CLUSTER ANALYSIS METHODS AND REGRESSION MODELLING

Assis.Prof/Dr. Mohamed A. Amin
Dean of Technical Management Institute
Al - Zafranyah - Bagdad - IRAQ

1.Introduction Cluster analysis as a 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 researches, we have to define the aim and the scope of empirical investigations. The exact defintion 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, second - spatial and third-time scope of empirical investigations. The definition of the scope of investigation involves many problems to be solved by statistitians .

The essential scope of the defintion 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.

It is the matter that in statistics and econometries the variable selection problem is videly discussed. Among many normal methods which have been proposed there exists an extensive group of cluster analysis methods. The problem of variable selection is one of the fieles where the cluster analysis methods and regression modelling meet.

Another problem is the construction of a model for some aggregates of variables. One possible appreach to the problem in to build some models for synthetic variables [2]. The creation of synthetic variable bases on application ofcluster analyais methods, especially liner ordering methods. It is the second common level of cluster analysis and regressien theory.

One of the basic conditions of correctness and effectiveness of regression modelling is construction of a model for homogenous statistical population. Only in this type of population, relations among variables under consideration will have statistically synonymous and stable character

Effective methods for obtaining such homogeneous populations are cluster analysis method

Problems of how to single out homegenous statistical unit subsets is the third field where cluster analysis and regressien modelling meet.

The fourth level is application of cluster analysis methods for verification and interpretion 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 frequentiy used in socio-economic researches in poland, is numerical taxonomy. We understand ''cluster analysis'' and ''numerical taxonomy'' or simply ''taxonomy'' as synonyms in this paper.

Econometrics is the most important application field of the regressien theory in socio-economic researches. Then 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 interpretion of the results of econometric modelling are discussed.

2.Taxonomic methods of homogeniztion of statistical population

The homogenity of statistical population is the basic condition if satisfactory results of regression modelling are to be obtaind, such homogenity can be understood as either spatial or time homogenity. It leads to the application of cluster analysis methods for obtaining such homogenous subsets of statistical units .[3]

The regression function is equivalent to the approximation function and it is possible to utilize some methods from the approximatien theory.[4] The basic problem of approximation is to define the function which describes the relation among the variables considered, based on information from the sample:

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

where:

$j=1,2,\dots,m$ - nmber of variables,

$i=1,2,\dots,n$ - number of observation.

In functions $h(X_1\ X_2\dots X_m)$ are undistinguishable if:

$$|\Psi(X_1\ X_2\dots X_m) - h(X_1\ X_2\ \dots\ X_m)|<\Delta, \qquad (2)$$

where:

$X_1\in[a_1,b_1\ ,\ X_2\in[a_2,b_2]\ \dots\ X_m\in[a_m,b_m]$;

$\Psi(X_1\ X_2\ \dots X_m)\in R_k$;

$h(X_1\ X_2\ \dots\ X_m)\ H_k$;$k\ \ n$;

Δ - tolerance(admissible error of approximation),

R_k - a set approximated function(first type of regressien),

H_k - a set approximating functions (second type of regression) .

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

$$P[\Psi(X_1X_2\ \dots\ X_m)-\Delta<Y<\Psi(X_1X_2\dots X_m+\Delta]\geqslant 1-\alpha, \qquad (3)$$

where:

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

Relation(3) can be more generally described as:

$$P[(Y\ X_1\ X_2\ \dots\ X_m)\in Q]\geqslant 1-\alpha, \qquad (4)$$

where :

Q – space limited by $V_1(X_1 X_2 \ldots X_m)$ and $V_2(X_1 X_2 \ldots X_m)$ also by $(a_1 a_2 \ldots a_m)$ and $(b_1 b_2 \ldots b_m)$.

In socio-economic researches 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 in the set spaces $Q\alpha_i$ in which for sequence $\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots$, where $Q < \alpha_i < 1$, The following relations are onserved:

$$P\left[(Y X_1 X_2 \ldots X_m) \in Q\alpha_i\right] = 1 - \alpha_i \qquad (5)$$

and

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

where:

$$(Y_i^{*} X_{i1}^{*} X_{i2}^{*} \ldots X_{im}^{*}) \in Q\alpha_i \ ;$$

$$(Y_i X_{i1} X_{i2} \ldots X_{im}) \in \bar{Q}\alpha_i \ ;$$

$\bar{Q}_i$ is a complement of $Q\alpha_i$.

The trace of distribution is defined as :

$$P\left[(Y X_1 X_2 \ldots X_m) \in Q\right] = 1 - \alpha , \qquad (7)$$

where α is near 0 .

For calculation of rhe distribution trace a variable can be utilized [2]. Some information about the trace of distribution can also be obtained by application of cluster analysis methods [6].

Application of cluster analysis methods leads to homogeneous subsets of observations which could be understood as a distribution trace. It ought to be of ellypsoidal shape in multidimensional space if the regression model is a liner one . For separation of an ellypsoidal distibution trace from a set of data, the deviation method proposed by pluta (1986) can be employed . In the first stage a est 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 better the situation .

Destimulant variable is a reverse to stimulant.

Then destimulants into stimulants are transformed by the formula :

$$X_{ij} = - Y_{ij} , \qquad (8)$$

where :

Y_{ij} – is i–observation of j variable which is recognized as a destimulant .

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

When the upper pole of the coordinates is shifted to the lower pole, it means that X_{ij} is transformed into u_{ij} according to the formula :

$$U_{ij} = x_{ij} - x_{oj} , \qquad (9)$$

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 :

$$R_i = \left[X_{i1}, X_{i2}, \ldots, X_{ij}, \ldots, X_{im}\right], \qquad (10)$$

where:

$$X_{ij} = X'_{oj} t_i , \qquad (11)$$

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

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

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

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

$$W^{*} = \left[(P_i - R_i)(P_i - R_i)^T\right]^{1/2} , \qquad (14)$$

where:

P – coordinates of the lower pole,

P_i – coordinates of the real data .

Values of M^{*} and W^{*} are presented on the graph, where M is the abscisa and W is the ordinate. This graph presents the distribution trace. If it is an ellypsoidal shap a linear regression function can be estimated, if it is not, ellypsoilal subsets must be separated by application of isomorphic subset procedure proposed by pluta (1986).

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(1972) has interesting properties. Let :

$$R = \left[r_{jk}\right] ; \quad j,k = 1,\ldots,m , \qquad (15)$$

be a correlation matrix among m preliminary proposed explanatory variables.

$$R_o = \left[r_{oj}\right] ; \quad j = 1,\ldots,m , \qquad (16)$$

be a vector of correlation coefficients between

an endogenous variable and the propsed explanatory variables.

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

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

where:
n - number of observations,

t_{α} - critical value of the student's distribution for n - 2 degrees of freedom at level of significance .

These variables for which $r_{0j} \leqslant r^{*}$ are eliminated from thevector R_{0} .

The variable for which r_{0j} has the highest value is chosen as the first explanatory variable. The following chosen variables are those, which are those, which are succesively maximally correlated with the endogeneous variable and insignificantly correlated with the previously chosen explanatory variables . This method has a a simple graph intrpretation and it can be-presented as a symetric, full and not-oriented graph.

Another method based on the correlation matrix, which also has a graph interpretation has been proposed by Bartosiewicz (1976). On the base of the correlation matrix R a graph G is constructed the propsed explanatory variables are nodes of the graph and significant correlation coefficients between variables are arrows of the graph. Those variables with nodes of graph with maximal number of connection with other nodes and also variables represented by isolated nodes of the graph belong to the optimal vector of explanatory variables .

4. Methods of synthetic variable construction

Problems of aggregation are very essential in socio-economic research, especially in construction of forecasting models, input-output analysis, operation research and cluster analysis. An original method of aggreation in economics, i.e. method of synthetic variable construction has been proposed by Hellwing(1968). On the base of Hellwing's proposition many methods of synthetic variable construction have been explained. They are based on the following criteria of classification of synthetic variables [7]:
- the way of allowance of stimulant and destimulant variable,
- the way of defining of the point of reference of coordinates,
- the way of normalization of variables,
- analytic from 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 piont of the coordinates 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 of destimulants can be adopted is a statistical pattern point.

There are 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 :

$$s_{i}^{(1)} = \sum_{j=1}^{m} \alpha_{j} x'_{ij} , \qquad (18)$$

$$s_{i}^{(2)} = \frac{1}{m} \sum_{j=1}^{m} \alpha_{j} x'_{ij} , \qquad (19)$$

$$s_{i}^{(3)} = \frac{\sum_{j=1}^{m} \alpha_{j}}{\sum_{j=1}^{m} \frac{\alpha_{j}}{x'_{ij}}} , \qquad (20)$$

where:
α_{j} - weight of variable j ,
x'_{ij} - normalized value of variable j .

B - multiplicative:

$$s_{i}^{(4)} = \prod_{j=1}^{m} (x'_{ij})^{\alpha_{j}} \qquad (21)$$

$$s_{i}^{(5)} = \left[\prod_{j=1}^{m} (x'_{ij})^{\alpha_{j}} \right]^{\frac{1}{\sum_{j} \alpha_{j}}} \qquad (22)$$

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

Among many possibilities of synthetic variable construction two of them are most popular standardized value method and development pattern method . synthetic variable in the standardzed value method can be defined Y the formula :

$$s_{i}^{(sv)} = \frac{1}{m} \sum_{j=1}^{m} \frac{x_{ij} - \bar{x}_{j}}{S_{j}} . \qquad (23)$$

In the pattern method it can be defined by:

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

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 Hellwig's stochastical dependence coefficient [9]. It is done by the formula :

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

where :

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

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

$q_j = p(Y=y_j)$; j=1,...,s ;

r - number of rows in the contingency table ,

s - number of columns ,

$$0 \leqslant S \leqslant 1 \qquad (26)$$

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

$$\Delta = \left[S_{ki} \right], \quad k, I = 1. \ldots, m \qquad (27)$$

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

REFERENCES

1- A. koutsoyiannis . (1983) Theory of econometrics (M. P.) .

2- Bartosiwicz. (1976) Econometrics (P. W. E)

3- G.C. chow (1985) Econometrics (M.H) .

4- G.S. Maddala (1977) Econometrics(E.H) .

5- M.J. Brennan (1973) Preface to econometrics.

6- Norman Draper (1981)Applied Regression analysis (j.W) .

7- R.Sokal& P.A. neath(1963) Numerical texanomy (U S A) .

8- S. Wheelwright(1977) forecosting methods . (J.W.) .

9- Therl (1971) principles of Econometrics (J.W).