Departament d'Economia Aplicada

Endogenous population subgroups: the best population partition and optimal number of groups

DOCUMENT DE TREBALL

Ambra Poggi

05.08

Universitat Autònoma de Barcelona

Facultat de Ciències Econòmiques i Empresarials

Aquest document pertany al Departament d'Economia Aplicada.

Data de publicació : **Maig 2005**

Departament d'Economia Aplicada
Edifici B
Campus de Bellaterra
08193 Bellaterra

Telèfon: (93) 581 1680
Fax:(93) 581 2292
E-mail: d.econ.aplicada@uab.es
http://www.ecap.uab.es

“Endogenous population subgroups:

the best population partition and optimal number of groups”

Ambra Poggi

University of “Piemonte Orientale”

Dept. of Economics (SEMEQ)

Via Perrone 18, 28100 Novara, Italy

ambra.poggi@eco.unipmn.it

Abstract

The aim of this paper is to suggest a method to find endogenously the points that group the individuals of a given distribution in k clusters, where k is endogenously determined. These points are the cut-points. Thus, we need to determine a partition of the N individuals into a number k of groups, in such way that individuals in the same group are as alike as possible, but as distinct as possible from individuals in other groups. This method can be applied to endogenously identify k groups in income distributions: possible applications can be poverty and polarization studies.

Keywords: stopping rule, optimal grouping, Gini index

JEL-Codes: I32, D30, C10

1. Introduction

The aim of this paper is to suggest a method to find endogenously the points that group the individuals of a given distribution in k clusters, where k is endogenously determined. These points are the cut-points. Thus, we need to determine a partition of the N individuals into a number k of groups, in such way that individuals in the same group are as alike as possible, but as distinct as possible from individuals in other groups.

This paper is motivated by the necessity to groups the population in different clusters to measure poverty, deprivation, social exclusion and polarisation. However, notice that the necessity to identify a certain number of groups in a given population exists not only in economics. In areas as medicine, psychology, soil science, ecology and taxonomy, the partition of the population into groups is necessary to make some inferences about property of "natural" groups (Krzanowsky and Lay, 1988).

When we wish to define k groups (and, therefore, to identify k-1 cut-points), we face two problems:

- the identification of the best subdivision of the population into a given number k of groups, and
- the determination of the best value of k (the optimal number of groups)

To solve the first problem, we need to formulate an objective function that quantifies the adequacy of a given partition of the population into k groups, and then to find the partition optimising this objective function. Various objective functions have been suggested in the literature, but we found particularly interesting the one proposed by Aghevliand Mehran (1981), successively applied to polarisation by Gradin (2000). Taking as given the number of groups, they proposed to minimise the differences within groups expressed as difference between the Gini index of the ungrouped population and the between-group Gini index. Since the Gini index of the ungrouped population is fixed, we only need to maximize of the between- group Gini index to get the best partition of the population in a given number of groups. This method is fully explained in section 2.

To solve the second problem, the usual approach adopted is to repeat the optimisation of the objective function for k=2,3,4,… groups, and to choose the value of k at which the final partition appears to be the “best”. This criterion is called stopping rule. Mariott (1971) and Krzanowsky and Lay (1988) proposed stopping rules based on the optimisation of within-group inequality within-group inequality. But, this criteria is not satisfactory since inequality is measured by within-group covariance that is not an additively decomposable measure. The main contribution of this paper to the literature is to propose a stopping rule base on the Gini index, an index normally used to analyse inequality in income distributions. Note that some examples are presented to show how the proposed general method and stopping rule work.

2. K endogenous population subgroups: a review.

In this section, we review a general method proposed by Aghevli and Mehran (1981) to identify the best subdivision of the population into a given number k of groups. The problem is the following: given data on a distribution, we wish to group the data into k groups in such way that differences are minimised within the groups and maximised between the groups. Differences can be measured by an inequality index. Therefore, we need some criteria to choose the adequate index.

We assume the number of groups existing in the population is given and equal to k. We consider a particular distribution *F* of the population over the bounded support *[a,b]*. Each individual *i* is represented by an attribute x_i. We have *n* individuals such that $x_1<x_2<\ldots<x_n$. We assume the existence of k groups. Thus, we wish to find endogenously the cut-points, y_1, $y_{2,\ldots}y_{k-1}$, that groups the population in k clusters such that $a<y_1<y_2<\ldots<y_{k-1}<b$ and n_j are the individuals in the j-th group, $[y_{j-1},y_j)$. The cut-points gives us a partition such that

$$n_1 \cup n_2 \ldots \cup n_k = n \qquad and \qquad n_1 \cap n_2 \ldots \cap n_k = \phi$$

Note that individual i belongs to the j-th group if, and only if, $x_i \in [y_{j-1},y_j)$. Moreover, note that the group construction implies no-overlap among group ranges.

Our aim is identify k groups in the population such that the dispersion internal to every group is minima. Thus, we need to minimise the sum of the internal group dispersions, and the internal group dispersion can be measured by the within-group inequality.

The overall dispersion can be expressed as a weighted sum of the dispersion values calculated from the subgroups plus a term capturing the between-group dispersion. Thus,

$$(1) \qquad I_{tot}=\Sigma_j^k \, q_j I_j + I_b$$

where I_j measures the inequality in group j, I_{tot} measures the overall inequality, I_b measures the between-group inequality, and q_j depends on the population and income share going to subgroup j and on the group position.

For a given distribution the overall inequality is fixed. Therefore, to minimise the sum of the internal group dispersions ($\Sigma_i q_j I_j$) is equivalent to maximise the dispersion between groups (I_b). In other words, minimising the within-group differences implies maximising the between-group differences. The best population subdivision in k groups is, therefore, computed by maximising the between group differences (I_b). Our objective function is the between-group dispersion measured by the between-group inequality. The partition into k groups that maximises the objective function is the optimal partition and it minimises the within-group dispersion.

For the implementation of the procedure to select the best partition, we need to choose an inequality measure. The latter has to be capable to be transformed in an additively decomposable index.

The decomposition of the overall inequality, in the sum of the group inequalities plus the between-group inequality, is possible using indices of the Generalised Entropy families and their monotonic transformations (Shorrocks, 1984). Thus, without imposing specific constrains, the only index, that can be used to measure dispersion, is an entropy index. However, the Gini index is decomposable, in sense of equation (1), when the group ranges do not overlap (Lambert-Aronson, 1993). Since we are interested in determining non-overlapping partitions, we can also use as dispersion measure some kind of indices not belonging to the Generalised Entropy family, but decomposable. In particular, we can use the Gini index without imposing further constrains. However, we cannot use grouping conditions based on the variance as proposed by Mariott (1971) and Krzanowski (1988).

The choice of the index to use is an important point in our analysis and can lead to some numerical differences. To made this choice we referee to the following requirement:

Requirement 1. The inequality index, I, has to be decomposable in sense of equation (1).

The indices satisfying Requirement 1, as seen above, are the ones of the Generalised Entropy family and the Gini index when the group ranges do not overlap. We choose to use the latter since it has already been used to group a population into different clusters by Aghevli and Meran (1981) and Gradin (2000). They minimised the within-group dispersions that are equal to the difference between the Gini index of the ungrouped distribution (G) and the between-group Gini index. It means to minimise the error due to grouping in the estimation of the Gini index from grouped data. Moreover, since G is fixed, to minimise the within-group dispersion implies to maximise the between-group Gini index.

Choosing as measure of inequality the Gini index, our problem reduces to find the k-1 cut-points, $y_1 \ldots y_{k-1}$, that maximises the between-group dispersion (G_b):

$$(2) \qquad \text{Max}\left\{ G_b(k) \right\}=\text{Max}\left\{(1/2n^2\mu)\, \Sigma_i^k \Sigma_j^k n_i n_j \,|\mu_i-\mu_j|\right\}$$

where μ_j is the mean of group $[y_{j-1},y_j)$ and n_j is the corresponding population share. We define $G^*_b(k)$ as the optimum value of $G_b(k)$ for a partition into k groups. In other words, $G^*_b(k)$ is obtained grouping the population in k groups using the optimal cut-points $y^*_1 \ldots y^*_{k-1}$.

Finally, note that the cut-points computed maximizing the between-group Gini index have some useful properties. First, if we multiply all the individual attributes by the same parameter, the cut-points of the new distribution are equal to the old cut-point multiplied by the above parameter. Second, the cut-points depend from the individual attributes, but they do not depend on the name of the individuals. Third, if we merge two or more identical population, we wish that the cut-points do not change.

3. Stopping Rule

In this section, we propose a stopping rule to determine the optimal number of groups, k*. The idea is to repeat the optimisation of the objective function for k=2,3,4,… groups, and to choose the value of k at which the final partition appears to be the "best". To determine the best final partition, we need to define a function, A_k, depending to the objective function (G^*_b) and from the number of groups (k). Such function should remain approximately constant over k for data from a uniform population. But, the optimal subdivision into k groups should provide a large increase in A_k if the data are from a population that is strongly grouped round k clusters. Hence, we suggest using A_k as basis for the stopping rule: the optimum value of k is the value that yields the maximum in A_k.

The main idea is the following. We can face two situations: a one-group distribution and a k-group distribution (k>1). On the border between these two situations, we find the uniform distribution. Thus, we need a function A_k able to tell us in which situation we are. For example, we should like A_k be less than zero if the distribution is one-group distribution, and be positive if we have a k-spike distribution. But, if the distribution is uniform, we should like A_k be equal zero for all k. Therefore, requiring A_k approximately zero for all k when the distribution is uniform, we obtain the following effects. First, if the dispersion in our distribution is smaller than the one in the uniform distribution, A_k will be smaller than zero: we face a one-group distribution. Second, if the dispersion in our distribution is bigger than the one in the uniform distribution, A_k will be bigger than zero: we face a k-group distribution (k>1).

We need to specify the following function:

$$A_k = A(G^*_b, k)$$

Axiom For uniform data, as k varies the function value A_k should remain constant.

Suppose $x_1, x_2, \ldots, x_n$ are uniformly distributed. If k=1, the between-group dispersion is equal zero, since we have only one group in the population: $G^*_1=0$. If k>1, the subdivision of the distribution into k groups is optimum when the groups have equal size, equal population share and $|\mu_j-\mu_{j+1}|= 1/k$ (property 4). Thus, we observe:

$G^*_1=0$

$G^*_2=2\ (1/ k^3)$ with k=2

$G^*_3= 2\ (1+1+2) / k^3$ with k=3

$G^*_4= 2\ (1+1+1+2+2+3) / k^3$ with k=4

…

$G^*_k= [2\ \Sigma_j^{k-1} j(k-j)] / k^3$

Hence, the subdivision of a uniform distribution into k groups, increases G^*_1 by the following term:

$$[2\ \Sigma_{j=1}^{k-1} j(k-j))] / k^3$$

For uniform data, this implies:

$$G^*_k - [(2\ \Sigma_j^{k-1} j(k-j)) / k^3] = G^*_1 = 0 \qquad \text{for all integer } k>1$$

Theorem The function A_k must have the following specification:

$$A_k= c\ [G^*_k - [(2\ \Sigma_j^{k-1} j(k-j)) / k^3]] \qquad \textit{with } c \in R_{++}$$

Then, we can define the following stopping rule.

Stopping Rule The optimal value of k is the value that maximises $A_{k.}$

4 Examples

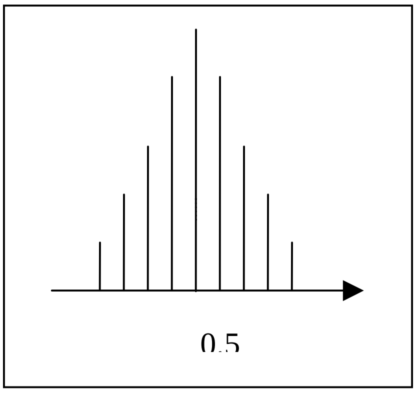
In this section, we show three simple distributions in order to illustrate how the proposed method works. In the first example, we consider one-spike distribution; in the second one, we study a two-spike distribution; and, in the third one, we analyse a three-spike distribution.

Example 1.

Let's consider the distribution

x=(0.1, 0.2, 02, 0.3, 0.3, 0.3, 0.4, 0.4, 0.4, 0.4, 0.5, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 0.9)

It is a one-group distribution as we can observe in the follow figure:



The results for the maximization of the between-group Gini index, for k=2 and k=3, are showed in the table below. We observe that A_k is less than zero for all k>1, which implies that our distribution is a one-group distribution.

.

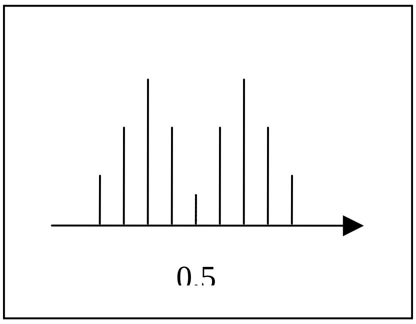
	cut-points		A_k
k=1	---		0
k=2	0.5		-0.09
k=3	0.4; 0.7		-0.0979

Example 2

Let’s consider the distribution

x=(0.1, 0.1, 0.1, 0.2, 02, 0.2, 0.2, 0.2, 0.3, 0.3, 0.3, 0.4, 0.5, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 0.8, 0.8, 0.8, 0.9, 0.9, 0.9)

It is a two-group distribution as we can observe in the follow figure:



The results for the maximization of the between-group Gini index, for k=2 and k=3, are shown in the following table. We observe that the stopping rule selects k=2. In other words, A_k is maximum for the subdivision of the distribution into two groups.

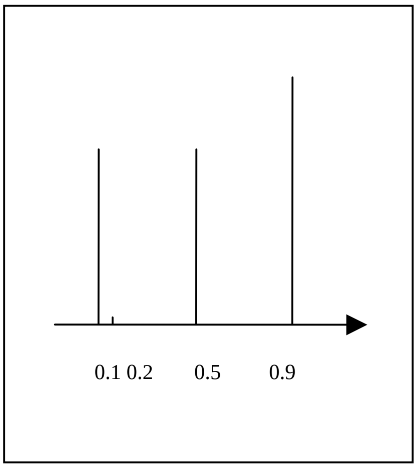
	cut-points	G_b	A_k
k=1	---	---	0
k=2	0.6	0.272	0.022
k=3	0.4; 0.8	0.30048	0.0044837

Example 3

Let’s consider the distribution

X	n times
0.1	13
0.2	1
0.5	13
0.9	20

It is a three-group distribution as we can observe in the follow figure:



The results for the maximization of the between-group Gini index, for k=2,3,4 are shown in the table below. We observe that the stopping rule selects k=3. In other words, A_k is maximum for the subdivision of the distribution into three groups.

	cut-points	G_b	A_k
K=1	---	0	0
K=2	0.9	0.2668	0.0168
k=3	0.5; 0.8	0.3253	0.0290
k=4	0.2;0.5;0.9	0.3264	0.0139

5. Conclusions

We proposed a new method to determine k-1 endogenous points that groups the population in k subgroups, where the number of groups (k) is endogenous.

The first part of the problem is, given data on a distribution, to group the data into k groups in such way that differences are minimised within the groups and maximised between the groups. Using the genral method proposed by Aghevliand Mehran (1981), we maximise the between-group Gini index finding the optimal partition of the distribution in k groups, with k given.

Our contribution is in the method that endogenously determines the optimal number of groups, k*. This method is called stopping rule. The idea is to repeat the optimisation of the objective function for k=2,3,4,… groups, and to choose the value of k at which the final partition appears to be the "best". To determine the best final partition, we defined a function, A_k, that remains approximately constant over k for data from a uniform population. We proposed to use A_k as basis for the stopping rule: the optimum value of k is the value that yields the maximum A_k.

Finally, note that results should never be accepted uncritically but should always be examined to make sure they are meaningful. Graphical analysis is useful to do so. Moreover, it should always be remembered that we cannot use our method to determine two, or more, groups in a unimodal distribution. Further research is necessary to extend this method to group individuals belonging to the same cluster in k sub-groups.

References

Aghevli, B.B., and Mehran, F. (1981) “Optimal grouping of income distribution data”, *Journal of American Statistical Association*, Volume 76, Issue 373

D’Ambrosio, C., Muliere, P., and Secchi, P. (2002) “The endogenous poverty line as a change-point in the income distribution (the principle of transfer revisited)”, 27th Conference IARIW, Sweden

Gradin, C. "Polarization by sub-population in Spain, 1973-91", *Review of Income and Wealth*, Series 46, No. 4, December 2000

Hey, J., and Lambert, P. "Relative deprivation and Gini coefficient: a comment", *Quartery Journal of Economics*, Vol. 95, Issue 3, Nov. 1980, pp. 567-573

Jenkins, S. and Lambert, P. (1993) “Ranking income distributions when needs differ”, *Review of Income and Wealth*, Series 39, No. 4

Krzanowski, W.J., and Lai, Y.T. (1988) “A criterion for determing the number of groups in a data set using sum of squares clustering”, *Biometrics*, 44, 23-34

Lambert, P.J., and Aronson, J.R. (1993) “Inequality decomposition analysis and Gini coeffiecient revisited”, *The Economic Journal*, 103, 1221-27

Mariott, F.H.C. (1971) “Practical problems in a method of cluster analysis”, *Biometrics*, 27, pp. 501-514

Shorrocks, A. (1984) “Inequality decomposition by population subgroups”, *Econometrica*, Vol. 52, No 6

Shorrocks, A. (1995) “Revisiting the Sen poverty index”, *Econometrica*, Vol.63

Últims documents de treball publicats

NUM	TÍTOL	AUTOR	DATA
05.07	ANÁLISIS DE LAS EMISIONES DE CO2 Y SUS FACTORES EXPLICATIVOS EN LAS DIFERENTES ÁREAS DEL MUNDO	Vicent Álcantara Emilio Padiila	Abril 2005
05.06	Descentralización del empleo: ¿compactación policéntrica o dispersión? El caso de la región metropolitana de Barcelona 1986-1996	Miguel Ángel García Ivan Muñiz	Abril 2005
05.05	Descentralización, integración y policentrismo en Barcelona	Ivan Muñiz/ Anna Galindo / Miguel Ángel García	Abril 2005
05.04	Knowledge, networks of cities and growth in regional urban systems	Joan Trullen / Rafael boix	Febrer 2005
05.03	Inequality in CO2 emissions across countries and its relationship with income inequality: a distributive approach	Emilio Padilla / Alfredo Serrano	Gener 2005
05.02	Environmental management problems, future generations and social decisions	Joan Pasqual / Emilio Padilla	Gener 2005
05.01	International inequalities in per capita CO2 emissions: a decomposition methodology by Kaya factors	Juan Antonio Duro / Emilio Padilla	Gener 2005
04.12	Eficiencia y equidad en la ubicación de bienes colectivos locales indivisibles	Joan Pasqual	Novembre 2004
04.11	Regional Income Inequalities in Europe: An Updated Measurement and Some Decomposition Results	Juan Antonio Duro	Octubre 2004
04.10	Caracterización de la privación y de la pobreza en Catalunya	Sara Ayllon / Magda Mercader / Xavier Ramos	Octubre 2004
04.09	Social exclusion mobility in Spain, 1994-2000	Ambra Poggi	Setembre 2004
04.08	Sources of Competitiveness in Tourist Local Systems	Rafael Boix / Francesco Capone	Setembre 2004
04.07	"WHO PARTICIPATES IN R&D SUBSIDY PROGRAMS?. The case of Spanish Manufacturing Firms"	J. Vicente BLANES / Isabel BUSOM	Agost 2004
04.06	Una aproximación sectorial a la localización industrial en Cataluña	Anna Matas Prat José Luis Roig Sabaté	Juny 2004
04.05	Firm Strategies in R&D: Cooperation and Participation in R&D Programs	Isabel Busom, Andrea Fernández-Ribas	Abril 2004