A C T A  U N I V E R S I T A T I S  L O D Z I E N S I S
FOLIA OECONOMICA 90, 1989

*Zdzisław Hellwig**, *Krzysztof Jajuga***

# APPLICATION OF CLASSIFICATION METHODS TO THE DETERMINATION OF LINEAR ECONOMETRIC MODEL DOMAIN

## 1. THE FORMULATION OF THE PROBLEM

In many econometric papers, where so much attention is paid to the formal aspects of econometric model building, the problem of econometric model domain determination is discussed very rarely. Looking over econometric papers, one can convince himself that not many authors consider the problem of restrictions to be imposed on the set of explanatory variables occurring in the model.

When a function is being defined, the following notation is used:

$$f: X \rightarrow Y,$$

where both sets, X and Y, should be strictly precised. Otherwise, the definition of the function f does not make sense.

Also the results of econometric model building should contain (except the form of the equation) the determination of model domain, that is, of such values of explanatory variables, which are allowed to be substituted in the equation of the model (for which econometric model is valid). Otherwise, the econometric model has no value for its user. Unfortunately, econometricians do not usually specify model domain and the substitution of absurd values may give absurd results.

* Professor at the Academy of Economics, Wrocław.
** Lecturer at the Academy of Economics, Wrocław.

The determination of econometric model domain is very important, particularly seeing that in mathematical programming much attention is paid to the problem of defining the restrictions with respect to which the extremum of decision funtion is sought.

This paper contains some simple proposals in this topic.

## 2. SOME CRITICAL REMARKS ON CLASSICAL ECONOMETRIC MODEL

It is known that from the formal point of view, econometric model may be considered as a regression equation. In linear econometric model theory it is assumed that regression equation contain as arguments three types of variables: jointly dependent variables, pre-determined variables and random components.

This is an example of a linear econometric model:

$$\begin{aligned} C_t &= \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 \tilde{W}_t + \varepsilon_{1t} \\ I_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + \varepsilon_{2t} \\ W_t &= \gamma_0 + \gamma_1 E_t + \gamma_2 E_{t-1} + \gamma_3 A_t + \varepsilon_{3t} \end{aligned} \tag{1}$$

where:

C - consumption,

I - investments,

W - wages in private sector,

P - profits,

K - capital resources at the end of the year,

E - private sector production.

$\tilde{W}$ - wages,

A - time variable.

In mathematical statistics the regression equation is of the form:

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_m X_m + \alpha_0 + Z_Y \tag{2}$$

and it is assumed that $X_1, X_2, \dots, X_m$ are random variables. On the other hand, in econometric model (1), it is usually assumed that explanatory variables are not random variables.

The concept of regression was introduced to statistics by Galton and developped by Pearson and Fisher. They dealt with the applications of statistics to biology. All their views were accepted almost without changes, also in econometrics. However, some of these concepts are not justifiable in econometrics. It means that in the process of econometric model building some rules should be obeyed. They are stated in the following points:

1) to formulate economic hypothesis;

2) to postulate certain analytical form of the model (e.g. the linear one);

3) to construct the potential list of explanatory variables;

4) to classify the variables;

5) to collect data and to specify their type (discrete, continuous, random, not random, dependent, independent);

6) to remove from the potential list some variables which do not comply with certain criteria;

7) to estimate the parameters of model;

8) to specify the range of the variation for explanatory variables;

9) to determine the domain of the model;

10) to verify certain hypotheses, for example of the lack of multicollinearity;

11) to calculate estimate errors and to check the significance of parameters;

12) to give the interpretation of parameters;

13) to subject the model to simulation and to check if fitted values obtained by means of a sample come from the same population.

Most of these rules (except point 9) are scrupulously obeyed by econometricians.

### 3. THE REVIEW OF SOME METHODS OF ECONOMETRIC MODEL DETERMINATION

We are going to present five different approaches to the problem of econometric model determination. Now we present four of them, the last one will be presented in Chapter 5.

### 3.1. The Product of Intervals, to Which Belong Empirical Data, that is "hypercube of econometric model validity"

Suppose that the following econometric model is being estimated:

$$Y = \sum_{j=1}^{m} \alpha_j X_j + Z \qquad (3)$$

To estimate, the n x m data matrix $\underline{X}$ is used (where n - number of observations, m - number of explanatory variables). Without the loss of generality we assume that the observations are centered.

Let us denote:

$$\xi_{1j} = \min_i x_{ij} \qquad \xi_{2j} = \max_i x_{ij} \qquad j = 1, \ldots, m$$

The domain in the form of hypercube is the product of the intervals:

$$[\xi_{11}, \xi_{21}] \times [\xi_{12}, \xi_{22}] \times \ldots \times [\xi_{1m}, \xi_{2m}]$$

Any point $[x_{01}, x_{02}, \ldots, x_{0m}]$ is admissible if it belongs to this hypercube, that is if:

$$\xi_{1i} \leqslant x_{0i} \leqslant \xi_{2i} \quad \text{for each } i \quad 1 \leqslant i \leqslant m.$$

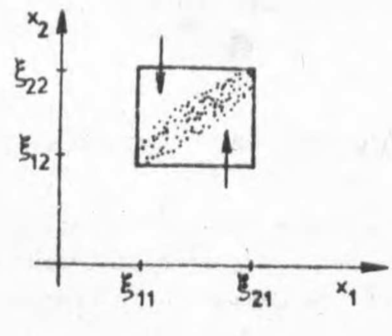


Fig. 1

In the case of high correlation of explanatory variables the model domain defined in such a way has too big "volume", that is it contains such areas, that the probability of belonging to these areas is approximately equal to 0. For $m = 2$, this case is illustrated in Figure 1.

Such areas of model domain are called the probabilistic gap. To reduce its volume, the procedure can be applied, where the econometric model is completed by the additional set of conditions.

### 3.2. Econometric Model With the Conditions on Explanatory Variables

Let us return to (4):

$$Y = \sum_{j=1}^{m} \alpha_j X_j + Z$$

Suppose that the variables $X_1, X_2, \ldots, X_m$ are numbered according to decreasing values of the square of correlation coefficient $\varrho_j^2 = \varrho^2 (X_j, Y)$. Thus the variable $X_1$ is preferential variable, that is the variable which may take any value from the interval $[\xi_{11}, \xi_{21}]$.

First we present the idea of the method for the case $m = 2$. Let us consider the regression of $X_2$ on $X_1$:

$$X_2 = \beta_{21} X_1 + Z_2 \tag{4}$$

Then the standard error of estimate for this regression is equal to:

$$\sigma_{2.1} = \sqrt{V(X_2)\,(1 - \beta_{21} \varrho_{12})} \tag{5}$$

where:

$V(X_2)$ - variance of the variable $X_2$,

$\varrho_{12}$ - correlation coefficient between $X_1$ and $X_2$.

Any point $[x_{01}, x_{02}]^T$ belongs to the domain, if:

1) $\xi_{11} \leqslant x_{01} \leqslant \xi_{21}$

2) $\beta_{21} x_{01} - t\sigma_{2.1} \leqslant x_{02} \leqslant \beta_{21} x_{01} + t\sigma_{2.1}$

where:

$$t = (\xi_{21} - \xi_{11}) / 2\sigma_1 \tag{6}$$

and $\sigma_1$ is the standard deviation of the variable $X_1$.

This procedure may be generalized. For $m = 3$ it can be presented as follows.

The regression of $X_3$ on $X_1$ and $X_2$ is determined:

$$X_3 = \beta_{31} X_1 + \beta_{32} X_2 + Z_3 \tag{7}$$

The standard error of estimate for this regression is equal to:

$$\sigma_{3.12} = \sqrt{V(X_3)(1 - \beta_{31} \varrho_{13} - \beta_{32} \varrho_{23})} \tag{8}$$

Any point $[x_{01}, x_{02}, x_{03}]^T$ belongs to the domain if:

1) $\xi_{11} \leqslant x_{01} \leqslant \xi_{21}$

2) $\beta_{21} x_{01} - t\sigma_{2.1} \leqslant x_{02} \leqslant \beta_{21} x_{01} + t\sigma_{2.1}$

3) $\beta_{31} x_{01} + \beta_{32} x_{02} + t\sigma_{3.12} \leqslant x_{03} \leqslant \beta_{31} x_{01} + \beta_{32} x_{02} +$
$+ t\sigma_{3.12}$

where t is given, as before, by (6).

The generalization of this procedure for any number of explanatory variables is straightforward. For m variables, it is to determine consecutively the regression of $X_2$ on $X_1$, $X_3$ on $X_1$ and $X_2$, and so on, finally the regression of $X_m$ on $X_1$, $X_2$, ..., $X_{m-1}$. For each regression the standard error of estimate is calculated.

For the considered point $[x_{01}, x_{02}, ..., x_{0m}]^T$ checking if it belongs to the domain is performed by turns for each its coordinate. After the verification of the condition $\xi_{11} \leqslant x_{01} \leqslant \xi_{21}$, for the rest of coordinates it is to verify, if they belong to the interval. The centre of this interval is determined from the proper regression equation and the range by means of standard error of estimate. Obviously, the point is admissible, that is it belongs to the domain, if all its coordinates belong to the proper intervals.

As it can be seen, the proposed procedure forces the user of econometric model to take into account the correlation of the variables in the process of substitution of the values of explanatory variables. However, this procedure has two main faults:

1) it may be numerically strenuous, since it requires not only the estimation of the model equations, but also the estimation of the equations specifying the restrictions imposed on the model (additionally m - 1 equations),

2) in contains some arbitrarity, since explanatory variables are numbered according to decreasing values of $\varrho_j^2$, but there are still other possible orders of variables (total number of orders is equal to m!).

How we are going to present the next method.

### 3.3. The Model Domain in the Form of Hypercube Determined by Principal Components

Let

$$\underline{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2m} \\ \dots & \dots & \dots & \dots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_m^2 \end{bmatrix}$$

denote the covariance matrix for explanatory variables.

By means of a well-known procedure the eigenvalues of this matrix are determined: $\lambda_1, \lambda_2, \dots, \lambda_m$.

Furthermore, let

$$\underline{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ u_{21} & u_{22} & \cdots & u_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ u_{m1} & u_{m2} & \cdots & u_{mm} \end{bmatrix}$$

denote the $m \times m$ matrix, where columns are the eigenvectors corresponding to these eigenvalues.

The transformation of data matrix $\underline{X}$ is performed, first by means of translation and then by rotation of axes.
Thus:

$$\underline{W} = (\underline{X} - \underline{M})\,\underline{U}$$

where:

$\underline{W}$ - $n \times m$ matrix of transformed data;

$\underline{M}$ - so called mean matrix, where all rows are equal to the mean vector of $\underline{X}$.

The two-dimensional case ($m = 2$) for presented procedure is illustrated in Figure 2.

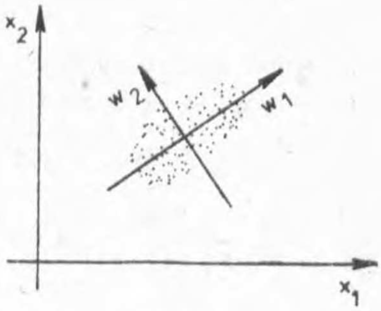


Fig. 2

Any point $[x_{01}, x_{02}]^T$ may be transformed and then it is to check if it belongs to a cube formed by the new pair of variables. Generally, any point $[x_{01}, x_{02}, \ldots, x_{0m}]^T$ belongs to the domain, if:

$$w_{Lj} \leqslant w_{0j} \leqslant w_{Uj} \quad \text{for each } j$$

where:

$w_{0j}$ - the jth coordirate of the point after transformation,

$w_{Uj} = \max_i w_{ij}$,

$w_{Lj} = \min_i w_{ij}$.

Instead of hypercube, one may use a hyperellipsoid.

## 3.4. Hyperellipsoidal Model Domain

To determine a hyperellipsoidal model domain the concept of so called ellipsoid of concentration is apllied.

For considered distribution it is given a m-dimensional mean vector $\underline{\mu}$ and m x m covariance matrix $\underline{\Sigma}$. The m-dimensional hyperellipsoid of concentration is a hyperellipsoid with a centre in $\underline{\mu}$ which has such a property, that the uniform distribution on this hyperellipsoid has the same first and second moments as considered distribution.

It can be proved (see [1]), that the equation of this hyperellipsoid is given by the formula:

$$|\underline{\Sigma}|^{-1} \sum_{i=1}^{m} \sum_{j=1}^{m} |\underline{\Sigma}_{ij}| (x_i - \mu_i)(x_j - \mu_j) - (m + 2) = 0$$

where:

$|\underline{\Sigma}|$ - the determinant of $\underline{\Sigma}$.

$|\underline{\Sigma}_{ij}|$ - the cofactor of (i, j)-th element of $\underline{\Sigma}$,

$\mu_i$ - the ith element of mean vector $\underline{\mu}$.

The considered point belongs to the domain if it belongs to this hyperellipsoid. It is easy to see that the axes of this hyperellipsoid agree with the principal components determined by the third method.

This procedure is justified when the joint distribution of variables is multinormal (or at least is such a distribution, whose equiprobability contours are hyperellipsoids). The experience indicates that such an assumption must not be accepted without the verification. In real applications we deal very often with distributions, which are not normal. Some of these situations will be presented below.

## 4. SPECIAL DISTRIBUTIONS, FOR WHICH THE NECESSITY OF MODEL DOMAIN DETERMINATION OCCURS

To make a graphic presentation possible, we limit ourselves to the two-dimensional case. In all presented situations the set of observations consists of two (of course, could be more) subsets. So the econometric model building should start from the determination of these subsets. Then the regression and its domain should be determined for each subset separately.

In each of Figures 3-8, except observations, the regression lines for each subset and the regression line for the whole set of observations are presented.

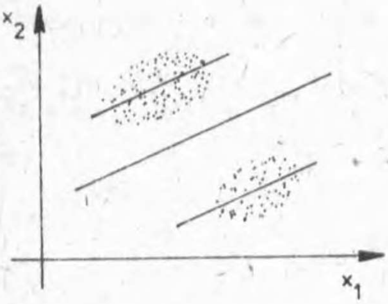


Fig. 3. Situation 1

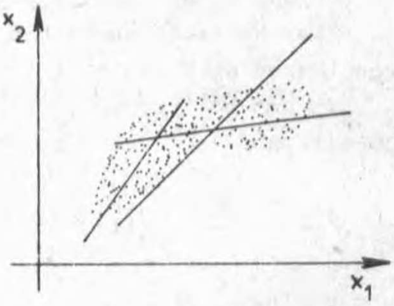


Fig. 4. Situation 2a

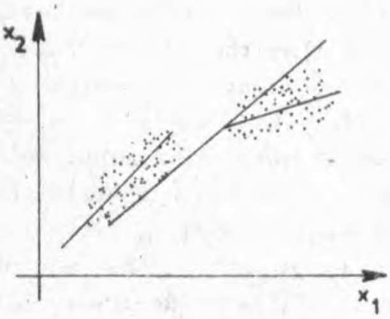


Fig. 5. Situation 2b

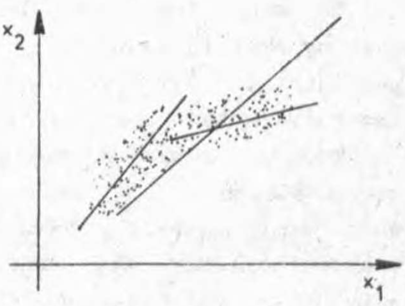


Fig. 6. Situation 2c


Fig. 7. Situation 3

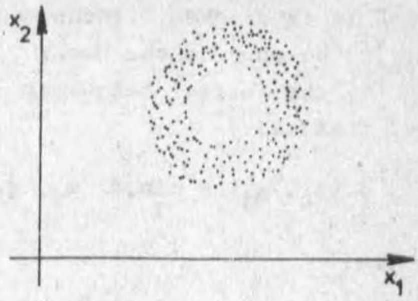


Fig. 8. Situation 4

In the situations 1, 2a, 2b, 2c, and 3 we deal with the observations drawn from the population with a distribution being a mixture of distributions. Note that in these cases the domain consists of two subsets. In the situation 3 only a part of the population is identified, and for the outliers, due to small number of observations, it is not possible to determine the regression, thus the arithmetic mean is the model.

The situation 4 is very difficult to recognize. It may suggest that the observations come from the population multinormally distributed. But it is not the case. However, in economic problems such a situation almost never occurs.

## 5. THE DENDRYT METHOD OF MODEL DOMAIN DETERMINATION

Now we are going to present very simple method. It allows us to cope with many situations occurring in real applications, particularly when the domain consists of several subsets.

The method is based on a well-known classification method, called Wrocław taxonomy method (or single linkage method), although it may be easily adapted to other classification methods.

Suppose given m-dimensional observations:

$$\underline{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^T, \quad i = 1, \ldots, n.$$

As a result of the construction and the partition of the dendryt, K classes of observations: $C_1, C_2, \ldots, C_K$ are obtained. Then it is checked if the considered point $\underline{x}_0 = [x_{01}, x_{02}, \ldots, x_{0m}]^T$ belongs to the domain. To solve this:

1) the nearest-neighbour distance of this point is determined, that is,

$$d(\underline{x}_0, \underline{x}_j) = \min_{l} d(\underline{x}_0, \underline{x}_l)$$

where:

$d(\underline{x}_0, \underline{x}_l)$ - the distance between the m-dimensional points $\underline{x}_0$ and $\underline{x}_l$.

Suppose that the nearest neighbour $\underline{x}_j$ belongs to the class $C_s$.

2) it is checked if:

$$d(\underline{x}_0, \underline{x}_j) \leqslant d_0$$

where:

$d_0$ - certain threshold value.

If this condition is fulfilled, then the considered point $\underline{x}_0$ belongs to the domain.

To determine the threshold value $d_0$, two approaches may be applied:

1. The max-min approach.

Here, the threshold value is the maximum nearest-neighbour distance in the class $C_s$, that is:

$$d_0 = \max_{\substack{i \\ \underline{x}_i \in C_s}} \min_{\substack{j \\ \underline{x}_j \in C_s}} d(\underline{x}_i, \underline{x}_j)$$

2. The frequency approach.

Here, the threshold value is determined by means of the method:

1) for each observation the nearest-neighbour distance is calculated:

$$d_i = d(\underline{x}_i, \underline{x}_{l_i}) = \min_{j} d(\underline{x}_i, \underline{x}_j)$$

2) for the nearest-neighbour distances the histogram of cumulative frequencies is determined,

3) as a threshold value such nearest-neighbour distance $d_i$ is taken, for which

$$d_i = \min \left\{ d_j \mid f(d_j) > 1 - \alpha \right\}$$

where:

$f(d_j)$ - cumulative frequency,

$\alpha$ - constant (e.g. 0.05).

REFERENCES

[1] C r a m e r H. (1946), *Mathematical Methods of Statistics*, Princeton University Press.

[2] G o l d b e r g e r A. (1972), *Teoria ekonometrii*, PWE, Warszawa.

*Zdzisław Hellwig, Krzysztof Jajuga*

ZASTOSOWANIE METOD KLASYFIKACJI
PRZY OKREŚLANIU DZIEDZINY LINIOWEGO MODELU EKONOMETRYCZNEGO

Artykuł zawiera opis pięciu podejść do problemu wyznaczania dziedziny liniowego modelu ekonometrycznego. Są to:

1) wartości zmiennych objaśniających należą do hiperszcześcianów,

2) wartości zmiennych spełniają pewne warunki skorelowania i istotności parametrów,

3) przypadek (1) wg metody głównych składowych,

4) wartości zmiennych są określane dla mieszanek rozkładów,

5) taksonomiczne metody określania dziedziny modelu.

Przez dziedzinę liniowego modelu rozumie się zbiór dopuszczalnych wartości zmiennych objaśniających.