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

*Krzysztof Jajuga**

ON REGRESSION ANALYSIS IN THE CASE OF HETEROGENEITY OF A SET OF OBJECTS

1. INTRODUCTION

Regression analysis is undoubtedly the most often used multivariate statistical method. Here, the dependence of a variable Y on a set of variables $X_1, X_2, \ldots, X_m$ is studied. The use of regression analysis is based on an assumption of homogeneity of a set of objects, for which the regression function is determined.

If stochastic approach is accepted, the homogeneity of a set of objects (observations) means that this set constitutes a random sample from a population where a random column vector $[Y, X_1, X_2, \ldots, X_m]'$ has a multivariate distribution, for example multinormal distribution.

However, in real applications (particularly, when the observations are given as cross-sectional data, but also for time series), this assumption is often not valid, that is, the heterogeneity of observations occurs. For example, when the relationship between production and employment is studied in certain branch of industry, the objects are enterprises of different size. Thus, the set of these objects may be highly heterogenous. So the form of the relationship may differ significantly for groups of enterprises. Similarly, heterogeneity may occur, when, for example, the objects are the countries of the world.

In all such cases, the assumption of homogeneity may be un-

* Lecturer at the Academy of Economics, Wrocław.

justifiable, the objects may come from different populations. Thus, it is reasonable to study the relationship for each class of objects separately.

In statistical and econometric papers some attention is paid to the problem of heterogenous set of objects. The general approach to determine linear regressions for different classes of observations separately is known as switching regressions. Here, it is assumed that the regression coefficients are constant over classes of observations but are different across different classes.

This type of models is used in statistical papers usually for time series. For example, it contains seasonality models [8], piecewise regression models with known join point [9]. In all these models it is assumed that the classification of objects (in this case objects are time units) in known. However, is real applications it is not a case. Usually, the classification of objects is not known a priori. When the time series are used, some methods to classify time units and then to determine switching regressions are proposed. They are given in [2], [4], [5], [6]. These methods consist in the determination of switching point (or points) for time series on the basis of time or dummy variable.

In this paper two methods of determining homogenous classes of observations of hyperellipsoidal shape are presented. These methods may be regarded as particular cases of switching regression models. However, they are applicable to both practical situations, when cross-sectional data or time series are given. In addition, they are to be used when the classification of observations is not known. Thus, in this sense they are more general than other methods proposed in statistical papers. One of the proposed methods is pure stochastic and is based on the mixture of distributions, the other one is distribution-free method.

2. DESCRIPTION OF THE HETEROGENEITY BY MEANS OF MIXTURES OF DISTRIBUTIONS

To determine the regression function in the case of heterogeneity of a set of observations, an assumption will be made.

We assume that a set of objects constitutes a random sample from a population, where a random vector $[Y, X_1, X_2, \ldots, X_m]'$ has a distribution being a mixture of multivariate distributions, for example, multivariate normal distributions.

So the density of a random vector $[Y, X_1, X_2, \ldots, X_m]'$ is given by the formula:

$$f(z) = f(y, x_1, x_2, \ldots, x_m) = \sum_{j=1}^{K} P_j f_j (y, x_1, x_2, \ldots, x_m) =$$

$$= (2\pi)^{-0.5(m+1)} \sum_{j=1}^{K} P_j |\Sigma_j|^{-0.5} \exp\left[-0.5\ (z - \mu_j)' \Sigma_j^{-1}\ (z - \mu_j)\right]$$

where:

K - number of mixture components, that is, number of homogenous classes in a set of objects;

μ_j - expected value of a random vector $[Y, X_1, X_2, \ldots, X_m]'$ for the jth component distribution;

Σ_j - covariance matrix of a random vector $[Y, X_1, X_2, \ldots, X_m]'$ for the jth component distribution;

P_j - jth mixing parameter.

Thus, it is easy to see, that regression, defined as conditional expected value, $E(Y|X_1, X_2, \ldots, X_m)$ is not a linear function. So it is unjustifiable to determine linear regression for whole set of observations. Instead, the homogenous classes of a set of observations, corresponding to particular component distributions, should be separated. Then, for each class, the linear regression may be used.

3. DETERMINING HOMOGENOUS CLASSES OF HYPERELLIPSOIDAL SHAPE BY THE MAXIMUM LIKELIHOOD METHOD

Obviously, to determine such homogenous classes, the methods for estimating the parameters of mixtures of multivariate normal distributions may be used. Unfortunately, serious difficulties occur.

First of all, it is to notice, that the number of parameters to estimate is usually large. It is equal to:

$$K[m + 1 + m + 1 + 0.5\ m\ (m + 1) + 1] - 1 =$$

$$= 0.5\ K(m^2 + 5m + 6) - 1$$

In the simpliest case, when $m = 1$ and $K = 2$, there are 11 parameters to estimate (4 expected values, 4 variances, 2 covariances and 1 mixing parameter). So the moment method of estimation is practically useless to estimate these parameters.

The most often used method of estimation is maximum likelihood method. However for the mixtures of multivariate normal distributions, the likelihood function (and, as a consequence, the necessary conditions to obtain extreme values of estimates) is so complicated (see e.g. [3]), that it is not possible to get the estimates analytically. They can be obtained only by means of numerical method. Here, an iterative algorithm, presented in [1], is used. This algorithm may be described in the following way:

1. Initial values of estimates are chosen (in any way):

$\hat{\mu}_j^0$, $j = 1, \ldots, K$

$\hat{\Sigma}_j^0$, $j = 1, \ldots, K$

$\hat{P}_j^0$, $j = 1, \ldots, K$

They have to satisfy the conditions:

$$\sum_{j=1}^{K} \hat{P}_j^0 = 1.$$

2. In the lth step ($l = 1, 2, \ldots$) of iterative procedure, the values of so called a posteriori probabilities are determined, using Bayes formula (for $i = 1, \ldots, n$; $j = 1, \ldots, K$):

$$\hat{p}^l(j|\mathbf{z}_i) =$$

$$= \frac{\hat{P}_j^{l-1} |\hat{\Sigma}_j^{l-1}|^{-0.5} \exp\left[-0.5(\mathbf{z}_i - \hat{\mu}_j^{l-1})'(\hat{\Sigma}_j^{l-1})^{-1}(\mathbf{z}_i - \hat{\mu}_j^{l-1})\right]}{\sum_{k=1}^{K} \left\{\hat{P}_k^{l-1} |\hat{\Sigma}_k^{l-1}|^{-0.5} \exp\left[-0.5(\mathbf{z}_i - \hat{\mu}_k^{l-1})'(\hat{\Sigma}_k^{l-1})^{-1}(\mathbf{z}_i - \hat{\mu}_k^{l-1})\right]\right.}$$

where:

$z_i = [y_i, x_{i1}, x_{i2}, \ldots, x_{im}]'$ denotes the ith observation of a random vector $[Y, X_1, X_2, \ldots, X_m]'$.

Then the estimates are calculated, using the following formulas (for $j = 1, \ldots, K$):

$$\hat{\mu}_j^l = \frac{\sum_{i=1}^{n} \hat{p}^l(j|z_i) z_i}{\sum_{i=1}^{n} \hat{p}^l(j|z_i)}$$

$$\hat{\Sigma}_j^l = \frac{\sum_{i=1}^{n} \hat{p}^l(j|z_i) (z_i - \hat{\mu}_j^l)(z_i - \hat{\mu}_j^l)'}{\sum_{i=1}^{n} \hat{p}^l(j|z_i)}$$

$$\hat{p}_j^l = n^{-1} \sum_{i=1}^{n} \hat{p}^l(j|z_i)$$

3. The iterative procedure, described above, is being continued untill the values of a posteriori probabilities (and, as a consequence, the estimates obtained in consecutive iterations) do not change significantly; for example, if the condition:

$$\max_{i,j} | \hat{p}^{r+1}(j|z_i) - \hat{p}^r(j|z_i) | < \varepsilon$$

is satisfied, where ε is small positive number.

Then, using obtained estimates, the assignment of all observations to the classes, is performed. To solve this, the following alternative conditions may be proposed:

the ith object (that is, the observation z_i) is assigned to the jth class, if:

a) $\hat{p}^r(j|z_i) = \max_l \hat{p}^r(l|z_i)$, or:

b) $d_{ij}^r = \min_l d_{il}^r$, where $d_{il}^r = (z_i - \hat{\mu}_l^r)' (\hat{\Sigma}_l^r)^{-1} (z_i - \hat{\mu}_l^r)$.

It is easy to show, that the condition b is the particular

case of the condition a, if the equality of covariance matrices and the equality of mixing parameters for each component of the mixture are assumed.

4. DETERMINING HOMOGENOUS CLASSES OF HYPERELLIPSOIDAL SHAPE BY THE DISTRIBUTION-FREE METHOD

To determine such homogenous classes, which correspond to equiprobability contours of multivariate normal distributions, a distribution-free method may be used. This method is based on the minimization of the following function:

$$L = \sum_{i=1}^{n} \sum_{j=1}^{K} f_{ij}^2 d_{ij}$$

where:

f_{ij} - so called degree of belongingness of the ith observation to the jth class,

d_{ij} - distance of the ith observation to the jth class, it is equal to:

$$d_{ij} = (\mathbf{z}_i - \mathbf{v}_j)' \mathbf{M}_j (\mathbf{z}_i - \mathbf{v}_j),$$

where:

$\mathbf{v}_j$ and $\mathbf{M}_j$ are respectively $(m + 1)$ - dimensional vector and positive definite $(m + 1) \times (m + 1)$ matrix, describing size and shape of the jth class.

The function L is linear with respect to $\mathbf{M}_j$ $(j = 1, \ldots, K)$, so the minimum of this function does not exist, since we can always change $\mathbf{M}_j$ in such a way, that the decrease of L is obtained. Thus additional conditions are introduced. There are the following:

$$|\mathbf{M}_j| = r_j, \qquad r_j > 0, \qquad j = 1, \ldots, K$$

$$\sum_{j=1}^{K} f_{ij} = 1, \qquad i = 1, \ldots, n$$

So finally we minimize the function:

$$L_0 = \sum_{i=1}^{n} \sum_{j=1}^{K} f_{ij}^2 (z_i - v_j)' M_j (z_i - v_j) - \sum_{i=1}^{n} \lambda_i \left(\sum_{j=1}^{K} f_{ij} - 1 \right) -$$

$$\sum_{j=1}^{K} \mu_j (|M_j| - r_j)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_m$, $\mu_1, \mu_2, \ldots, \mu_K$ are Lagrange multipliers.

Also in this case extreme values may be obtained by means of numerical methods. It is implemented by the following iterative algorithm (see [3]):

1. Initial values of degrees of belongingness f_{ij}^0 ($i = 1, \ldots, n$; $j = 1, \ldots, K$), as well as positive numbers $r_1, r_2, \ldots, r_K$ are chosen (in any way). They have to satisfy the conditions:

a) $0 < f_{ij}^0 < 1, \qquad i = 1, \ldots, n; \quad j = 1, \ldots, K$

b) $\sum_{j=1}^{K} f_{ij}^0 = 1, \qquad i = 1, \ldots, n$

2. In the lth step ($l = 1, 2, \ldots$) of the iterative procedure, it is to determine:

a) the values describing size and shape of classes, using formulas ($j = 1, \ldots, K$):

$$v_j^l = \frac{\sum_{i=1}^{n} (f_{ij}^{l-1})^2 z_i}{\sum_{i=1}^{n} (f_{ij}^{l-1})^2}$$

$$P_j^l = \frac{\sum_{i=1}^{n} (f_{ij}^{l-1})^2 (z_i - v_j^l)(z_i - v_j^l)'}{\sum_{i=1}^{n} (f_{ij}^{l-1})^2}$$

$$M_j^l = \left[r_j |P_j^l| \right] \frac{1}{m+1} (P_j^l)^{-1}$$

b) the distances of the objects to each class, using formula:

$$d_{ij}^1 = (z_i - v_j^1)' M_j^1 (z_i - v_j^1), \quad i = 1, \ldots, n; \quad j = 1, \ldots, K$$

c) new values of degrees of belongingness, for each object, according to the rule:

- if such k exists, that $d_{ik}^1 = 0$, then:

$$f_{ij}^1 = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}, \quad j = 1, \ldots, K;$$

- if for each k, $d_{ik}^1 \neq 0$, then:

$$f_{ij}^1 = \frac{(d_{ij}^1)^{-1}}{\sum_{r=1}^{k} (d_{ir}^1)^{-1}}, \quad j = 1, \ldots, K.$$

3. The iterative procedure is continued untill the values of degrees of belongingness do not change significantly, for example, if the condition:

$$\max_{i,j} | f_{ij}^{r+1} - f_{ij}^r | < \varepsilon$$

is satisfied, where ε is a small positive number.

Similarly as for the maximum likelihood method, the assignment of objects to classes is made. To solve this, the following condition is used:

the ith object (that is, the observation z_i) is assigned to the jth class, if:

$$f_{ij}^r = \max_l f_{il}^r$$

5. EXAMPLES

As an illustration of described methods, four examples will be presented. In all examples the set of objects is heterogenous and several homogenous subsets of this set can be distinguished.

To make a graphic presentation of examples possible, m = 1 was assumed. In each example the homogenous classes were determined by means of both described methods. The number of classes was known and fixed. For all examples the same classifications were obtained for both methods, that is, the maximum likelihood method and the distribution-free method. Then for each homogenous class the linear regression function was determined. To make comparisons, linear regression function for the whole set of objects was also determined. For each regression function the determination coefficients were calculated. They measure the goodness-of--fit of each regression function. The results are presented below and on the Figures 1, 2, 3 and 4. These Figures 1-4 contain the observations (the two-dimensional points corresponding to the observations are numbered by integer numbers) and the regression lines.

E x a m p l e 1, n = 35, K = 3.

The classification is as follows:

class 1: objects 1-10,
class 2: objects 11-25,
class 3: objects 26-35.

Regression functions and determination coefficients for each class:

$$\hat{Y}^1 = 1.9423\ X + 5.8077, \qquad R_1^2 = 0.8243$$

$$\hat{Y}^2 = 0.9747\ X - 1.1910, \qquad R_2^2 = 0.9622$$

$$\hat{Y}^3 = 0.4905\ X - 3.9192, \qquad R_3^2 = 0.9308$$

and for whole set of objects:

$$\hat{Y} = -0.5196\ X + 11.8093, \qquad R^2 = 0.2774$$

E x a m p l e 2, N = 50, K = 2.

The classification is as follows:

class 1: objects 1-12, 14-23, 39,
class 2: objects 13, 24-38, 40-50.

Regression functions and determination coefficients for each class:

$$\hat{Y}^1 = 1.1242\ X - 1.1553, \qquad R_1^2 = 0.9468$$

$$\hat{Y}^2 = -1.1887\ X + 23.1388, \qquad R_2^2 = 0.9056$$

and for whole set of objects:

$$\hat{Y} = 0.0706\ X + 10.0351, \qquad R^2 = 0.0034$$

E x a m p l e 3, n = 40, K = 2.
The classification is as follows:
class 1: objects 1-18,
class 2: objects 19-40.
Regression functions and determination coefficients for each class:

$$\hat{Y}^1 = 1.0024\ X - 0.6834, \qquad R_1^2 = 0.9468$$

$$\hat{Y}^2 = -0.7330\ X + 17.0680, \qquad R_2^2 = 0.6681$$

and for the whole set of objects:

$$\hat{Y} = 0.0565\ X + 5.7116, \qquad R^2 = 0.0110$$

E x a m p l e 4, n = 24, K = 2.
The classification is as follows:
class 1: objects 1-12,
class 2: objects 13-24.
Regression functions and determination coefficients for each class:

$$\hat{Y}^1 = 2.2161\ X + 0.5793, \qquad R_1^2 = 0.6931$$

$$\hat{Y}^2 = 0.4623\ X + 3.1261, \qquad R_2^2 = 0.7795$$

and for the whole set of objects:

$$\hat{Y} = -0.0736\ X + 11.0238, \qquad R^2 = 0.0053$$

In these examples it is easy to see the usefulness of both methods to determine the homogenous classes of hyperellipsoidal shape. Due to determining linear regressions for each class separately, considerable improvement of goodness-of-fit is achieved.

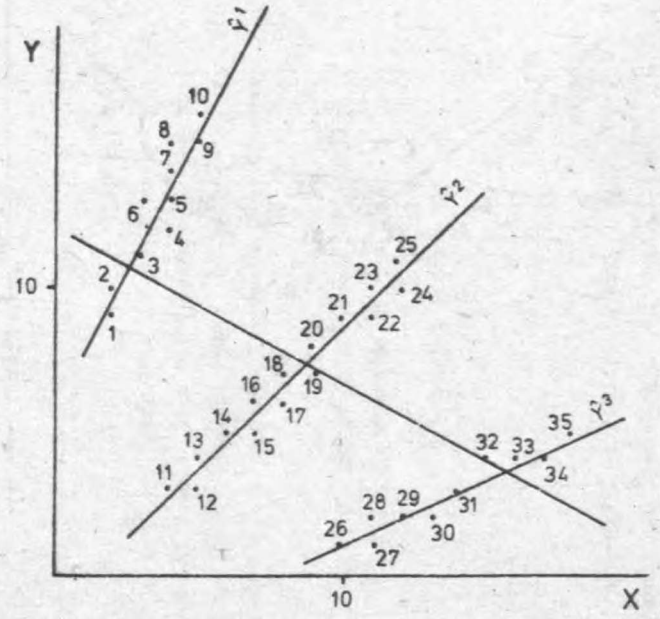

Fig. 1. Observations and regression lines for example 1

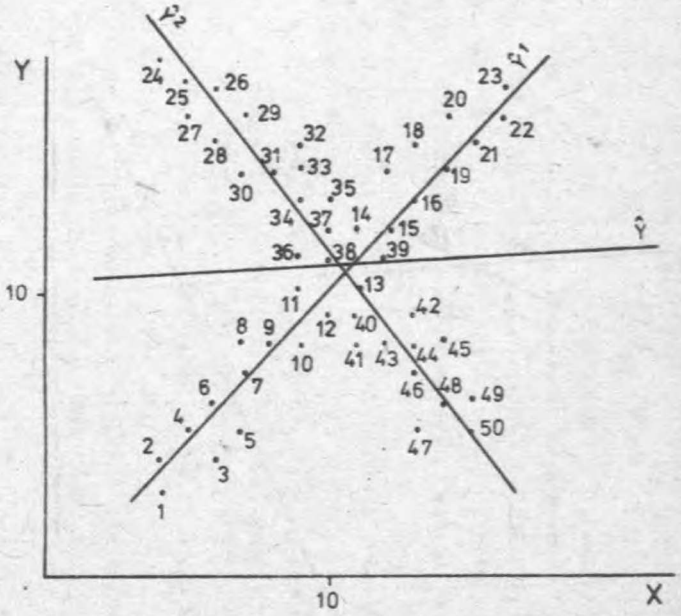

Fig. 2. Observations and regression lines for example 2

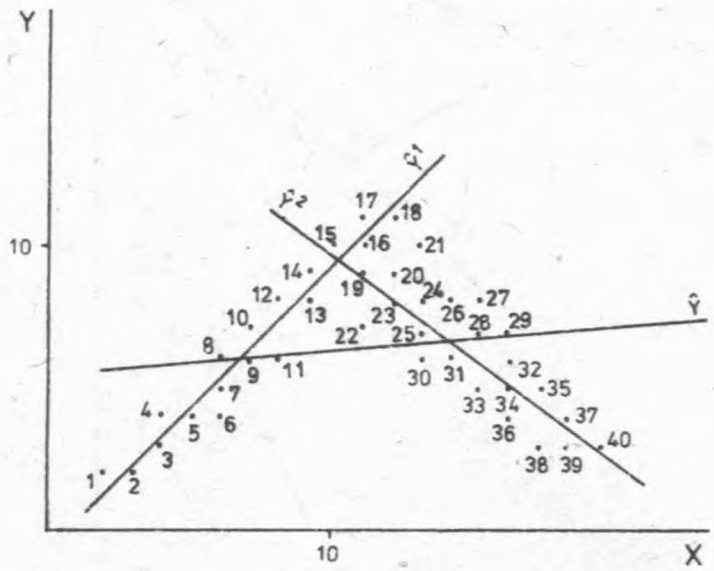

Fig. 3. Observations and regression lines for example 3

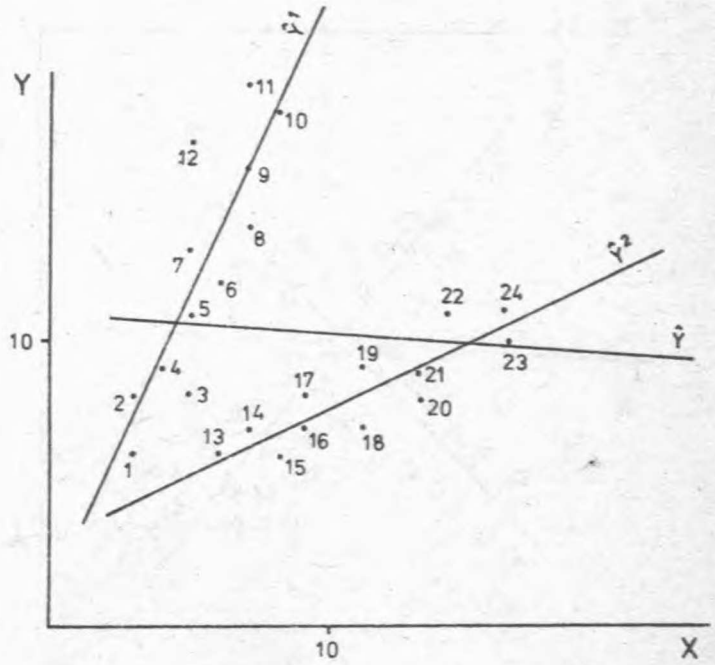

Fig. 4. Observations and regression lines for example 4

REFERENCES

[1] Bezdek J. C., Dunn J. C. (1975), *Optimal Fuzzy Partitions: a Heuristic for Estimating the Parameters in a Mixture of Normal Distributions*, IEEE Transactions on Computers, 24, 835-838.

[2] Brown R., Durbin J., Evans J. (1975), *Techniques for Testing the Constancy of Regression Relationships Over Time*, J. Roy. Statist. Soc., Ser. B, 149-163.

[3] Duda R. O., Hart P. E. (1973), *Pattern Classification and Scene Analysis*, Wiley, New York.

[4] Farley J., Hinich M. (1970), *Testing for a Shifting Slope Coefficient in a Linear Model*, J. Amer. Statist. Assoc., 65, 1320-1329.

[5] Farley J., Hinich M., McGuire T. (1975), *Some Comparisons of Tests for a Shift in the Slopes of a Multivariate Linear Time Series Model*, J. Econom., 3, 297-318.

[6] Goldfeld S., Quandt R. (1973), *The Estimation of Structural Shifts by Switching Regressions*, Ann. Econ. Soc. Measur., 2, 475--485.

[7] Gustafson D. E., Kessel W. C. (1978), *Fuzzy Clustering with a Fuzzy Covariance Matrix*, Scientific Systems Inc., Cambridge.

[8] Kmenta J. (1971), *Elements of Econometrics*, Macmillan, New York.

[9] Poirier D, (1976), *The Economics of Structural Change*, North Holland, Amsterdam.

Krzysztof Jajuga

UWAGI O ANALIZIE REGRESJI
W PRZYPADKU NIEJEDNORODNEGO ZBIORU OBIEKTÓW

W artykule prezentuje się metodologię badań statystycznych w rozumieniu analizy regresji dla przypadku, gdy zbiór obiektów, będących przedmiotem badania, nie jest jednorodny. Proponuje się dwie metody pozwalające określić homogeniczne klasy o hiperelipsoidalnym kształcie. Wszystkie rozważania są ilustrowane czterema przykładami.