OUTLIERS IN REGRESSION

Dagmar Blatná

Introduction

A observation that is substantially different from all other ones can make a large difference in the results of regression analysis**.** Outliers occur very frequently in real data, and they often go unnoticed because nowadays much data is processed by computers, without careful inspection or screening. Outliers may be a result of keypunch errors, misplaced decimal points, recording or transmission errors, exceptional phenomena such as earthquakes or strikes, or members of a different population slipping into the sample.

Outliers and leverage

Outliers play important role in regression. It is common practice to distinguish between two types of outliers. Outliers in the response variable represent model failure. Such *observations are called* ***outliers****. Outliers with respect to the predictors are called* ***leverage*** **points**. They can affect the regression model, too. Their response variables need not be outliers.

In regression it helps to make a distinction between two types of leverage points: good and bad. ***A good leverage point*** is a point that is unusually large or small among the *X* values but is not a regression outlier. That is, the point is relatively removed from the bulk of the observation but reasonably close to the line around which most of the points are centered. A good leverage point has limited effect on giving a distorted view of how majority of points are associated. Good leverage points improve the precision of the regression coefficients.

A bad leverage point is a point situated far from the regression line around which the bulk of the points are centered. Said another way, a bad leverage point is a regression outlier that has an *X* value that is an outlier among *X* values as well (it is relatively far removed from the regression line). Bad leverage point has grossly effect estimate of the slope of the regression line if an estimator with a small breakdown point is used. Bad leverage points reduce the precision of the regression coefficients.

Outliers are always identified with respect to a specific benchmark or null model. Numerous difficulties can arise during the outlier identification stage. The most notorious one is the masking effect. If there are several outliers grouped close together in a region of the sample space far away from the bulk of the data, most nonrobust outlier detection methods fail to identify these observation as outliers. In other words, the outliers mask one onother. Leverage points do not necessarily correspond to outliers.

Observations whose inclusion or exclusion result in substantial changes in the fitted model (coefficients, fitted values) are said to be influential.

We are mostly concerned with ***regression outliers***, that is, cases for which $(x_{k_1},...,x_{k_p},y_k)$ deviates from the linear relation followed by the majority of the data, taking into account both the explanatory variable and the response variable simultaneously. A leverage point is then still defined as a point $(x_{k_1},...,x_{k_p},y_k)$ for which $(x_{k_1},...,x_{k_p})$ is outlying with respect to the $(x_{i_1},...,x_{i_p})$ in the data set.

Detecting Influential Observations

Many numerical and graphic diagnostics for detecting outliers and influential cases on the fit have been suggested.

Numerical diagnostics

Diagnostics are certain quantities computed from the data with the purpose of pinpointing influential points, after which these outliers can be removed or corrected. When there are only one a single outlier, some of these methods work quite well by looking at the effect of deleting one point at a time.

Unfortunately, it is much more difficult to diagnose outliers when there are several of them, and diagnostics for such multiple outliers are quite involved and often give rise to extensive computation.

The robust regression is extremely useful in identifying outliers. The least median of squares (*LMS*) procedure and the least trimmed squares (*LTS*) regression are reliable data analytic tools that may be used to discover regression outliers both in simple and multivariate situations. Certain robust methods can withstand leverage points, whereas others cannot, and that some diagnostics allow us to detect multiple outliers, whereas others are easily masked.

For identifying outliers and leverage points some measures can be used.

Rousseeuw and van Zomeren (1990) suggest using the *LMS* estimator to detect regression outliers .This method begins by computing ***the residuals associated with LMS regression***

$$s = 1.4826\left(1+\frac{5}{n-p-1}\right)\sqrt{M_r}\ , \tag{1}$$

where M_r is the median of $r_1^2,...,r_n^2$, the squared residuals, p is the number of predictors.

The point $(y_i, x_{i1},...,x_{ip})$ is labeled a regression outlier if the corresponding standardized residual is large. In particular, Rousseeuw and van Zomeren label the *i*-th vector a regression outlier if $(|r_i|/s) > 2.5$.

The ordinary or simple residuals (observed - predicted values) are the most commonly used measures for detecting outliers.

Standardized Residuals are the residuals divided by the estimates of their standard errors. They have mean 0 and standard deviation 1. There are two common ways to calculate the standardized residual for the *i*-th observation. Studentized residuals are a type of standardized residuals that can be used to identify outliers. One uses the residual mean square error from the model fitted to the full dataset (internally studentized residuals). The other uses the residual mean square error from the model fitted to the all of the data except the *i*-th observation (externally studentized residuals). The externally standardized residuals follow at *t* distribution with *n*-*p*-2 df.

The studentized residuals are a first means for identifying outliesrs. Attention should be paid to studentized residuals that exceed +2 or -2 and get even more concerned about residuals that exceed $|2|$ and even yet more concerned about residuals that exceed $|3|$.

Now lets look at the leverage's to identify observations that will have potential great influence on regression coefficient estimates. Generally, a point with leverage greater than (2k+2)/*n* should be carefully examined, where *k* is the numer of predictors and *n* is the number of observations. (In our example this works out to (2.4+2)/28=0.35.)

Some measures combine information on the residuals and leverage and they are general measures of influence.

The robust distance is defined as

$$RD(x_i) = \sqrt{[x_i - \mathbf{T(X)}]^T \mathbf{C(X)}^{-1}[X_i - \mathbf{T(X)}]} \tag{2}$$

where **T(X)** and **C(X)** are the robust location and scatter matrix for the multivariates.

One classical method to identify leverage points is inspects the use of ***the Mahalanobis distances*** MD_i to find outliers x_i:

$$MD_i = \sqrt{(x_i - \mu).\mathbf{C}^{-1}(x_i - \mu)^T} \tag{3}$$

where **C** is the classical sample covariance matrix. In classical linear regression, the diagonal elements h_{ii} of the *hat* matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{4}$$

are used to identify leverage points. The *i*-th leverage $h_i = H_{ii}$ is the *i*-th diagonal element of the hat matrix ***H.*** Rousseeuw and Van Zomeren (1990) report the following monotone relationship between the h_{ii} and MD_i

$$h_{ii} = [((MD_i)^2)/(n\text{-}1)] + [1/n] \tag{5}$$

and point out that neither the MD_i nor the h_{ii} are entirely safe for detecting leverage points reliably. Multiple outliers do not necessarily have large MD_i values because of the *masking effect*.

Rousseeuw and Leroy (1987) suggest using $h_i > 2p/n$ and $MD_i^2 > \chi^2_{p-1;0.95}$ as benchmarks for leverages and Mahalanobis distances.

The ***Cook's distance*** is defined

$$CD_i = (p\sigma^2)^{-1}(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})^T(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}}) \tag{6}$$

where σ^2 is estimator of the error variance

$$\sigma^2 = \sum_{i=1}^{n} r_i^2 / n - p. \tag{7}$$

Cook's distance for the *i*-th observation is based on the differences between the predicted responses from the model constructed from all of the data and the predicted responses from the model constructed by setting the *i*-th observation aside. For each observation, the sum of squared residuals is divided by (*p*+1) times the Residual Mean Square from the full model. Some analysts suggest investigating observations for which Cook's distance is greater than 0.5. The lowest value that Cook's D can assume is zero. The conventional cut-off point is 4/n.

Generally, when the statistics CD_i, h_i and MD_i are large, case *i* may be an outlier or influential case.

Cook's distance, leverages, and Mahalanobis distance can be effective for finding influential cases when a single outlier exist, but can fail if there are two or more outliers. Nevertheless, these numerical diagnostics combined with plots such as residuals versus fitted values and fitted values versus the response are probably the most effective techniques for detecting cases that affect the fitted values when the multiple linear regression model is a good approximation for the bulk of the data.

DFITS$_i$ is the scaled difference between the predicted responses from the model constructed from all of the data and the predicted responses from the model construced by setting the *i*-th observation aside. It is similar to Cook's distance. Unlike Cook's distance, it does not look at all of the predicted values with the *i*-th observation set aside. Some analysts suggest investigating observations for which |DFITS$_i$| is greater than $2\sqrt{(p+1)/(n-p-1)}$. Cook's D and DFITS give similar answers.

Graphic diagnostics

In the simple regression model, one can make a plot of the (x_i, y_i) , which is called a scatterplot, in order to visualize the data structure. Many people will argue that regression outliers can be discovered by looking at the least squares residuals. Unfortunately, this is not true when the outliers are leverage points. If one would apply a rule like "delete the points with largest *LS* residuals", then the "good" points would have to be deleted first. Often, influential points remain hidden, because they do not always show up in the usual LS residual plot.

LS fit can masks the bad points. The *LS* residuals associated with outliers even may lie within a horizontal band. Because of this effect, the interpretation of a residual plot corresponding to the *LS* estimator is dangerous. Residual plots corresponding to robust estimators (*LMS* or *LTS)* are even more useful in problems with several variables.

A scatter plot of *x* versus *y* is used to visualize the conditional distribution *y*|*x*. For the simple linear regression model, by far the most effective technique for checking the assumption of the model is to make ***a scatterplot of x versus Y and residual plot of x versus*** r_i**.** Departure from linearity in the scatterplot suggests the simple linear regression model is not adequate. Points in the residual plot should scatter about the line *r* = 0 with the pattern. If curvature is present or if the distribution of the residuals depends on the value of *x*, then the simple linear model is not adequate.

In the multiple regression model with large p it is not sufficient to look at each variable separately or even at all plots of pairs of variables. The identification of outlying $(x_{i_1},...,x_{i_p})$ is more difficult problem.

To vizualize outliers and leverage points several plots can be used:

- a ***residual plot*** is a plot of a variable W_i versus the residuals r_i. Typically W_i is a linear combination of the predictors.

- ***a forward response plot*** is a scatterplot of the fitted values $\hat{Y}_i$ versus the response Y_i.

- ***a regression diagnostic plot*** is a plot of the standardized residuals of robust regression (*LMS* or *LTS*) versus the robust distances RD_i proposed by Rousseeuw and Van Zomeren (1990). Two horizontal lines corresponding to residual values of +2.5 and -2.5 are useful to distinguish between small and large residuals, and one vertical line corresponding to the $\sqrt{\chi^2_{n;0.975}}$ is used to distinguish between small and large distances.

- ***a plot of the standardized residuals versus their index,***

- ***a plot of the standardized residuals versus fitted values,***

- ***a Normal Q-Q plot of the standardized residuals,***

- ***a Distance-Distance plot***, introduced by Rousseeuw and van Zomeren (1990), displays the robust distances versus the classical Mahalanobis distances. The horizontal and vertical lines are drawn at values equal to the cutoff which defaults $\sqrt{\chi^2_{n,0.975}}$. Points beyond these lines can be considered outliers.

- ***a leverage versus residual-squared plot*** shows the leverage by the residual squared and looks for observations that are simultaneously high on both of these measures. An observation that both has a large residual and large leverage is potentially the most influential. Using residual squared instead of residual itself, the graph is restricted to the first quadrant and the relative positions of data points are preserved. This is a quick way of checking potential influential observations and outliers at the same time.

Example

The data set contains information on 28 members and candidate countries of the EU (year 2002) where the response variable is *internet users per 100 population* (y) and the regressors are: *GDP per capita in PPS* (x_1), *gross domestic expenditure on research and development in per cent* (x_2), *total youth educational attainment level* (x_3), *expenditure on information technology as % of GDP* (x_4). Source of data: Eurostat, CSU.The results were obtained using statistical software packages *SAS 9.1* and *S-Plus 6.2*.

As the first step each regressor was analyzed separately using robust regression method *LTS* with diagnostic tools : *Mahalanobis distance, robust MCD distance, standardized robust residuals.* Results of diagnostics of outliers and leverage points are presented in Table 1.

Table 1 Diagnostics of outliers and leverage points in simple regression

Regressor	*Leverage point*	*Outlier*
GDP per capita in PPS (x_1)	No 7 (Luxemburg)	No 7 (Luxemburg), No 18(Estonia)
youth educational attainment level (x_2),	No 10 (Portugal) No 28 (Malta)	-
expenditure on research and development (x_3)	No 3 (Finland) No 15 (Sveden)	-
expenditure on information technology (x_4)	-	-

LTS regression yields the multiple regression equation in form

$$\hat{y} = 32.2414 + 0.1499x_1 - 0.4377x_2 + 0.3447x_3 + 10.9252x_4$$

Tables 2 and 3 display results of regression diagnostics (outliers and leverage points) based on *LTS* estimates and diagnostics summary.

Table 2 Robust diagnostics

The ROBUSTREG Procedure
Diagnostics

Obs	Mahalanobis Distance	Robust MCD Distance	Standardized Robust Leverage	Residual	Outlier
3	2.1203	4.6762	*	0.4658	
5	2.7368	4.7845	*	0.0430	
7	3.9928	10.1284	*	-0.4542	
10	3.4669	4.8992	*	0.4559	
15	3.0743	6.8864	*	-0.1575	
18	2.2751	2.2826		5.0246	*
25	1.5552	2.1303		3.3943	*

Table 3 Diagnostics summary

Diagnostics Summary

Observation Type	Proportion	Cutoff
Outlier	0.0769	3.0000
Leverage	0.1923	3.3382

R-Square for LTS Estimation

R-Square	0.9316

The comparison of results of simple regression of each separately regressors (Table 1) and results of multiple regression (see Table 2) shows they are fairly different because of possible masking effect. Observation 18 was identified as outlier by using simple regression $y - x_1$ but not in multiple case. Observation 25 was not identified as outlier in simple regression at all. So observation 5 was identified as leverage point only in case of multiple regression.

From possible graphical diagnostics tools a plot of the standardized residuals versus fitted values is presented. in Figure 1.

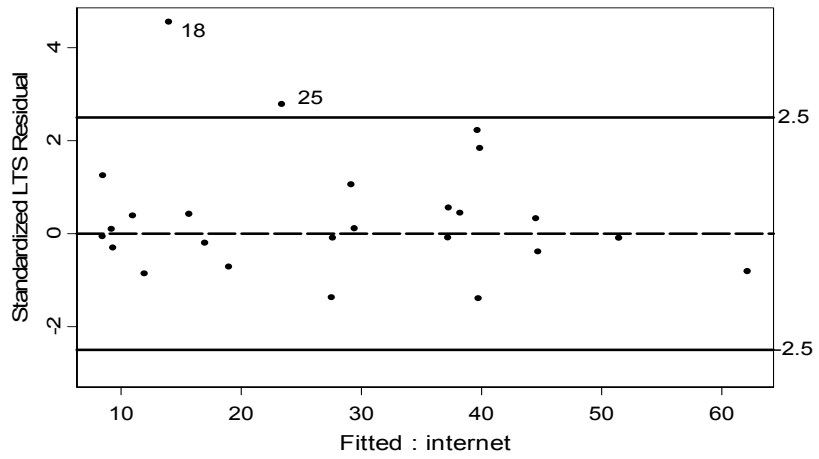

Figure 1

Figure 1 gives evidence of the presence of outlying observations, because two points fall behind the band. Observations 18 (Malta) and 25 (Slovinia)) are identified as outliers.

Figure 2 presented ***a regression diagnostic plot*** (a plot of the standardized residuals of robust regression *LTS* versus the robust distance). We can see that observation 18 (Malta) and 25 (Slovenia) are identified as outliers and observation 3 (Finland), 5 (Ireland), 7 (Luxembourg), 10 (Portugal) and 15 (Sveden) are identified as leverage points. In our example non observation is outlier and leverage point at the same time.

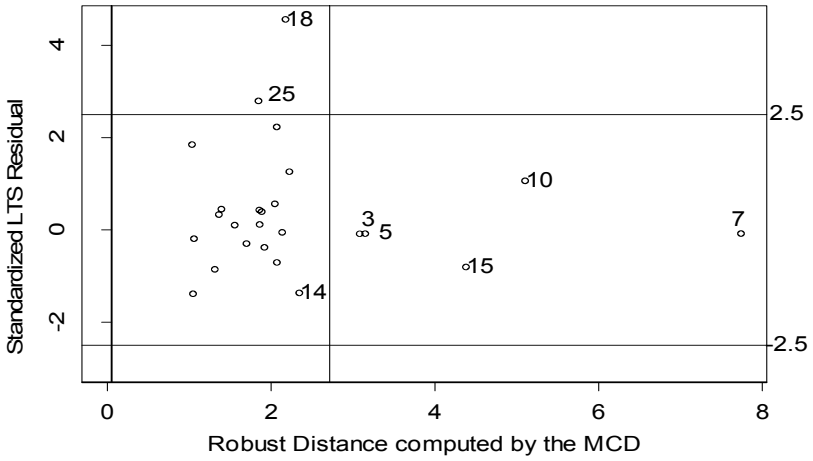

Figure 2

References

[1] Blatná, D.: *Robust Regression*. In: Applications of mathematics and statistics in economy. Wroclav 2005 (in print).

[2] Dallal, G. E.: *Regression Diagnostics*. http://www.tufts.edu/gdallal/

[3] Hampel,F.R.- Ronchetti,E.M.- Rousseeuw, P.J.- Stahel, W.A.: *Robust Statistics. The Approach Based on Influence Functions.* J.Wiley, N.York 1986. ISBN 0-471-82921-8.

[4] Huber, P.J.: *Robust Statistics*. John Wiley, New York 1981.

[5] Olive,D.: *Applied Robust Statistics*. Preprint M-02-006., http://www.math.siu.edu/

[6] *Regression with SAS*. http:www.ats.ucla.edu/stat/sas

[7] *Robust regression*. http://en.wikipedia.org/wiki/Robust_regression.

[8] Rousseeuw, P.J.- Leroy,A.M.: *Robust Regression and Outlier Detection*. J.Wiley, New Jersey 2003. ISBN 0-471-48855-0.

[9] Rousseeuw,P.J.- van Zomeren, B. C.: *Unmasking Multivariate Outliers and Leverage Points.* Journal of the American Statistical Association *85,1990. 633-639*.

[10] SAS OnlineDoc™.Version 8.

[11] *The home of the S-PLUS statistical software package*. http://www.insightful.com/

[12] Wilcox,R.R.: *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press. London 1999. ISBN 0-12-751545-3.

[13] Wilcox,R.R.: *Fundamentals of Modern Statistical Methods*. Springer, New York 2001, ISBN 0-387-95157-1.

Dagmar Blatná

University of Economics Prague
Department of Statistics and Probability
W. Churchill Sq. 4
130 67 Prague 3
Czech Republic
blatna@vse.cz