Wrocław University of Economics

EKONOMICKÁ FAKULTA
UNIVERZITA MATEJA BELA V BANSKEJ BYSTRICI

VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE
VŠE

Conference Proceedings
Full Text Papers

edited by
Zofia Rusnak and Beata Zmyślona

17th AMSE

Applications of Mathematics and Statistics in Economics

International Scientific Conference | Poland • 27-31 August 2014

Layout

Martin Boďa, Beata Zmyślona, Grzegorz Rogoziński

Front page design

Grzegorz Rogoziński

CD cover design

Beata Dębska

Articles published in the form submitted by the authors

ISBN 978-83-7695-421-9

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
53-345 Wrocław, ul. Komandorska 118/120
www.ue.wroc.pl
Sprzedaż książek tel./fax 71 36-80-602
e-mail: econbook@ue.wroc.pl www.ksiegarnia.ue.wroc.pl

Contents

17th AMSE
APPLICATIONS OF MATHEMATICS AND STATISTICS IN ECONOMICS
International Scientific Conference | Poland • 27-31 August 2014

APPLICATION OF ROBUST REGRESSION IN AN ANALYSIS OF THE INTERNET ACCESS IN THE EUROPEAN COUNTRIES

DAGMAR BLATNÁ
University of Economics, Prague, Faculty of Statistics and Informatics, Department of Statistics and Probability, W.Churchill sq. 4, 130 67 Prague 3, Czech Republic
email: blatna@vse.cz

Abstract

Individuals Regularly Using the Internet (IRUI) is one of the indicators of the information society representing the computer literacy of a country's population. The values of this indicator depend on many different economic factors of the general economic background, employment, innovation and research, science and technology. The values of these indicators vary greatly between the European countries and, consequently, the occurrence of outliers can be expected in an IRUI analysis. In such a case, the classical statistical approach – the least squares method may be highly unreliable, robust regression methods constituting an acceptable and useful tool that can be employed for detecting influential observations as well. The aim of this paper is to demonstrate the applicability and advantages of robust regression methods in an analysis of the European countries' actual economic data. The results obtained by using classical linear least squares and robust regression analysis are compared. The economic IRUI analysis, however, was not the main focus of the present paper.

Key words: *robust regression, LS regression, outliers, leverage points, internet access*

DOI: 10.15611/amse.2014.17.02

1. Introduction

The main goal of this paper is to demonstrate the applicability and advantages of robust regression methods in an analysis of the European countries' actual economic data. Indicator IRUI (Individuals Regularly Using the Internet) is one of the indicators of the information society representing the computer literacy of a country's population. This Internet access is monitoring by the Digital Agenda for Europe (DAE), one of seven flagships initiatives under Europe 2020, the EU's strategy to deliver smart sustainable and inclusive growth. The DAE aims to reboot Europe's economy and help Europe's citizens and businesses to get the most out of digital technologies. The aim is to deliver sustainable economic and social benefits from a Digital Single market based on fast and ultra fast internet and interoperable applications, with broadband access. The DAE has set three targets related to broadband access, two of which refer to broadband coverage: all homes should have access to broadband of at least a basic quality by 2013, and all homes should have access to high-speed broadband of at least 30 Mbps by 2020. The appropriate level of Internet access of individuals is prerequisite for the fulfillment of these targets. The values of this indicator depend on many different economic factors of the general economic background, employment, education, innovation and research, science and technology. The values of these indicators vary greatly between the European countries and, consequently, the occurrence of outliers can be expected in an Internet access analysis. In such a case, the classical statistical approach – the least squares method (LS) may be highly unreliable, robust regression methods constituting an acceptable and useful tool that can be employed for detecting influential observations as well. The economic internet usage analysis not being its main objective.

The remainder of the paper is organized as follows. Section 2 gives description of the analyzed variables and data file. In section 3, we introduce a robust method of regression analysis, robust diagnostic tools and model selection criteria. Subsequently results are presented and commented in section 4. A final section summarizes the results obtained.

2 Analyzed variables and the data set

The level of the Internet access can be characterized by a few indicators. In our paper, we analyse that of IRUI (Individuals Regularly Using the Internet[1]). IRUI is expressed as a percentage of individuals in the 16-74 age group with at least once a week frequency of using the Internet, IRUI being one of the indicators of the information society that represents the computer literacy of a country's population.

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. A complete list of the indicators employed in the analysis is given in the appendix to this paper, calculations being performed by means of SAS 9.2 and S-Plus 6.2statistical software.

3 Methodology

3.1 Robust regression

The aim of a regression analysis is to find a good estimate of unknown regression coefficients from the observed data. The usual estimator of regression coefficients comes from the method of ordinary least squares, LS being an optimal regression estimator under the sets of assumptions on the distribution of the error term (normality, homoskedasticity, independence of the errors) and predicted variables.

Robust regression provides an alternative to LS regression that works with less restrictive assumptions. The primary purpose of a robust regression technique is to fit a model that describes information in the majority of the data. In particular, it provides much better regression coefficient estimates when outliers are present in the data. Outliers violate the assumption of normally distributed residuals in LS regression.

It is a common practice to distinguish between two types of outlying observations in regression analysis, those in the response variable representing model failure. Such observations are called outliers in the *y*-direction or vertical outliers, those with respect to the predictors being labeled as leverage points. The leverage point is defined as $(x_{k_1},...,x_{k_p},y_k)$, for which $(x_{k_1},...,x_{k_p})$ is outlying with respect to $(x_{i_1},...,x_{i_p})$ in the data set. Outliers that bias the parameter estimates are called bad leverage points, whereas outliers lying along the predicted model are called good 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.

In the analysis, LS, LTS, MM and RWLS regression methods have been applied.

[1] "Regular use" means at least once a week (i.e. every day or almost every day or at least once a week)

First, let us briefly mention the principles of the selected robust methods. Two regression methods were employed. 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 number of observations. The usual choice $h \approx 0.75n$ yields the breakdown point of 25 %; (see Hubert, Rousseeuw, van Aelst (2008).

LTS regression with a high breakdown point is a reliable data analytic tool thatcan be used to detect vertical outliers, leverage and influential points (observations whose inclusion or exclusion result in substantial changes in the fitted model) in both simple and multivariate settings. A more detailed description is available in, e.g., Ruppert, Carroll (1980), Rousseeuw (2003), Chen (2002), or Hubert, Rousseeuw, Van Aelst (2008).

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), Chen (2002) or Rousseeuw (2003)). At the first stage, an initial regression estimate is computed; it is consistent, robust, with a high breakdown point but not necessarily efficient. In our analysis, two methods of initial estimates are used (LTS and S regression). 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 regression parameters based on a proper redescending ψ-function. A more detailed description of robust regression methods is available in Chen (2002), Rousseeuw (2003), Yohai (1987), SAS and SPLUS manuals. Tukey's bisquare loss function was employed.

Reweighted least squares (RWLS) regression minimizes the sum of squared residuals multiplied by the weights w_i, which are determined from the LTS solution. The effect of the weight staking just the values of 0 or 1 is the same as in the cases for which w_i equals zero that are deleted. Therefore, RWLS can be seen as ordinary LS on a "reduced" data set consisting of only those observations that received non-zero weights.

3.2 Identification of outliers, leverage and influential points

The following numerical and graphic diagnostics for detecting vertical outliers, leverage points and influential observations have been applied (more in detail see, e.g.,Rousseeuw (1984), Rousseeuw, van Zomeren (1990), Rousseeuw (2003), Olive (2002), Chen (2002)):

- Residuals associated with LTS regression,
- Standardized residuals (the residuals divided by the estimates of their standard errors, the mean equaling 0 and standard deviation 1).
- Studentized residuals (a type of standardized residuals follows at t-distribution with n-p-2 Df). Attention should be paid to studentized residuals that exceed ± 2.5 (or ± 2.0).
- The robust distance (Mahalanobis distance)
- Diagnostic plots are provided as fundamental data mining graphical tools that quickly identify outliers and determine whether they have an impact on classical estimates. To visualize vertical outliers and leverage points, the following tools were used: a regression diagnostic plot (that of the standardized residuals of robust regression vs. robust distances $RD(x_i,)$), the standardized residuals plot vs. their index, a Normal Q-Q plot of the standardized residuals and a plot of kernel estimates of the residuals' density.

3.3 Model selection methods

In the case of classical LS regression – the classical R-squared, the results of significance (t and F) tests as well as the diagnostics of residuals' normality are applied. As for robust regression, the decision which of the candidate models is to be preferred is based on the following robust diagnostic selection criteria: robust index of determination, robust deviance, significance robust tests (robust t-test, robust F-tests**,** robust Wald test), Robust Akaike's Information Criterion (AICR), Robust Bayesian information criterion (BICR) and Robust Final Prediction Error (RFPE); the above criteria are dealt with in, e.g.,Chen (2002b), Ronchetti (1985), Hampel, Ronchetti, Rousseeuw, Stahel (1996) or SAS and S-Plus manuals.

4 Results and discussion

Due to an enormous output of our analysis, it was impossible to present all tables and graphs. For the IRUI dependent variable, only two models distinguished from the statistical point of view are presented. For each model – fitting results, a numerically robust diagnostic of outliers and leverage points, graphic identification of outliers (a diagnostic graph), goodness-of-fit robust tests and a plot of kernel estimates of residuals' density are presented.

The first model includes explanatory variables Comparative price level (CPL) and Persons with upper secondary or tertiary education attainment (PUSE). Both LS and robust diagnostics identified five leverage points, none of them, however, being also a vertical outlier). Numerical robust diagnostic is shown in Table 1.

Table 1 Robust diagnostics (IRUI~CPL+ PUSE model)

Observation	Mahalanobis distance	Robust MCD distance	Leverage	Stand. robust residual	Outlier
8 Greece	0.7417	2.9253	*	-2.3937	
9 Spain	1.3499	4.4792	*	-0.096	
11 Italy	1.2279	3.9351	*	-1.7993	
17 Malta	2.9831	8.7156	*	1.8825	
21 Portugal	2.6809	8.0036	*	-0.1308	

Source: the author

In such a case, regression parameters both of LS and robust models are similar to each other. All models are presented in Table 2, the model fitting results being shown in Table 3. Because of the absence of vertical outliers, we can consider the LS model fully appropriate. Both the explanatory variables have a positive influence on IRUI, partial regression coefficients being statistically significant at a 3% level at least (see Table 3).

Table 2 LS and robust regression fits (IRUI~CPL+ PUSE model)

Model	Regression fit	R-sq
LS	-6.6367 + 0.4742 CPL + 0.3708 PUSE	0.6003
MM/S	-6.0208 + 0.4746 CPL + 0.3642 PUSE	0.5424
MM/LTS	-10.8565 + 0.4877 CPL + 0.4121 PUSE	0.5258
RWLS	-6.6367 + 0.4741 CPL + 0.3708 PUSE	
LTS	-8.8214 + 0.5022 CPL + 0.4008 PUSE	0.7722

Source: the author

Table 3 Model IRUI~CPL+ PUSE fitting results

Method	Parameter	Value of regr.coeff.	Standard error	t-value	Pr(>| t|) (p-value)	Wald test (Chi-sq)	P(>Chi) (p-value)
LS	Intercept	-6.6367	13.0501	-0.5081	0.6157		
MM/LTS	Intercept	-10.8565	15.0193			0.52	0.4698
MM/S	Intercept	-6.0208	16.6794	-0.3610	0.7213		
RWLS	Intercept	-6.6367	13.0501			0.26	0.6111
LS	CPL	0.4742	0.0852	5.565	0.0000		
MM/LTS	CPL	0.4877	0.0912			28.63	0.0001
MM/S	CPL	0.4746	0.1080	4.3953	0.0002		
RWLS	CPL	0.4742	0.0852			30.97	0.0001
LS	PUSE	0.3708	0.1228	3.0185	0.0059		
MM/LTS	PUSE	0.4121	0.1419			8.44	0.0037
MM/S	PUSE	0.3642	0.1572	2.3169	0.0294		
RWLS	PUSE	0.3708	0.1228			9.11	0.0025

Source: the author

In the second model with four exploratory variables GERD (Gross domestic expenditure on R&D), TEA (Tertiary educational attainment), HICP (Harmonized Indices of Consumer Prices) and HBA (Households with broadband access), robust diagnostics reveal four vertical outliers (15 Luxembourg, 18 the Netherlands, 24 Slovakia and 27 the United Kingdom) and thirteen leverage points (see Table 4). One observation (18 the Netherlands) is a vertical outlier and leverage point simultaneously. This observation is thus identified as an influential point. Classical diagnostics reveal only leverage points, no vertical outliers (see Figure 1). In such a case, differences between classical and robust models are anticipated. For both LS and robust regression fitting models, see Table 5.

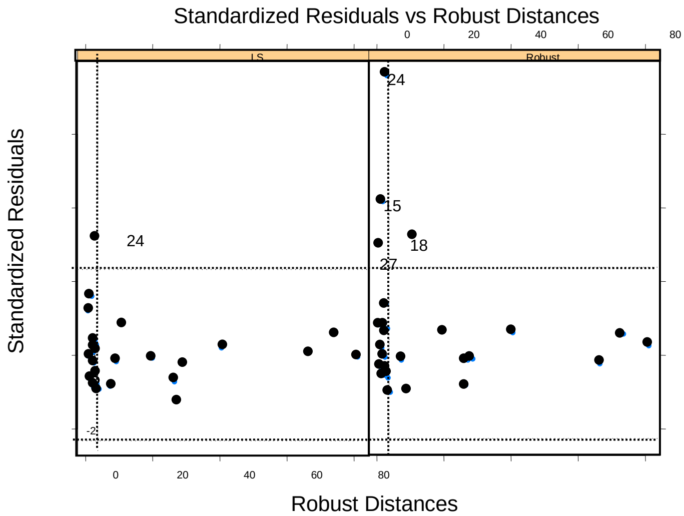

Figure 2 Diagnostic Plot (IRUI ~GERD + TEA + HICP + HBA model)
Source: the author

Table 4 Robust diagnostics (IRUI ~GERD + TEA + HICP + HBA model)

Observation	Mahalanobis distance	Robust MCD distance	Leverage	Stand. robust residual	Outlier
2 Bulgaria	2.7368	15.1077	*	0.7759	
5 Germany	1.8445	2.5091	*	-0.5125	
6 Estonia	1.7464	7.4710	*	-0.1618	
7 Ireland	2.5680	4.7208	*	1.7335	
8 Greece	1.5037	3.7754	*	-0.5506	
13 Latvia	2.7659	12.5959	*	0.0044	
14 Lithuania	1.9750	8.0509	*	-0.8913	
15 Luxembourg	1.3512	1.7339		5.8299	*
16 Hungary	1.7746	9.1245	*	0.5271	
17 Malta	3.1495	7.6023	*	-0.4367	
18 Netherlands	1.8884	5.5402	*	4.4749	*
22 Romania	2.7565	14.2269	*	-0.4465	
23 Slovenia	0.6481	2.5752	*	-0.5498	
24 Slovakia	1.8057	2.0969		10.0342	*
25 Finland	2.6515	4.162	*	-0.8605	
27 United Kingdom	0.8343	1.3322		4.3653	*

Source: the author

As you can see from Tables 5 and 6, all exploratory variables have a positive influence on IRUI. Two regression parameters of the LS model are not significant. Multimodality of the kernel estimate of residuals' density plot (see Figure 2) confirms the presence of an influential point. The kernel estimate of residuals' density of LS fit is bias (it is not centered around zero). Owing to the detected influential point and non-normality of LS residuals' density, the robust model will be given preference. Goodness-of-fit tests for the robust MM model are presented in Table 7.

Table 5 Models fitting results ((IRUI ~GERD + TEA + HICP + HBA)

Method	Parameter	Value of regr.coeff.	Standard error	t-value	Pr(>|t|) (p-value)	Wald test (Chi-sq)	P(>Chi) (p-value)
LS	Intercept	-23.7976	25.8689	-0.9159	0.3676		
MM/LTS	intercept	-45.7121	15.9947			8.19	0.0043
MM/S	intercept	-43.8249	11.6330	-3.7673	0.0011		
LS	GERD	2.0617	2.1476	0.9600	0.3475		
MM/LTS	GERD	4.4500	1.3577			14.74	0.0010
MM/S	GERD	5.3701	1.0086	5.3241	0.0000		
LS	TEA	0.1414	0.1593	0.8875	0.3844		
MM/LTS	TEA	0.1996	0.0988			4.08	0.0434
MM/S	TEA	0.1706	0.0721	2.3676	0.0271		
LS	HICP	0.2481	0.1836	1.3510	0.1904		
MM/LTS	HICP	0.4064	0.1126			13.02	0.0003
MM/S	HICP	0.4160	0.0820	5.0732	0.0000		
LS	HBA	0.8712	0.1654	5.2668	0.0000		
MM/LTS	HBA	0.8031	0.1020			61.98	0.0000
MM/S	HBA	0.7325	0.0768	9.5428	0.0000		

Source: the author

Table 6 LS and robust regression fits (IRUI ~GERD + TEA + HICP + HBA)

Model	Regression fit	R-sq
LS	-23.7976 + 2.0617 GERD + 0.1414 TEA + 0.2481 HICP + 0.8712 HBA	0.8324
MM/LTS	-45.7121 + 4.4500 GERD + 0.1996 TEA + 0.4064 HICP + 0.8031 HBA	0.7426
MM/S	-43.8249 + 5.3701 GERD + 0.1706 TEA + 0.4160 HICP + 0.7325 HBA	0.7084
RWLS	-45.8790 + 4.1428 GERD + 0.1989 TEA + 0.3998 HICP + 0.8319 HBA	-
LTS	-52.1931 + 5.3082 GERD + 0.1000 TEA + 0.4825 HICP + 0.7779 HBA	0.9706

Source: the author

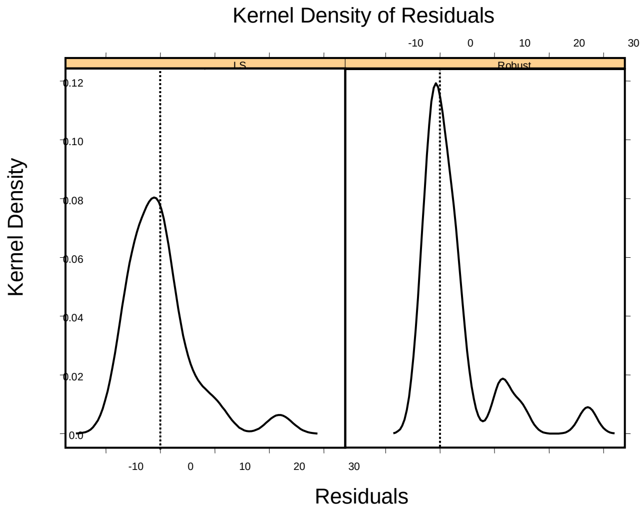

Figure 2 Kernel estimate of residuals' density (IRUI ~GERD +TEA+HICP+HBA model)
Source: the author

Table 7 Goodness-of-fit tests for robust MM model

Outliers	R-sq.	AICR	BICR	Deviation	RFPE
Leverage points 2,5,6,7,8,12,13,14,16,17,18,22,23,25 Vertical outliers 15, 18, 24,27	0.7426	20.979	34.005	439.10	18.022

Source: the author

5 Conclusion

The level of the Internet access can be characterized by a few indicators. IRUI (Individuals Regularly Using the Internet) is expressed as a percentage of individuals in the 16-74 age group with at least once a week frequency of using the Internet, IRUI being one of the indicators of the information society representing the computer literacy of a country's population. The adequate level of Internet access of individuals is prerequisite for the fulfillment of the objectives set out in the Digital Agenda for Europe, which is one of seven flagships initiatives under Europe 2020, the EU's strategy to deliver smart sustainable and inclusive growth. The values of IRUI depend on many different economic factors of the general economic background, employment, education, innovation and research, science and technology. Two distinguished models from a statistical point of view were presented.

In an analysis of real economic data, vertical outliers, leverage points and influential points are supposed to occur. In such a case, the application of the least square regression (LS) could lead to incorrect results, robust regression methods being a useful analytical tool. Robust regression with a high breakdown point (LTS) can detect influential points as well.

Robust techniques provide the results similar to LS regression when the data are linear with normally distributed errors. When vertical outliers are not identified in the data, errors being normally distributed, LS regression is a fully appropriate method and should be preferred. This conclusion is demonstrated by the model with exploratory variables Comparative price level (CPL) and Persons with upper secondary or tertiary education attainment. Both the explanatory variables have a positive influence on IRUI (see IRUI~CPL+ PUSE model).

However, regression coefficient estimates can differ markedly when the data contain significant vertical outliers and influential points. In such cases, robust regression techniques should be preferred. It is evident that an improper use of the classical least square regression model with significant variables without the corresponding identification of outliers and assessment of residual normality can lead to the acceptance of incorrect LS models. This conclusion is observed in the model with exploratory variables Gross domestic expenditure on R&D, Tertiary educational attainment, Harmonized Indices of Consumer Prices and Households with broadband access. All explanatory variables have a positive impact on IRUI (see IRUI ~GERD +TEA+HICPA+HBA model).

The economic IRUI analysis, however, was not the main focus of the present paper.

Acknowledgements

The support of the grant Internal Grant Agency of the University of Economics, Prague (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

1. CHEN,C. 2002 Robust Regression and Outlier Detection with the ROBUSTREG procedure. SUGI Paper, SAS Institute Inc., Cary, NC.,2002 http://www2.sas.com/proceedings/sugi27/p265-27.pdf
2. EUROPEAN COMMISSION Documents and Working papers (2000-2014)
3. HAMPEL, F.R., RONCHETTI, E. M., ROUSSEEUW, P. J., STAHEL, W. A. 1996. Robust Statistics. The Approach Based on Influence Functions. New York: J. Willey, 1986.
4. HUBERT, M., ROUSSEEUW, P.J., VAN AELST. 2008. High-Breakdown Robust Multivariate Methods. Statistical Science 2008, 23 (1), pp.92-119.
5. OLIVE, D.J. 2002. Applications of robust distances for regression. Technometrics.2002, 44(1), pp.64-71.
6. RONCHETTI, E. 1985, Robust Model Selection in Regression, Statistics & Probability Letters, 2008,3, pp.21–23.

7. ROUSSEEUW, P.J. 1984. Least median of squares regression. Journal of the American Statistical Association ,1984. 79(388), pp.871 -880.
8. ROUSSEEUW, P. J., LEROY, A. M. 2003. Robust Regression and Outlier Detection. NewJersey: J.Willey.2003.
9. ROUSSEEUW, P.J., VAN ZOMEREN, B.C.1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association,1990, 85(411), pp.633-639.
10. RUPPERT,D., CARROLL, R.J. 1990. Trimmed Least Squares Estimation in the Linear Model. Journal of the American Statistical Association, 1990, 75, pp.828-838.
11. SAS 9.2. Help and documentation.
12. SOMMER, S., HUGGINS, R. M. 1996. Variable Selection Using the Wald Test and a Robust Cp, Applied Statistics, 45, 1996, pp.15–29.
13. S-PLUS 6 Robust Library User's Guide. 2002 Insightful Corporation, Seatle, Washington. 2002
14. YOHAI, V.J. 1987. High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics,1987, 15(20), pp.642-656.
15. ZAMAN, A., ROUSSEEUW, P.J., ORHAN, M. 2000. Econometric applications of high-breakdown robust regression techniques. Economics Letters, Vol. 71,2000, pp.1-8.

List of indicators

CPL	Comparative Price Level (% of GDP)
ELE	Early leavers from education and training (%)
ER	Employment rate (%), age group 20-64
GERD	Gross domestic expenditure on R&D (% of GDP)
GDPc	GDP per capita in Purchasing Power Standards (PPS)
GDP.gr	Gross Domestic Product (growth)
HBA	Households with broadband access (%)
HICP	Harmonized Indices of Consumer Prices - Annual average rate of change (%)
HRST	Human Resources in Science and Technology (%)
ILCS	Individuals' level of computer skills (%)
LLL	Participation in education and training (%)
LPH	Labour productivity per hour worked
LPP	Labour productivity per person employed
LTU	Long-term unemployment (Annual average, in % of active population);
NNI	Net national income (% of GDP
PUSE	Persons with upper secondary or tertiary education attainment (%), 25-64 years
RULC	Real unit labour cost growth - Percentage change on previous period
TEA	Tertiary educational attainment (%)r, age group 30-34