POZNAŃ UNIVERSITY
OF ECONOMICS

Stages of the Convergence in the Developed European Economies

Edited by
Małgorzata Kokocińska

Poznań University of Economics Press
POZNAŃ 2012

Chapter 4

THE APPLICATION OF BAYESIAN METHODS IN STUDIES ON ECONOMIC GROWTH IN REGIONS

The considerations on the issue of economic convergence and inequality in European regions presented in the previous chapters, facilitate a better understanding of the issues of economic growth and convergence, and add to the extensive literature around the world on this subject. The results of the empirical studies discussed in chapters 2 and 3 provide grounds for the logical continuation of considerations in the field of analysis of the sources of economic growth in regions. It is one of the most significant challenges of contemporary theory of economics and economic policy to accurately identify the factors influencing the pace of economic growth. This task becomes particularly important when analyzing the regions of the European Union.

The literature on the subject, e.g. Sala-i-Martin et al. [2004] and J.C. Cuaresma et al. [2008], encompasses a range of studies that refer to various factors and groups of factors responsible for the processes of economic growth. These studies provide the foundation for the considerations below. There is consent in the literature that methods developed on the basis of Bayesian econometrics are generally applicable in the analysis of such a complex economic phenomenon as the determination of the sources of economic growth.

The purpose of this chapter is to present the Bayesian methods applied in the research into the factors responsible for regional economic growth. It presents the assumptions of the Bayesian pooling approach in regression models, the elements of Bayesian inference, and the details of the MC^3 algorithm applied in successive stages of this study. The chapter also discusses the issue of the collinearity of explanatory variables. These issues constitute the theoretical core of empirical studies on convergence in the EU regions, conducted in the following chapter.

4.1. The Bayesian pooling approach in linear regression models

This section concerns selected elements of Bayesian inference in a linear regression model. It presents the estimation of the model parameters, comparison of competitive specifications, and Bayesian pooling approach [cf. Osiewalski and Steel 1993]. In this case, the Bayesian pooling approach means the averaging of posterior distributions of the parameters that are of interest, weighted by the posterior probabilities of individual specifications. In the case of regression models, this method is referred to as BMA (Bayesian Model Averaging) in the literature on the subject and has found numerous applications in such fields of science as medicine, sociology and economics [cf. Hoeting, Madigan, Raftery and Volinsky 1999; Steel 2011]. Another essential element discussed in this section is the presentation of a numerical procedure referred to as MC^3 (Markov Chain Monte Carlo Model Composition), which is a special case of the Monte Carlo method, based on Markov chains (Markov Chain Monte Carlo – MCMC) and applied in the BMA method.

Bayesian inference, along with the MC^3 algorithm, allows the selection of the most likely combination of independent variables (i.e. factors responsible for the processes of economic growth) from a very large set of variables to be made, as well as the calculation of the explanatory power of all the interesting models and their ranking from the most to the least likely one, and the averaging of posterior estimations (including the mean and variance), weighted with the posterior probability of the models.

Why is it necessary to apply a Bayesian pooling approach and numerical techniques for a simple regression model if the estimation results can be obtained analytically, provided that prior distributions are selected accurately[20]? The answer is very simple. When the number of independent variables in a regression model is very large, it is very time consuming, or virtually impossible, to compute all the possible combinations of these variables. It also frequently turns out that the model with the greatest explanatory power has small posterior probability. When we focus on this one model exclusively, we actually ignore a vast amount of additional information provided by other models whose total posterior probability can be very high.

Madigan and Raftery [1994] additionally indicated that popular methods of choosing variables applied in the classical approach can lead to different selections of independent variables, and thus to different conclusions.

[20] Cf. Zellner [1971].

Let us consider, for example, a regression model with three potential independent variables X_1, X_2, X_3. In this case, we have $L=2^3=8$ linear combinations of independent variables. They can be listed as follows:

$$M_1 : y = \alpha_0 + e, \quad M_2 : y = \alpha_0 + \alpha_1 X_1 + e,$$

$$M_3 : y = \alpha_0 + \alpha_2 X_2 + e, \quad M_4 : y = \alpha_0 + \alpha_3 X_3 + e,$$

$$M_5 : y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + e, \quad M_6 : y = \alpha_0 + \alpha_1 X_1 + \alpha_3 X_3 + e,$$

$$M_7 : y = \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 + e, \quad M_8 : y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + e.$$

Let us assume that a random element has normal distribution. Given conjugate prior distributions, the estimation of the parameters in all the above-listed models can be performed analytically, without the need for numerical methods[21]. The assumptions adopted also allow the analytical calculation of the explanatory power of competitive models to be performed, and for the most probable posterior one to be determined.

In the classical approach, a typical procedure for regression model construction involves the estimation of the parameters of the model followed by the rejection of insignificant variables so that a single accurate model is designed. In this approach, the uncertainty related to the explanatory value of a model is ignored, which means that we are unable to estimate its probability. The difference between Bayesian inference and the classical approach is, among other things, that the former takes into account the uncertainty related to the selection of a model by means of calculating its posterior probability. Let us assume that model five (M_5) obtained the highest explanatory power with posterior probability amounting to 0.3. All the remaining models had a smaller explanatory power, but their total likelihood amounted to 70%. If we analyze only one model we will ignore an abundance of additional information included in the remaining models. That is why it is sometimes necessary to apply the Bayesian pooling approach involving, among other things, the averaging of parameter estimations and their posterior distributions weighted with the posterior probabilities of individual specifications. We can make conclusions on all the interesting values not only on the basis of a single model, but all the models, consistent with their explanatory power.

[21] The concept of conjugate distributions says that if the prior distribution of a parameter we are interested in belongs to a given family of distributions, then its posterior distribution also belongs to the same family for any size of an *n* sample and any number of observations. An ideal family of distributions is one allowing an easy point estimation of a parameter to be obtained, and one flexible enough to easily express initial information.

When a set of potential independent variables comprises 30 elements, the number of possible combinations increases to as many as $L = 2^{30} = 1\,073\,741\,824$. Assuming that computing each combination takes only one second, it would take as long as 34 years to calculate all of them! Therefore we need a more efficient algorithm to compute combinations, in order to focus on the most likely variants and ignore those with negligible posterior probability. That is the purpose of the MC^3 algorithm, developed by Madigan and York [1995].

4.2. The elements of Bayesian inference

Let θ signify the vector of parameters, which is the object of an examination. Let us also assume that the initial information concerning this vector can be expressed by means of prior density. Let us subsequently consider an econometric model where the observation vector $y = (y_1,...,y_N)$ has a probability distribution expressed by the density function $p(y|\theta)$. The foundation of Bayesian inference concerning the vector of parameters q is a well-known Bayesian formula:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \propto p(\theta)p(\theta|y) \tag{1}$$

where $p(\theta|y)$ stands for the posterior density distribution, describing a researcher's 'final' knowledge of the parameter q, computed on the basis of initial (prior) knowledge and derived from the sample; $p(y)$ stands for the density of marginal distribution of the observation vector y, expressed for a continuous random variable as $p(y) = \int p(\theta)p(y|\theta)d\theta$; $p(y|\theta)$ is the sample density, which determines the degree of confidence concerning the values assumed by an examined phenomenon, given a set value of the parameter q. It corresponds to the probability function, i.e. $l(\theta;y) = p(y|\theta)$, or density treated as a function of the parameter given a fixed outcome.

The initial information on the parameters of a sample model is thus expressed in the prior distribution $p(q)$, whereas the information from the sample is included in the sample density $p(y|\theta)$. The final result obtained in Bayesian inference, in contrast to classical methods, is not a point estimation of a parameter, but its entire distribution. The equivalence of point estimation of a parameter known in classical inference is provided by the measures of the central tendency and the dispersion of posterior distributions, such as a median, expected value, variance and interquartile deviation.

In Bayesian inference likelihood can be used to express information on more than just parameters. Another fundamental issue concerns the estimation of the explanatory power of the econometric model and the calculation of its posterior probability.

Let us consider a set of mutually exclusive and competitive models $M_1,..., M_m$ and the corresponding prior probabilities $\Pr(M_1),...,\Pr(M_m)$, whose total amounts to one, i.e. $\sum_{r=1}^{m}\Pr(M_r)=1$. The posterior probability of any model M_i can be computed in the following manner on the basis of the Bayesian formula:

$$\Pr(M_i \mid y)=\frac{\Pr(M_i)\,p(y \mid M_i)}{\sum_{r=1}^{m}\Pr(M_r)\,p(y \mid M_r)}. \tag{2}$$

Formula (2) allows the posterior probability of every model $M_r\ (r=1,...,m)$ to be calculated provided that we know the density of marginal distribution $P(y \mid M_r)$. In a linear regression model density can be calculated analytically (cf. formula 10). In more complex models numerical integrations are necessary, which can significantly complicate the computation process [Osiewalski 2001].

As already mentioned, the BMA method consists in the averaging of posterior distributions of interesting parameters, weighted by the posterior probabilities of individual specifications. Let us describe the method in a more formal manner.

Let us assume that a researcher is interested in the parameter ψ, which is a common element of all competitive models. Since we know the posterior probability of each model, the following density of posterior distribution can be a source of information:

$$p(\psi \mid y)=\sum_{r=1}^{m}\Pr(M_r \mid y)\,p_r(\psi \mid y, M_r). \tag{3}$$

Density $p(\psi \mid y)$ is therefore obtained by means of the weighted averaging of individual densities of posterior distributions $p_r(\psi \mid y)$, weighted by the posterior probabilities of competitive models. Selected moments of posterior distribution can be averaged analogically:

$$E\left(\psi^s \left| y\right.\right) = \sum_{r=1}^{m} \Pr\left(M_r \left| y\right.\right) E_r\left(\psi^s \left| y, M_r\right.\right), \tag{4}$$

where s stands for the order of the moment $(s = 1,2,...)$.

4.3. Regression model and its estimation

Let us assume that we have data derived from $i = 1,...,N$ objects[22]. The vector of observations $y = (y_1,...,y_N)$ refers to the response variable. Let us also assume that we have *K* potential explanatory variables related to a response variable. The matrix with dimensions $N \times K$ contains observations on the explanatory variables. Let M_r stand for $r = 1, ..., m$ regression models, where *m* stands for a maximum number of combinations of independent variables, i.e. $m = 2^K$.

The considered regression model has the following form:

$$y = \alpha l_N + X_r \beta_r + \varepsilon, \tag{5}$$

where l_N means an $N \times 1$ vector of ones, X_r is an $N \times k_r$ matrix related to model M_r, and containing some (or all) columns of matrix X, β_r is a $k_r \times 1$ vector of structural parameters, α is an intercept coefficient, common for all regression models, a random component ε is a vector of dimensions $N \times 1$ and normal distribution $N\left(0, h^{-1} I_N\right)$, parameter h is an inverse variance of random component, i.e. $h = \frac{1}{\sigma^2}$, and the symbol I_N stands for an identity matrix of size *N*.

Let us assume that we have initial information on regression coefficients β_r, and some knowledge on common parameters, i.e. h and α:

$$\beta_r \left| h \sim N\left(0_{k_r}, h^{-1}\left[g_r X_r' \; X_r\right]^{-1}\right)\right. \tag{6}$$

and

$$p(h) \propto \frac{1}{h}, \; p(\alpha) \propto 1. \tag{6a}$$

[22] For more on the estimation of parameters in a linear regression model of algorithm MC^3 cf. G. Koop [2003].

Symbol $N(a,B)$ stands for a multidimensional normal distribution with mean a, and variance B, g_r stands for a constant defined as follows:

$$g_r = \begin{cases} \dfrac{1}{K^2} & \text{dla} \quad N \le K^2 \\ \dfrac{1}{N} & \text{dla} \quad N > K^2 \end{cases} \tag{7}$$

Using Bayesian formula (1) we obtain the posterior distribution of the parameters we are interested in. It can be demonstrated that in this case, the posterior distribution of the vector of regression coefficients β_r is a multivariate Students-*t* distribution with the following vector of means:

$$E(\beta_r \mid y, M_r) = \left[(1+g_r)X_r' X_r\right]^{-1} X_r' y \tag{8}$$

It can be easily observed that for small values, i.e. when $g_r \to 0$, the posterior expected values are very close to those that can be obtained by means of the classical method of least squares.

The matrix of posterior covariance has the following form:

$$\operatorname{var}(\beta_r \mid y, M_r) = \frac{N s_r^2}{N-2}\left[(1+g_r)X_r' X_r\right]^{-1}, \tag{9}$$

where $s_r^2 = \dfrac{\dfrac{1}{g_r+1} y' P_{X_r} y + \dfrac{g_r}{g_r+1}(y-\bar{y}l_N)'(y-\bar{y}l_N)}{N}$.

Given the above-mentioned assumptions, the sample density after the analytical integration of parameters in model r is as follows:

$$p(y \mid M_r) \propto \left(\frac{g_r}{g_r+1}\right)^{\frac{k_r}{2}} \left[\frac{1}{g_r+1} y' P_{X_r} y + \frac{g_r}{g_r+1}(y-\bar{y}l_N)'(y-\bar{y}l_N)\right]^{-\frac{N-1}{2}}, \tag{10}$$

where $P_{X_r} = I_N - X_r\left(X_r' X_r\right)^{-1} X_r'$.

On the basis of formulas (8) and (9), the posterior point estimations of regression coefficients can be calculated, while formulations (10) and (2) allow the posterior probability of each potential combination of explanatory variables to be calculated. More information on Bayesian inference in single-equation models in Polish is provided by J. Osiewalski [1991].

4.4. The MC^3 algorithm

Let us now discuss the foundations of the MC^3 algorithm. It facilitates easy 'capturing' of the models with the greatest explanatory power. Its main task is to sample in the regions where the most likely models occur, while neglecting the areas with the least likely models. The MC^3 algorithm, developed by Madigan and York [1995], is a special case of a numerical procedure, referred to in the literature on the subject as the Metropolis-Hastings method, which in turn is a special case of the Monte Carlo method, based on Markov chains. Before discussing the details of the MC^3 algorithm, let us first consider the foundations of the MCMC methods and the Metropolis-Hastings procedure. In the Polish literature Osiewalski [2001], Pajor [2003] and Marzec [2008], among others, discuss the MCMC methods in detail.

The fundamental idea of the MCMC methods is as follows: given a certain posterior distribution with density $p(\theta|y)$ a Markov chain $(\theta^1, \theta^2, ...)$ is generated whose stationary distribution is posterior distribution. Thus we generate d initial samples and test their convergence with the stationary distribution (burn-in period). After convergence occurs, further stages $(\theta^{d+1}, \theta^{d+2}, ...)$ involve the samples from the distribution $p(\theta|y)$ and can serve the purpose of approximation of its characteristics.

In order to estimate the parameters in the Metropolis-Hastings algorithm, samplings from the proposal density $f(\theta^*|\theta^{i-1})$ are used to generate a Markov chain. A general outline of the Metropolis-Hastings algorithm is as follows:

1. Starting point θ^0 is selected
2. For successive iterations $i = 1, 2, \ldots$:
   - value θ^* is sampled from a proposal density $f(\cdot)$,
   - $\theta^* \leftarrow f(\theta^*|\theta^{(i-1)})$,
   - acceptance probability $\alpha(\theta^{(i-1)}, \theta^*)$ is calculated,
   - $\theta^{(i)} = \theta^*$ is assumed with the likelihood $\alpha(\theta^{(i-1)}, \theta^*)$ as well as $\theta^{(i)} = \theta^{(i-1)}$ with the likelihood $1 - \alpha(\theta^{(i-1)}, \theta^*)$.
3. After the chain's convergence occurs, the characteristics of posterior distribution are calculated on the basis of iterations performed.

The appropriate selection of a proposal density is a crucial element of the Metropolis-Hastings algorithm. An ideal proposal density coincides in shape and location with the posterior distribution. A good proposal density should meet the following criteria:

- it can be easily sampled from,
- acceptance probability can be easily calculated,

– each successive iteration is relatively distant from the previous one (it ensures better convergence with the stationary distribution),
– successive iterations are accepted relatively frequently (otherwise the Markov chain remains in one point for too long).

In the MC^3 algorithm a chain of models is generated instead of the values of parameters. The general idea of the MCMC does not fundamentally change. Let us mark the model sampled (and accepted) in the i iteration of the chain as M^i. The successive steps of the MC^3 algorithm can be presented in the following manner:

1. The starting model M^i $(i=1)$ is adopted.
2. For $i=2$, from the set of models containing: (a) model $M^{(i-1)}$ sampled and accepted in the previous step; (b) all models derived as a result of the elimination of a single independent variable from the model $M^{(i-1)}$; (c) all models derived as a result of the addition of a single independent variable from the model $M^{(i-1)}$; a candidate model M^* is sampled (with uniform likelihood).
3. If the posterior probability of a sampled model is significantly higher, it replaces the former model $M^* = M^{(i-1)}$. Otherwise we return to step 2 and adopt $M^{(i)} = M^{(i-1)}$.

The formal presentation of the acceptance probability is as follows:

$$\alpha\left(M^{(i-1)}, M^*\right) = \min\left\{\frac{p\left(y|M^*\right)p\left(M^*\right)}{p\left(y|M^{(i-1)}\right)p\left(M^{(i-1)}\right)}, 1\right\}, \tag{11}$$

where densities $p\left(y|M^*\right)$ i $p\left(y|M^{(i-1)}\right)$ are calculated from the formula (10). When all models have the same prior probability, the formula (10) is reduced to a simpler form:

$$\alpha\left(M^{(i-1)}, M^*\right) = \min\left\{\frac{p\left(y|M^*\right)}{p\left(y|M^{(i-1)}\right)}, 1\right\}. \tag{12}$$

The posterior probability of models M_r, i.e. $\Pr\left(M_r|y\right)$, can be very easily approximated with the following formula:

$$\Pr\left(M_r|y\right) \approx \frac{\text{the number of samplings where model } M_r \text{ is accepted}}{\text{total number of samplings}}. \tag{13}$$

Posterior characteristics of selected parameters can be obtained by means of weighted averaging of individual posterior distributions (formula (3)) or Rao and Blackwell approach [cf. Koop 2003].

Summing up, it should be emphasized that Bayesian inference provides tools that describe the uncertainty related to the selection of a model in a strictly probabilistic manner. The above-mentioned MC^3 algorithm is in turn an efficient technique providing for sampling in the areas where the most likely models occur, while neglecting those where models with very small explanatory power emerge.

4.5. The issue of the collinearity of explanatory variables

Since the study involved a database of a relatively large size and adopted the classical assumption of a linear econometric model (linear multiple regression), independent variables cannot be collinear[23]. Two issues have to be borne in mind [Welfe 2003, p. 140]:

– collinearity is the characteristic of the set of data we possess and not of the variables themselves,
– there always occurs a correlation between economic variables, the only problem is the degree of this correlation.

Let X stand for the observation matrix on independent variables and Y – for the vector of observation on response variable. The OLS estimator is derived as a result of the following transformation:

$$a = \left(X^T X\right)^{-1} X^T Y \,. \tag{14}$$

If independent variables generate a linear combination (are strictly collinear), the determinant of square matrix $X^T X$ equals zero and there is no matrix that would be inverse to it $\left(X^T X\right)^{-1}$, meaning that there is no OLS estimator.

The problem of collinearity is slightly different when independent variables do not form a (strict) collinear combination, but are strongly correlated. The determinant of square matrix $X^T X$ then approximates zero, resulting in a distorted value of the variance matrix – the covariance of the estimates of structural parameters, determined using the formula:

[23] Cf. Maddala [2006], pp. 318-321; Gruszczyński et al. [2009], pp. 57-59; Gruszczyński et al. [2003], pp. 41-43; Gajda [2004], pp. 96-100; Greene [2003], p. 14; Davidson, MacKinnon [2004], p. 99; Kufel [2011], p. 64; Borkowski et al. [2003], p. 26; Welfe [2003], pp. 139-152.

$$D^2(a) = S_e^2\left(X^T X\right)^{-1}, \tag{15}$$

where S_e^2 is residual variance (the estimation of random component variances). The distortion of the variance-covariance matrix produces distorted *t*-Student statistics which serve the purpose of inferring on the significance of the influence of individual independent variables on the response variable, and the deformed estimates of structural parameters of the model. This is a paramount econometric issue which may lead to the following cognitive errors [Welfe 2003, p. 141]:

– failing to reject the hypothesis of the insignificance of a given independent variable when it actually exerts influence on the response variable,
– measure of the force of influence the independent variables exerted on the dependent variable.

In the case of collinearity, the variables (a variable) which generate this problem in a data set have to be rejected.

It follows from the above that the issue of collinearity is an essential issue related to the analysis of multiple regression. To verify whether this issue concerns the analyzed set of data, the method of matrix conditional number (CN), developed by Belsley, Kuh and Welsch (BKW) (presented in Belsley et al. [1980]), can be applied. It is a broadly applied measure allowing the variables which generate the issue of collinearity in a given database to be identified.

Another frequently applied measure of collinearity is variance inflation factor (VIF), based on the analysis of coefficients of multiple correlation between independent variables, although both instruments are criticized [Maddala 2006, pp. 318-321].

The procedure in the BKW method is as follows [Welfe 2003, p. 145]:

– transformation matrix: $X^* = P\left(X^T X\right)P$ is determined, where:

$$P = \begin{bmatrix} \dfrac{1}{\sqrt{x_{1(i)}^T x_{1(i)}}} & 0 & \cdots & 0 \\ 0 & \dfrac{1}{\sqrt{x_{2(i)}^T x_{2(i)}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{1}{\sqrt{x_{k(i)}^T x_{k(i)}}} \end{bmatrix}, \tag{16}$$

where *k* is a number of independent variables,

– eigenvalues of matrix X^* are determined,
– CN measure is determined as:

$$\mathrm{CN} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \ , \tag{17}$$

where the index value above 20 is considered an indication of the potential presence of collinearity.

The study presented in the following chapter of this book was carried out on the basis of the original software[24] allowing the MC^3 algorithm to be extensively applied. The implementation of the BKW method in a MATLAB environment, licensed by GNU GPL, was created by James P. LeSage (Department of Economics, University of Toledo; http://www.business.txstate.edu/users/jl47/). The BMA solution presented in this study applies this very code. The function implementing a CN measure is automatically activated each time the order of matrix X is smaller than k.

The example of the application of the BKW method (Figure 28) results from the analysis obtained by means of this original software, and is an initial step in the empirical application of the MC^3 algorithm.

The BKW procedure produced a matrix, where the first column contains the values of CN conditional indicators (arranged in a non-decreasing order), whereas the rows contain determined Pearson coefficients of linear correlation between individual variables, which is demonstrated in Figure 31. This arrangement of results means that collinear variables need to be sought in the last rows, and the number of variables to eliminate can be determined on the basis of the conclusion printed below the results of the BKW procedure, beginning with "Rząd macierzy wynosi…" ("The order of matrix amounts to"), circled in red in Figure 31.

There are $39 - 37 = 2$ collinear variables in the presented example. The numbers of these variables should be sought in the resulting matrix header. For example, on the basis of the results in Figure 31, variables ranging from 34 to 39 form a block of collinear variables, therefore one (any) of them should be deleted[25]. If independent variables are not collinear, the software displays a window where basic parameters that control its course can be set.

[24] The authors of the source code of the MATLAB software are: Marcin Błażejowski from Toruń School of Banking and Jacek Kwiatkowski from Nicolaus Copernicus University in Toruń.

[25] The decision on which variable to delete should always be made on substantive grounds. For instance, one can delete a variable with the least intuitive interpretation, as a factor to influence the values of response variable Y.

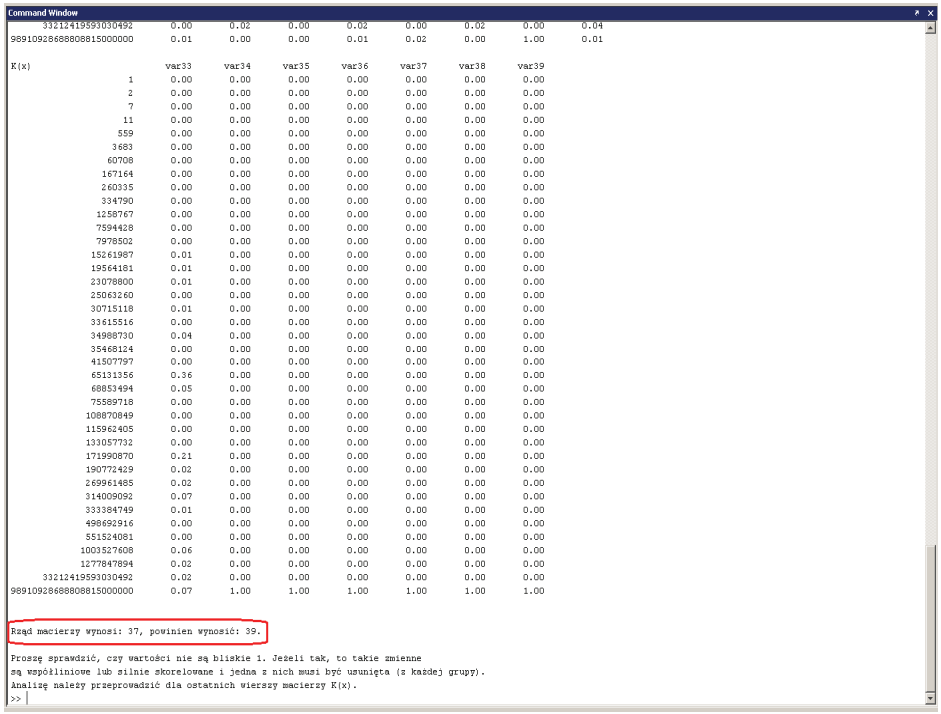

Figure 31. Application of BKW method
Source: author's own analysis

The above methods and theoretical tools serve the purpose of classification of the models and variables that are most likely to influence the pace of economic growth in the regions of the European Union, conducted in chapter 5.