RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS
No. 66 2009
Towards Information-based Welfare Society

Dagmar Blatná
University of Economics, Prague, Czech Republic

EUROPEAN COUNTRIES ANALYSIS USING ROBUST REGRESSION METHODS

Abstract

European countries can be characterized by indicators of general economic background, employment, innovation and research, science and technology. Values of these indicators are varying among European countries. The most used statistical tool for analyzing dependences is the regression analysis. The classical statistical approach – the least squares method (*LS*) may be highly unsatisfactory in the presence of outliers which can be supposed in analysis of European countries data. In such a case robust regression is acceptable and useful tool. The paper proves that the estimates of regression coefficients obtained by using a robust regression method can be significantly different from the ones obtained in the case of classical regression. The differences in results are significant namely in the cases where outliers and leverage points are identified. Some regression models suitable both from the point of view of goodness-of-fit test and satisfying t-tests and chi-square tests for individual parameters of regression models for *Labour productivity per person employed* are presented.

1. Introduction

The goal of this paper is pointing out how robust regression methods can be used in the analysis of real European countries data. Three situations would be distinguished. Either outliers are identified by robust regression only or by both classical and robust regression or are not identified.

2. Data set and variables analyzed

The set of analyzed data contains the information about 29 European countries (27 members of the European Union, one candidate country (Turkey) and one non-EU member (Norway). The data were obtained from Eurostat. The analysis is based on the data of the year 2006 (the last year for which the data for analyzed countries were available).

The following indicators from different economic fields have been taken into account:

- *indicators of general economic background (economy and finance)*:
  - gross domestic product per capita in Purchasing Power Standards (*GDP*),
  - real GDP growth rate – percentage change (*gGDP*),
  - labour productivity per person employed (GDP in PPS per person employed) (*LP*)
  - total investment (% of GDP) (*INV*),
  - total state aid – % of GDP (*SA*);
- *indicators of prices:*
  - inflation rate (annual average rate of chance) (*IR*),
  - comparative price levels;
- *indicators of employment (labour market)*:
  - total employment rate – % (*EM*),
  - unemployment rate – total – unemployed persons as a share of the total active population (*UN*);
- *indicators of energy:*
  - electricity prices [euro per kWh] (*EP*),
  - energy intensity of the economy – gross inland consumption of energy divided by GDP (*EN*];
- *indicators of science and technology (information society statistics):*
  - gross domestic expenditure on R&D – % of GDP (*GE*),
  - ICT expenditure – % of GDP – communication expenditure (*ICT*),
  - ICT expenditure – % of GDP –information technology expenditure (*IT*),
  - turnover from innovation (% of innovative enterprises) (*IN*),
  - export of high technology product as a share of total exports (*HT*),
  - percentage of households having access to the Internet at home (*INT*);
- *indicators of labour market and education:*
  - total population having completed at least upper secondary education (*ED*),
  - early school leavers – total (*EX*).

In all cases, labour productivity per person employed (GDP in PPS per person employed) (*LP*) was considered as dependent variable, rest of indicators were taken as explanatory variables.

3. Regression analysis

A regression analysis is the most commonly used statistical tool for analyzing the dependences. The classical statistical approach – the least squares method (*LS*) may be highly unsatisfactory due to the presence of outliers which can be supposed in the analysis of the European countries data. In such a case, the robust regression is an acceptable and useful tool because it provides a good fit to the bulk of the data and exposes the outliers quite clearly.

Robust regression techniques are an important complement to classical least squares (*LS*) regression. Robust techniques provide results similar to *LS* regression when the data are linear with normally distributed errors. However, the results can differ significantly when the errors do not satisfy the normality conditions or when the data contain significant outliers.

It is common practice to distinguish between two types of outliers. Outliers in the response variable represent model failure. Such *observations are called **outliers**. Outliers with respect to the predictors are called **leverage points***. In regression it helps make a distinction between two types of leverage points: good and bad. ***A good leverage point*** is a point that is unusually large or small among the X values but is not a regression outlier. That is, the point is relatively removed from the bulk of the observation but reasonably close to the model around which most of the points are centred. A good leverage point has limited effect on giving a distorted view of how majority of points are associated. Good leverage points improve the precision of the regression coefficients. ***A bad leverage point*** is a point situated far from the regression model around which the bulk of the points are centred. In other words, a bad leverage point is a regression outlier having *X* value that is an outlier among *X* values as well (it is relatively far removed from the regression fit).

First, let us briefly mention the principle base of selected robust methods.

The least trimmed squares (*LTS*) estimators are obtained by minimizing

$$\sum_{i=1}^{h} r_{(i)}^{2} ,$$

where $r_{(\mathrm{i})}$ is the i-th order statistic among the squared residuals written in the ascending order,

$$h=[n/2]+[(p+1)/2]$$

and $[x]$ denotes the largest integer which is less or equal to x.

The *MM*-estimates are defined by a three-stage procedure. At the first stage an initial regression estimate is computed – it is consistent, robust, has high breakdown-point but is not necessarily efficient. At the second stage, an *M*-estimate of the errors scale is computed using residuals based on the initial estimate. Finally, at the third stage, an *M*-estimate of *MM* estimates is a combination of high breakdown value estimation and efficient estimate of the regression parameters based on a proper redescending ψ-function is computed.

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

A robust regression with high breakdown point *LTS* can be used to detect outliers, leverage points and influence points (the observations whose inclusion or exclusion result in substantial changes in the fitted model – coefficients, fitted values).

4. Identification of outliers

Many numerical and graphic diagnostics for detecting outliers and influential cases on the fit can be used (see e.g. [4]). In our paper the following ones have been used:

- *residuals associated with LTS regression;*
- *standardized residual*s (the residuals divided by the estimates of their standard errors; they have mean 0 and standard deviation 1);
- *studentized residual*s (a type of standardized residuals follow 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* defined as

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

  where T(X) and C(X) are the robust location and scatter matrix for the multivariates;
- *diagnostic plot*s are provided as a fundamental data mining graphical tools for quick identifying of an outlier and determining whether or not outliers have influence on the classical estimate. In the simple regression model, one can make a scatter plot, in order to visualize the data structure. In the multiple regression model with large *p* it is not sufficient to look at each variable separately or even at all plots of pairs of variables. The identification of outlying $(x_{i_1},...,x_{i_p})$ is more difficult problem.

To visualize the outliers and leverage points several diagnostic plots were proposed:

- a residual plot (a plot of a variable W_i versus the residuals r_i),
- a forward response plot (a scatter plot of the fitted values $\hat{Y}_i$ versus the response Y_i),
- a regression diagnostic plot (a plot of the standardized residuals of robust regression versus the robust distances $RD(x_i)$),
- a plot of the standardized residuals versus their index,
- a plot of the standardized residuals versus fitted values,
- a Normal Q-Q plot of the standardized residuals,
- a Distance-Distance plot (displays the robust distances versus the classical Mahalanobis distances),
- a leverage versus residual-squared plot.

5. Methods of model selection

When we need to decide which of two linear models with different independent variables to use, or which of several alternative linear models with different sets of independent variables to use we need robust test statistics and robust model selection criteria.

Robust index of determination R-squared for *M* (or *MM*) regression is defined as

$$R^2 = \frac{\sum_{i=1}^{n} \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right) - \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T \hat{\beta}}{\hat{s}}\right)}{\sum_{i=1}^{n} \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right)} = 1 - \frac{\sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T \hat{\beta}}{\hat{s}}\right)}{\sum_{i=1}^{n} \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right)} \tag{2}$$

where $\hat{\beta}$ is the *M* (or *MM*) estimator of β, $\hat{\mu}$ is the *M* (or *MM*) estimator of location, and $\hat{s}$ is the *M* (or *MM*) estimator of the scale parameter in the full model.

Robust deviance is defined as the optimal value of the objective function on the σ^2-scale:

$$D = 2(\hat{s})^2 \sum \rho\left(\frac{y_i - x_i^T \hat{\beta}}{\hat{s}}\right) \tag{3}$$

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

Significance robust tests of variables for determining which of two candidate models is preferred.

Robust t-test (*t*-statistics and *p*-value of the robust coefficient estimates for the robust fit are themselves robust because they are computed using a robust covariance matrix for the parameter estimates).

Robust Wald's test (*chi*-statistics and *p*-value of the robust coefficient estimates for the robust fit are computed using a robust covariance matrix for the parameter estimates).

Robust F-tests

$$F = 2\frac{n-p}{p-q}\sum_{i=1}^{n}\left[\rho\left(\frac{y_i - x_{q,i}^T \hat{\beta}_q}{\hat{s}_q}\right) - \rho\left(\frac{y_i - x_{p,i}^T \hat{\beta}_p}{\hat{s}_p}\right)\right], \tag{4}$$

where $\hat{\beta}$, β are robust *M* or *MM* estimates of regression parameters in full model and sub-model using an appropriate bounded function ρ ($q<p$), $\hat{s}_{p,}$ $\hat{s}_q$ are robust estimates of standard deviations in full model and sub-model.

Robust selection information criteria.

Robust Akaike's Information Criterion (AICR) defined as

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

where $r_{i;p}$ are regression residuals connected with *M*-estimate of parameters, σ is robust estimate of σ, p is the number of parameters. The best one model has the lowest *AICR* value.

Robust Bayesian information criterion (BICR) (sometimes also named the Schwarz information criterion). The formula for the *BICR* is defined as

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

The model with the smallest *BICR* value is chosen.

Robust Final Prediction Error (RFPE) is generalization to the AIC to robust model. For a p-dimensional model of p predictor variables, *RFPE* is calculated as

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

where $r_i = y_i - x_{p.i}^T\hat{\beta}$ and $\Psi=\rho'$ is the derivative of the loss function. When considering a variety of model choices with respect to several different choices of predictor variables, the model with the smallest value of *RFPE* is preferred.

6. Results of regression analysis of dependent variable labour productivity per person employed

The following regression methods have been applied in our analysis:

- least squares methods (*LS*),
- least trimmed squares regression (*LTS*),
- *MM* – regression,
- reweighted least squares (*RWLS*).

Software SAS 9.1 and S-PLUS 6.2 have been used in our analysis.

For analysis of dependent variable *PL* many of different combinations of explanatory variables were computed using above mentioned selected regression methods.

In the case of the classical *LS* regression, the classical *R*-square and the results of significance of *t*-tests and *F*-tests are used. In the case of the robust regression, the decision which of candidate model may be preferred is based on the robust diagnostic selection criteria mentioned above.

As an example, the results of dependence of the *labour productivity per person employed* (*GDP* in PPS per person employed) (*LP*) on combination of the explanatory variables *ICT* and *PL* are illustrated. This model belongs to proper ones from all points of view and satisfies the recommended ways for model selection. In this case one outlier and six leverage points were detected by using *LTS* regression (the summary of the robust diagnostic is shown in Table 1, for another diagnostic criteria see Table 5).

Table 1. Robust diagnostics

Observation	Robust MCD Distance	Leverage point	Stand. Robust Residual	Outlier
2 Bulgaria	5.2756	*	−0.1523	
4 Denmark	3.0000	*	3.0085	*
6 Estonia	6.1326	*	0.9011	
12 Cyprus	7.6344	*	0.6327	
19 Poland	3.9943	*	-1.2779	

Source: Občianske Združenie Financ, 2007, pp. 16-22.

As we can see, one observation (4 Denmark) is identified both as a leverage point and an outlier. Similar result can also be seen from the graphical outlier detection tools – Standardized Residuals vs. Robust Distances Plot (see Fig. 1).

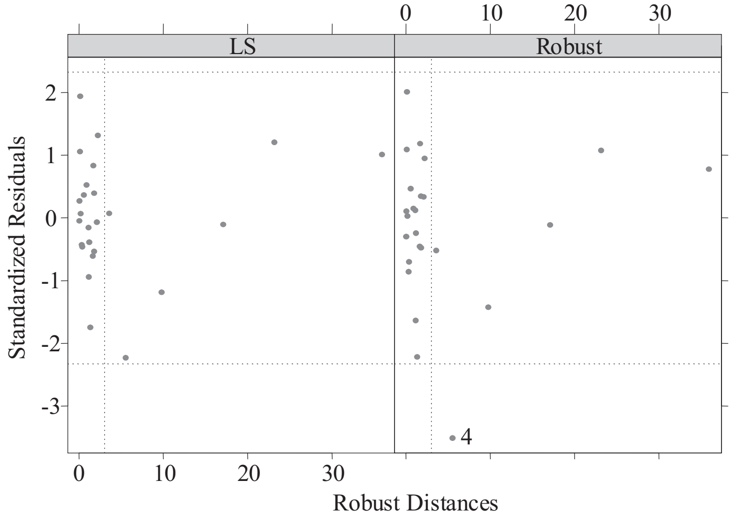

Fig. 1. Standardized Residuals vs. Robust Distances

Points outside the horizontal lines are regarded as residual outliers, and points to the right of the vertical line are leverage points. In our case, the *LS* fit produces no residual outliers, whereas the robust fit produces one outlier and six leverage points. One point (4) for the robust fit is both outlier and leverage point. The interpretation is that this point has substantial influence on the *LS* fit. (This example also illustrates the problem of outliers masked in the *LS* fit). In the case when outliers are identified, the difference between the regression *LS* fit and the robust fit can be anticipated. In Table 2, the model fitting for the example above (*LP~ ICT +PL*) is presented. One can find here the estimates of regression parameters obtained with the use of robust *LTS* method, *RWLS* and classical *LS* method.

Table 2. Model fitting results for *labour productivity per person employed* (robust *LTS*, robust *RWLS* and classical *LS* method)

Method	Parameter	Value of regression coefficient	Standard error	*t*-value	Pr(>|*t*|) (*p*-value)	Wald test (Chi-sq)	P(>Chi) (*p*-value)
LTS	Intercept	58.0363	29.3079	1.9802	0.0603		
RWLS	Intercept	58.1788	20.3251			12.12	0.0005
LS	Intercept	81.2465	21.9939	3.6941	0.0013		
LTS	*ICT*	−7.8344	3.6819	−2.1278	0.0448		
RWLS	*ICT*	−7.8836	2.2647			12.12	0.0005
LS	*ICT*	−9.9435	2.5224	−3.9421	0.0007		
LTS	*PL*	0.7493	0.1813	4.1334	0.0004		
RWLS	*PL*	0.7480	0.1359			30.29	<0.0001
LS	*PL*	0.5682	0.1426	3.9840	0.0006		

Source: own calculations.

Obviously, the estimates of regression parameters obtained by classical and robust methods are different both for intercept and partial regression parameters. Here *t*-value denotes the test statistic related to individual *t*-tests and Chi-sq connected with Wald test. Finally, the symbols Pr(>$|t|$) and P(>Chi) express the minimal significance level, where the null hypothesis can be rejected. Thus, all values estimated are significant at 5% level.

R-squares are 0.8990 for LS fit and 0.6879 for robust fit. However, in respect of existing bad leverage points and outliers, the use of a robust model is recommended.

If no outliers and bad leverages are identified, *LS* and robust regressions should provide similar results. As an example, we present the dependence of *LP* on *ICT* and *UN* (see Table 3). In this case, results of classical *LS* regression are quite satisfactory. Further, one can see from Table 3 that the results obtained by robust regression are very close to the ones obtained by classical *LS* regression if there are no outliers. We do not mention estimations made with the use of *RWLS* method – they are the same as the estimations made with the use of *LTS* method.

Table 3. Model fitting results for *labour productivity per person employed* (robust *MM*, and classical *LS* method)

Method	Parameter	Value of regression coefficient	Standard error	*t*-value	Pr(>|t|) (*p*-value)	Wald test (Chi-sq)	P(>Chi) (*p*-value)
MM	Intercept	178.7478	10.0201			318.23	<.0001
LS	Intercept	178.1943	9.2358	19.2938	0.0000		
MMt	*ICT*	−17.7370	1.9492			82.81	<.0001
LS	*ICT*	−17.5317	1.6991	−10.318	0.0000		
MM	*UN*	−2.1825	1.1203			3.80	0.0514
LS	*UN*	−2.1681	1.0421	−2.0805	0.0493		

Source: own calculations.

Further, very close coincidence in results obtained with the use of robust and classical regressions arises even in cases, when the same outliers are identified both by robust and classical diagnostic tools. This case can be demonstrated on the example of dependence *LP* ~ *IN*, where outliers 8 (Greece) and 25 (Sweden) have been identified. The results are compiled in Table 4.

Table 4. Model fitting results for *labour productivity per person employed* (robust *LTS*, robust *RWLS* and classical *LS* method)

Method	Parameter	Value of regression coefficient	Standard error	*t*-value	Pr(>|t|) (*p*-value)	Wald test (Chi-sq)	P(>Chi) (*p*-value)
Robust	Intercept	−81.6731	23.2301	−3.5158	0.0019		
RWLS	Intercept	−81.6832	17.1435			22.70	<.0001
LS	Intercept	−71.5944	23.8497	−3.0019	0.0064		
Robust	*IN*	7.4764	1.0026	7.4574	0.0000		
RWLS	*IN*	7.4770	0.7387			102.45	<.0001
LS	*IN*	7.2092	1.0315	6.9894	0.0000		

Source: own calculations.

To describe the dependence of the *labour productivity per person employed* (*LP*) in the European countries, other regression models can be used. Some of them are compiled in Table 5 presenting the results of both the *LS* and robust fit and goodness-of-fit tests for the robust models.

In Table 5, only regression model suitable both from point of view of goodness-of-fit test and satisfying t and chi-square tests for individual parameters are presented. Only the last three more complicated models obtained using backward stepwise selection with *RFPE* contain some parameters with non-significant tests.

Table 5. Another satisfactory regression models acceptable from the point of view of parameters significance tests for the variable *labour productivity per person employed*

Robust outliers (RO) Leverage points (LP)	Model		Regression equation	*R-sq.*	*AICR*	*BICR*	*RFPE*
RO: 4, 15, 24 LP: 4, 9, 20, 27	1	*MM*	$-10.016+1.266PC-0.971EM+0.71ED$	0.690	21.78	31.60	17.31
		LS	$30.46+1.171PL-1.048EM+0.340ED$	0.715			
RO: 4, 15 LP: 2, 3, 6, 9, 14, 20, 21, 27	2	*MM*	$-26.185+0.924PL+0.562ED-0.018EN$	0.719	22.16	31.61	14.51
		LS	$16.335+0.764PL+0.287ED-0.027EN$	0.725			
RO: 25 LP: 2, 6, 8, 12, 19, 25	3	*MM*	$66.431-11.569ICT+3.113IN$	0.676	18.08	24.50	15.29
		LS	$85.258-13.568ICT+2.692IN$	0.863			
RO: 4 LP: 2, 6, 12, 19	4	*MM*	$58.036-0.834ICT+0.749PL$	0.688	20.60	27.19	14.71
		LS	$81.246-9.945ICT+0.568PL$	0,898			
RO: 4 LP: 2, 6, 9, 12, 18, 19, 25	5	*MM*	$21.595-7.21ICT+0.793PL+0.366ED$	0.744	19.44	28.53	13.03
		LS	$44.661-9.26ICT+0.614PL+0.367ED$	0.911			
RO: 4, 14 LP: 2, 5, 6, 9, 12, 14, 18, 19, 22, 23	6	*MM*	$9.933-5.394ICT+0.902PL+0.394ED-12.287SA$	0.718	18.56	30.95	18.70
		LS	$53.01-8.925ICT+0.589PL+0.317ED-5.458SA$	0.903			
RO: 15 LP: 3, 4, 7, 9, 20, 21, 27	7	*MM*	$1.837+1.295PL+0.872ED-6.147EX-0.983EM$	0.725	20.91	33.11	16.35
		LS	$21.829+1.440PL+0.674ED-11.044EX-0.807EM$	0,792			
RO: 14, 11, 22 LP: 3, 4, 5, 6, 8, 9, 11, 14, 18, 19, 22, 23, 24, 25	8	*MM*	$-48.804-1.053EM+1.051ED+3.324IN-1.184HT+1.295PL-10.346EX$	0.724	17.67	35.21	21.16
		LS	$-31.379-0.825EM+0.882ED+2.066IN-0.537HT+1.147PL-7.332EX$	0.919			

Source: own calculations.

7. Conclusions

To select an acceptable regression method, the following way recommended in literature can be used: compare the *LS* and robust *MM*-estimate, if there is a significant difference, use robust regression method with high breakdown point (*MM*). Another applicable recommendation is to use backward stepwise variable selection with *RFPE* for selecting variables included in the final model. *RFPE* is computed at each step, and a variable is eliminated only if *RFPE* goes down.

With a view to existing outliers and bad leverages, it is recommended to prefer robust regression model against classical one in most cases. Unambiguous selection of suitable model describing the dependence of *labour productivity per person employed* in European countries on selected set of explanatory variables is impossible unless we prefer only one criterion for the selection of suitable model. All resulting robust regression models presented can be considered as satisfactory.

References

[1] Antoch J., Ekblom H., Víšek J.A., *Robust Estimation in Linear Model.XploRe Macros*, http://www.quantlet.de/codes/rob/ROB.htlm 1999.

[2] Blatná D., "Robust model selection criteria", [in:] *Applications of Mathematics and Statistics in Economy*, Univerzita Mateje Bela, Banská Bystrica 2007, Občianske Združenie Financ, 2007, pp. 16-22.

[3] Blatná D., "Outliers in regression. Trutnov 30.08.2006 - 3.09.2006", [in:] *AMSE 2006* [CD-ROM], KSTP VŠE, Praha 2006, pp. 1-6.

[4] Blatná D., "Robust regression", Prace Naukowe Akademii Ekonomicznej nr 1162, *Application of Mathematics and Statistics in Economics*, AE, Wrocław 2007, pp. 19-29.

[5] Blatná D., *Robust Regression in Analysis of Internet Access in European Countries*, Aplimat, Slovak University of Technology Bratislava 2008, pp. 1053-1061.

[6] Olive D., *Applied Robust Statistics*, preprint M-02-006, http://www.math.siu.edu/.

[7] *Regression with SAS*, http://www.ats.ucla.edu/stat/sas.

[8] *Robust regression*, http://en.wikipedia.org/wiki/Robust_regression.

[9] Rousseeuw P.J., Leroy A.M., *Robust Regression and Outlier Detection*, J. Wiley, New Jersey 2003.

[10] SAS 9.1.3 Help and Documentation.

[11] S-PLUS 6 Robust Library. User's Guide.

[12] *The home of the S-PLUS statistical software package*, http://www.insightful.com/.

[13] Wilcox R.R., *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, London 1999.

ANALIZA KRAJÓW EUROPEJSKICH ZA POMOCĄ METOD REGRESJI ODPORNEJ

Streszczenie

Kraje europejskie scharakteryzowane za pomocą wskaźników ekonomicznych, takich jak zatrudnienie, innowacje, badania naukowe, technologia, analizowane są za pomocą regresji odpornej. W pracy wykazano, że współczynniki regresji odpornej mogą się istotnie różnić od współczynników uzyskanych zwykłą metodą. Różnice powodowane są obserwacjami odstającymi. Odpowiednie modele regresji rozpatrzone są ze względu na wydajność pracy w przeliczeniu na jednego zatrudnionego.