GREENHOUSE GAS EMISSIONS ANALYSIS IN EU COUNTRIES USING ROBUST REGRESSION APPROACH

Dagmar Blatná

Abstract

The aim of this paper is to demonstrate the possibilities and advantages of robust regression applications. The analyzed indicator – Greenhouse Gas Emissions (GGE) – is one of the headline indicators being tracked within the EU 2020 strategy. The GGE in the European Union countries depends on numerous economic indicators and relevant factors. The values of these indicators vary between the EU countries and, consequently, the occurrence of outliers can be envisaged in an analysis of greenhouse gas emissions. In such a case, the robust regression methods represent useful tools for analyzing dependencies. High breakdown-point robust regression methods allow to detect regression outliers, leverage points and influential observations as well. In terms of the level of the analyzed indicator, the set of the EU states is divided into two significantly different groups – the Eurozone and non-Eurozone countries, the regression analysis being performed for both the whole set of EU countries and the sub-groups.

Key words: robust regression, outliers, leverage points, greenhouse gas emissions

JEL Code: C39, C13, C52

Introduction

The GGE (Greenhouse Gas Emissions) is one of the headline indicators being tracked within the EU 2020 strategy for smart, sustainable and inclusive growth. In the area of sustainable growth, the Resource-efficient Europe initiative was established. For 2020, the EU has made a unilateral commitment to reduce overall greenhouse gas emissions from its 28 member states by 20 % compared to 1990 levels. This indicator shows total man-made emissions of the so-called Kyoto basket of greenhouse gases. It presents annual total emissions in relation to those observed in 1990. The aggregate greenhouse gas emissions are expressed in units of CO2 equivalents. In 2014, the European Commission proposed the domestic 2030 greenhouse gas reduction target of at least 40 % compared to 1990 together with other main building blocks of the 2030 policy framework for climate and energy policies.

Greenhouse gas emissions in the European countries depend on numerous indicators of general economic background – the level of economic development and activity, science and technology, the rate of employment, price level, etc. The values of these indicators vary between the EU countries and, consequently, the occurrence of outliers can be envisaged in the analysis of greenhouse gas emissions. In such a case, the classic statistical approach – the least squares method (LS) – may be highly unreliable, the robust regression methods representing acceptable and useful tools. The aim of the present paper is to demonstrate the applicability and advantages of robust regression in the European GGE analysis based on 2012 data , the economic and enviromental GGE analysis not beeing its main objective.

1 Methodology

Robust regression provides an alternative to LS regression that works under less restrictive assumptions. The primary purpose of a robust regression technique is to fit a model that describes the information contained in the majority of data, allowing much better regression coefficient estimates, particularly when outliers are present in the data.

The main analytic tool employed here is MM regression, MM-estimates (proposed by Yohai (1987)) combining a high breakdown point with good efficiency. MM regression is defined by a three-stage procedure (for details, see (Yohai, 1987) or (Rousseeuw, Leroy, 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, an M-estimate of 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. In the analysis, Tukey's bisquare loss function

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

has been employed, where e denotes the residuum, the tuning constant k equaling 4.685 for the bisquare loss function. A more detailed description of robust regression methods is available in (Rousseeuw, Leroy, 2003), (Yohai, 1987), SAS and SPLUS manuals.

In order to identify vertical outliers, leverage and influential points (observations whose inclusion or exclusion result in substantial changes in the fitted model), the least trimmed squares (LTS) regression with a high breakdown point has been used. The 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. The usual choice $h \approx 0.75n$ yields the breakdown point of 25 %; see (Hubert, Rousseeuw, Van Aelst, 2008). In this paper, the residuals associated with LTS regression and the robust distance for outlier identification has been employed. A more detailed description of the LTS regression method is available in, e.g. (Ruppert, Carroll, 1980), (Rousseeuw, Leroy, 2003) or (Hubert, Rousseeuw, Van Aelst, 2008).

So as to quickly visualize vertical outliers and leverage points, the regression diagnostic plots (those of the standardized residuals of robust regression vs. robust distances RD (x_i,)) have been used as well. Horizontal broken lines are placed at +2.5 and -2.5 and the vertical line at the cut-offs 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 the horizontal lines are regarded as vertical outliers; see (Rousseeuw, Van Zomeren, 1990).

2 Analysis Results and Discussion

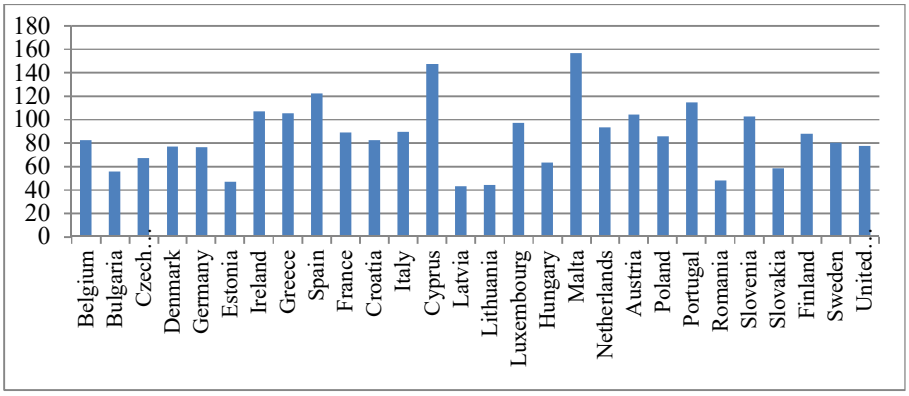
All analyses are based on 2012 data, calculations being performed by means of SAS 9.2 and S-Plus 6.2 statistical software and EXCEL. All the data as well as indicator definitions have been adopted from the Eurostat database[1]. The economic indicators employed in the analysis are given in an appendix to this paper.

Greenhouse gas emissions (GGE) differ greatly between the European countries, their values in the research year (2012) varying from 42.92 % (Latvia) to 156.9 % (Malta); see the graphical representation in Fig. 1. In terms of the level of the reference indicator, the set of the European countries can be divided into two groups – the Eurozone and non-Eurozone countries. The GGE difference is significant at the 5% level. The standard two-sample t-test

[1] http:// ec.europa.eu/data/database

and Wilcoxon rank-sum test output is available in Tab. 1. The following regression analysis is performed both for all the EU member states and for the two sub-groups of countries.

Fig. 1: Greenhouse gas emissions in the EU countries in 2012



Data Source: European Environment Agency. Author's elaboration.

Tab. 1: Summary statistics and two-sample tests for the difference of GGE levels

	Eurozone	Non-Eurozone	Standard two-sample t-test (not assuming equal variances)
Number	19	9	
Average	93.2447	70.9622	t = 2.6598
Variance	977.146	168.799	d.f. 26
Stnd. deviation	31.2593	12.9923	p-value = 0.0132
Maximum	42.92	47.96	
Minimum	156.90	85.85	Exact Wilcoxon rank-sum test
Stnd.skewness	0.2631	-0.8632	W = 320
Stnd. kurtosis	0.0274	-0.4234	p-value = 0.0284

Source: Author's calculation

2.1 Regression analysis for the set of 28 EU countries

For the regression analysis with the dependent variable GGE, the selected indicators from different economic fields have been taken into account as explanatory variables. Many regression models having been calculated – the fitting results, numerically robust diagnostics of outliers and leverage points, graphic identification of outliers (a diagnostic graph), goodness-of-fit robust tests and a plot of kernel residual density estimates were obtained for each model. The decision which of the candidate models is to be preferred is based on robust diagnostic selection criteria – the robust index of determination (Rsq.), robust deviance (D), significance robust tests (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, e.g. in (Ronchetti, 1985), (Ronchetti, 1997), (Hampel, Ronchetti, Rousseeuw, Stahel, 1996) or SAS and S-Plus manuals.

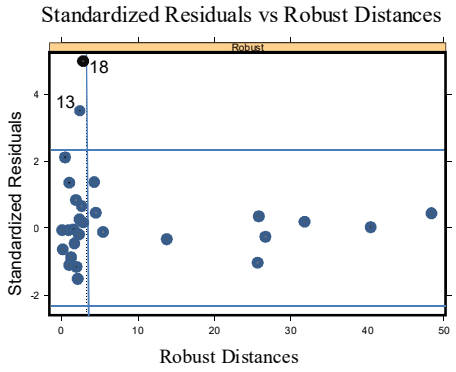
As an example, the results of the GGE dependence on the combination of ED (energy dependence), GDPG (GDP growth) and HICP (harmonized indices of consumer prices) explanatory variables are shown. This model belongs to the acceptable ones in all respects, satisfying the recommended ways for the model selection. In this model, robust diagnostic tests reveal ten leverage points and two vertical outliers (13 Cyprus, 18 Malta). The summary of robust diagnostics and fitting results are indicated in Tab. 2 and 3, respectively. For other diagnostic criteria, see Tab. 4. Multimodality of the kernel estimate of the residual density plot (see Fig. 3) confirms the presence of outlier points.

Tab.2: Robust diagnostics (GGE~ED +GDPG model)

Observation	Mahalanobis distance	Robust MCD distance	Leverage	Stand. robust residual	Outlier
2 Bulgaria	1.9834	6.9175	*	0.1889	
4 Czech Republic	2.3096	2.6414	*	0.6092	
6 Estonia	2.1264	6.2889	*	0.2666	
7 Ireland	1.7850	2.6060	*	0.3860	
8 Greece	3.0794	4.6448	*	-0.2387	
13 Cyprus	1.7101	1.8089		3.1691	*
14 Latvia	2.7004	8.5573	*	0.3170	
15 Lithuania	2.2850	6.2014	*	-0.9118	
17 Hungary	2.2339	6.3403	*	-0.2225	
18 Malta	1.8637	1.9272		4.4535	*
21 Poland	1.2772	2.5063	*	1.2323	
23 Romania	2.4176	7.7987	*	0.0499	

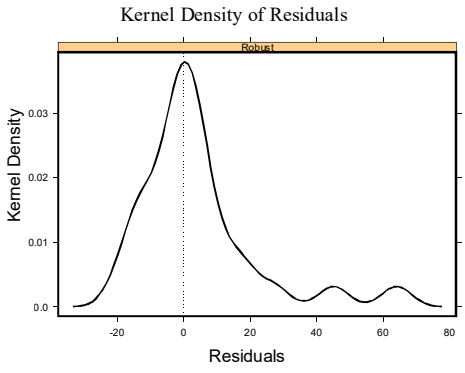
Source: EUROSTAT data, author's calculation

Fig. 2: Diagnostic plot (GGE~ED +GDPG+ HICP model)



Source: EUROSTAT data, author's elaboration

Fig. 3: Kernel estimate of residuals' density (GGE~ED +GDPG+HICP model)



Source: EUROSTAT data, author's elaboration

Tab. 3: GGE~ ED+ GDPG + HICP model fitting results

Parameter	Value.	Std. error	t-value	Pr(>|t|)
Intercept	198.9071	33.2998	5.9732	0.0000
ED	0.2201	0.1037	2.1225	0.0480
GDPG	-3.6856	1.2048	-3.0598	0.0054
HICP	-1.0416	0.2563	-4.0636	0.0004

Source: EUROSTAT data, author's calculation

The index of determination R-sq. of this model equals 0.596. As you can see from Tab. 3, all the partial regression coefficients are statistically significant (at a 0.05% level). In the EU countries, a higher energy dependence of the country and a lower grows of HDP and lower HICP, are connected with a higher level of greenhouse gas emissions.

Other acceptable robust regression MM models supplement by goodness-of-fit tests are shown in Tab.4. As you can see, in all acceptable models, The GDP grows is included as a significant exploratory variable.

Tab. 4: Goodness-of-fit tests of acceptable robust regression models

Outliers Leverage points	MM regression models	Rsq.	AICR	BICR	D	RFPE
O: 18,13 L:2,4,6,7,8,14,15,17,21,23	198.91+0.22 ED-3.685 GDPG -1.042 HICP	0.596	23.113	33.184	4146.6	17.274
O:**18** L: 4,6,8,14,15,18	62.079 +0.373 ED -5.928 GDPG	0.537	21.627	28.300	8271.7	11.828
O:9,13,18 L: 2,6,7,8,14,15,17,21,23	214.662 -3.712 GDPG -1.092 HICP	0.544	23.429	30.845	4749.5	14.155
O:13, 18 L:2,3,6,8,14,15,21,23,25	98.917 -0.092 EI -4.352 GDPG	0.518	22.417	29.592	6396.8	15.376
O:18 L: 6.8.13.14.15,	31.060 +398.609 EPHC - 4.978 GDPG	0.604	22.615	29.189	5082.9	12.604

Bold type indicates influential points. Numbers of countries identified as outliers: 18 Malta, 13 Cyprus, 9 Spain
Source: EUROSTAT data, author's calculation

2.2 Regression analysis for groups of the EU countries

2.2.1 Eurozone countries set

The acceptable robust regression models calculated for the Eurozone countries are presented in Tab. 5. In all models, 18 Malta was identified by robust diagnostics as an outlier observation, in one of them being an influential point at the same time (an outlier and leverage point simultaneously). The predictor growth of GDP (GDPG) was included in all acceptable models both for all EU countries (see Tab. 4) and the Eurozone set (see Tab. 5). The higher grows of GDP is connected with lower values of greenhouse gas emissions in the Eurozone of EU countries .

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

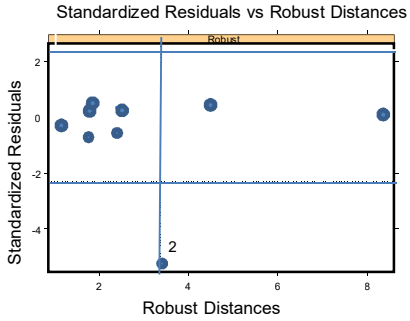
Outliers Leverage points	MM regression models	Rsq.	AICR	BICR	D.	RFPE
O: 18 L: 6,7,8,14,15	173.761+0.325 ED - 4.898 GDPG -0.879 HICP	0.5432	15.109	23.639	4887.6	12.314
O: 18, L: 6,8,13,14,15	36.776 + 376.291 EPHC - 6.356 GDPG	0.6208	14.057	23.331	3416.4	10.128
O: 13,18 L: 6,8,14,15,24,25	109.064 – 4.188 GDPG – 0.119 EI	0.5334	15.351	21.966	4549.8	11.504
O: **18**,8 L: 13,18	63.175 - 9.204 GDPG + 250.632 EPIC	0.4678	15.317	21.027	4877.4	13.817

Bold type indicates influential points. Numbers of countries identified as outliers: 18 Malta, 13 Cyprus, 8 Greece
Source: EUROSTAT data, author's calculation

2.2.2 Non-Eurozone countries

One of the acceptable models included HICP, CPL and HTE exploratory variables. As we can see from a graphical outlier detection tool – the Standardized Residuals vs Robust Distances Plot (see Fig. 4), one observation (3 Czech Republic) is identified simultaneously both as a leverage point and an outlier, thus being found as an influential point. Goodness-of-fit test outputs for acceptable robust models with the dependent GGE variable for the set of non-Eurozone countries are presented in Table 6.

Fig. 4: Diagnostic plot (GGE~ HICP+CPL+HTE model)



Source: EUROSTAT data, author's elaboration

Tab. 6: Goodness-of-fit tests of acceptable robust regression models – non-Eurozone countries

Outliers Leverage points	MM regression models	Rsq.	AICR	BICR	D.	RFPE
O: **3** L: 3,4,17	297.327 -1.475 HICP – 0.279 CPL +0.640 HTE	0.7112	5.395	13.456	230.37	4.493
O: - L: 3,4	184.693 +0.165 ED -1.919 GDPG -0.932 HICP	0.5544	4.792	10.912	208.05	6.363
O: 3 L: 4,27,28	266.904-1.376 HICP -0.208 CPL	0.6167	6.687	11.066	290.22	4.759

Bold type indicates influential points. The number of a country identified as an outlier: 3 Czech Republic
Source: EUROSTAT data, author's calculation

As you can see from Tab. 6, for the Non-Eurozone countries' set, different combinations of exploratory variables were included. In all acceptable models, the HICP is included as a significant exploratory variable. Higher value of HICP is connected with lower value of greenhouse gas emissions in the Non-Eurozone countries.

Only one combination of predictors in models calculated for the whole set of 28 EU countries was accepted in the case of sub-group countries – namely the model with ED, GDPG and HICP explanatory variables. However, as can be seen from Tables 4–6, the values of the partial regression coeficients differ.

Conclusion

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 LS regression might lead to incorrect results, robust regression methods having proved more useful analytical tools; those with a high breakdown point (LTS) allowing to detect influential points as well.

For the regression analysis with the dependent variable GGE, the selected indicators from different economic fields have been taken into account as explanatory variables. The regression analysis being performed for both the whole set of EU countries and the sub-groups (the Eurozone and non-Eurozone countries). In all acceptable models for the dependent GGE variable in the EU countries, outliers have been identified, the application of robust regression thus being the best solution.

Only one combination of predictors in models calculated for the whole set of 28 EU countries was accepted in the case of sub-group countries – namely the model with ED, GDPG and HICP explanatory variables. However, as can be seen from Tables 4–6, the values of the partial regression coeficients differ, thus justifying the division of the EU countries into two sub-groups. Some other of acceptable robust regression models for the dependent GGE variable, suitable in term of goodness-of-fit tests, are presented as well. In all acceptable models both for all EU countries and the Eurozone set, in all acceptable models, the GDP grows is included as a significant exploratory variable. The higher grows of GDP is connected with lower values of greenhouse gas emissions in the Eurozone of EU countries. For the Non-Eurozone countries' set, different combinations of exploratory variables is included. In all acceptable models, the HICP is a significant exploratory variable. Higher value of HICP is connected with lower value of greenhouse gas emissions.

The aim of this paper was to demonstrate the advantage and applicability of robust regression in analysis of greenhouse gas emissions in the EU countries. It was to be borne in mind that an economic or environmental GGE analysis was not the primary focus of the present paper.

Appendix. List of indicators used in the presented models

EI	Energy intensity of the economy
ED	Energy dependence (%)
EGRS	Electricity generated from renewable sources
EPHC	Electricity prices for household comsumers
EPIC	Electricity prices for indutrial comsumers
CPL	Comparative Price Level
GDPG	Gross Domestic Product (growth) y/y change
GGE	Greenhouse gas emissions, base year 1990
HICP	Harmonized Indices of Consumer Prices - Annual average rate of change (%)
THE	High-tech export

Acknowledgment

This paper was processed with contribution of long term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague.

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. (2008). High-Breakdown Robust Multivariate Methods. *Statistical Science* 2008, *23* (1), 92-119.

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.*

Yohai, V. J. (1987). High Breakdown-point and High Efficiency Robust Estimates in

Regression. *The Annals of Statistics.*, *15.*(2.), 642-656.

S-PLUS. "S-PLUS 6 Robust Library User's Guide."

SAS. "SAS 9.2 Help and documentation."

Contact

Dagmar Blatná

University of Economics, Prague

Faculty of Informatics and Statistics

W. Churchill sq. 4

130 67 Prague 3

Czech Republic

blatna@vse.cz