ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 123, 1992

*Edward Nowak**

ON REGRESSION ANALYSIS UNDER HETEROGENOUS OBSERVATIONS

1. INTRODUCTION

Regression modelling of interdependencies between economic phenomena consists in analysing and processing of statistical information related to some distinguished variables. The information can have the form of dynamic or cross-section series. If methods of regression analysis apply to data having the form of cross section series then the objects of investigation, to which observations correspond, should constitute a homogenous set in a settled sense. However, in practice we often deal with sets of heterogenous objects. For example, when analysing a set of forms of a given region we deal with private cooperative and state farms. On the other hand, in a given industrial branch we can distinguish between enterprises of different size and enterprises of different type of production.

If there is a supposition that objects of investigation form a heterogenous set in a settled sense then we ought to divide this set into typological groups comprising similar units and build regression models for each group separately. In different typological groups comprising homogenous units, there can exist different interdependencies amongst distinguished phenomena: endogenous variable and explanatory variables.

Construction of regression model when the set of objects is heterogenous is a multi-stage procedure. Therefore, the fol-

[*] Associate Professor at the Academy of Economics, Wrocław.

lowing stages of the model construction can be distinguished here:

1) specification of endogenous variables and explanatory variable,

2) distinguishing of subsets comprising homogenous units,

3) estimation of the model parameters,

4) verification of the model.

Although, all stages of the model construction (except stage 2) are similar to those of classical regression model yet the model builder faces several specific problems. This comes from the fact that\ on one hand we deal with units of investigation which form a kind of unity, and on the other hand, these units are very often heterogenous and constitute subsets of similar objects. All objects of the set under investigation share certain common properties and apart from this we often observe certain specific properties characteristic of only these units which belong to a given subset.

All the problems mentioned here should be taken into account when a regression model under heterogenous observations is constructed.

2. CLASSIFICATION OF OBJECTS IN REGRESSION INVESTIGATION

Let

$$\Omega = (O_1, O_2, \ldots, O_T) \tag{1}$$

denote a set of objects of investigation and Y denote an endogenous variable and $X_1, X_2, \ldots, X_k$ explanatory variables.

As a result of conducted measurements we have (T x 1) variate vector of observations of endogenous variable having the form

$$\underline{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_T \end{bmatrix},$$

where y_t (t = 1, 2, ..., T) denotes value of Y variable in O_t object and (T x K) variate of the form:

$$\underline{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \cdot & \cdot & \cdot & \cdot \\ x_{T1} & x_{T2} & \cdots & x_{TK} \end{bmatrix} \tag{3}$$

where x_{tk} $(t = 1, 2, ..., T;\ k = 1, 2, ..., K)$ denotes value of X_k variable in O_t object.

Now, the problem consists in dividing set Ω into G subsets $A_1, A_2, ..., A_G$ (also called typological groups or classes) so that the following conditions are satisfied

$$\bigcup_{g=1}^{G} A_g = \Omega. \tag{4}$$

$$A_g \neq \emptyset \quad (g = 1, 2, ..., G) \tag{5}$$

$$A_g \cap A_h = \emptyset \quad (g, h = 1, 2, ..., G,\ g \neq h). \tag{6}$$

These are sufficients conditions for presentation of classification of objects $O_1, O_2, ..., O_T$. Apart from these conditions the considered subsets should have the following properties:

- degree of similarity of objects belonging to different typological groups should be the smallest;
- degree of similarity of objects belonging to the same typological groups should be the greatest.

An important question connected with classification of objects for regression modelling is the evaluation of the objects´ homogeneity. Therefore two types of classification can be distinguished:

- classification which is exogenous in relation to endogenous variable and explanatory variables,
- classification which is endogenous in relation to endogenous variable and explanatory variables.

In the first case it is assumed that division of the set of objects into typological groups was made on the basis of external (in relation to endogenous variable and explanatory variables) information, and in this sense the division is given a priori. Then, properties of the modelled system are the basis for distinguishing groups of similar objects.

Let us illustrate such situation with a few examples.

A classical example is construction of regression model when the objects of investigation are territorial (administrative) units of the country. In this case such a given a priori classification is the division of country into regions. Typological groups are then identified with these regions.

Another example can be found in [11] where relations between demand for services and factors determining the demand (with division into voivodeships) is analysed. In the study quoted above three groups of voivodeships were distinguished:

- a group comprising voivodeships where surplus of supply over demand is observed;
- a group including voivodeships characterized by a relative balance of supply and demand;
- a group comprising voivodeships where surplus of demand over supply is observed.

Another example of this kind would be the analysis of relation between the level of total costs and the size of production of electric power and heat in thermal power stations. Three kinds of thermal power stations can be distinguished here: power plants, heat and power generating plants and heating plants. Due to the fact that these plants differ in the character of their production it is advisable to conduct regression investigation for each kind of thermal power station separately. If the second approach is followed then classification of units of investigation is made on the basis of statistical information concerning endogenous variable and explanatory variables.

In classical taxonomic analysis measures of objects' similarity e.g. Euclidean distances or urban distances etc. are determined as various differences of standardized values of variables accepted for a description of the classified units. These measures define the degree of similarity of each pair of objects. If two points in a multivariate space of observations (representing two objects) are close to each other they are more similar than points which are distant from each other. This kind of classification approach is not applicable to regression modelling.

Such a case demands the use of classification procedures which allow, on the basis of the distinguished subsets of observations, to build regression models best adjusted to the empirical data. The kind of procedures should be based on the idea of similarity in the sense of relations occuring between endogenous variable Y and explanation variables $X_1, X_2, \ldots, X_k$.

Typological groups ought to be distinguished in such a way that the relations between endogenous variable and explanatory variables are different in different subsets, and relations between these variables are similar within the same group.

Proposals of this kind of procedures are presented amongst others, in the following studies: Bekker et al. (1975), Jajuga (1985), Kowerski (1986), Pluta (1986).

Now a classification procedure worked out by Kowerski will follow. In original version it was presented for regression with one explanatory variable. Here, we present a generalized version of the procedure for the case of many explanatory variables. All K-element combinations of points are considered and hiperplanes containing these points are determined. Next, distances of the remaining T-K points from the determined hiperplanes are calculated. The distance of point from the hiperplane is measured by probability in the sense of relation. A matrix of distances (in the sense of relation) containing $\binom{T}{K}$ rows and T-K columns is obtained in this way. Particular rows correspond to all the possible hiperplanes determined by points contained in K-element subset of the T-element set.

The next step is to determine the so called boundary distance which is used as a criterion of evaluation whether the distance of the point from the straight line is significant or not. We put 0 in the appropriate place if on the level of the boundary distance an insignificance of distance is observed. Otherwise we put 1. In this way matrix of zeroes and ones is obtained. Of course, all rows in which only ones have been observed can be eliminated at once because for K arbitrary points it is always possible to draw a hiperplane K-1 of independent variables.

The final division of set of objects into typological groups is obtained as a result of the application of an algorithm of vector's elimination (cf. Chomątowski, Sokołowski, 1978).

3. ESTIMATION OF REGRESSION MODEL PARAMETERS UNDER HETEROGENEITY OF OBJECTS

When objects of investigation constitute a heterogenous set then separate regression models for the distinguished typological groups are constructed. Each model's parameters are estimated separately on the basis of statistical data concerning endogenous variable and explanatory variables in a given group. Regression models estimated for particular typological groups explain variability of endogenous variable resulting from characteristics of these groups. However, regression model estimated on the basis of statistical data on all objects of the set under investigation explains variability of endogenous variable resulting from the properties of the whole set of objects.

As it was mentioned above, in case of heterogeneity of objects a regression model should be built in such a way that it takes into account both properties of the whole set and the specific properties of the particular typological groups. This applies, first of all, to estimation of model's parameters. Examples of solution satisfying this postulate can be found in the literature of the subject.

The idea of one class of such solutions presented in Bartosiewicz et al. (1982), Bekker et al. (1975) and Pluta (1986) consists in estimating regression model parameters for a given typological group on the basis of statistical information. Additionally, we make use of set data on other groups accepted the with appropriate weights which depend on the level of similarity between these groups and the given group. This is realized directly in Bekker et al. (1975) and Pluta (1986). Observations coming from different groups are weighted and indicators of similarity between the distinguished group and other groups take the role of weights. Bartosiewicz et al. (1982) proposed a different procedure. It consists in multiplication of observations in the remaining groups proportionally to the level of similarity to the distinguished group for which the regression model is built.

In the paper entitled "Regression investigations of productivity under heterogeneity of objects" we proposed a two-stage procedure of such estimation of models´ parameters which takes into account both properties of the whole set of objects and the specific properties of particular groups. At the first stage we estimate parameters of regression model defining dependency between endogenous variable and explanatory variables for the´whole set of objects of investigation. Next, on the basis of the model thus estimated, theoretical values of endogenous variable are calculated. At the second stage, the model parameters are estimated (for each typological group separately) on the basis of set of data twice as big as the original group size. These parameters are as follows: original values of endogenous variable and explanatory variable in the group, and additionally theoretical values of endogenous variable calculated for global regression model and once more original values of explanatory variables.

The characteristic feature of this kind of procedure is a "cautious" estimation of parameters in typological groups. Moreover, it enables estimation of regression model parameters for small-size typological groups due to the fact that the set of observations used for estimation of group models´ parameters is increased.

4. SOME PROBLEMS OF REGRESSION MODEL VERIFICATION UNDER HETEROGENEITY FOR OBSERVATIONS

A model built on the basis of heterogenous data, just as any estimated classical regression model, undergoes verification which aims at evaluation of the model´s practical application. In this respect two problems need to be considered:

- evaluation of similarity of results of model´s parameters estimation for (different) typological groups,
- evaluation of the model´s adjustment to empirical data.

Assume that regression models, estimated for typological groups, contain identical explanatory variables.

Let $\underline{\alpha}_g$ (g = 1, 2, ..., G) denote vector of structural parameters of the A_g group model; $\underline{\alpha}_g$ (g = 1, 2, ..., G) - vector of estimates of structural parameters of the same group.

The problem of similarity between the estimation of results for structural parameters of A_g and A_h models may be treated as verification of the following hypotheses:

$$H_o: \underline{\alpha}_{-g} = \underline{\alpha}_{-h} \qquad (g, h = 1, 2, \ldots, G \qquad g \neq h) \tag{7}$$

against alternative hypotheses on

$$H_1: \underline{\alpha}_{-g} \neq \underline{\alpha}_{-h} \qquad (g, h = 1, 2, \ldots, G \qquad g \neq h). \tag{8}$$

If there are no reasons to reject H_o hypothesis then structural parameters of A_g and A_h-group models can be accepted as similar. If, on the other hand, H_o hypothesis must be rejected then structural parameters of A_g and A_h group are recognized as dissimilar.

Let us consider the problem of adjustment of the whole set of group models to empirical data now. Adjustment of models which belong to particular typological groups, can be evaluated by means of such classical goodness-of-fit measures as e.g.: residual variances $S_g^2(\underline{e})$ $(g = 1, 2, \ldots, G)$ and determination coefficients R_g^2 $(g = 1, 2, \ldots, G)$. Similar goodness-of-fit measures based on the already mentioned measures may be created for the whole set of group models (cf. Rozin, 1979). Let T_g $(g = 1,2,\ldots,G)$ denote the size of particular typological groups and $\sum_{g=1}^{G} T_g = T$. Residual variance for the whole set of regression group 1 models can be defined as a weighted mean of residual variances of models for typological groups:

$$S^2(\underline{e}) = \sum_{g=1}^{G} S_g^2(\underline{e}) \frac{T_g}{T} \tag{9}$$

Determination coefficient R^2 for the whole set of regression, group models can be determined in a similar way as a weighted mean of determination coefficients for typological groups:

$$R^2 = \sum_{g=1}^{G} R_g^2 \frac{T_g}{T} \tag{10}$$

Other properties of the model i.e.:

- quality of estimation of regression models' structural parameters in typological groups;

- properties of random deviations in the whole set of group regression models can undergo verification too; for this kind of investigation, all the methods, which apply to classical regression models, are used directly.

REFERENCES

Bartosiewicz S., (1977), *O pewnej modyfikacji metod wyboru predyktant*, "Przegląd Statystyczny", nr 1.

Bartosiewicz S., Dziechciarz J., Nowak E., Pluta W. (1982), *Problem separowalności zbiorów obiektów i zbiorów cech. Autonomiczne funkcje quasi-regresji*, [in:] *Zastosowanie technik wielowymiarowej analizy porównawczej w dynamicznych i przekrojowych badaniach ekonomicznych*, Research work, R. III. 9, Akademia Ekonomiczna, Wrocław (typescript).

Bekker A. W., Jagolnicer M. A., Kolokolov A. A. Gladkikh B. A. (1975), *Raspoznavanie obrazov pri postroeni ekonomiko-statisticheskikh modelej*, Nauka, Nowosybirsk.

Chomątowski S., Sokołowski A. (1978), *Taksonomia struktur*, "Przegląd Statystyczny", nr 2.

Guzik B. (1978), *Dobór zmiennych do modelu segmentowego*, "Przegląd Statystyczny", nr 3.

Hellwig Z. (1965), *Aproksymacja stochastyczna*, PWE, Warszawa 1965.

Hellwig Z. (1983), *Wyznaczanie parametrów regresji w warunkach skąpej informacji*, "Zeszyty Naukowe Politechniki Szczecińskiej", nr 236.

Jajuga K. (1985), *Regresja rozmyta. Analiza zależności między zmiennymi w warunkach niejednorodności zbioru obiektów*, "Przegląd Statystyczny", nr 4.

Kowerski M. (1984), *Kilka uwag na temat analizy zjawisk ekonomicznych na podstawie modeli ekonometrycznych opartych na danych przekrojowych*, "Wiadomości Statystyczne", nr 4.

Kudrycka I. (1984), *Problemy i metody modelowania ekonometrycznego*, PWN, Warszawa.

Metody badania usług rynkowych, (1982), ed. K. Zając, PWE, Warszawa.

Nowak E. (1984), *Problemy doboru zmiennych do modelu ekonometrycznego*, PWN, Warszawa.

Nowak E. (1984), *Regresyjne badania efektywności produkcji w warunkach niejednorodności zbioru obiektów*, "Ruch Prawniczy, Ekonomiczny i Socjologiczny", nr 3.

Nowak E. (1986), *Wyznacza.ie parametrów modelu ekonometrycznego z koincydencją*, "Przegląd Statystyczny", nr 3.

Pluta W. (1986), *Wielowymiarowa analiza porównawcza w modelowaniu ekonometrycznym*, PWN, Warszawa.

Rozin B. B. (1979), *Teoria rozpoznawania obrazów w badaniach ekonomicznych*, PWN, Warszawa.

Edward Nowak

O ANALIZIE REGRESJI W WARUNKACH NIEJEDNORODNOŚCI OBSERWACJI

Prezentowany artykuł poświęcono analizie regresji w warunkach, gdy zbiór obserwacji jest niejednorodny w ustalonym sensie. Wtedy należy przeprowadzić podział tego zbioru na jednorodne podzbiory i budować modele regresyjne dla wyodrębnionych podzbiorów.

Można wskazać na dwa podejścia do zagadnienia klasyfikacji obserwacji w badaniach regresyjnych. Pierwszym rodzajem jest klasyfikacja egzogeniczna w stosunku do zmiennej objaśnianej i zmiennych objaśniających. Wtedy podstawą do wyodrębnienia grup jednorodnych obserwacji są merytoryczne właściwości modelowanego systemu. Drugim rodzajem jest klasyfikacja endogeniczna w stosunku do zmiennej objaśnianej i zmiennych objaśniających. Wtedy podstawą podziału zbioru obserwacji są informacje statystyczne dotyczące analizowanych zmiennych. Procedury klasyfikacji powinny opierać się na idei podobieństwa w sensie relacji zachodzących między zmienną objaśnianą a zmiennymi objaśniającymi.

Podczas weryfikacji modelu powinien być uwzględniony fakt, że równania regresji wyznaczane dla podzbiorów obserwacji składają się na model zjawiska. Ważnym zagadnieniem jest tutaj ocena podobieństwa wyników oszacowania parametrów modeli dla grup typologicznych. Badaniom powinny podlegać także:

- dopasowanie modelu do danych empirycznych,
- własności odchyleń losowych modelu.

Badania te powinny być przeprowadzane zarówno odrębnie dla wydzielonych grup typologicznych, jak i dla całego zbioru obserwacji.