Application of Robust Regression and Bootstrap in Productivity Analysis of GERD Variable in EU27

Dagmar Blatná[1] | *University of Economics, Prague, Czech Republic*

Abstract

The GERD is one of Europe 2020 headline indicators being tracked within the Europe 2020 strategy. The headline indicator is the 3% target for the GERD to be reached within the EU by 2020. Eurostat defines "GERD" as total gross domestic expenditure on research and experimental development in a percentage of GDP. GERD depends on numerous factors of a general economic background, namely of employment, innovation and research, science and technology. The values of these indicators vary among the European countries, and consequently the occurrence of outliers can be anticipated in corresponding analyses. In such a case, a classical statistical approach – the least squares method – can be highly unreliable, the robust regression methods representing an acceptable and useful tool. The aim of the present paper is to demonstrate the advantages of robust regression and applicability of the bootstrap approach in regression based on both classical and robust methods.

Keywords

LTS regression, MM regression, outliers, leverage points, bootstrap, GERD

JEL code

C19, C49, O11, C13

INTRODUCTION

GERD represents total gross domestic expenditure on research and experimental development (R&D) as a percentage of GDP (Eurostat), R&D expenditure capacity being regarded as an important factor of the economic growth. GERD is one of Europe 2020 indicator sets used by the European Commission to monitor headline strategy targets for the next decade –A Strategy for Smart, Sustainable and Inclusive Growth (every country should invest 3% of GDP in R&D by 2020). GERD comprises expenditure of four institutional sectors of production – business enterprise, government, higher education and private non-profit establishments. Expenditure data involve the research funds allocated in the national territory, regardless of their source.

Generally GERD depends on various elements of a general economic background, such as the employment, innovation, research, science and technology. Both GERD and the above indicators' values

[1] University of Economics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. Email: blatna@vse.cz.

vary among the European countries and, consequently, the occurrence of outliers can be envisaged in the EU countries' GERD analysis.

A classical statistical approach to regression analysis – the least squares method (LS) – can be highly unsatisfactory due to the presence of outliers that are likely to occur in an analysis of any real data. In such a case, robust regression becomes an acceptable and useful tool, since it provides a good fit to the bulk of the data, the outliers being exposed clearly enough. The aim of this paper is to verify the applicability of the robust regression and bootstrapping (resampling) technique based on both LS and robust regression, the economic GERD analysis not being its main objective.

1 LITERATURE

Robust regression techniques are rarely used in economic analysis; only a few applications can be found in the available literature. Zaman, Rousseeuw, Orhan (2001), for instance, applied a high breakdown robust regression method to three linear models, having compared regression statistics for both the LS technique used in the original paper and the robust method. The authors eventually recommended that robust techniques should be used to avoid the confusion effect of "bad" leverage points leading to a significant bias of the regression results. Finger, Hediger (2007) promoted the application of robust instead of LS regression for the estimation of agricultural and environmental production function and Colombier (2009) also estimated the growth effects of OECD fiscal policies having employed robust methods.

Numerous analyses of R&D expenditure have been made on the basis of different criteria such as the source of funds, field of science, type of costs, economic activity, enterprise size class, socioeconomic objectives, regions, etc. Guellec (1997, 2001) dealt with the cause of fluctuations in investments in R&D and the connection between GERD and productivity growth. Kroll, Zenker (2009) looked into the development of R&D expenditure at a regional level and Zhang (2006) published the results of an empirical analysis of national energy R&D expenditures. Since the launch of Europe 2020 strategy, a lot of studies, papers and reports have been released. Commenting on the strategy, some of them make relevant remarks regarding the 3% target for the GERD indicator to be reached within the EU by 2020. Albu (2011), for example, investigated to what extent the EU members complied with the R&D investment targets set by Europe 2020 strategy, their actual spending being below 2% of GDP on average and only three member states reporting the R&D expenditure ratio to be higher than 3% of GDP. Dachs (2012) analyzed an economic impact of the internationalization of business investments in R&D, Spišáková (2013) examinined the influence of the economic crisis on the achievement of Europe 2020 target in the R&D area.

2 METHODOLOGY

2.1 The principle of robust regression

Robust regression techniques are an important complement to the classical least squares (LS) regression method. Robust techniques produce results similar to LS regression when the data are linear with normally distributed errors. The results, however, can differ significantly when the errors do not satisfy normality conditions or when the data contain outliers. Robust regression is an alternative to LS regression when the data are contaminated with outliers or influential observations. It can be used for detecting influential observations as well.

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 outliers in the y-direction or vertical outliers, those with respect to the predictors being labelled 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. 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 selected robust methods used in our analysis. In robust regression, an important role is played by the breakdown point which is the fraction of "bad" data that the estimator can tolerate without being affected to an arbitrarily large extent. Having a zero breakdown value, even a small proportion of deviant observations can cause systematic distortions in LS regression estimates. Two regression methods with a high breakdown point were employed. The least trimmed squares (LTS) estimator (proposed by Rousseeuw 1984)) is obtained by minimizing $\sum_{i=1}^{h} r_{(i)}^2$, where $r_{(i)}^2$ the i-th order statistic among the squared residuals written in the ascending order, h is the largest integer between $[n/2]+1$ and $([n/2]+[(\mathrm{p}+1/2)])$, 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 that can 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), Fox (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. 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 by means of the formula:

$$\sum_{i=1}^{n} \mathbf{x}_i \psi\left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right) = 0\,, \tag{1}$$

where $\hat{\sigma}$ stands for a robust estimation of the residual standard deviation (calculated in the 2nd step) and $\psi = \rho'$ is the derivation of the proper loss function ρ. A more detailed description of robust regression methods is available in Chen (2002), Rousseeuw (2003), Fox (2002), Yohai (1987), SAS and SPLUS manuals. Due to SAS and S-PLUS software used in the analysis, Tukey's bisquare loss function was employed:

$$\rho(e) = \begin{cases} \dfrac{k^2}{6}\left\{1 - \left[1 - \left(\dfrac{e}{k}\right)^2\right]^3\right\} & \text{for } |e| \le k \\ \dfrac{k^2}{6} & \text{for } |e| > k \end{cases}, \tag{2}$$

where e means residuum, the tuning constant $k = 4.685$ for the bisquare loss function.

2.1.1 Identification of outliers, leverage and influential points

Extensive numerical and graphical diagnostic methods for detecting outliers and influential observations can be used. For more details, see, e.g. Rousseeuw, Van Zomeren (1990), Rousseeuw (2003), Fox (2002), Olive (2002), Chen (2002). In this paper, the following methods have been employed:

– *Residuals associated with LTS regression;*
– *Standardized residuals* (the residuals divided by the estimates of their standard errors, the mean and standard deviation equalling 0 and 1 respectively);
– *Studentized residuals* (a type of standardized residuals follows at t distribution with n-p-2 Df), attention being paid to studentized residuals that exceed ± 2.5 (or ± 2.0);

- *The robust distance* defined as:

$$RD(x_i) = \sqrt{[x_i - \mathbf{T(X)}]^T \, \mathbf{C(X)}^{-1} [x_i - \mathbf{T(X)}]}\,, \tag{3}$$

where $\mathbf{T}(\mathbf{X})$ is the robust location estimates vector and $\mathbf{C}(\mathbf{X})$ is the scatter matrix for the matrix of covariates;

- *Diagnostic plots* provided as fundamental data mining graphical tools for quick identification of an outlier, determine whether outliers have influence on classical estimates. In order to visualize vertical outliers and leverage points, the following plots were used:
  - *regression diagnostic plot* (a plot of standardized residuals of robust regression versus robust distances $RD(x_i,)$),
  - *plot of standardized residuals versus their index,*
  - *normal Q-Q plot of standardized residuals* and
  - *plot of kernel estimate of residuals´ density.*

2.2 The principle of bootstrap in regression

The bootstrap was introduced by Efron (1979). Bootstrapping is a general approach to statistical inference based on replacement of the true sampling distribution for a statistic by resampling from the original observed data (the original sample of size *n*). Bootstrap technique assumes only finite values of some moments, but hardly any restricting assumptions about the underlying probability distribution. It replaced classical methods' assumptions with complex calculations for the correctness assessment of a relationship found within a particular sample. The fundamental element of bootstrap is a bootstrap sample. The resampling procedure in regression brings *R* artificial samples of *n* pairs of observations from the data in the original observed sample. For bootstrapping pairs in regression models, the bootstrap sample is selected by simple random sampling observations (i.e. the response value and the corresponding vector of independent regressor variables) without replacement. Then standard errors, confidence intervals and the bias of bootstrap parameter estimates are calculated. The bias is estimated by the difference between an average bootstrapped value of the regression coefficient and its original-sample value. The bootstrap percentile interval (*EP*) is based on empirical quantiles of the bootstrap regression coefficients b_b^*, while the bias-corrected, accelerated percentile interval (BC_a) with correction factors for lower and upper percentiles is grounded on the jackknife values of the statistic β (see, e.g. Cole (1999), DiCiccio, Efron (1996), Freedman (1981), Efron (1993, 2000), Stine (1990). The resampling distribution of the regression coefficients is then constructed empirically by resampling from the sample.

In the bootstrap regression procedure, the least squares (LS) method is often used to estimate the parameters of regression models. It is, however, extremely sensitive to outliers and non-normality of errors. The robust bootstrapping method replaces the classical bootstrap mean and standard deviation with robust estimates, using robust regression estimates with a high breakdown point. In our analysis, MM-regression with initial LTS estimates has been used. The bootstrap is not used for regression parameters estimation, being a tool for the acquisition of confidential intervals and bias regression parameters estimation.

3 RESULTS AND DISCUSSION

The following regression methods have been employed in an analysis of the GERD in EU27 countries:

- least squares regression (LS),
- least trimmed squares regression (LTS),
- MM-regression (MM),
- bootstrap regression based on the LS method (B),
- bootstrap regresion based on robust MM-regression (RB).

The analysis is based on 2010 data, calculations being performed by means of SAS 9.2 and S-Plus 6.2 statistical software. All the data as well as indicator definitions have been adopted from the Eurostat database.[2] The economic indicators employed in the analysis are given in the appendix to this paper.

The GERD (total gross domestic expenditure on research and experimental development as a percentage of GDP) is one of Europe 2020 headline indicators being tracked within the Europe 2020 strategy. The headline indicator is the 3 % target for the GERD to be reached within the EU by 2020. "This target has succeeded in focusing attention on the need for the both the public and private sectors to invest in R&D but it focuses on input rather than impact" (see European Commision, 2010, p. 8). From this point of view, GERD is consider as a dependent variable in the analysis.

For the GERD as the dependent variable, numerous linear regression models have been tested using the least squares linear regression (LS) and robust MM-regression. Identification of vertical outliers, leverage points and influential points was performed using LTS regression. SAS uses the default value $h = [(3n + p + 1)/4]$. For $n = 27$ and $p = 3$ or $p = 4$, we get $h = 21$, and the corresponding breakdown point of about 21–25%. The existence of vertical outliers or leverage points in the model can be quickly identified from the robust diagnostic plot, LS diagnostics being on the left and robust diagnostics on the right side. Horizontal broken lines are located at +2.5 and –2.5 and the vertical line is located at the cutoffs of $\pm\sqrt{\chi^2_{p-1;0.975}}$, 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. In the case of classical LS regression, the classical index of determination (*R*-squared) and the results of significance *t*-tests and *F*-tests (at a significance level of 5%) were used. In the case of robust regression, the decision which of the alternative models should be preferred was based on robust diagnostic selection criteria: the robust index of determination (*R*-squared), significance robust Wald and *F*-test and robust selection information criteria –Robust Akaike's Information Criterion (*AICR*), Robust Bayesian Information Criterion (*BICR*) and Robust Final Prediction Error (*RFPE*), (see e.g. Hampel (1983), Hampel, Ronchetti, Rousseeuw, Stahel (1996), Ronchetti (1985), Sommer, Huggins (1996), SAS and SPLUS manuals). In both LS and robust regressions, the normality of residuals was also taken into consideration to determine which model ought to be preferred. Numerous regression models, using the set of indicators (predictors) available from the Eurostat database, have been computed. For regression models that fulfill the aforementioned criteria, both classical and robust bootstrapping regression were applied as well. In the analysis, only models with two or three regressors were fully acceptable. 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.

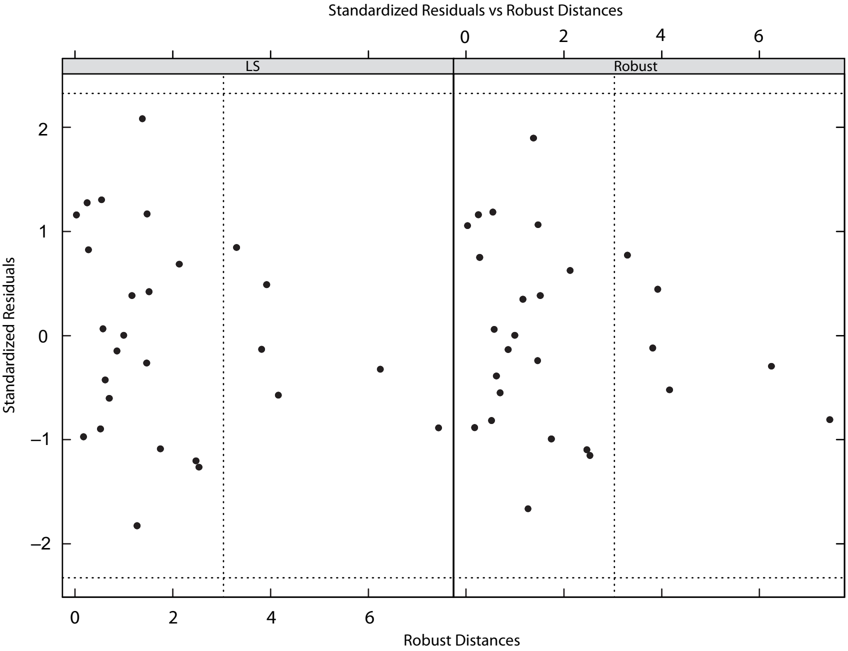
In the presented models the following predictors have been included:

CPL	Comparative Price Level (EU27 = 100%);
ER	Employment rate total (the ratio of employed persons aged 20–64 and the total population of the same age group;
HICP	Harmonised indices of consumer prices (2005 = 100);
IRUI	Individuals regularly using the Internet (in percent; frequency of Internet access: once a week);
LPH	Labour productivity per hour worked;

In the first model that includes explanatory variables CPL and IRUI, both LS and robust diagnostics identified six leverage points, none of them, however, being also an vertical outlier (see Figure l).

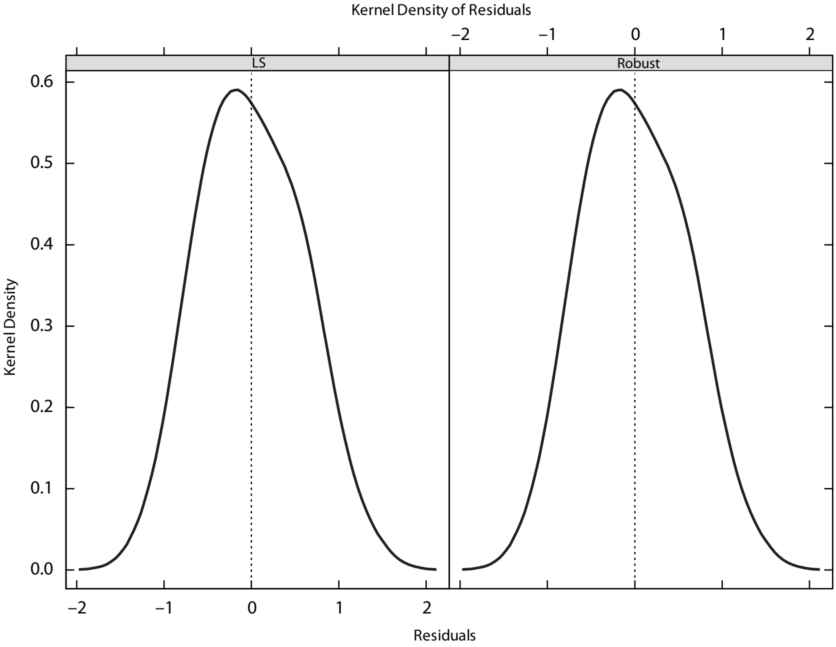
[2] <http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database> and / or <http://apl.czso.cz/pll/eutab/html>.

Figure1 Diagnostic Plot (GERD~CPL+IRUI model)



Source: Author's own elaborations

Figure 2 Kernel estimate of residuals' density (GERD~CPL+IRUI model)



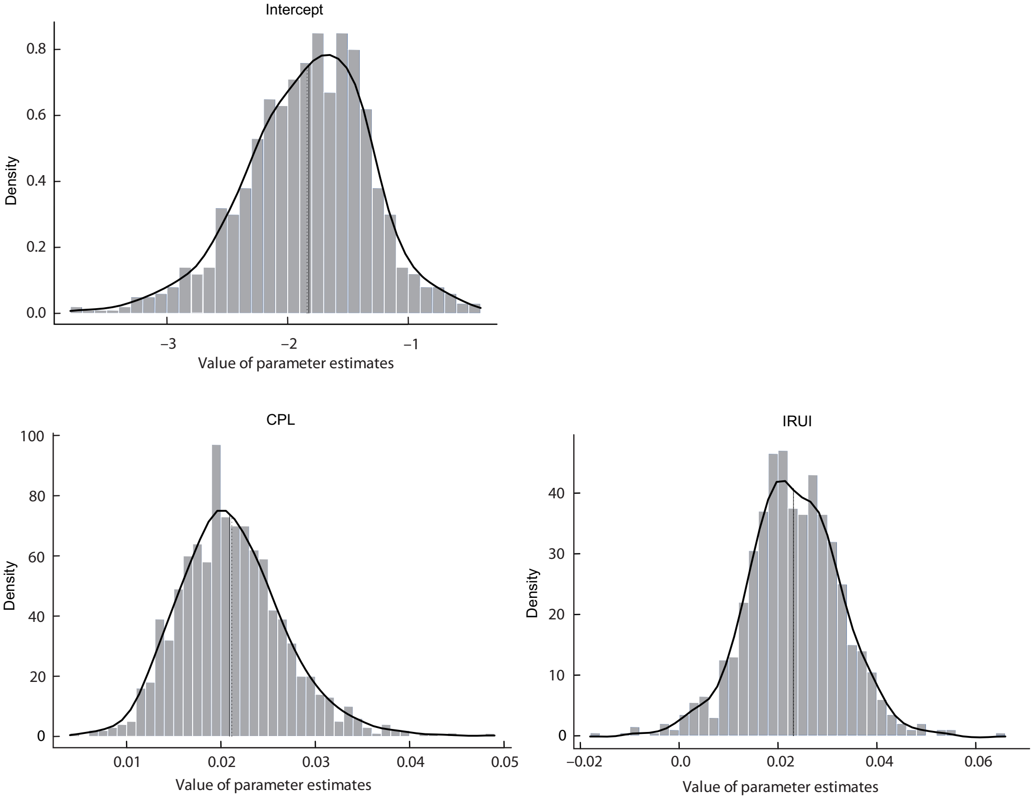
Source: Author's own elaborations

Table 1 Classical and robust bootstrap regression, LS and MM regression for GERD ~ CPL + IRUI model

		Observed	Bias	Mean	SE	95% EP	95% BCa
B R = 1000	Interc.	−1.8247	−0.0003	−1.8250	0.5101	−2.909; −0.906	−3.0039; −0.979
	CPL	0.0209	0.0004	0.0213	0.0053	0.0112; 0.032	0.0112; 0.032
	IRUI	0.0231	−0.0005	0.0226	0.0096	0.003; 0.042	0.0035; 0.0423
RB R = 1000	Interc.	−1.8247	−0.0003	−1.8250	0.5101	−2.909; −0.906	−3.0040; −0.979
	CPL	0.0209	0.0004	0.0213	0.0053	0.0112; 0.032	0.0112; 0.0321
	IRUI	0.0231	−0.0005	0.0226	0.0096	0.003; 0.042	0.0035; 0.0422
		Parameter	**SE**	**t**	**p-value**	**95% conf. interval**	
LS R-sq. 0.6629	Interc.	−1.8247	0.5175	−3.526	0.0017	−2.8928; −0.7566	
	CPL	0.0209	0.0065	3.2344	0.0035	0.0076; 0.0343	
	IRUI	0.0231	0.0099	2.3347	0.0283	0.0027; 0.0435	
MM R-sq. 0.5724	Interc.	−1.8247	0.6539	−2.790	0.0102	−2.8391; −0.7566	
	CPL	0.0209	0.0081	2.5927	0.0160	0.0082; 0.0336	
	IRUI	0.0231	0.0123	1.8834	0.0718	0.0037; 0.0425	

Source: Data EUROSTAT, author's own calculations

Figure 3 Histograms for classical replications of regression coefficients for GERD ~ CPL + IRUI model (R = 1 000)



Source: Author's own elaborations

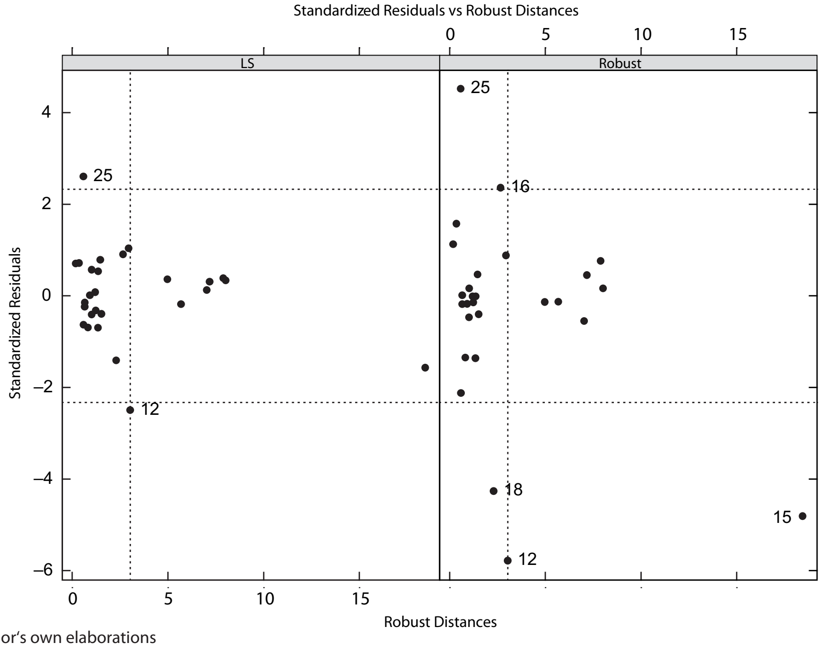
Since no vertical outliers are identified, the LS and MM-regression models are identical (see Table 1), classical and robust bootstraps provide the results very close to the values of the estimated regression coefficients of LS and MM-regression fits. Kernel estimates of residuals' density are almost normal but are not centred around zero both for LS and MM regression models (see Figure 2). Classical bootstrap provides the lowest standard errors and the narrowest confidence intervals of the estimated regression coefficients; they are even narrower than LS ones (for any regression coefficients). The bias is a difference between an average bootstrapped value of the regression coefficient and its original sample value. Histograms of regression coefficients' estimates are adequately symmetric in both bootstrap methods, robust bootstrap, however, providing broader confidence intervals. Histograms of regression coefficients' estimates for classical bootstrap see in Figure 3.

Due to the absence of vertical outliers, both classical regression and classical bootstrap are fully appropriate in the model with explanatory CPL and IRUI variables. The dependence can be expressed in the form:

$$\text{GERD} = -1.8247 + 0.0209\ \text{CPL} + 0.0231\ \text{IRUI}. \tag{4}$$

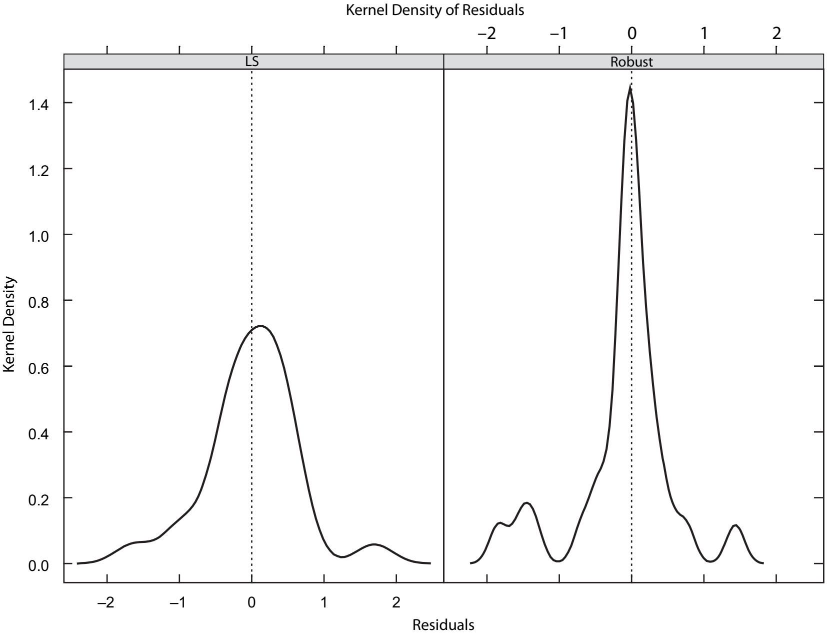
The index of determination R-sq. equals 0.6629. Both the explanatory variables have a positive influence on GERD, the partial coefficients being statistically significant at a 3% level at least. Comparative price levels (CPL) indicie the ratio between purchasing power parities (PPPs) and the market exchange rate in a particular country. The ratio is calculated in relation to the EU average (EU27 = 100). If the CPL index for a country is higher/lower than 100, the country concerned is relative expensive/cheap compared to the EU average. CPL is a measure of a nominal convergence. IRUI expresses the percentage of individuals regularly using the internet; it is one of indicators of information society expressing computer literacy of a country. In the EU countries, both a higher CPL value and a higher computer literacy, are connected with a higher expenditure on R&D. This conclusion is in general conformity with the European Commission recommendations in the area of "smart growth" promotion in the EU.

Figure 4 Diagnostic Plot (GERD ~ ER + LPH model)



Source: Author's own elaborations

Figure 5 Kernel estimate of residuals' density (GERD ~ ER + LPH model)



Source: Author's own elaborations

Table 2 Classical and robust bootstrap regression, LS and MM regression for GERD ~ ER + LPH model

		Observed	Bias	Mean	SE	95% EP	95% BCa
B R = 1000	Interc.	–5.4990	–0.0645	–5.5638	1.8579	–9.355; –1.974	–9.0248; –1.5642
	ER	0.0915	0.0001	0.0917	0.0309	0.0318; 0.153	0.0282; 0.1514
	LPH	0.0090	0.0007	0.0097	0.0045	0.0020; 0.019	–0.0000; 0.018
RB R = 1000	Interc.	–7.0419	0.0366	–7.005	2.0862	–10.98; –1.422	–8.8610; 1.6918
	ER	0.1108	0.0001	0.1109	0.0322	0.0215; 0.172	–0.0229; 0.1383
	LPH	0.0126	–0.0008	0.0118	0.0044	0.0007; 0.016	–0.0011; 0.0159
		Parameter	**SE**	**t**	**p-value**	**95% conf. interval**	
LS R-sq. 0.5639	Interc.	–5.4990	1.6750	–3.2830	0.0031	–8.9560; –2.0421	
	ER	0.0916	0.0265	3.4565	0.0021	0.0367; 0.1463	
	LPH	0.0090	0.0041	2.2213	0.0360	0.0006; 0.0173	
MM R-sq 0.5380	Interc.	–7.0419	1.6958	–4.1525	0.0004	–8.5971; –3.7757	
	ER	0.1108	0.0271	4.0953	0.0004	0.0664; 0.1435	
	LPH	0.0127	0.0045	2.8188	0.0095	0.0009; 0.0122	

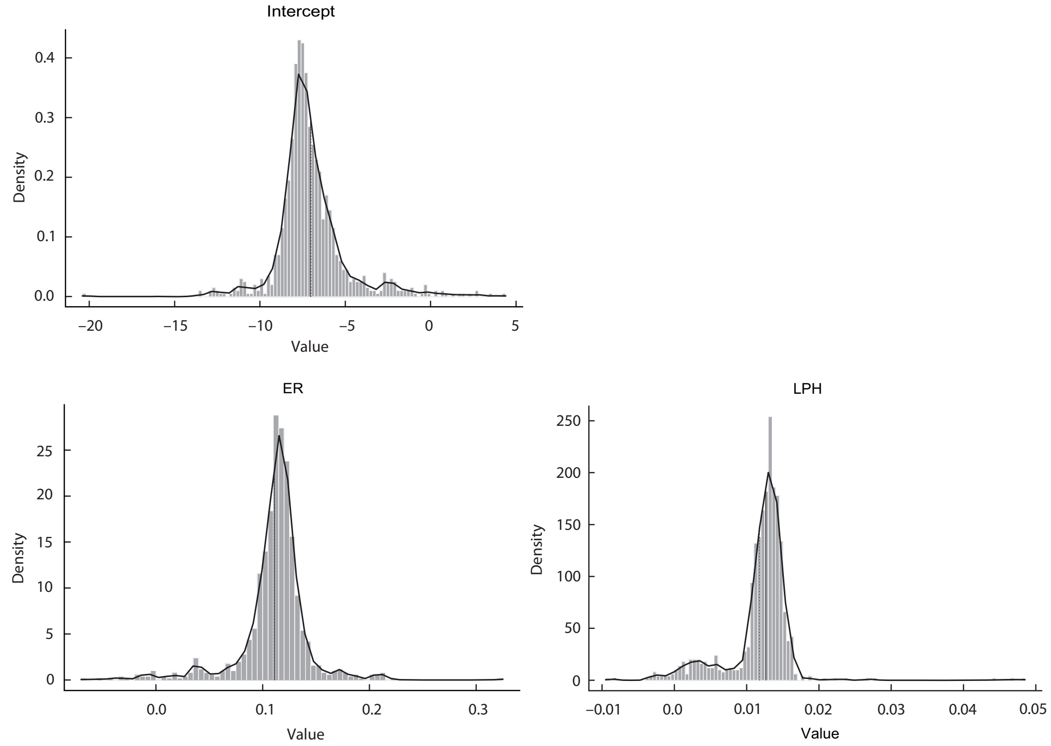
Goodness-of-fit tests for robust MM model	AICR	BICR	RFPE
	22.53	29.958	24.258

Source: Data EUROSTAT, author's own calculations

The last model includes exploratory variables ER and LPH. This model is quite distinct from the previous ones. Robust diagnostics reveal four vertical outliers (12 Cyprus, 15 Luxembourg, 18 Netherlands, 25 Finland) and seven leverage points. Two observations (12 Cyprus, 15 Luxembourg) are vertical outliers and leverage points simultaneously. These observations are thus identified as influential points. Classical diagnostics reveal only two vertical outliers and seven leverage points, none of them being identified as an influential point (see Figure 4). In such a case, the differences between classical and robust models are anticipated. For fitted values, see Table 2.

Multimodality of the kernel estimate of residuals' density plot (see Figure 5) confirms the presence of outlier points. The same is apparent from histograms of the regression coefficient estimates obtained by robust bootstrapping (Figure 6). Robust bootstrap provides tightly concentrated and markedly heavy-tailed distributions as a consequence of the existence of outliers. Robust bootstrap can be used as well, despite providing slightly biased estimates. It has to be taken into account, however, that the regression coefficients have higher standard errors and wider confidence intervals than those in the MM model (see Table 2).

Figure 6 Histograms for robust replications of regression coefficients for GERD – ER + LPH model (R = 1000)



Source: Author's own elaborations

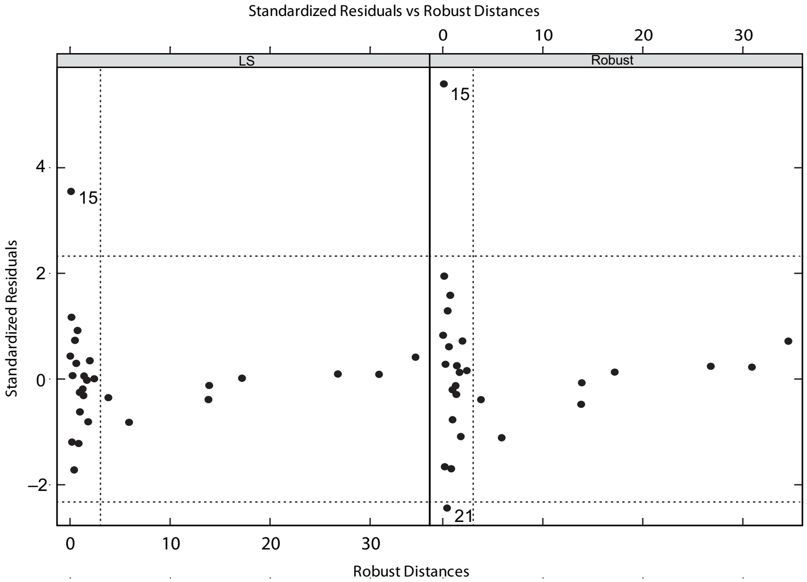
Due to the existence of influential points, the model estimated by robust regression has to be preferred. It is obvious that improper use of the classic LS regression model with significant variables without adequate identifications of outliers and testing of the normality of residuals, can lead to the acceptance of an incorrect LS model.

The exploratory variable ER (employment rate) is an indicator of labour market conditions. An increasing employment rate can lead to a decline in the percentage of GDP destined for unemployment and social security benefits, thus creating prerequisites for an increase in the proportion of GDP spent on research and development. LPH (labour productivity per hour worked) is intended to give a picture of the produktivity of national economies expressed in relation to the European Union average. If the index of a country is higher than 100, this country's level of GDP per hour worked is higher than the EU average. LPH is then a measure for the economic activity. The high level of economic activity and better working conditions are prerequisites for increasing the ratio of R&D expenditure. This could be expressed by the robust model:

$$GERD = -7.0419 + 0.1108\ ER + 0.0127\ LPH. \tag{5}$$

In the economic literature, the GERD indicator is more frequently perceived as a factor of labour productivity growth. In the analysed period (2010), the value of the Pearson correlation coefficient between GERD and LPH was 0.5888, the value of the robust correlation coefficient being 0.4744. We presented one of suitable regression model with regressors GERD and HICP (harmonised indices of consumer prices). In this model, both LS and robust diagnostics reveal the same vertical outlier (15 Luxembourg) and seven leverage points (see Figure7). Robust diagnostics identify another vertical outlier (21 Portugal). None of them is an influential point. Multimodality of the robust regression kernel estimate of residuals' density (see Figure 8) validates the presence of outlier points.

Figure 7 Diagnostic Plot LPH ~ GERD + HICP model

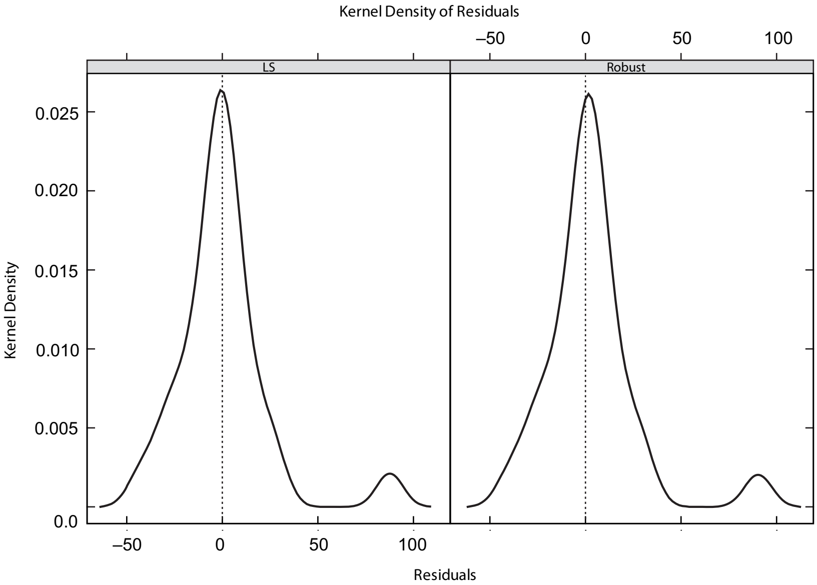


Source: Author's own elaborations

For the results of fits see Table 3. As far as GERD and HICP regressions with LPH as a dependent variable, the regression coefficients of both regressors are statistically significant (at a 5% level). The statistically significant regression coefficients indicate a positive influence of the ratio of R&D expenditure and a negative influence of inflation on labour productivity per hour worked. The resulting model estimated by robust regression has a form of:

$$LPH = 305.3371 + 11.4531\ GERD - 2.0088\ HICP. \tag{6}$$

Figure 8 Kernel estimate of residuals' density LPH ~ GERD + HICP model



Source: Author's own elaborations

Table 3 Classical and robust bootstrap regression, LS and MM regression for LPH ~ GERD + HICP model

		Observed	Bias	Mean	SE	95% EP	95% BCa
B R = 1000	Interc.	315.3248	3.7909	319.116	54.4330	224.47; 448.696	224.442; 448.661
	GERD	11.2297	0.4133	11.643	4.3786	3.0265; 20.3203	0.4397; 19.0223
	HICP	–2.0699	–0.0349	–2.105	0.4004	–3.0265; –1.4023	–2.9392; –1.3644
RB R = 1000	Interc.	305.3371	26.3859	331.723	190.089	71.232; 808.863	50.293; 707660
	GERD	11.4531	0.7014	12.154	9.411	–5.5374; 35.064	–3.7961; 37.1874
	HICP	–2.0088	–0.2294	–2.238	1.643	–6.5055; –0.2633	–5.6710; –0.1185
		Parameter	**SE**	**t**	**p-value**	**95% conf. interval**	
LS R-sq. 0.5605	Interc.	315.3248	76.1644	4.1401	0.0004	158.1292; 472.5204	
	GERD	11.2297	6.0305	1.8622	0.0749	–1.2166; 23.676	
	HICP	–2.0699	–0.6059	–3.4164	0.0023	–3.3204; –0.8195	
MM R-sq 0.6117	Interc.	305.3371	69.4729	4.3951	0.0002	197.0488; 406.8435	
	GERD	11.4531	5,4362	2.1068	0.0458	2.8750; 19.3544	
	HICP	–2.0088	0.5514	–3.6429	0.0023	–2.8062; –1.1466	
Goodness-of-fit tests for robust MM model					AICR	BICR	RFPE
					22.2319	29.1046	18.3399

Source: Data EUROSTAT, author's own calculations

CONCLUSIONS

The GERD (total gross domestic expenditure on research and experimental development as a percentage of GDP) is one of Europe 2020 headline indicators being tracked within the Europe 2020 strategy. The headline indicator is the 3% target for the GERD to be reached within the EU by 2020.

GERD is composed of expenditure of four institutional sectors of production (business enterprise, government, higher education and private non-profit organizations). The EU countries are distinct in their structure of GERD and the ways of increasing the ratio of R&D expenditure, depending on their economic policies. In general, the value of GERD is closely linked with the country's economic development, labour market conditions and computer literacy of the population. The economic GERD analysis, however, was not the main focus of the present paper.

The statistical conclusions are not based exclusively on the results produced in this paper, but also on economic theories and research findings of the GERD variable analysis that are not explicitly referred to.

When the vertical outliers are not identified in the data, errors being normally distributed, classical LS regression is a fully appropriate method and should be preferred. In such a case, classical bootstrap regression provides even more accurate estimates of the regression parameters (with smaller standard errors and narrower confidence intervals) than LS regression. Classical bootstrap outstrips robust methods in all cases when the vertical outliers are not identified and errors are normally distributed regardless of the existence of leverage points. This conclusion was demonstrated in the GERD ~ CPL+ IRUI model.

In models with detected vertical outliers, robust regression ought to be preferred since it produces the best results. Problems with the outliers in bootstrap regression can be resolved using robust bootstrap methods. Robust bootstrap in such cases gives results similar to robust regression, but the confidence intervals are wider than the robust regression ones. This conclusion is relevant when the outliers in both x-direction (leverage points) and in y-direction (vertical outliers) are detected. With an increasing outlier's proportion, the accuracy of bootstrap estimates of the regression parameters declines. This conclusion is observed in LPH ~ GERD + HICP model.

In cases where more vertical outliers and leverage points are detected, robust regression should be preferred. The bootstrap distribution may be a rather poor estimator of the regression estimates' distribution. These results are relevant for both classical and robust bootstrap because of the proportion of the outliers in bootstrap samples which can be higher than that in the original dataset. Outlying and non-outlying observations have the same chance of belonging to any bootstrap sample and, consequently the proportion of outliers in a bootstrap sample can be even larger than the fraction of outliers that can be tolerated by robust estimates. Thus the distributions of the regression parameters have heavy tails, the confidence intervals of the regression parameters being wide. This conclusion is manifested by the results of the GERD ~ ER + LPH model.

To sum up, the findings of this study indicate that in situations when the vertical outliers are identified, robust regression with a high breakdown point ought to be given preference. It is evident that improper use of the classical LS regression model with significant variables without corresponding identifications of outliers and assessment of residual normality can lead to the acceptance of an incorrect LS model.

References

ALBU, N. Resarch and Development spending in the EU: 2020 growth strategy in perspective. *SWP Working Paper FG 1. SWP,* Berlin, 2011/Nr. 08.

CHEN, C. Robust Regression and Outlier Detection with the ROBUSTREG procedure [online]. *SUGI Paper,* SAS Institute Inc., Cary, NC., 2002. <http://www2.sas.com/proceedings/sugi27/p265-27.pdf>.

COLE, S. R. Simple bootstrap statistical inference using the SAS system. *Computer Methods and Programs in Biomedicine*, 1999, 60, pp. 79–82.

COLOMBIER, C. Growth Effects of Fiscal Policies: An Application of Robust Modified M-Estimator. *Applied Economics,* Vol. 41, Issue 7, 2009, pp. 899–912.

DACHS, B., KAMPIK, F., SCHERNGELL, T., ZAHRADNIK, G., HANZL-WEISS, D., HUNYA, G., FOSTER, N., LEITNER, S., STEHRER, R.,URBAN, W. *Internationalisation of business investments in R&D and analysis of their economic impact.* Luxembourg: Publications Office of the European Union, 2012.

DICICCIO, T. J., EFRON, B. Bootstrap confidence intervals. *Statistical Science,* 1996, 11(3), pp. 189–212.

EFRON, B. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 1979, 7, pp. 1–26.

EFRON, B. The bootstrap and Modern Statistics. *JASA,* 2000, 95(452), pp. 1293–1296.

EFRON, B., TIBSHIRANI, R. J., eds. *An Introduction to the Bootstrap.* New York: Chapman&Hall/CRC, 1993.

EUROPEAN COMMISSION. *EUROPE 2020. A European Strategy for smart, sustainable and inclusive grows* [online]. Brussels, COM, 2010. <http://ec.europe.eu/eu2020/pdf>.

FINGER, R., HEDIGER, W. The Application of Robust Regression to a Production Function Comparison – the Example of Swiss Corn. *The Open Agriculture Journal,* 2008, 2, pp. 90–98.

FOX, J. *Bootstrapping Robust regression models: Appendix to an R and S-PLUS. Companion to applied Regression* [online]. 2002. <http://cran.r-project.org/doc/contrib/FoxCompanion/appendix-bootstrapping.pdf>.

FREEDMAN, D. A. Bootstrapping regression models. *The Annals of Statistics,* 1981, 9(6), pp. 1218–1228.

GUELLEC, D., IOANNIDIS, E. Causes of Fluctuations in R&D Expenditures. A Quantitative Analysis. *OECD Economic Studies,* No. 29, 1997/II.

GUELLEC, D., POTTELSBERGHE DE LA POTTERIE, B. R&D and Productivity Growth: Panel Data Analysis of 16 OECD Countries. *OECD Economic Studies,* No. 33, 2001/II.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J., STAHEL, W. A. *Robust Statistics. The Approach Based on Influence Functions.* New York: J. Willey, 1986.

HAMPEL, F. R. Some Aspects of Model Choice in Robust Statistics. In *Proceedings of the 44th Session of the ISI*, Book 2, 1983, pp. 767–771.

HUBERT, M., ROUSSEEUW, P. J., VAN AELST. High-Breakdown Robust Multivariate Methods. *Statistical Science*, 2008, 23(1), pp. 92–119.

KROLL. H., ZENKER, A., SCHUBERT, T. *An Analysis of the Development of R&D Expenditure at Regional Level in the Light of the 3% Target.* Luxembourg: Publications Office of the European Union, 2009.

OLIVE, D. J. Applications of robust distances for regression. *Technometrics.* 2002, 44(1), pp. 64–71.

RONCHETTI, E. Robust Model Selection in Regression. *Statistics & Probability Letters,* 1985, 3, pp. 21–23.

ROUSSEEUW, P. J. Least median of squares regression. *Journal of the American Statistical Association*, 1984, 79(388), pp. 871–880.

ROUSSEEUW, P. J., LEROY, A. M. *Robust Regression and Outlier Detection.* New Jersey: J. Willey, 2003.

ROUSSEEUW, P. J., VAN ZOMEREN, B. C. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association,* 1990, 85(411), pp. 633–639.

RUPPERT, D., CARROLL, R. J. Trimmed Least Squares Estimation in the Linear Model. *Journal of the American Statistical Association,* 1990, 75, pp. 828–838.

SALIBIAN-BARRERA, M., ZAMAR, R. H. Bootstrapping robust estimates of regression. *The Annals of Statitics,* 2002, 30(2), pp. 556–582.

SAS 9.2 Help and documentation.

SOMMER, S., HUGGINS, R. M. Variable Selection Using the Wald Test and a Robust Cp. *Applied Statistics,* 1996, 45, pp. 15–29.

SPIŠÁKOVÁ, E. Influence of the Economic Crisis on the Fulfillment of the Target of Strategy Europe 2020 in the Area of Research and Development. In RAJMUND, M., eds. *Financial Aspects of Recent Trends in the Global Economy,* ASERS Publishing, 2013(1), No 1.

S-PLUS 6 Robust Library User's Guide. Seatle, Washington: Insightful Corporation, 2002.

STINE, R. An introduction to bootstrap methods, examples and ideas. *Sociological Methods and Research,* 1990, 18(2–3), pp. 243–291.

YOHAI, V. J. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics,* 1987, 15(20), pp. 642–656.

ZAMAN, A., ROUSSEEUW, P. J., ORHAN, M. Econometric applications of high-breakdown robust regression techniques. *Economics Letters,* Vol. 71, 2000, pp. 1–8.

ZHANG, J. T. An empirical Analysis for National Energy R&D Expenditures. *International Journal of Global Energy Issues,* 2006, Vol. 26, No 1–2, pp. 141–159.

ANNEX: LIST OF INDICATORS

CPL Comparative Price Level (EU-27 =100%); (tec00120),

ER Employment rate total (the ratio of employed persons aged 20–64 and the total population of the same age group (t2020_10); (tsdec420),

GERD	Gross domestic expenditure on R&D (total gross domestic expenditure on research and experimental development as a percentage of GDP; (t2020_20), (tsdec320),
GGD	General government debt (percentage of GDP); (tsdde410),
HBA	Households with broadband access to the Internet (percentage of all households); (tin00073),
HICP	Harmonised indices of consumer prices (2005 = 100); (tec0027),
HRST	Human Resources in Science and Technology (percentage of active population aged 25–64 years; (tsc00025),
HTE	High-tech exports; (tin00140),
ILCS	Individuals' level of computer skills (in percent) (tsdsc470),
IR	Inflation rate (HICP); (tec00118),
IRUI	Individuals regularly using the Internet (in percent; frequency of Internet access: once a week); (tin00091),
LLL	Life-long learning (participation in education and training; percentage of people aged 25–64); (tsdsc440),
LPH	Labour productivity per hour worked; (tec00117),
LPP	Labour productivity per person employed; (tec00116),
LTU	Long-term unemployment, total (annual average; percentage of active population); (tsdsc330),
PUSE	Persons with upper secondary or tertiary education attainment (in percent), 25–64 years; (tps00065),
REER	Real effective exchange rate (index, 2005 = 100); (tsdec330),
SRE	Share of renewables in gross final energy consumption (tsdcc110);
UR	Unemployment rate, total (percentage of the labour force); (tsdec450),
TEA	Tertiary educational attainment, age group 30–34 (t2020_41),
TEAT	Tertiary educational attainment, age group 25–64 (tps00065).