Control

and Cybernetics

VOL. 18 (1989) No. 1

Cluster analysis methods and regression modelling

by

JÓZEF POCIECHA
KAZIMIERZ ZAJĄC

Academy of Economics
Institute of Statistics, Econometric and Computer Science
ul. Rakowicka 27
31-510 Kraków, POLAND

The paper discusses applications of cluster analysis methods in the process of regression model building. There are four fields of possible aplications of cluster analysis: division of a heterogeneous set of observations into homogeneous subsets, selection of a suitable set of explanatory variables, construction of synthetic variables, and similarity analysis of different regressions. The review and discussion of results in this subject, recently published in Polish, has been presented.

1. Introduction

Cluster analysis as a science of quantitative methods of homogeneous group formation, can be easily and fruitfuly applied to the regression modelling process. At the first stage of regression model building, especially in socio-economic research, we have to define the aim and the scope of empirical investigations. The exact definition of the aim implies a concrete class of the model. The definition of the scope has a triple character. First, it is a definition of essential or substantial, second — spatial and third — temporal scopes of empirical investigations. The definition of the scope of investigation involves many problems to be solved by statisticians.

The essential scope of the definiton of regression modelling depends on the specification of endogenous and exogenous variables. The specification of the variables can be based on an adequate theory. On the other hand, there are some formal requirements with regard to variables. The variable selection problem is widely discussed in statistics and econometrics, see [4] or [9]. Among many formal methods which have been proposed there exists an extensive group of cluster analysis methods. The problem of variable selection is one of the fields where the cluster analysis methods and regression modelling meet.

Another problem is the construction of a model for some aggregates of variables. One possible approach to the problem is to build some models for synthetic variables, see [17]. The creation of a synthetic variable is based on an application of cluster analysis methods, especially linear ordering methods. It is the second common level of cluster analysis and regression theory.

One of the basic conditions for the correctness and the effectiveness of regression modelling is the construction of a model for a homogeneous statistical population. Only in this type of population, relations among variables under consideration will have a statistically unambiguous and stable character. Effective methods for obtaining such homogeneous populations are cluster analysis methods. Problems of how to single out homogeneous statistical subsets is the third field where cluster analysis and regression modelling meet.

The fourth level is the application of cluster analysis methods for verification and interpretation of the results of regression modelling. Indices of taxonomic similarity could be applied to the measurement of similarity among structural parameters or stochastical structure parameters of the estimated models.

An alternative name of cluster analysis, more frequently used in socio-economic researches in Poland, is numerical taxonomy. We will use ,,cluster analysis" and ,,numerical taxonomy" or simply ,,taxonomy" as synonyms in this paper. The regression model will be interpreted as an econometric model in this paper. Econometric models built with the aid of taxonomic methods are also called taxonometric models.

In the following parts of the paper taxonomic methods of homogenization of statistical population, taxonomic procedures of explanatory variable selection, methods of synthetic variable construction and taxonomic methods of verification and interpretation of the results of econometric modelling are discussed. The discussion will base both on theoretical results and empirical investigations in socio-economic sciences recently published in Poland.

2. Taxonomic methods of homogenization of a statistical population

The homogeneity of a statistical population is the basic condition if satisfactory results of regression modelling are to be obtained. Such homogeneity can be understood as either spatial or time homogeneity. It leads to the application of cluster analysis methods for obtaining such homogeneous subsets of statistical units.

The classical example of econometric models built for homogeneous spatial unit subsets i.e. regions was presented by Podolec and Zając, [15]. Grabiński, [3], shows how to utilize cluster analysis methods in order to obtain homogeneous sets of time intervals. The more general review is done by Pociecha, Podolec, Sokołowski and Zając, [14].

The regression function is equivalent to the approximation function and it is possible to utilize some methods from the approximation theory, see e.g. [6]. The basic problem of approximation is to define the function which describes the relation among the variables considered, based on information from the sample:

$$(y_1\ x_{11}\ x_{21}\ \ldots\ x_{m1}),\ (y_2\ x_{12}\ x_{22}\ \ldots\ x_{m2}),\ \ldots$$
$$\ldots,\ (y_n\ x_{1n}\ x_{2n}\ \ldots\ x_{mn}) \tag{1}$$

where:

$j = 1, 2, \ldots, m$ — numbers of variables,
$i = 1, 2, \ldots, n$ — numbers of observations.

In stochastic approximation procedure it is assumed that all functions $h\ (X_1\ X_2\ \ldots\ X_m)$ are undistinguishable if:

$$|\Psi(X_1\ X_2\ X_m) - h\ (X_1\ X_2\ \ldots\ X_m)| < \Delta \tag{2}$$

where:

$$X_2 \in [a_1, b_1],\ X_2 \in [a_2, b_2],\ \ldots,\ X_m \in [a_m, b_m];$$
$$\Psi\ (X_1\ X_2\ \ldots\ X_m) \in R_k;$$
$$h\ (X_1\ X_2\ \ldots\ X_m) \in H_k;\ k < n;$$

Δ — tolerance (admissible error of approximation),
R_k — a set of approximated functions (first type of regression),
H_k — a set of approximating functions (second type of regression).

The stochastic approximation problem is to find a family of R_k—functions which satisfy the relation:

$$P\ [\Psi\ (X_1\ X_2\ \ldots\ X_m) - \Delta < Y < \Psi\ (X_1\ X_2\ \ldots\ X_m + \Delta] \geqslant 1 - \alpha, \tag{3}$$

where:

$1 - \alpha$ — the likelihood of approximation.

Relation (3) can be more generally described as:

$$P\ [(Y\ X_1\ X_2\ \ldots\ X_m) \in Q] \geqslant 1 - \alpha, \tag{4}$$

where: Q — space limited by $v_1\ (X_1\ X_2\ \ldots\ X_m)$ and $v_2\ (X_1\ X_2\ \ldots\ X_m)$, and also by $(a_1\ a_2\ \ldots\ a_m)$ and $(b_1\ b_2\ \ldots\ b_m)$.

In socio-economic research it is difficult to analyse multidimensional distributions and for that reason it is convenient to define the spectrum and trace of distribution. The spectrum of distribution is the set of spaces Q_{α_i} in which for a sequence $\alpha_1, \alpha_2, ..., \alpha_i, ...,$ where $Q < \alpha_i < 1$, the following relations are observed:

$$P\,[Y\,X_1\,X_2\,...\,X_m) \in Q_{\alpha_i}] = 1 - \alpha_i \tag{5}$$

and

$$f(y_i^*\,x_{i1}^*\,x_{i2}^*\,...\,x_{im}^*) > f(y_i\,x_{i1}\,x_{i2}\,x_{im}), \tag{6}$$

where:

$$(y_i^*\,x_{i1}^*\,x_{i2}^*\,...\,x_{im}^*) \in Q_{\alpha_i};$$

$$(y_i\,x_{i1}\,x_{i2}\,...\,x_{im}) \in \overline{Q}_{\alpha_i};$$

$\overline{Q}_{\alpha_i}$ is a complement of Q_{α_i}.

The trace of distribution is defined as:

$$P\,[(Y\,X_1\,X_2\,...\,X_m) \in Q] = 1 - \alpha. \tag{7}$$

where α is near 0.

For the calculation of the distribution trace a distance variable can be utilized, see e.g. [16]. Some information about the trace of distribution can also be obtained by aplication of cluster analysis methods, [11]. Application of cluster analysis methods leads to homogeneous subsets of observations which could be understood as a distribution trace. It ought to be of ellipsoidal shape in multidimensional space if the regression model is a linear one. For separation of an ellipsoidal distribution trace from a set of data, the deviation method proposed by Pluta, [11], can be employed.

In the first stage a set of variables must be divided into two subsets:

— stimulant variables,

— destimulant variables.

Such a variable is recognized as a stimulant variable for which the greater the value the higher the position of the object under consideration. Destimulants are transformed into stimulants by the formula:

$$x_{ij} = -\,y_{ij}, \tag{8}$$

where: y_{ij} — is i-th observation of j-th variable which is recognized as a destimulant.

In the next stage the upper and lower „pole" of the set of observation is defined. Coordinates of the upper and lower pole are maximal and minimal values of variables, respectively.

Then, the initial point of the coordinates is shifted to the lower pole, i.e. x_{ij} is transformed into u_{ij} according to the formula:

$$u_{ij} = x_{ij} - x_{oj}, \tag{9}$$

where: x_{oj} — are coordinates of the lower pole.

A line through the lower and upper pole is drawn and this is named the axis of the set. Deviations of observations from the axis of the set inform us about the shape of the distribution trace. Subsequently, perpendicular projections of observations on the axis of the set are calculated. The coordinates of these projections are:

$$R_i = [x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im}], \tag{10}$$

where:

$$x_{ij} = x'_{oj}\, t_i, \tag{11}$$

x'_{oj} — coordinates of the upper pole,

$$t_i = \frac{\sum_{j=1}^{m} x'_{oj} u_{ij}}{\sum_{j=1}^{m} (x'_{oj})^2} \tag{12}$$

Then the measures M^* and W^* are defined and calculated as follows:

$$M^* = [(R_i - P)(R_i - P)^{\mathrm{T}}]^{1/2} \tag{13}$$

$$W^* = [(P_i - R_i)(P_i - R_i)^{\mathrm{T}}]^{1/2} \tag{14}$$

where: P — coordinates of the lower pole, P_i — coordinates of the real data.

Values of M^* and W^* are presented in the diagram, where M is the abscissa and W is the ordinate. This graph presents the distribution trace. If it is of an ellipsoidal shape a linear regression function can be estimated, if it is not, ellipsoidal subsets must be separated by application of isomorphic subset procedure proposed by Pluta, [11].

3. Taxonomic selection of variables

The most universal measure of taxonomic similarity among variables is the correlation coefficient. Among many taxonomic procedures of selection of variables, based on the correlation matrix the method proposed by Pluta, [10], has interesting properties. Let:

$$R = [r_{jk}]; \quad j, k = 1, ..., m, \tag{15}$$

be a correlation matrix among m preliminarily proposed explanatory variables,

$$R_o = [r_{oj}]; \quad j = 1, ..., m, \tag{16}$$

be a vector of correlation coefficients between an endogenous variable and the proposed explanatory variables.

In the next stage, correlation coefficients are tested using the following statistics:

$$r^* = \left[\frac{t_\alpha^2}{t_\alpha^2 + n - 2}\right]^{1/2}, \tag{17}$$

where:

n — number of observations,

t_α — critical value of the Student's distribution for n-2 degrees of freedom at α level of significance.

Those variables for which $r_{oj} \leqslant r^*$ are eliminated from the vector R_o.

The variable for which r_{oj} has the highest value is chosen as the first explanatory variable. The following chosen variables are those, which are succesively maximally correlated with the endogenous variable and insignificantly correlated with the previously chosen explanatory variables. This method has a simple graph interpretation and it can be presented as a symmetric, full and not-oriented graph.

Another method based on the correlation matrix, which also has a graph interpretation has been proposed by Bartosiewicz, [1]. On the basis of the correlation matrix R a graph G is constructed. The proposed explanatory variables are nodes of the graph and significant correlation coeficients between variables are arrows of the graph.

4. Methods of synthetic variable construction

Problems of aggregation are essential in socio-economic research, especially in construction of forecasting models, input-output analysis, operation research and cluster analysis. An original method of aggregation in economics, i.e.

method of synthetic variable construction has been proposed by Hellwig, [7]. On the basis of Hellwig's suggestion many methods of synthetic variable construction have been established. They are based on the following criteria of classification of synthetic variables, discussed in more detail in [4]:

— the way of allowance of stimulant and destimulant variable,
— the way of defining the point of reference of coordinates,
— the way of normalization of variables,
— analytic form of the aggregative function,
— weighting system of importance of variables.

Among the methods of construction of synthetic variables there are procedures based either only on stimulants or only on destimulants or on both types of variables. The reference point of the coordinates used for development of socio-economic models could be chosen on the basis of expert's opinions, international comparisons and in a statistical way. A hypothetical point with maximal observed values of stimulants and minimal observed values for destimulants can be adopted as a statistical pattern point.

There are the following methods of normalization of variables:

1) rank method, which changes the observed values into their ranks,
2) quotient transformations,
3) standardizations,
4) unitarizations.

The following functions can be chosen as aggregative functions:

A — additive:

$$S_i^{(1)} = \sum_{j=1}^{m} \alpha_j \, x'_{ij}, \tag{18}$$

$$S_i^{(2)} = \frac{1}{m} \sum_{j=1}^{m} \alpha_j \, x'_{ij}, \tag{19}$$

$$S_i^{(3)} = \frac{\sum\limits_{j=1}^{m} \alpha_j}{\sum\limits_{j=1}^{m} \frac{\alpha_j}{x'_{ij}}} \tag{20}$$

where:

α_j — weight of variable j,
x_{ij} — normalized value of variable j.

B — multiplicative:

$$S_i^{(4)} = \prod_{j=1}^{m} (x'_{ij})^{\alpha_j} \tag{21}$$

$$S_i^{(5)} = \left[\prod_{j=1}^{m} (x'_{ij})^{\alpha_j}\right]^{\frac{1}{\sum \alpha_j}} \tag{22}$$

Most frequently every variable is assumed to carry the same weight.

Among many possibilities of synthetic variable construction (24 variants are, for instance, presented in [4]) two of them are most popular: standardized value method and the pattern of development method. A synthetic variable in the standardized value method can be defined by the formula:

$$S_i^{(sv)} = \frac{1}{m}\sum_{j=1}^{m} \frac{x_{ij} - \bar{x}_j}{S_j}. \tag{23}$$

In the pattern method it can be defined by:

$$S_i^{(dp)} = \left[\frac{1}{m}\sum_{j=1}^{m} (x'_{ij} - x'_{oj})^2\right]^{1/2} \tag{24}$$

where:

x'_{ij} — standardized value of x_{ij},
x_{oj} — standardized value of the pattern.

Another approach to aggregation of „simple" variables is formation of a homogeneous aggregated variable by application of the Hellwig's stochastical dependence coefficient. This coefficient was defined by Hellwig, [8], and modernized by Czerwiński, [2]. Its properties and possibility of application in socio-economic research has been discussed by Pociecha, [12]. The stochastical dependence coefficient has been taken by Pociecha, [13], in order to choose the optimal vector of explanatory variables. The coefficient is defined by the formula:

$$\delta = \left[\frac{1 - \sum_{ij} min\,(p_{ij}, p_i, q_j)}{1 - max\,(\sum_i p_i^2, \sum_j q_j^2)}\right]^{1/2} \tag{25}$$

where:

$p_{ij} = P(X = x_i, Y = y_j);$

$p_i = P(X = x_i); \quad i = 1, ..., r;$

$q_j = P(Y = y_j); \quad j = 1, ..., s;$
r — number of rows in the contingency table,
s — number of columns,
$0 \leqslant \delta \leqslant 1$

For a variable, which is supposed to be an aggregate of m elementary variables, the matrix Δ of stochastical dependence can be calculated:

$$\Delta = [\delta_{kl}], \quad k, l = 1, ..., m \tag{27}$$

Any cluster analysis method could be applied for obtaining homogeneous subsets of elements of matrix Δ. Those subsets could be recognized as homogeneous aggregative variables.

5. Similarity analysis of regression models

In the econometric modelling process a great number of models can be constructed. These differ with respect to the analytical form or spatial and time scope investigations. In that situation there arises the problem of comparisons of econometric analysis results.

The useful tool of similarity analysis of econometric modelling results is application of cluster analysis methods, as proposed by Wydymus, [17, 18]. Let X be a three-dimensional matrix:

$$X = [x_{ijt}], \qquad \begin{array}{l} i = 1, ..., n; \\ j = 1, ..., m; \\ t = 1, ..., l; \end{array} \tag{28}$$

where:

n — number of observations,
m — number of variables,
l — number of time units.

For each object and each variable a trend model can be constructed. If it is a linear model:

$$\hat{y}_{ijt} = a_{oij} + a_{1ij}t, \tag{29}$$

we obtain a matrix:

$$A_o = [a_{oij}], \tag{30}$$

which is a matrix of expected initial values of the parameters. Then we can apply any cluster analysis method to obtain homogeneous groups of estimates. In such a way we can analyse a matrix:

$$A_1 = [a_{1ij}]. \tag{31}$$

The same type of analysis could be done for any nonlinear model as power or exponential trend model.

When for each object a descriptive model is constructed and estimated, it is based on cross-section and time series data. Then, matrix of estimates is obtained:

$$B = [b_{ij}]; \quad i = 1, ..., n; \quad j = 1, ..., m. \tag{32}$$

The simplest way of similarity analysis in such a situation is the analysis of signs of the estimates.Then, matrix B is changed into matrix:

$$S = [sign_{ij}], \tag{33}$$

and estimates with the same sign are looked for in it.

Another aproach is to change matrix B into a zero-one matrix:

$$T = [t_{ij}], \tag{34}$$

where:

$t_{ij} = 1$ when the estimate is statistically significant,
$t_{ij} = 0$ when the estimate is insignificant.

Significances could be tested using the Student's t-test. The estimates similar with respect to their significance are looked for.

The same type of methodology can be applied for similarity analysis of stochastical structure parameters. It is easy to extend the proposed approach to similarity analysis of any other type of econometric models.

References

[1] Bartosiewicz S. Ekonometria (Econometrics, in Polish). Warszawa, PWE, 1976.

[2] Czerwiński Z. O mierze zależności stochastycznej (On the Measure of Stochastical Dependence, in Polish). *Przegląd Statystyczny,* 2, 1970.

[3] Grabiński T. Dynamiczne modele taksonomii (Dynamic Models of Taxonomy, in Polish). Ph.D. thesis, Akademia Ekonomiczna w Krakowie, Kraków, 1975.

[4] Grabiński T., Wydymus S., Zeliaś A. Metody doboru zmiennych w modelach ekonometrycznych (Methods of Variable Selection in Econometric Models, in Polish). Warszawa, PWN, 1982.

[5] Grabiński T., Wydymus S., Zeliaś A. Metody prognozowania rozwoju społeczno-gosodarczego (Forecasting Methods of Socio-Economic Development, in Polish). Warszawa, PWE, 1983.

[6] Hellwig Z. Aproksymacja stochastyczna (Stochastic Approximation, in Polish). Warszawa, PWE, 1965.

[7] Hellwig Z. Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr (Procedure

for Evaluating the Skilled Manpower Data and the Typology of Countries by Means of the Taxonomic Method, in Polish). *Przegląd Statystyczny*, 4, 1968.

[8] HELLWIG Z. On the Measurement of Statistical Dependence. *Zastosowanie Matematyki*, vol. X, 1969.

[9] NOWAK E. Problemy doboru zmiennych do modelu ekonometrycznego (Problems of Selecting Variables for an Econometric Model, in Polish). Warszawa, PWN, 1984.

[10] PLUTA W. Metoda doboru zmiennych objaśniających w modelach symptomatycznych (Method for Explanatory Variables Selection in Symptomatic Models, in Polish). *Przegląd Statystyczny*, 2, 1972.

[11] PLUTA W. Wielowymiarowa analiza porównawcza w modelowaniu ekonometrycznym (Multidimensional Comparative Analysis in Econometric Modelling, in Polish). Warszawa, PWN, 1986.

[12] POCIECHA J. O zastosowaniu współczynnika zależności stochastycznej (On the Application of Stochastic Dependence Coefficient, in Polish). Problemy Statystyczne i Demograficzne, Prace Komisji Socjologicznej nr 33, PAN, Oddział w Krakowie, Kraków, 1976.

[13] POCIECHA J. O metodzie wyboru optymalnego wektora zmiennych objaśniających w przypadku stosowania słabych skal pomiaru (Some Problems of Choice of the Optimal Vector of the Explanatory Variables in the Case of Weak Scales of Measurement, in Polish). Studia z zakresu zastosowań metod ilościowych w ekonomii, demografii i socjologii, Prace Komisji Socjologicznej nr 40, PAN, Oddział w Krakowie, 1977.

[14] POCIECHA J., PODOLEC B., SOKOŁOWSKI A., ZAJĄC K. Metody taksonomiczne w badaniach społeczno-ekonomicznych (Taxonomic Methods in Socio-Economic Research, in Polish). Warszawa, PWN, 1988.

[15] PODOLEC B., ZAJĄC K. Ekonometryczne metody ustalania rejonów konsumpcji (Econometric Methods of Establishing the Consumption Regions, in Polish). Warszawa, PWE, 1978.

[16] TRYBUŚ G. Zmienna losowa dystansowa (The Distance Random Variable, in Polish). Prace Naukowe AE we Wrocławiu nr 173, Wrocław, 1981.

[17] WYDYMUS S. Metody wielowymiarowej analizy rozwoju społeczno-gospodarczego (The Method of Multi-Dimensional Analysis of the Socio-Economic Development, in Polish). Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, seria specjalna: monografie nr 62, Kraków, 1984.

[18] WYDYMUS S. Porównywanie wyników taksonometrii wielokryterialnej (Comparing the Results of Multicriterion Taxonometrics, in Polish). Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, nr 243, Kraków, 1987.

Received, May 1988.

Metody analizy skupień i modelowanie regresyjne

W artykule rozważa się zastosowania analizy skupień w procesie budowy modeli regresyjnych. Zwrócono uwagę na cztery dziedziny potencjalnych zastosowań analizy skupień: 1. podział niejednorodnego zbioru obserwacji na podzbiory jednorodne; 2. wybór odpowiedniego zbioru zmiennych wyjaśniających; 3. tworzenie zmiennych syntetycznych; i 4. analiza podobieństwa różnych regresji. Przedstawiono przegląd ostatnio opublikowanych prac z literatury polskiej przedmiotu, wraz z komentarzami.

Методы кластерного анализа и регрессионное моделирование

В статье рассматривается применение кластерного анализа в процессе разработки регрессионных моделей. Обращено внимание на четыре области возможных применений кластерного анализа: 1) разделение неоднородного множества наблюдений на однородные подмножества; 2) выбор соответствующего множества поясняющих переменных; 3) образование синтетических переменных; и 4) анализ подобия разных регрессий. Представлен обзор, вместе с комментариями, опубликованных за последнее время работ в польской предметной литературе.