HIGH-BREAKDOWN ROBUST REGRESSION IN ANALYSIS OF INTERNET USAGE IN EUROPEAN COUNTRIES HOUSEHOLDS

Dagmar Blatná

Abstract

Robust regression methods are acceptable and useful tools for analyzing dependences in data sets with outliers. High-breakdown point regression methods can detect regression outliers, leverage points and influential observations as well. The broadcast internet access of hoseholds (BIAH) is one of the indicators of the information society. Its value depends on the level of economic development, education, employment rate and other relevant factors. The Internet penetration rate as well as the above mentioned criteria vary greatly in the European countries and, consequently, the occurrence of outlying observations can be anticipated in the corresponding analysis. Improper use of the classical least square (LS) regression models with significant variables without accompanying identification of outliers and the assessment of residual normality can lead to the acceptance of an incorrect LS model. Research results obtained by using both high-breakdown point robust regressions and a classical LS regression analysis are being compared in the present paper.

Key words: robust regression, high-breakdown point, outliers, leverage points, influential points

JEL Code: C39, C19, C49

Introduction

The aim of this paper is to demonstrate the applicability and advantages of robust regression methods with a high-breakdown point in an analysis of the European countries' real economic data. As a consequence of great variability of the indicators analyzed, the occurrence of outliers can be anticipated in the corresponding analysis. In such a case, a classical statistical approach – the least squares method – can be highly unreliable, the robust regression methods (namely those with a high-breakdown point) representing an acceptable and useful tool.

1 Methodology

It is a common practice to distinguish between two types of outlying observations in the regression, those in the response variable representing a model failure. Such observations are called either outliers in the *y*-direction or vertical outliers, those with respect to the predictors being labelled as leverage points. Regression outliers (influential points) are the cases for which $(x_{k_1},...,x_{k_p},y_k)$ deviates from the linear relation followed by the majority of the data, both the explanatory and response variable being taken into account simultaneously.

First, let us briefly mention the principles of the robust method used.

MM-estimates (proposed by Yohai,1987) combine a high-breakdown point with good efficiency (approximately 95% to LS under the Gauss-Markov assumption). MM regression is defined by a three-stage procedure (for details, see Yohai, 1987) or Rousseeuw, 2003). At the first stage, an initial robust high breakdown regression estimate is computed; it is consistent, robust, but not necessarily efficient. At the second stage, an M-estimate of the error scale is computed, using residuals based on the initial estimate. Finally, at the third stage, an M-estimate of the regression parameters based on a proper redescending ψ-function is computed. The breakdown value of the MM-estimate is determined by that of its initial estimate taken in the first step, the term "MM" referring to the fact that more than one M-estimates are used for the final estimate calculation.

The least trimmed squares (LTS) estimator (proposed by Rousseeuw, 1984) is obtained by minimizing $\sum_{i=1}^{h} r_{(i)}^2$, where $r_{(i)}^2$ is the *i*-th order statistic among the squared residuals written in the ascending order, *h* is the largest integer between $[n/2]+1$ and $([n/2]+[(p+1)/2])$, *p* is the number of predictors (including an intercept) and *n* is the sample size. The usual choice $h \approx 0.75n$ yields the breakdown point of 25 %; (see Hubert & Rousseeuw & Van Aelst, 2008). The LTS estimate has an asymptotic breakdown point equal to 50 %, but has relatively low efficiency when all observations satisfy the regression model with normal errors. Despite being highly resistant, its efficiency is so low that it is not appropriate as a self-contained, stand-alone estimator. The LTS estimate plays the role of an initial estimate in MM-regression. LTS residuals can also be used effectively in outlier diagnostics. A more detailed description is available in, e.g., (Ruppert & Carroll, 1980), (Rousseeuw, 2003) or (Hubert, &Rousseeuw & Van Aelst, 2008).

S estimation is a high-breakdown value method introduced by Rousseeuw and Yohai (1984), minimizing robust M-estimate of residuals' dispersion. Its breakdown point can also attain 50 % but its efficiency is higher than that of the LTS estimate. S-estimates have essentially the same asymptotic performance as regression M-estimation. The S-estimate plays an analogous role to the LTS estimate, being used as an initial estimate in MM-regression.

LTS and S estimates may be appropriate only when they are supposed just to ensure resistance, not making inference about the population.

Numerical and graphic diagnostic methods for detecting outliers, leverage points and influential observations can be employed. In this paper, the following ones have been used: Residuals associated with LTS regression, Standardized residuals, Studentized residuals (a type of standardized *t* distribution residuals with *n*-*p*-2 Df), Robust distance, Diagnostic plots, Normal Q-Q plot of the standardized residuals and Plot of kernel density of residuals.

In order to select a proper regression model, the following diagnostic tools and selection information criteria were used: significance robust *t*-, *F*- and Wald tests, Robust Akaike's Information Criterion (*AICR*), Robust Bayesian Information Criterion (*BICR*) and Robust Final Prediction Error (*RFPE*); the above criteria being dealt with in, e.g., (Ronchetti, 1985), (Hampel & Ronchetti & Rousseeuw & Stahel, 1996) or SAS and S-Plus manuals.

2 Results of Analysis and Discussion

The level of broadcast internet access of households (BIAH) is one of the indicators of the information society being constructed as a percentage of households with broadband Internet access. The population considered is aged between 16 and 74, the households with at least one member within this age range being included. The analysis is based on 2010 data of 27 EU countries. All the data as well as indicator definitions have been adopted from the Eurostat database. Different economic indicators have been used as explanatory variables, calculations being performed by means of SAS 9.2 and S-Plus 6.2 statistical software.

BIAH values in the European countries depend on numerous factors of economic development, such as economic activity, employment rate, education, social background, etc.

For the BIAH as the dependent variable, a few linear regression models – namely MM, LTS, S and, for better comparison, LS regression – have been tested using robust regression methods with high breakdown point. Identification of vertical outliers, leverage points and influential points was performed using LTS regression. The list of indicators employed is given in the appendix to this paper. The selected models – mutually different from the statistical point of view – are presented, the occurrence and variety of outliers being crucial for their choice. In all tables, t denotes the test statistic related to individual *t*-tests, *p*-value expresses the minimal significance level where the null hypothesis can be rejected, *R-sq*. denoting the index of determination.

The results of the BIAH dependence on the combination of CPL and PUSE explanatory variables are presented. This model is satisfactory, being broadly consistent with the selection criteria. Identification of vertical outliers, leverage points and influential points was performed using LTS regression. In the given case, the LS fit produces no residual outliers, the robust fit, on the other hand, producing one outlier and six leverage points. As we can see, one observation (17 Malla) is identified both as a leverage point and a vertical outlier (see Tab. 1). Graphical outlier detection tools indicate a similar outcome (see Fig. 1). The horizontal broken lines are located at +2.5 and -2.5 and the vertical line at cut-off points $\pm\sqrt{\chi^2_{p-1;0.975}}$, where p is the number of predictors. The points lying to the right of the vertical line are leverage points, those lying above or below horizontal lines are regarded as vertical outliers.

Tab.1: Robust diagnostics (BIAH ~CPL+ PUSE model)

Observation	Mahalanobis distance	Robust MCD distance	Leverage	Stand. robust residual	Outlier
2 Bulgaria	1.7783	2.8611	*	-1.2882	
4 Denmark	2.1521	2.5996	*	-0.4063	
8 Greece	0.7417	3.5492	*	-1.3895	
9 Spain	1.3499	5.3064	*	0.4852	
11 Italy	1.2279	4.5846	*	-0.7399	
17 Malta	2.9831	9.9364	*	3.5523	*
21 Portugal	2.6809	9.4802	*	1.0288	
22 Romania	1.4550	3.1847	*	-1.7913	

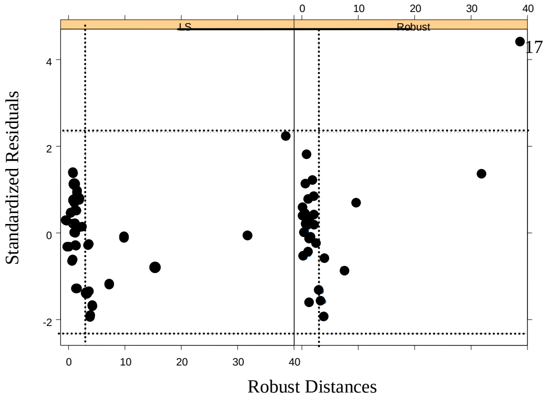
Source: data EUROSTAT, author's calculation

In Tab. 2, model fittings are presented. Estimates of the regression coefficient differ somewhat across the robust methods, the results obtained of two MM regression methods depending on the employed method of initial estimates. Due to the same reason, the results of reweighted least squares regression methods are different as well. Neither LTS nor S estimates can be used as self-contained estimates. The LS fit is quite different from robust fits. Owing to the existence of influential points, the model estimated by robust regression has to be preferred. This result is confirmed by the non-normality of the LS model residuals. Multimodality of the kernel estimate of residuals' density plot (see Fig. 2) validates the presence of outlier points.

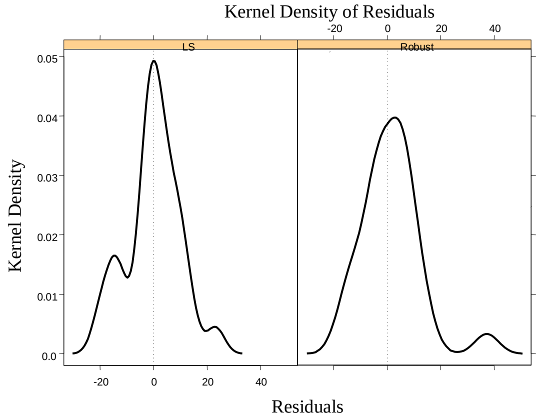
Fig. 1: Diagnostic Plot (BIAH ~CPL + PUSE model)

Fig. 2: Kernel estimate of residuals' density (BIAH ~CPL + PUSE model)

	Parameter	Value of regression coefficient	Standard error	t-value	Pr(>|t|) (p-value)	Wald test (Chi-sq)	P(>Chi) (p-value)
MM/LTS	intercept	-17.6107	15.3966			1.31	0.2527
MM/S	intercept	-21.6334	19.4678	-1.1112	0.2775		
LTS	intercept	-1.9912					
S	intercept	-13.5460	0.4875			0.69	0.4066
LS	intercept	4.8592	13.3468	0.3641	0.7190		
MM/LTS	CPL	0.5099	0.0840			36.81	0.0001
MM/S	CPL	0.5357	0.1098	4.8786	0.0001		
LTS	CPL	0.4248					
S	CPL	0.4875	0.0897			29.53	0.0001
LS	CPL	0.4730	0.0871	5.4283	0.0000		
MM/LTS	PUSE	0.3855	0.1446			7.11	0.0077
MM/S	PUSE	0.4018	0.1779	2.2586	0.0333		
LTS	PUSE	0.3080					
S	PUSE	0.3643	0.1514			5.79	0.0161
LS	PUSE	0.1426	0.1256	1.1350	0.2676		



Source: data EUROSTAT, author's elaboration



Source: data EUROSTAT, author's elaboration

Tab. 2: Model BIAH ~CPL+ PUSE fitting results

Source: data EUROSTAT, author's calculation

In another presented model, the one with exploratory ER, IRUI and HTE variables, the robust diagnostic reveals four vertical outliers and nine leverage points, three of them (2 Bulgaria, 17 Malta, 22 Romania) being influential points. Classical LS diagnostic did not reveal vertical outliers, only leverage points (see Fig. 3 and Tab. 3). This example also illustrates the problem of outliers masked in the LS fit. If the outliers are identified, the difference between the regression LS fit and the robust fit can be anticipated.

Tab. 3: Robust diagnostics (BIAH ~ER+IRUI+HTE model)

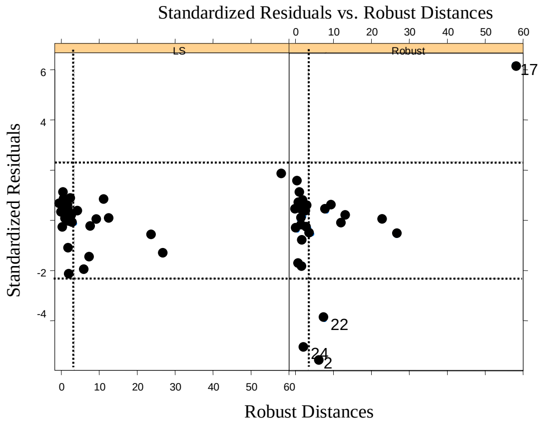
Observation	Mahalanobis distance	Robust MCD distance	Leverage	Stand. robust residual	Outlier
2 Bulgaria	1.646	2.5902	*	-6.3614	*
7 Ireland	1.1145	4.0873	*	0.5889	
10 France	1.0944	3.7394	*	0.2057	
12 Cyprus	3.0530	3.7955	*	0.3490	
15 Luxembourg	2.2816	6.5578	*	-0.8846	
16 Hungary	2.0443	6.3043	*	-0.0977	
17 Malta	3.2020	9.8823	*	7.0291	*
21 Portugal	1.9284	3.5386	*	-0.2699	
22 Romania	2.1011	2.4010	*	-4.5087	*
24 Slovakia	1.8614	1.9895		-6.1922	*

Source: data EUROSTAT, author's calculation

Fig. 3: Diagnostic Plot

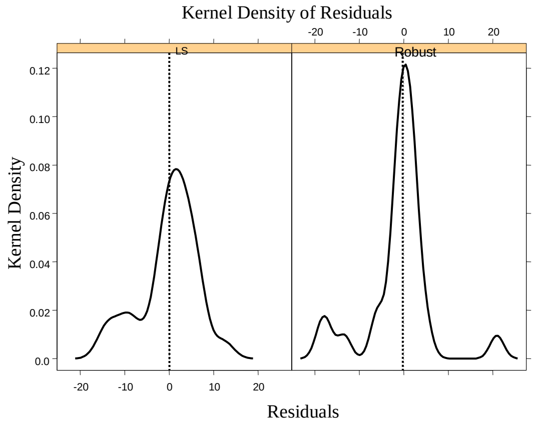
Fig. 4: Kernel estimate of residuals'

(BIAH ~ER+IRUI+HTE model)



Source: data EUROSTAT, author's elaboration

density (BIAH ~ER+IRUI+HTE) model)



Source: data EUROSTAT, author's elaboration

See Tab. 4 for the results of regression fits. Due to the existence of influential points, the differences of regression fits obtained using robust regression and classical LS regression are more distinct. Partial regression coefficients of the LS model are not statistically significant (at a 5% level). As you can see from Fig. 4, residuals of the LS model are not normally distributed, the density estimate of residuals for the robust fit is very compact and centered on zero in the central region, distinct bumps indicating the presence of outliers. The model estimated by robust regression has to be preferred. It is obvious that improper use of the classical LS regression model with significant variables without adequate identifications of outliers and testing of the normality of residuals can lead to the acceptance of an incorrect LS model.

Tab. 4: Model BIAH ~ER+IRUI+HTE fitting results

	Parameter	Regression coefficient	Standard error	t-value	Pr(>|t|) (p-value)	Wald test (Chi-sq)	P(>Chi) (p-value)
MM/LTS	intercept	-11.6013	10.1752			1.30	0.2542
MM/S	intercept	-10.3977	8.8118	-1.1800	0.2501		
LTS	intercept	-8.7758					
S	intercept	-10.2007	8.9302			1.3	0.2553
LS	intercept	-10.9929	17.8313	-0.6165	0.5436		
MM/LTS	ER	0.3609	0.1782			4.10	0.0428
MM/S	ER	0.3875	0.1548	2.5042	0.0198		
LTS	ER	0.3519					
S	ER	0.3681	0.1560			5.56	0.0183
LS	ER	0.2565	0.3115	0.8234	0.4189		
MM/LTS	IRUI	0.7791	0.0699			124.29	0.0000
MM/S	IRUI	0.7316	0.0633	11.5563	0.0000		
LTS	IRUI	0.7587					

S	IRUI	0.7512	0.0627			143.68	0.0000
LS	IRUI	0.7805	0.1184	6.5933	0.0000		
MM/LTS	HTE	-0.2890	0.1201			5.79	0.0161
MM/S	HTE	-0.2650	0.1046	-2.5323	0.0186		
LTS	HTE	-0.2968					
S	HTE	-0.3232	0.1305			6.13	0.0133
LS	HTE	0.1637	0.1834	0.8924	0.3814		

Source: data EUROSTAT, author’s calculation

Other acceptable robust regression MM models supplemented by goodness-of-fit tests are shown in Tab. 5.

Tab. 5: Goodness-of-fit tests of acceptable robust regression models

Outliers	Robust MM model	R-sq.	AICR	BICR	Deviation
17	-17.611+0.510 CPL + 0.386 PUSE	0.4272	21.800	28.510	1882.31
2,17,22,24	-11.601+0.361 ER +0.779 IRUI-0.289 HTE	0.6512	22.258	33.383	592.440
2,12,**22**	55.26 + 0.498 TEA - 2.525 LTU	0.3778	22.505	29.965	2055.56
2,17	28.165+ 9.448 GERD + 0.372 TEA	0.5286	22.241	29.783	1259.01
2,12,17	-62.184 – 0.561 TEA + 1.504 ER	0.4303	22.278	30.166	1753.78

Source: data EUROSTAT, author’s calculation. Highlighted in bold represents influential points.

Conclusion

As assumed, the European countries' data contain outliers and influential points. Therefore, in general, robust models are more applicable than the classical LS regression model. The nature of the performance indicator of the EU countries may be the cause of common statistical methods leading to incorrect conclusions due to the existence of outlying observations. The robust MM method with high breakdown should be taken into consideration and eventually preferred. Both LTS and S estimates play the role of initial estimates in MM-regression and thus cannot be used as self-contained final ones.

In the BIAH (the level of broadcast Internet access of households) analysis in the EU countries, several robust regression models can be taken into account. The applicability and advantages of high-breakdown robust regression methods in the analysis of the Internet use were corroborated. However, for the choice of a final model describing the dependence of households with broadband Internet access on the selected set of explanatory variables, the more general economic outlook is necessary.

Appendix. List of indicators in the presented models used

BIAH	Level of Broadcast Internet Access of Households
CPL	Comparative Price Level
ER	Employment Rate, age group 20-64
GERD	Gross Domestic Expenditure on R&D
GDP	GDP per capita in Purchasing Power Standards (PPS)
HICP	Harmonized Indices of Consumer Prices - Annual average rate of change (%)
HTE	High-Tech Export
IRUI	Individuals Regularly Using the Internet (percentage of individuals)
LTU	Long-Term Unemployment
PUSE	Persons with Upper Secondary or Tertiary Education Attainment (%),25-64 years
TEA	Tertiary Educational Attainment, age group 30-34

Acknowledgment

The support of the grant VŠE IGA 128/2014 Consequences of assumption violations of classical statistical methods and the possible use of alternative statistical techniques in economic applications is gladly acknowledged.

References

Hampel, F.R. & Ronchetti, E.M. & Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* New York: J. Willey.

Hubert, M. & Rousseeuw, P.J.& Van Aelst, S. (2008). High-Breakdown Robust Multivariate Methods. *Statistical Science* 2008, *23* (1), 92-119.

Olive, D.J. (2002). Applications of robust distances for regression. Technometrics. 2002,44(1),pp.64-71.

Ronchetti, E. (1997). Robustness Aspects of Model Choice. *Statistica Sinica*, *7*, 327-338.

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

Rousseeuw,P.J. & Leroy, A. M. (2003). *Robust Regression and Outlier Detection.* New Jersey: J Willey.

Rousseeuw, P.J. & Van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*(*411*) ,633- 639.

Ruppert, D. & Carroll, R.J. (1990). Trimmed Least Squares Estimation in the Linear Model. *Journal of the American Statistical Association, 75, 828-838.*

Sommer,S. & Huggins, R.M. (1996). Variable Selection Using the Wald Test and a Robust Cp, Applied Statistics, 45, 1996, pp.15–29.

Yohai, V. J. (1987). High Breakdown-point and High Efficiency Robust Estimates in Regression. *The Annals of Statistics.*, *15.*(2.), 642-656.

Zanan, A. & Rousseeuw, P. J. & Orhan, M. (2001). Econometric Applications on High-Breakdown Robust Regression Technique. *Economic Letters*, *71*, 1-8.

Contact

Dagmar Blatna

University of Economics, Prague

Faculty of Informatics and Statistics

W. Churchill sq. 4

130 67 Prague 3

Czech Republic

blatna@vse.cz