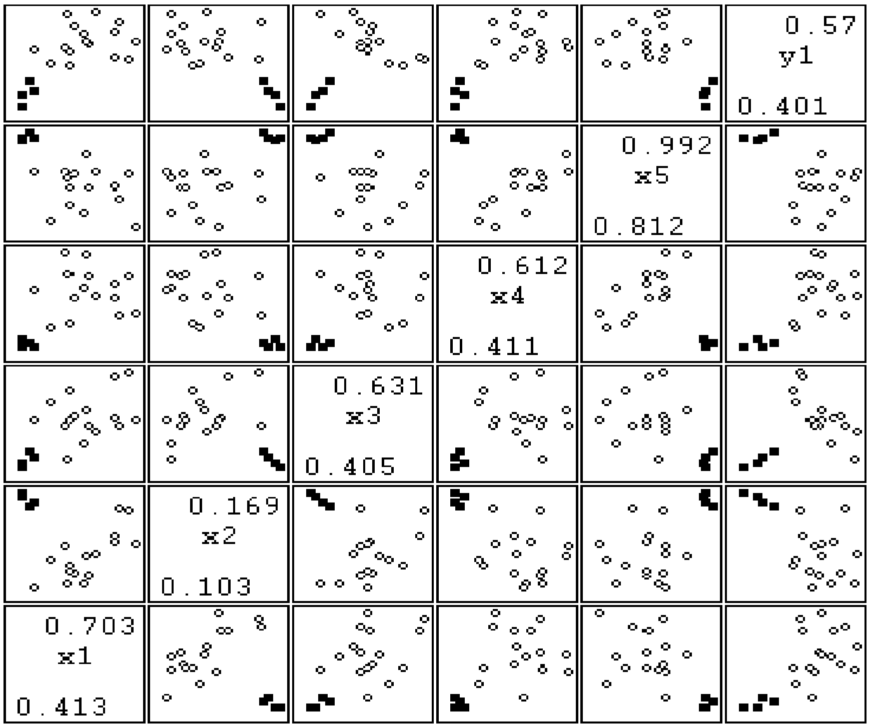

Figure 1. Scatterplot matrix illustrating the Specific Wood Gravity data. Points belonging to the FOUR subset are highlighted

Rousseeuw and Leroy show in their book, that the traditional methods, like the leverages, Mahalanobis Distances, ordinary and studentized residuals, also robustified by the leaving–one–out technique do not indicate for any atypical or influential data vectors. The DFFITS methods points to quite other observations (no. 7, 11, 12) as influential. Only, when using the robust Least Median (LM) method Rousseeuw & Leroy have found that the indicated four data vectors are very specific (are big outliers) and are very influential for the regression.

However, the LM method is very computer intensive. Cook and Hawkins (1990, quoted after Hadi & Velleman [14]) needed for the Wood Gravity data 57,000 samples to obtain by the LM algorithm the sought solution.

Then, the LM algorithm does not find always all the hidden outliers.

Briefly speaking, it does not solve all the problems and does not answer all the questions.

Since then the Wood Gravity data set has received much attention and was discussed, among others, by [1, 2, 16, 14], who proposed other, not so computer intensive methods.

One may get some idea about this data set by looking at the scatterplot matrix presented in Figure 1. When looking at this figure one may note, that the points belonging to the subset FOUR held a special position. They are marked in Figure 1 by filled squares (remind that these are the points specially modified by Rousseeuw & Leroy).

3 Finding outliers using the grand tour method

The used hereafter grand tour method is fully described in [6, 8]. Here we will only say, that it works making sequentially fast projections of the multivariate data cloud.

The data are at the begin standardized to obtain centered and reasonably scattered projections. Then we start to execute the algorithm, which proceeds in the following steps.

1. Generate the rotation plane $<UOV>$ passing through the center 0 and two randomly generated points U, V located on the unit hypersphere $\in R^p$, the p dimensional feature space.
2. Using U, O, V calculate the rotation matrix $\mathbf{A}$.
3. Using the matrix $\mathbf{A}$ rotate the data matrix $\mathbf{X}$ obtaining rotated coordinates of the data points $\in R^p$ (rotation is equivalent to make a transformation).
4. Project the rotated points onto the plane $<\tilde{x}_1, \tilde{x}_2>$ spanned by the horizontal ($\tilde{x}_1$) and vertical ($\tilde{x}_2$) axes of the screen.
5. Draw in the plane $<\tilde{x}_1, \tilde{x}_2>$ a concentration ellipse calculated from coordinates $\{\tilde{x}_{i1}, \tilde{x}_{i2}\}$ of the visible (in the projection plane) point projections $\tilde{P}_i$, $(i = 1, \ldots, n)$. The ellipse may be drawn as a robustified one.
6. Notify, which points are found outside the concentration ellipse.

The steps 1–6 are carried out continuously several hundred times. The points found outside the concentration ellipse are suspected to be atypical (i.e. outliers). The fact of being noticed beyond the border of the concentration ellipse is recorded in a neighboring count plot.

Figure 2 shows three snapshots from a run of the described algorithm – when considering the Wood Gravity data and taking all the 6 variables.

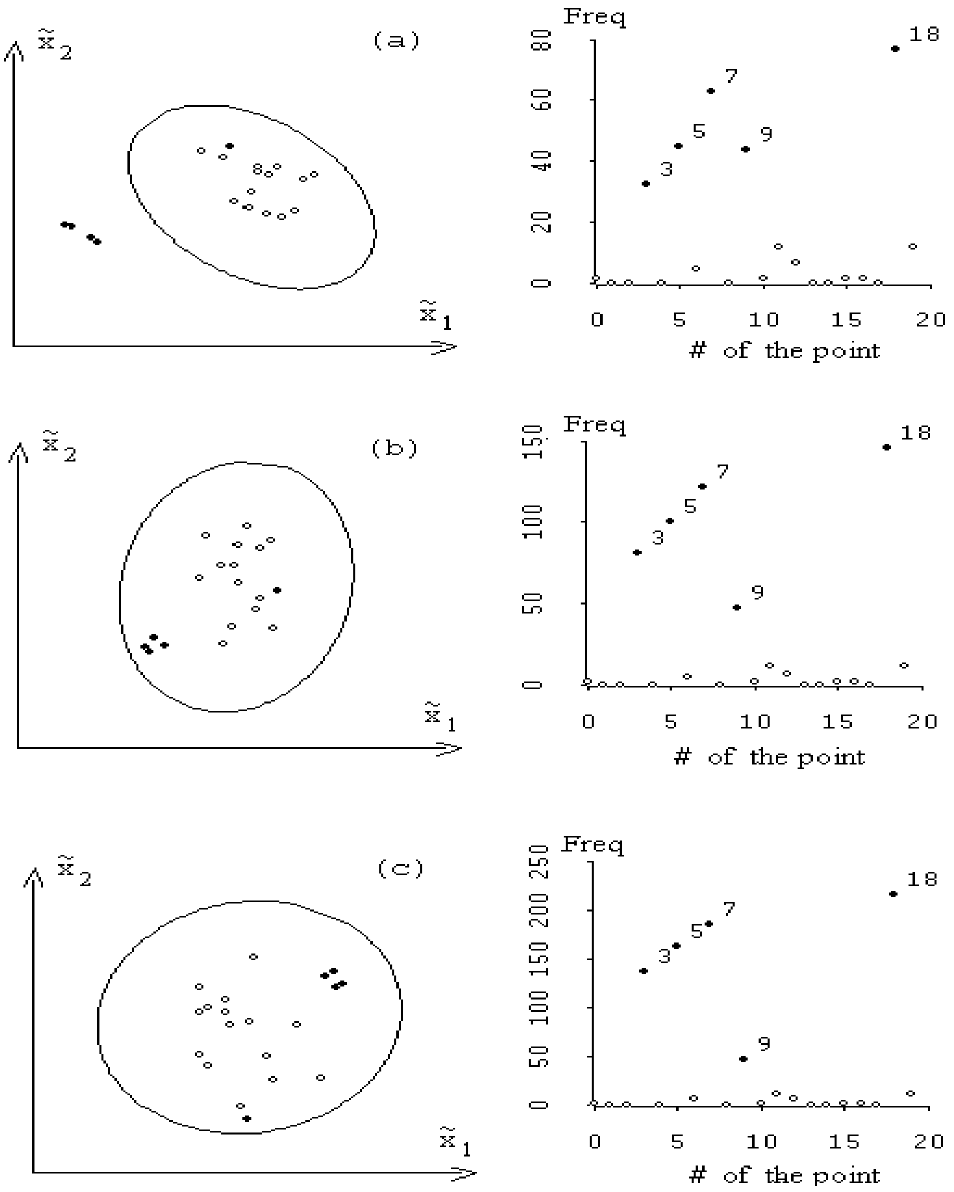

Figure 2. Three snapshots from the grand tour projections obtained for the Wood Gravity data.
Left: Projection planes $<\tilde{x}_1, \tilde{x}_2>$ containing point projections, with robustified concentration ellipse superimposed.
Right: Count plots recording how many times the given point has been notified outside the concentration ellipse.
Notice, that **the numeration of points starts in the plots from 0!**

The left plots of Figure 2 show the point projections visible in the projection plane $<\tilde{x}_1, \tilde{x}_2>$ — in 3 selected phases of run of the grand tour algorithm.

In the plot (a) we see just, how the points belonging to the FOUR are far apart from the main cloud of the data.

In the plot (b) they keep still an outstanding position, however their separation from the other data points is very mediocre (this may, in a sense, correspond to their position visible in the scatterplot matrix shown in Figure 1).

In the part (c) the subset FOUR is in fact absorbed in the entire data cloud.

The count plots visible in the right part of Figure 2 show, that really the points belonging to the FOUR were notified quite frequently outside the concentration ellipse.

4 Regression in presence of outliers

We know already that the analysed data set contains 4 serious outliers. Will they be discovered by an analysis of the residuals?

In the following we present the results of the regression analysis carried out in 3 variants:

I. Entire data set.

II. Dropping from the entire data set the FOUR subset.

III. Adding to the model special variable accounting the outliers.

In the first (I) variant we calculate the regression in the traditional way using the model:

$$y = b_0 + b_1 x_1 + \ldots + b_5 x_5 + e,$$

using for the estimation all 20 data vectors.

In the second (II) variant we use the same model, however for the estimation of the parameters $b_0, b_1, \ldots, b_5$ we drop from the entire data set the FOUR subset, and carry out the estimation with the 16 remaining data vectors.

In the third (III) variant we consider the following model:

$$y = b_0 + b_1 x_1 + \ldots + b_5 x_5 + b_\gamma \gamma + e,$$

with γ being an artificial variable defined as follows ($i = 1, \ldots, n$):

$$\begin{aligned} \gamma_i &= 1, \text{ for } i \in FOUR, \\ &= 0, \text{ for } i \notin FOUR. \end{aligned}$$

The variable γ accounts for the special effect of the four outliers. We know from the graphical analysis that they belong to one group – thus we introduce only one additional variable. If the outliers were subdivided into several distinct subgroups then we would use for each outlier group a separate additional variable (whether the outliers appear in several subgroups, may be seen in the grand tour visualization).

The goodness of fit of these 3 models is shown below:

Variant	R^2	$\hat{\sigma}$	df	F
I. Entire data	0.808	0.0241	14	11.81
II. Dropping FOUR	0.958	0.0074	10	46.00
III. Additional variable γ	0.9435	0.0136	13	36.17

Looking at the results above we state firstly a high value of R^2 (the squared multiple correlation coefficient). The F values indicate for a highly significant fit of all the assumed regression models.

It is striking that after dropping the FOUR subset the goodness of fit improves much (the residual standard deviation $\hat{\sigma}$ drops from 0.0241 to 0.0074).

Introducing addition variable γ (variant III) improves also the fit, as compared to the basic (I) model; however we do not attain the goodness of the second (II) model.

Let us look more closely at the residuals as put against the fitted values. They are shown in Figure 3.

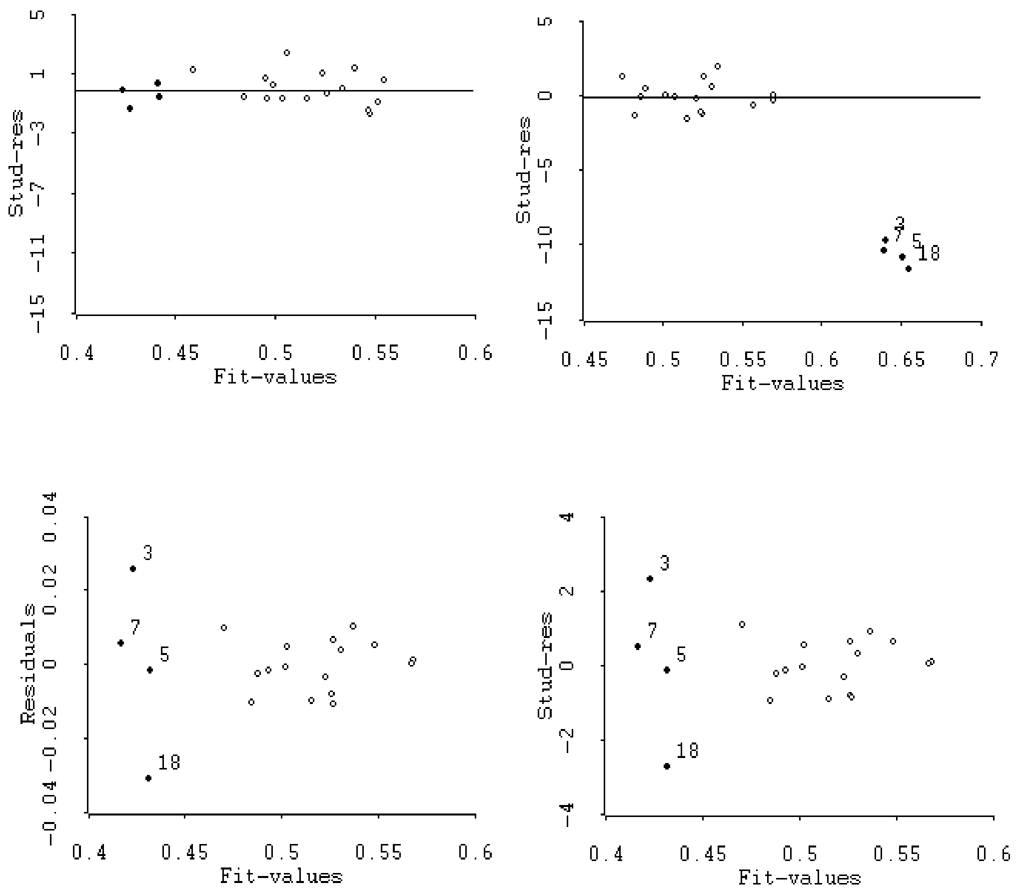

Figure 3. Studentized residuals against fitted values for variants I and II of the calculations (upper plots), and for variant III (lower plots). In the lower left plot also the ordinary residuals are shown

The Studentized residuals put against fitted values – are shown in Figure 3, upper two plots. The left plot in Figure 3 shows the results obtained when executing variant I. The studentized residuals for the FOUR subset does not differ whatsoever from the remaining data points.

However, the residuals evaluated from regression obtained in the II variant of calculations show a striking difference for the FOUR and the remaining data points, which may be seen the upper right plot of Figure 3.

The lower plots in Figure 3 show the ordinary and the studentized residuals put against fitted values obtained when using the variant III with the additional variable γ accounting for the outliers. Here the residuals for the FOUR (especially for the points #3 and #18) differ from the others, which indicates for an inadequacy of the model.

5 Final Remarks

The grand tour method permits quite easy to identify atypical multivariate observations. This might be of great help for the data analyst, who may then use a regression model accounting for the outliers.

Other method of making predictions (calculate the fitted values) in presence of outliers is provided by neural networks. Especially useful are here neural networks with radial basis activation functions (r.b. network) [7]. However, neural network does not permits for identifying atypical data vectors and is deficient in this respect.

For the regressional calculations the XLispStat module R-code [9] was used. It permits easily for deleting indicated subsets of data vectors and recalculating the modified models. The Figures 1 and 3 were obtained using that module.

The grand tour evaluations were done also by a module in XLispStat [18], so were also obtained the plots shown in Figure 2.

References

[1] Atkinson, A.C., Stalactite plots and robust estimation for the detection of multivariate outliers. In: *New Directions in Statistical Data Analysis and Robustness*, Eds. S. Morgenthaler, E. Ronchetti, and W.A. Stahel, Basel, Birkhäuser, 1993.

[2] Atkinson, A.C., & Mulira, H.M., The stalactite plot for the detection of outliers. *Statistics and Computing*, 3, 27–35, 1993.

[3] Atkinson, A.C., Fast very robust methods for the detection of multiple outliers. *JASA*, 89, 1329–1339, 1994.

[4] Barnett, V., & Lewis, T., *Outliers in Statistical Data*. 3rd Edition, Wiley, Chichester, 1994.

[5] Bartkowiak, A., Some basics for detecting multivariate outliers in regressional context. *Biocybernetics and Biomedical Engineering*, 17, 57–83, 1997.

[6] Bartkowiak A., Szustalewicz A., Detecting outliers by a grand tour. *Machine Graphics & Vision*, 6, 487–505, 1997.

[7] Bartkowiak A., Szustalewicz A., Predictions in presence of outliers: which tools to choose? *Intelligent Information Systems VII, Proceedings of the Workshop held in Malbork, Poland, June 15–19, 1998*, 267–270, 1998.

[8] Bartkowiak, A. & Szustalewicz, A., Watching steps of a grand tour implementation, *Machine Graphics & Vision*. Manuscript 1–18, submitted, 1998.

[9] Cook R.D., Weisberg S., *An Introduction to Regression Graphics.* Wiley, New York, 1994.

[10] Gnanadesikan, R., *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York, 2nd Edition (1st edition 1977), 1997.

[11] Hadi, A.S., Identifying multiple outliers in multivariate data. *J. R. Statist. Soc.,* B, 54, 761–771, 1992.

[12] Hadi, A.S. & Simonoff, J.S., Procedures for the identification of multiple outliers in linear models. *JASA*, 88, 1264–1272, 1993.

[13] Hadi, A.S., A modification of a method for the detection of outliers in multivariate samples. *J. R. Statist. Soc., B,* 56, 393–396, 1994.

[14] Hadi, A.S. & Velleman, P., BACON: Blocked adaptive computationally–efficient outlier nominators. Manuscript 1–23, 1998.

[15] Rousseeuw P.J., Leroy A.M., *Robust Regression and Outlier detection*. Wiley, N.Y. 1987.

[16] Rocke, D.M. & Woodruff, D.L., Identification of outliers in multivariate data. *JASA*, 91, 1047–1061, 1996.

[17] Swayne D.F., Cook D., Buja A., Interactive dynamic graphics in the X Windows system with a link to S. In: *ASA Proceedings of the Section on Statistical Graphics*, 1–8, 1994.

[18] Tierney, L., *LISP–STAT, an Object–Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley, New York, 1990.