18th AMSE Applications of Mathematics and Statistics in Economics
International Scientific Conference | Jindřichův Hradec, 2-6 September 2015

APPLICATION OF ROBUST REGRESSION METHODS IN AN ANALYSIS OF THE EUROPEAN COUNTRIES' SHARE OF RENEWABLE ENERGY IN GROSS FINAL ENERGY CONSUMPTION

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

Abstract
The renewable energy share in gross final energy consumption (SRE) is one of the indicators within the Europe 2020 strategy for smart, sustainable and inclusive growth. Its values depend on the general economic background, price level, energy dependence and other factors. The SRE as well as the above mentioned economic factors vary greatly in the European Union countries and, consequently, the occurrence of outlying observations can be anticipated in a respective analysis. In such a case, robust regression methods represent an acceptable and useful analytic tool, high-breakdown point methods allowing the detection of regression outliers, leverage points and influential observations as well. In terms of SRE level, the EU countries can be divided into three significantly different groups. The analysis, based on 2012 data, has been performed for both the whole set of EU countries and the subgroups. The aim of this paper is to verify the applicability of the robust regression in an analysis of the European countries' SRE, the economic and environmental SRE analysis not being its main objective.

Key words: *robust regression, outliers, leverage points, renewable energy.*

1. Introduction

In March 2010, the European Commission presented a ten-year plan – the Europe 2020 strategy for smart, sustainable and inclusive growth. The Commission proposed five headline targets to be achieved by 2020. The Europe 2020 program was divided into seven flagship initiatives supporting the main targets. In the Resource-efficient Europe initiative, three aims were set. The second energy and climate headline target of the Europe 2020 strategy is to increase the share of renewable energy in gross final energy consumption (SRE) up to 20 % by 2020. In 2014, the European Commission proposed an objective of raising the share of renewable energy to at least 27 % of the energy consumption by 2030.

The European renewable energy target has been broken down into national targets that reflect differences in resource bases and economic wealth.

The indicator of renewable energy share is defined as a share of renewables in gross final energy consumption, referring to the quantity of energy consumed within a country's territory (Eurostat data). The energy sources taken into consideration are hydro, geothermal, wind and solar power, biomass and the biodegradable fraction of waste.

The SRE values depend on the general economic background, price level, energy dependence and other factors. The share of renewable energy as well as the above mentioned economic factors differ greatly in the European Union countries, and thus the

occurrence of outlying observations can be anticipated in an analysis. In such a case, robust regression methods serve as an acceptable and useful analytic tool.

Robust regression techniques are rarely used in economic analysis; only a few applications can be found in the available literature. Zaman et al. (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. Finger and 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.

The aim of this paper is to verify the applicability of the robust regression in an analysis of the European countries' SRE, the economic and environmental SRE analysis not being its main objective.

The rest of the paper is organized as follows. Section 2 provides the methodology, introducing robust regression methods, outlier identification and model selection criteria. The outcomes of the analysis are presented, commented on and summarized in sections 3 and 4, respectively.

2. Methodology

A regression analysis is the most commonly used statistical tool for analyzing the dependences. 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). LS being an optimal regression estimator under the sets of assumptions on the distribution of the error term (normality, homoscedasticity, independence of the errors) and predicted variables. A classical statistical approach to regression analysis 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.

Robust regression analysis provides an alternative to the least-squares regression when fundamental assumptions are undermined by the nature of the data. The main purpose of robust regression is to provide resistant results in the presence of outliers. Robust regression limits the influence of outliers, thus achieving the coveted stability.

Two regression methods with a high breakdown point have been employed. Yohai MM-estimator is a special type of M-estimation. MM-estimation is a combination of high breakdown value and efficient estimation. MM regression is defined by a three-stage procedure; for details, see (Yohai, 1987) or (Rousseeuw and Leroy, 2003). At the first stage, an initial regression estimate (LTS or S-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

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

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

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

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

LTS regression with a high breakdown point is a reliable data analytic tool for the detection of vertical outliers, leverage and influential points (observations whose inclusion or exclusion results in substantial changes in the fitted model). 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, et al. (2008). A more detailed description is available in, e.g. Ruppert and Carroll (1980), Rousseeuw (2003) or Hubert et al. (2008).

In this paper, the following methods for the detection of outliers and leverage points have been employed:

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

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

where $\mathbf{T(X)}$ is the robust location estimate vector and $\mathbf{C(X)}$ is the scatter matrix for the matrix of covariates; more details see in (Olive, 2003).

The existence of vertical outliers or leverage points in the model can be quickly identified from the graphic outlier detection tool – the robust diagnostic plot (Standardized Residuals vs Robust Distances Plot). 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 regarded as leverage points, those lying above or below horizontal lines as vertical outliers, see e.g. in (Rousseeuw and van Zomeren, 1990), Chen (2002).

The decision which of the alternative robust regression models should be preferred was based on robust diagnostic selection criteria: the robust index of determination (*R*-squared), significance robust *t*, Wald and *F*-test, robust deviance (*D*) and robust selection information criteria – Robust Akaike's Information Criterion (*AICR*), Robust Schwarz-Bayesian Information Criterion (*BICR*) and Robust Final Prediction Error (*RFPE*), see e.g Stahel (1996), Ronchetti (1985), Sommer and Huggins (1996), SAS and SPLUS manuals. The normality of residuals was also taken into account to determine which model is to be preferred. It was considered using graphic tools – the Kernel density of residuals plot and Normal Q-Q plot.

Formulas for the above mentioned criteria are defined as

$$D = 2\left(\hat{s}\right)^2 \sum \rho\left(\frac{y_i - x_i^T \hat{\beta}}{\hat{s}}\right), \tag{4}$$

where $\hat{\beta}$ is the *MM*-estimator of β and $\hat{s}$ is that of the scale parameter in the full model,

$$AICR = 2\sum_{i=1}^{n} \rho(r_{i;p}) + \alpha p = 2\sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T \hat{\beta}}{\hat{\sigma}}\right) + \alpha p, \tag{5}$$

$$BICR = 2\sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T \hat{\beta}}{\hat{\sigma}}\right) + p\ln(n), \tag{6}$$

where $r_{i;p}$ are regression residuals connected with an *M*-estimate of parameters, $\hat{\sigma}$ is a robust estimate of σ, and p the number of parameters.

$$RFPE = \sum_{i=1}^{n} \rho\left(\frac{y_i - x_{p,i}^T \hat{\beta}_p}{\hat{s}_p}\right) + p\frac{\frac{1}{n}\sum_{i=1}^{n} \psi^2\left(\frac{r_i}{\hat{s}}\right)}{\frac{1}{n}\sum_{i=1}^{n} \psi'\left(\frac{r_i}{\hat{s}}\right)}, \tag{7}$$

where $r_i = y_i - x_{p.i}^T \hat{\beta}_p$ and $\psi = \rho'$ is the derivative of the loss function.

3. Analysis Results and Discussion

The analyzed indicator SRE is defined as a share of renewables in gross final energy consumption, referring to the quantity of energy consumed within a country's territory.

The analysis was based on 2012 data, calculations being performed through 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[1].

The share of renewable energy in the gross final energy consumption in 2012 ranged from 51.0 % in Sweden to 2.7 % in Malta (for graphical display, see Figure 9). Most differences stem from variations in natural resources, mostly in the potential for building hydropower plants and the availability of biomass.

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

Due to great SRE variability, the set of 28 EU countries can be divided into three relatively homogeneous groups, the main criterion being the geographic location (N=North, M=Middle, S=South). The results of ANOVA and multiple range tests are shown in Table 1 and 2, respectively.

Table 1. ANOVA Table

Source	Sum of squares	Df	Mean square	F-ratio	p-value
Between gross	1974.21	2	987.103	15.6	0:0001
Within groups	1638.19	25	65.527		
Total	3612.39	27			

Source: author's calculations

Table 2. Multiple range test

Set	Count	Average	Variance	Contrast	Difference	± limits
N	12	10.492	57.743	N-M	-21.942*	8.336
M	10	32.433	112.027	N-S	-4.568	7.138
S	6	15.060	49.209	M-S	17.373*	8.609
Total	28	16.825	133.792	*denotes a significant difference		

Source: author's calculations

Robust regression both for the whole set of 28 EU countries and the three subgroups was performed. The results are presented in sections 3.1 and 3.2.

3.1 Robust regression analysis for 28 EU countries set

Several robust regression models with SRE as a dependent variable and the set of exploratory variables were computed. In the presented models the following predictors have been included:

ED	Energy dependence %
EGRS	Electricity generated from renewable sources %
EI	Energy intensity of the economy
EPHC	Electricity prices for household consumers
HICP	Harmonized Indices of Consumer Prices-Annual average rate of change (%)
IA	Internet access at home (% of households)
GDPG	Gross Domestic Product (growth) y/y change

The model with Energy dependence (ED) and Electricity generated from renewable sources (EGRS) predictors is an acceptable one in all respects, satisfying the recommendations for the model selection as well. The energy dependence (in %) indicates the scope to which an economy relies on imports so that it can meet its energy needs. The indicator is calculated as net imports divided by the sum of gross inland energy consumption plus bunkers. Electricity generated from renewable sources (EGRS) is calculated as its ratio to the gross national electricity consumption in a given calendar year, the proportion of

electricity produced from renewable energy sources to the national electricity consumption thus being measured.

The robust diagnostics model (see Table 3) reveals six outliers and the same number of leverage points, two observations being both outliers and leverage points simultaneously (13 Cyprus and 27 Sweden), i.e. influential points. The same information can be drawn from graphical identification, namely the robust regression graph; see Figure 1. The model fitting results are presented in Table 4.

Table 3. Robust diagnostics (SRE~ED +EGRS model)

Observation	Mahalanobis distance	Robust MCD distance	Leverage	Stand. robust residual	Outlier
4 Denmark	2.2765	3.4617	*	-0.1696	
6 Estonia	1.6769	1.7477		2.8763	*
13 Cyprus	1.7548	2.8894	*	2.6289	*
14 Latvia	1.3234	2.2781		2.5909	*
15 Lithuania	1.0920	1.8104		2.6708	*
18 Malta	1.9677	3.2753	*	0.0511	
20 Austria	2.6742	4.2210	*	-0.2427	
22 Portugal	1.9243	2.7247	*	0.0975	
26 Finland	0.4749	0.9370		2.6742	*
27 Sweden	2.2501	4.0745	*	2.9162	*

Source: EUROSTAT data, author's calculation

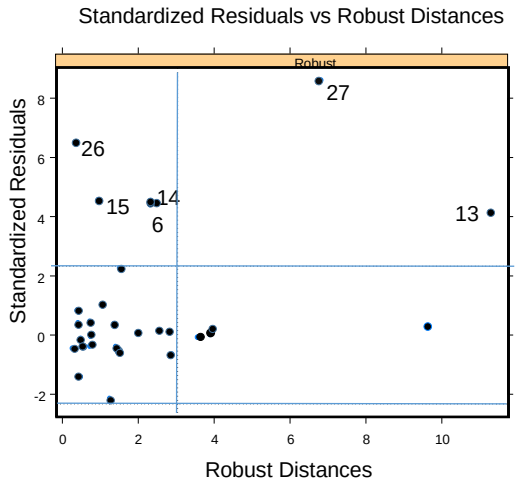

Figure 1. Diagnostic plot (SRE~ED + EGRS model)
Source: EUROSTAT data, author's elaboration

Table 4. Fitting results – SRE – ED + EGRS model

Parameter	Value of regr.coeff.	Standard error	t-value	Pr(>|t|) (p-value)	Wald test (Chi-sq)	P(>Chi) (p-value)
intercept	7.608	1.924	3.954	0.000	18.29	0.000
ED	-0.069	0.025	- 2.784	0.010	9.070	0.003
EGRS	0.439	0.042	10.438	0.000	127.46	0.000

Source: EUROSTAT data, author's calculations.

The index of determination R-sq. of this model equals 0.520. As you can see from Table 4, both the partial regression coefficients are statistically significant (even at a 0.01% level). In the EU countries, a lower energy dependence of the country and a higher proportion of electricity produced from renewable energy sources, are connected with a higher share of renewable energy in the gross final energy consumption. This conclusion is in general conformity with the European Commission recommendations in the area of the Resource-efficient Europe initiative in the EU.

For the kernel estimation of residuals' density, see Figure 2. Multimodality of the density validates the presence of outliers. This model meets all the above mentioned criteria.

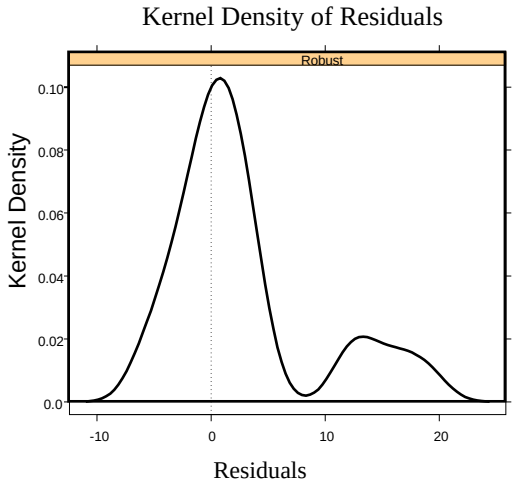

Figure 2. Kernel estimate of residuals'density (SRE~ED + EGRS model)
Source: EUROSTAT data, author's elaboration

Other regression models suitable in terms of goodness-of-fit tests as well as satisfying t and chi-sq. tests for individual parameters are presented in Table 5.

In all regression fits, the explanatory EGRS was included. The statistically significant partial regression coefficients indicate a positive influence of the proportion of electricity produced from renewable energy sources on value of share of renewable energy in the gross final energy consumption. In this same direction influence energy intensity and HICP (harmonized indices of consumer prices). On the other hand, the higher level of electricity prices for household consumers and higher energy dependence are connected with lower proportion of electricity produced from renewable energy sources.

Table 5. Acceptable robust regression models and robust goodness-of-fit tests

Outliers Lev. points	Robust fit	*R-sq*	*AICR*	*BICR*	*D*	*RFPE*
O: 6,14,15,26,**27** L:4,13,18,20,22,**27**	7.608– 0.069 ED + 0.439 EGRS	0.520	23,506	31.276	726.12	21.529
O: **6,**1526,**27** L:2,**6**,14,20,23,**27**	-1.396 +0.510 EGRS +0.019 EI	0.526	22.628	30.183	628.589	21.723
O : 26,**27** L:2,6,7,14,15,17,23, **27**	-36.352 + 0.472 EGRS + 0.328 HICP	0.583	21.970	29.578	513.522	19.612
O: 6,14,**15**,26,**27** L:6,9,13,14,**15**,20, 22,26,**27**	14.684– 86.162 EPHC + 0.442 EGRS	0.517	22.833	30.520	638.056	26.471
O: 6,14,**15,**26,**27** L:4,13,14,**15**,16,18, 20,**27**	13.570 –0,098 IA – 0.071 ED - 0.424 EGRS	0.552	22.372	33.451	657.873	19.857

Source: EUROSTAT data, author's calculations. Bold type highlights influential points.

3.2 Robust regression for subgroups of countries

The model regression fits for the subgroups of countries are presented in Table 6. In terms of the identification of outliers, the situation was fundamentally different. In the robust regression models for the whole EU set (see Table 5), a lot of outliers and influential points were identified. As far as subgroups are concerned, only in the S set an influential point was diagnosed (13 Cyprus). Robust diagnostics reveal an outlier for the N set as well, this one not being simultaneously a leverage point. In the regression fits for the M set, an outlier was not identified, the robust and classical LS regression fits being the same. For the M set, therefore, LS fits can be considered satisfactory, whereas for S and N sets the robust regression method is to be applied.

Table 6. Acceptable robust regression models for subgroup countries' sets

Outliers /Lev. points	Set	Robust fit	*R-sq*	*AICR*	*BICR*	*D*	*RFPE*
O:4 Denmark L: 6 Estonia	N	45.749 + 0.469 EGRS – 0.207 HICP	0.667	22.833	30.520	638.05	3.582
O:- L:-	N	15.785 +0.500 EGRS	0.586	4.600	6.484	118.85	3.614
0 L:10 France, 20 Austria	M	17.068 + 0.456 EGRS – 105.329 EPHC	0.776	9.831	14.389	21.800	5.713
O:- L: Austria	M	2.861 + 0.4344 EGRS	0.490	8.733	11.378	67.258	6.135
O:**13** Cyprus L: 13Cyprus, 18 Malta	S	13.544 + 0.346 EGRS – 0.107 ED	0.778	5.839	10.530	39.768	4.956
O:22 Portugal L :2 Bulgaria,23 Romania	S	20.249 + 0.223 GDPC – 0.362 ED	0.717	7.311	12.879	31.020	4.863
O:13 Cyprus L:13Cyprus, 18 Malta	S	15.268 + 0.390 ERGS – 82.740 EPHC	0.588	7.725	11.260	55.897	8.965

Source: EUROSTAT data, author's calculations. Bold type highlights influential points.

Table 7 presents robust fits with the same explanatory variables calculated for the whole EU countries' set and for the subgroups. None of the explanatory variable combinations has

provided an acceptable regression model for all groups of countries. Even in situations when the same predictors are included in different sets, the regression equations differ from each other. This confirms the suitability of the division of the EU countries into relatively homogeneous groups.

Table 7. Comparison of robust fits for the EU set and subgroups

Outliers/Leverage points	set	Robust fit	*R-sq.*
O: 6,14,15,26,**27** L:4,13,14,16,18,20,22,27	EU	7.608 + 0.439 EGRS – 0.069 ED	0.520
O:**13** / L: 13,18	S	13.544 + 0.346 EGRS – 0.107 ED	0.778
O: 26,**27** L:2,6,7,14,15,17,23,27	EU	-36.352 + 0.472 EGRS + 0.328 HICP	0.583
O: 4 / L: 6	N	45.749 + 0.469 EGRS – 0.207 HICP	0.667
O: 6,14,**15**,26,**27** L:6,9,13,14,15,20,22,26,27	EU	14.684 – 86.162 EPHC + 0.442 EGRS	0.517
O:- / L: 10,20	M	17.068 + 0.456 EGRS – 105.329 EPHC	0.776
O:**13** / L:13,18	S	15.268 + 0.390 ERGS – 82.740 EPHC	0.588

Source: EUROSTAT data, author's calculations. Bold type highlights influential points.

4. Conclusion

The share of renewable energy in gross final energy consumption is one of the indicators in the area of sustainable growth dealt with in the Europe 2020 strategy. The SRE is an important indicator for the assessment of targets' achievements regarding renewable sources of energy.

The SRE varies greatly in the European Union countries and, consequently, the occurrence of outlying observations was confirmed in the respective regression analysis, outliers and influential points having been identified in the two subgroups as well. In such situations, the robust regression methods represent an acceptable and useful analytic tool. In analyses when outliers were not identified, the robust and classical LS regression fits being the same. So, for the M set, LS fits can be considered satisfactory.

In all regression fits, the explanatory EGRS was included. The statistically significant partial regression coefficients indicate a positive influence of the proportion of electricity produced from renewable energy sources on value of share of renewable energy in the gross final energy consumption. This conclusion is in general conformity with the European Commission recommendations in the area of the Resource-efficient Europe initiative in the EU. In the same direction influence energy intensity and HICP (harmonized indices of consumer prices). On the other hand, the higher level of electricity prices for household consumers and higher energy dependence are connected with lower proportion of electricity produced from renewable energy sources in the European Union countries.

Several acceptable regression models have been calculated. For the final selection of a suitable model describing the European SRE dependence on the chosen explanatory variables, a thorough follow-up economic assessment is necessary.

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. COLOMBIER, C. Growth Effects of Fiscal Policies: An Application of Robust Modified M-Estimator. In Applied Economics, vol. 41, iss. 7, 2009, pp. 899-912.
3. EUROPEAN COMMISSION Documents and Working papers (2000-2014)
4. Finger, R., HEDIGER, W. 2008. The Application of Robust Regression to a Production Function Comparison – the Example of Swiss Corn. In The Open Agriculture Journal, 2008, 2, pp. 90-98.
5. HUBERT, M., ROUSSEEUW, P.J., VAN AELST. 2008. High-Breakdown Robust Multivariate Methods. In Statistical Science 2008, vol. 23, iss. 1, pp.92-119.
6. OLIVE, D.J. 2002. Applications of robust distances for regression. Technometrics.2002, vol. 44, iss. 1, pp. 64-71.
7. RONCHETTI, E. 1985, Robust Model Selection in Regression, Statistics & Probability Letters, 2008, 3, pp. 21–23.
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. In Journal of the American Statistical Association, 1990, vol. 85, iss. 411, pp. 633-639.
10. RUPPERT, D., CARROLL, R.J. 1990. Trimmed Least Squares Estimation in the Linear Model. In 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, Seattle, Washington. 2002
14. YOHAI, V.J. 1987. High breakdown-point and high efficiency robust estimates for regression. In The Annals of Statistics, 1987, vol. 15, iss. 20, pp. 642-656.
15. ZAMAN, A., ROUSSEEUW, P.J., ORHAN, M. Econometric applications of high-breakdown robust regression techniques. In Economics Letters, vol. 71, 2000, pp.1-8.