Centro de Estudios Económicos
www.colmex.mx
El Colegio de México, A.C.

Serie documentos de trabajo

A CLUSTERING PROCEDURE FOR THE ESTIMATION OF ECONOMETRIC MODELS WITH SYSTEMATIC PARAMETER VARIATION

Carlos M. Jarque
Secretaría de Programación y Presupuesto

DOCUMENTO DE TRABAJO

Núm. VIII - 1983

A CLUSTERING PROCEDURE FOR THE ESTIMATION OF ECONOMETRIC MODELS WITH SYSTEMATIC PARAMETER VARIATION

By

CARLOS M. JARQUE

Secretaría de Programación y Presupuesto

ABSTRACT

We suggest a two stage procedure for the estimation of models with systematic parameter variation. In the *first stage,* the observations would be classified by the use of a criterion suggested, aiming at the formation of groups of homogeneous parameter values (regimes). The application of clustering techniques in the present setting is indicated. In the *second stage,* the econometric estimation of the regimes would follow.

RESUMEN

En este trabajo sugerimos un procedimiento bietápico para la estimación de modelos econométricos con variación sistemática en los parámetros. En la *primera etapa,* las observaciones se clasifican utilizando un criterio propuesto el cual permite formar grupos con valores homogéneos de los parámetros. En la *segunda etapa* se realizaría la estimación econométrica de los parámetros de grupo.

Address:
Dirección General de Estadística
Instituto Nacional de Estadística, Geografía e Informática
Insurgentes Sur 795 -- PH
México, D.F.

A CLUSTERING PROCEDURE FOR THE ESTIMATION OF ECONOMETRIC MODELS WITH SYSTEMATIC PARAMETER VARIATION

By

CARLOS M. JARQUE

Secretaría de Programación y Presupuesto

Documento de Trabajo

1983

A CLUSTERING PROCEDURE FOR THE ESTIMATION OF ECONOMETRIC MODELS WITH SYSTEMATIC PARAMETER VARIATION

BY

CARLOS M. JARQUE

1. INTRODUCTION

Assume that observations on a *cross-section* of, say N individuals, are available and that a regression model is written as

$$y_i = x_i'\beta + u_i \qquad i = 1,\ldots,N, \qquad (1)$$

where y_i is the i'th observation on the endogenous variable Y; x_i is a K by 1 vector representing the i'th observation on K fixed regressors $X_1,\ldots,X_K$; u_i is the i'th unobservable disturbance and β is a K by 1 vector of unknown parameters.

Under (1), and provided no functional relation exists between the regressors, the k'th element in β, say β_k, would be interpreted as the partial derivative of Y with respect to X_k - irrespective of i. In practice, there may be reasons to believe the increase in Y, due to a unit increase in X_k, is not the same for all the individuals in the cross-section. Furthermore, it may even be thought that each individual reacts in its own particular way to an increase in X_k, i.e., that each has 'its own value' of β_k. To account for this *parameter variation*, β could be replaced by β_i in (1) giving

$$y_i = x_i'\beta_i + u_i \qquad i = 1,\ldots,N. \qquad (2)$$

Without additional assumptions, it is not possible to proceed any further due to the fact that - in (2) - there are NK parameters to be estimated (apart from those related to disturbance terms) and only N observations. Various assumptions can be made to overcome this problem.

For instance, one may assume parameter variation only occurs in the coefficient associated with the intercept term, and therefore introduce variation through the use of *dummy variables*. Other approaches include *Random Coefficient Models* (e.g. see Hildreth and Houck (1968) and Swamy (1971)); *Switching Regressions* (e.g. see Goldfeld and Quandt (1973, 1976)); *Segmented Polynomial Regressions* (e.g. see Hudson (1966) and Gallant and Fuller (1973)); *Piecewise Regressions* (e.g. see McGee and Carleton (1970)) and *Spline Regression Models* (e.g. see Poirier (1976)).

Here we assume we can specify a set of p variables $Z_1,\ldots,Z_p$ that affect the value of the vectors $\beta_{i.}$[1] Also, that $z_i' = (z_{i1},\ldots,z_{ip})$ is known for all the N individuals in the cross-section, where z_{ij} is the value of Z_j for individual i. Further, we assume to have

$$\beta_i = F(z_i) + \varepsilon_i \qquad i = 1,\ldots,N, \tag{3}$$

where $F(z_i) = (F_1(z_i),\ldots,F_K(z_i))'$; $F_k(z_i)$ denotes a non-stochastic function which is equal to the expectation of the k'th element of β_i given z_i; and $\varepsilon_i = (\varepsilon_{i1},\ldots,\varepsilon_{iK})'$ is a K by 1 random vector with zero expectation and VCM given by $E[\varepsilon_i\varepsilon_j'] = \Omega$ if $i = j$ and 0 otherwise. We refer to parameter variation of the kind specified in (3) as *Systematic Parameter Variation* (SPV).

In this paper we concentrate on the SPV model given by equations (2) and (3). In Section 2 we comment on a test for SPV, and in Section 3 we suggest a *two stage* estimation procedure that may be used when there is evidence of parameter variation. The *first stage* of our estimation procedure is presented in Section 4. The *second stage* is discussed in Section 5. A numerical exercise is included in Section 6, comparing three estimation procedures under various forms of SPV. Other possible approaches to the problem of *parameter variation* and some concluding remarks are found in Section 7.

2. TESTING FOR SYSTEMATIC PARAMETER VARIATION

We note the existence of ε_i in (3) makes β_i a random vector; also, that if an element in x_i is constant for all i, u_i would not be distinguishable from the varying intercept and it could be subsumed into the latter. We assume our regression model contains an intercept and (without loss of generality) omit the term u_i in (2).

If $F(z_i)$ were known, substitution of (3) into (2) would yield an equation amenable to econometric analysis; and its estimation could be carried out - for example - by the use of nonlinear procedures. Unfortunately, in general $F(z_i)$ would be unknown. If we estimated equation (1) - neglecting the SPV - problems would arise because of functional misspecification. We may, therefore, be interested in testing the existence of SPV. We have noted $F(z_i)$ is in general unknown. So, to derive a test for SPV we proceed under the presumption that existence of SPV may be detected (hopefully in many cases) by assuming linearity, i.e., by setting

$$F(z_i) = \Gamma \begin{bmatrix} 1 \\ z_i \end{bmatrix} = \begin{bmatrix} \gamma_{1o} & \gamma_1' \\ \gamma_{2o} & \gamma_2' \\ \vdots & \vdots \\ \gamma_{Ko} & \gamma_K' \end{bmatrix} \begin{bmatrix} 1 \\ z_i \end{bmatrix} , \qquad (4)$$

and testing $H_o: \gamma_1' = \ldots = \gamma_K' = 0$, where γ_{ko} is a scalar and $\gamma_k' = (\gamma_{k1},\ldots,\gamma_{kp})$ is a 1 by p vector of coefficients, for $k = 1,\ldots,K$.

When we substitute (4) into our model we obtain

$$y_i = x_{i+}'\gamma + u_{i+} , \qquad (5)$$

where $x'_{i+} = (x'_i, (z'_i \otimes x'_i))$, $\gamma = Vec\{\Gamma\}$, $u_{i+} = x'_i \varepsilon_i$, $\otimes$ denotes Kronecker product, and $Vec\{\cdot\}$ is the *vector operator* such that, if A is an n by r matrix given by $A = (a_1, \ldots, a_r)$, $Vec\{A\}$ is an nr by 1 vector equal to $(a'_1, \ldots, a'_r)'$.

The disturbances u_{i+} in (5) are heteroscedastic but, nevertheless, we can easily test for SPV by using the heteroscedasticity-consistent test suggested by White (1980 , p.820). For this problem the test statistic is

$$F_{SPV} = \hat{\gamma}'R'[R(X'_+X_+)^{-1}(\sum_{i=1}^{N} \hat{u}^2_{i+} x_{i+} x'_{i+})(X'_+X_+)^{-1}R']^{-1}R\hat{\gamma} ,$$

where $\hat{\gamma} = (X'_+X_+)^{-1}X'_+y$, X_+ is an N by $(K+Kp)$ matrix with i'th row given by x'_{i+}; $R = [0; I_{Kp}]$ is a Kp by $(K+Kp)$ matrix, and $\hat{u}_{i+} = y_i - x'_{i+}\hat{\gamma}$. Under H_o: $\gamma_1 = \ldots = \gamma_K = 0$ (and provided regular conditions are satisfied) F_{SPV} would be asymptotically distributed as $\chi^2_{(Kp)}$. H_o would be rejected for large values of F_{SPV}. If H_o were accepted, we could say there is lack of evidence of SPV and use results of the usual regression model, or results of the random coefficient model (e.g., see Hildreth and Houck (1968)). If H_o is rejected, we may follow two approaches: *Firstly*, we may suppose $\Gamma(1, z'_i)'$ is a reasonably good approximation to $F(z_i)$, and regard $x'_{i+}\hat{\gamma}$ as the estimated model; *Secondly*, we may use the estimation procedure suggested in the next Section. (In Section 6 this suggested procedure is found to perform better than the first procedure in terms of goodness-of-fit).

3. A TWO STAGE ESTIMATION PROCEDURE

We assume each element of $F(z)$ is a 'smooth function' over the region of interest in the loose sense that, for values of $z = (Z_1, \ldots, Z_p)'$ that are 'close', the values of $F(z)$ would also be 'close'. The

motivation for the approach of this section is the idea that if the N individuals are *classified* into say L groups, so that within a group h the values of z_i are 'close', then - by smoothness of $F(z)$ - the values $F(z_i)$ would be 'close' for members of that group and could be approximated by the group mean.

Now we introduce necessary definitions. Let I_h be the subset of the set of integers $\{1,2,\ldots,N\}$ that defines group h for a given classification; D_{ih} be a dummy variable that takes the value 1 if $i \in I_h$ and 0 if not; and N_h be the number of individuals in group h, for $h = 1,\ldots,L$. In addition, observe (3) and define

$$\beta(h) = \sum_{i \in I_h} F(z_i)/N_h \ . \tag{6}$$

Similarly, define $\bar{\beta} = \sum_{i=1}^{N} F(z_i)/N$ and note that, using (6), this is

$$\bar{\beta} = \sum_{h=1}^{L} \frac{N_h}{N} \beta(h) \ . \tag{7}$$

We refer to the vectors $\beta(1),\ldots,\beta(L)$ as *regimes*, and to $\bar{\beta}$ as the *macroparameter*.

In terms of $\beta(h)$ our original model (given by (3) and (2) without u_i) may be written as

$$y_i = x_i'\beta_i \tag{8}$$

with

$$\beta_i = \beta(h) + v_{ih} \qquad i \in I_h;\ h = 1,\ldots,L \ , \tag{9}$$

where $v_{ih} = F(z_i) - \beta(h) + \varepsilon_i$. It may be shown that $E[v_{ih}] = E_i E_{\varepsilon/i}[v_{ih}] = 0$ for all i, h, and we use Δ_h to denote $E[v_{ih}v'_{jh}]$ for $i = j$. It is also interesting to note that equation (9) may be written as

$$\beta_i = \sum_{h=1}^{L} \beta(h)D_{ih} + v_i \qquad i = 1,\ldots,N , \tag{10}$$

where $v_i = \sum_{h=1}^{L} v_{ih}D_{ih}$.

The first term in the *RHS* of (10) is a vector containing K step-functions, given by

$$F^s(z_i) = \sum_{h=1}^{L} \beta(h)D_{ih} ,$$

and may be regarded as an approximation to $F(z_i)$. Our proposal is to estimate the model using $F^s(z_i)$, i.e., to use (10) rather than (3). This estimation problem may be more specifically stated as - how to classify the N individuals into L groups, and - how to estimate the *regimes* - so the resulting step-functions $F^s(z_i)$ are the 'best' approximation to the elements in $F(z_i)$. To estimate $F^s(z_i)$ we may proceed in two stages. In the *first stage*, L and $I_1, I_2, \ldots, I_L$ (and hence D_{ih}) would be determined by the use of an appropriate classification or clustering criterion. This is discussed in Section 4. In the *second stage* the parameter vectors $\beta(1),\ldots,\beta(L)$ and $\bar{\beta}$ (i.e., the *regimes* and the *macroparameter*) would be estimated by the use of existing econometric procedures. This is illustrated in Section 5. □

Before concluding this section, we note the problem of classification of individuals has been referred to in the econometric literature as a *'sample separation problem'*. Various authors - although in perhaps different contexts - have commented on this. For instance, Kooyman

(1976, p.127) states that observations should be divided into groups "that are homogeneous in respect of value of the parameters" and notes that - unfortunately - subdivision "is in most cases subjective". Similarly, Poirier (1976, p.155) considers the choice of sample partition, and notes the 'difficulty of the problem'. In turn, Chenery and Syrquin (1975, p.162) state that

> "splitting the sample and estimating separate patterns for the subgroups may contribute to a better analysis". They note the classification "should rely as much as possible on theoretical arguments". Also that "clustering techniques may be useful in suggesting ways to quantify theory-based group factors, and (that) its applicability to this problem should be further studied".

The approach presented in the next section (which uses clustering techniques) provides a 'less-subjective' solution to the problem of econometric sample separation. We hope our results lead to the solution of similar problems (e.g., determination of knot location in spline functions).

4. FIRST STAGE: Clustering of Individuals

The *first stage* of our estimation procedure deals with a classification problem, and for this it may seem natural to use *Cluster Analysis.*[2] Cluster Analysis is a generic term applied to a set of classification techniques. A classification, as generally understood, allocates individuals or entities to initially undefined groups or clusters, so that entities in a cluster are in some sense close to one another.

In the previous section it was said that if z_i was 'close' to z_j, it would be assumed the conditional expectation of β_i, given z_i, would be 'close' to that of β_j given z_j. The term 'close' was left undefined.

Of course a definition is required, and for it a distance measure has to be given. There are many of these. For example, Cormack (1971) presents ten different ones which have been proposed by several authors. It is not the intention to review these here. The important point to note is that various distance measures are used as optimizing criteria in existing clustering algorithms (e.g., see Bolshev (1969), Everitt (1974), Ball (1971) or Hartigan (1975)) and that these criteria have arisen in many fields (e.g. Biology, Psychology, Anthropology and Physics). In general, different clustering criteria would provide different classifications. Although many might seem appealing for the purpose of classification, to this stage it is not clear how these criteria relate to estimation aims of the models here considered.

In Subsection 4.1 a clustering criterion is suggested which is derived within an *econometric estimation framework*. This is obtained by maximizing the "Overall Relative Explanatory Power" of $F^s(z_i)$ to the conditional expectation of β_i given z_i, i.e., $F(z_i)$.[3]

In Subsection 4.2 several indicators are given for the determination of the number of groups in which the individuals should be classified.

4.1 Determination of $I_1, I_2, \ldots, I_L$[4]

For the purpose of obtaining our clustering criterion we shall use equation (4), i.e., we shall set $F(z_i)$ equal to $\Gamma(1, z_i')'$. This amounts to taking a Taylor-series expansion of $F_k(z_i)$, and neglecting all the nonlinear terms in the derivation of the clustering criterion. It has been assumed that $F_k(z_i)$ is a smooth function, and our use of equation (4) is based on the presumption that the 'optimum classification' should not be too sensitive to departures from linearity. □

We *first consider the case* $p = 1$, i.e., the case where β_i depends on a single variable, say Z_1. Then, β_i would be given by (set $p = 1$ in (4) and substitute the result in (3))

$$\beta_i = \begin{bmatrix} \gamma_{1o} & \gamma_{11} \\ \gamma_{2o} & \gamma_{21} \\ \vdots & \vdots \\ \gamma_{Ko} & \gamma_{K1} \end{bmatrix} \begin{bmatrix} 1 \\ z_{i1} \end{bmatrix} + \varepsilon_i \ , \tag{11}$$

and the k'th element of β_i would be given by

$$\beta_{ik} = F_k(z_{i1}) + \varepsilon_{ik} \ , \tag{12}$$

where $F_k(z_{i1}) = \gamma_{ko} + \gamma_{k1} z_{i1}$, for $i = 1,\ldots,N$ and $k = 1,\ldots,K$. It may be shown the variance explained by the regression of β_{ik} on $F_k(z_{i1})$ is equal to

$$R^2_{(1,k)} = \frac{\gamma^2_{k1} \, V[Z_1]}{V[\beta_{ik}]} \ , \tag{13}$$

where $V[Z_1] = \sum_{i=1}^{N} (z_{i1} - \bar{z}_1)^2/N$ and $\bar{z}_1 = \sum_{i=1}^{N} z_{i1}/N$. (For proof see Proposition 1 in Appendix and set $p = 1$).

Now we assume the individuals are classified into L groups, and that $F_k(z_{i1})$ is approximated by a step-function $F^s_k(z_{i1})$ with value $\beta_k(h)$ for all the N_h individuals in group h. In this case we may write (see (10))

$$\beta_{ik} = F^s_k(z_{i1}) + v_{ik} \ , \tag{14}$$

where $F_k^s(z_{i1}) = \beta_k(1)D_{i1} + \ldots + \beta_k(L)D_{iL}$ for $i = 1,2,\ldots,N$ and $k = 1,\ldots,K$. It may be shown (see Proposition 2 in Appendix and set $p = 1$) that for (14) the explained variation of the regression, say $R^2_{(2,k)}$, is given by

$$R^2_{(2,k)} = \frac{\gamma^2_{k1} \sum_{h=1}^{L} \frac{N_h}{N} (\bar{z}_{h1} - \bar{z}_1)^2}{V[\beta_{ik}]} , \qquad (15)$$

where $\bar{z}_{h1}$ is the h'th group mean of Z_1.

We take the ratio of $R^2_{(2,k)}$ to $R^2_{(1,k)}$ as a measure of "Relative Explanatory Power" (this term is used by Aigner, Goldberger and Kalton (1975)). The measure refers to the explanatory power of the step-function approximation $F_k^s(z_{i1})$ made to $F_k(z_{i1})$, and is more formally defined as

$$\mathcal{R}^2_k = \frac{R^2_{(2,k)}}{R^2_{(1,k)}} = \frac{\sum_{h=1}^{L} \frac{N_h}{N} (\bar{z}_{h1} - \bar{z}_1)^2}{V[Z_1]} \qquad k = 1,2,\ldots,K. \qquad (16)$$

We may see $\mathcal{R}^2_k$ is also the squared 'correlation coefficient' between $F_k^s(z_{i1})$ and $F_k(z_{i1})$, and the complement of 'information loss' due to a step-function approximation when (12) is true. An alternative expression for $\mathcal{R}^2_k$ is obtained by using the identity

$$\sum_{h=1}^{L} \frac{N_h}{N} (\bar{z}_{h1} - \bar{z}_1)^2 = V[Z_1] - \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (z_{hi1} - \bar{z}_{h1})^2 , \qquad (17)$$

where z_{hi1} is the value of Z_1 for the i'th individual in group h, giving

$$\mathcal{R}^2_k = 1 - D/V[Z_1] , \qquad (18)$$

with

$$D = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (z_{hil} - \bar{z}_{hl})^2 \ . \tag{19}$$

The average of the "Relative Explanatory Power" coefficients, i.e., the average of the K squared 'correlation coefficients' $\mathcal{R}^2_1,\ldots,\mathcal{R}^2_K$, may be taken as a measure of "Overall Relative Explanatory Power".[5] This is

$$\bar{\mathcal{R}}^2 = \frac{1}{K} \sum_{k=1}^{K} \mathcal{R}^2_k = 1 - D/V[Z_1] \ , \tag{20}$$

and a criterion suggested for the classification of the individuals is to find $I_1,\ldots,I_L$ such that $\bar{\mathcal{R}}^2$ is maximized. In this case, $\bar{\mathcal{R}}^2$ is independent of Γ and given that $V[Z_1]$ is fixed, the "Overall Relative Explanatory Power" will be maximized when D is minimized. (In fact, minimizing D implies maximizing $\mathcal{R}^2_k$ for each $k = 1,\ldots,K$). This is equivalent to minimizing the within group sum of squares of Z_1, and we may therefore use the procedure of Singh (1975) or the clustering algorithms of Hartigan (1975, Chapter 4), MacQueen (1967), Sparks (1973) or Ward (1963) [see also Jarque (1982, Section 3.4)]. The resulting classification shall be denoted by C^* and referred to as the *optimum classification.* □

Now we consider the *more general case where* $p > 1$. Here we have $\beta_i = \Gamma(1, z_i')' + \varepsilon_i$, and the k'th element of β_i would be given by

$$\beta_{ik} = F_k(z_i) + \varepsilon_{ik} \ , \tag{21}$$

where $F_k(z_i) = \gamma_{ko} + \gamma_k' z_i$ for $i = 1,\ldots,N$ and $k = 1,\ldots,K$. It can be shown (see Proposition 1 in Appendix) that in this case

$$R^2_{(1,k)} = \frac{\gamma_k' \Sigma \gamma_k}{V[\beta_{ik}]} \ ,$$

where

$$\Sigma = \sum_{i=1}^{N} (z_i - \bar{z})(z_i - \bar{z})'/N$$

is the VCM of the Z-variables and $\bar{z} = \sum_{i=1}^{N} z_i/N$. Similarly, (see Proposition 2 in Appendix) the regression of β_{ik} on $F_k^s(z_i) = \beta_k(1)D_{i1} + \ldots + \beta_k(L)D_{iL}$ gives

$$R^2_{(2,k)} = \frac{\gamma_k' B \gamma_k}{V[\beta_{ik}]} ,$$

where

$$B = \sum_{h=1}^{L} (N_h/N)(\bar{z}_h - \bar{z})(\bar{z}_h - \bar{z})'$$

and

$$\bar{z}_h = \sum_{i \in I_h} z_i/N_h .$$

Therefore, the "Relative Explanatory Power" referring to the k'th element of β_i is equal to

$$\mathcal{R}^2_k = \frac{R^2_{(2,k)}}{R^2_{(1,k)}} = \frac{\gamma_k' B \gamma_k}{\gamma_k' \Sigma \gamma_k} . \tag{22}$$

The quantities $\mathcal{R}^2_k$, B, N_h and $\bar{z}_h$ depend on a given classification C, and to emphasize this in what follows we shall write them as $\mathcal{R}^2_k(C)$, B(C), $N_h(C)$ and $\bar{z}_h(C)$. Using this notation we obtain that the "Overall Relative Explanatory Power" is given by[6]

$$\mathcal{R}^2(C) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{R}^2_k(C) = \frac{1}{K} \sum_{k=1}^{K} \frac{\gamma_k' B(C) \gamma_k}{\gamma_k' \Sigma \gamma_k} . \tag{23}$$

As for the $p = 1$ case, the clustering criterion would be to find the classification C^* that maximizes $\mathfrak{R}^2(C)$. □

First, it is interesting to note that if $K = p$ and each element in β_i is determined by only one Z-variable, i.e., Γ is given by

$$\Gamma_D = \begin{bmatrix} \gamma_{1o} & \gamma_{11} & 0 & \cdots & 0 \\ \gamma_{2o} & 0 & \gamma_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \gamma_{po} & 0 & 0 & \cdots & \gamma_{pp} \end{bmatrix},$$

then $\mathfrak{R}^2(C)$ reduces to a form which is independent of Γ and equal to

$$\mathfrak{R}^2(C) = \frac{1}{p} \sum_{j=1}^{p} \frac{[B(C)]_{jj}}{[\Sigma]_{jj}},$$

where $[A]_{ij}$ denotes the i,j'th element of the matrix A. The use of identity (17) for each variable Z_j, in the expression for $\mathfrak{R}^2(C)$, reduces this to $\mathfrak{R}^2(C) = 1 - (1/p)\Phi(C)$, with

$$\Phi(C) = \frac{1}{N} \sum_{j=1}^{p} \sum_{h=1}^{L} \sum_{i=1}^{N_h(C)} \frac{(z_{hij}(C) - \bar{z}_{hj}(C))^2}{V[Z_j]}, \qquad (24)$$

and where $z_{hij}(C)$ and $\bar{z}_{hj}(C)$ denote, respectively, the value of Z_j for the i'th individual in group h, and the h'th group mean of Z_j, when the classification is C. Therefore, we see that maximizing $\mathfrak{R}^2(C)$ is equivalent to minimizing $\Phi(C)$. Hartigan (1975, Chapter 4), MacQueen (1967), Sparks (1973) and Ward (1963) provide *algorithms* for the minimization of functions such as $\Phi(C)$. These algorithms may clearly be used for the computation of the optimum classification C^* by feeding them with the data

$z_{hij}(C)/(V[Z_j])^{\frac{1}{2}}$ (see equation (24)). □

When we try to extend the results to a more general form for Γ, problems are encountered since $\mathcal{R}^2(C)$ would now depend on the last p columns in Γ, i.e., the vectors $\gamma_1', \ldots, \gamma_K'$ (see (4)). However, these may be estimated by the use of OLS on the equation resulting from substitution of (21) into $y_i = x_i'\beta_i$. The estimator we obtain is $\hat{\gamma} = Vec\{\hat{\Gamma}\} = Vec\{((\hat{\gamma}_{1o}, \hat{\gamma}_1')', \ldots, (\hat{\gamma}_{Ko}, \hat{\gamma}_K')')'\} = (X_+'X_+)^{-1}X_+'y$, where X_+ is defined in Section 2.

It seems natural to proceed to find C^* such that $\hat{\mathcal{R}}^2(C)$ is maximized, where $\hat{\mathcal{R}}^2(C)$ is equal to $\mathcal{R}^2(C)$ (see (23)) but replacing γ_k by $\hat{\gamma}_k$, i.e.,

$$\hat{\mathcal{R}}^2(C) = \frac{1}{K} \sum_{k=1}^{K} \frac{\hat{\gamma}_k' B(C) \hat{\gamma}_k}{\hat{\gamma}_k' \Sigma \hat{\gamma}_k} . \tag{25}$$

Define

$$W(C) = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h(C)} (z_{hi}(C) - \bar{z}_h(C))(z_{hi}(C) - \bar{z}_h(C))' , \tag{26}$$

and recall the identity $B(C) = \Sigma - W(C)$. Now substitute this last expression for $B(C)$ into (25), obtaining

$$\hat{\mathcal{R}}^2(C) = 1 - \frac{1}{NK} \left\{ \sum_{k=1}^{K} \sum_{h=1}^{L} \sum_{i=1}^{N_h(C)} \frac{(\hat{\beta}_{hik}(C) - \bar{\hat{\beta}}_{hk}(C))^2}{V[\hat{\beta}_k]} \right\} , \tag{27}$$

where $\hat{\beta}_{hik}(C) = \hat{\gamma}_{ko} + \hat{\gamma}_k' z_{hi}(C)$,

$\bar{\hat{\beta}}_{hk}(C) = \hat{\gamma}_{ko} + \hat{\gamma}_k' \bar{z}_h(C)$

and

$$V[\hat{\beta}_k] = \hat{\gamma}_k' \Sigma \hat{\gamma}_k \ .$$

We therefore see that $\hat{R}^2(C)$ is maximized when the term enclosed in { } in (27) is minimized; and, for this, we may again use the *clustering algorithms* mentioned previously, but - this time - feeding them with the data

$$\hat{\beta}_{hij}(C)/(V[\hat{\beta}_j])^{\frac{1}{2}}.^7$$ □

In summary, to maximize the "Overall Relative Explanatory Power" of the approximation $F^S(z_i)$ made to $F(z_i)$, the N individuals should be classified such that (19) is minimized when $p = 1$, and (24) is minimized when $\Gamma = \Gamma_D$. For the more general case of a "non-diagonal" Γ, the criterion is to maximize (25). Having found $I_1, I_2, \ldots, I_L$ the econometric estimation of the *regimes* would follow.

4.2 Determination of L

In Subsection 4.1 it was assumed L was known. The proper choice of L is important given that it will partly determine how good the approximation to $F(z_i)$ is. The number of observations N will restrict the value of L, due to a requirement on the minimum number of observations per group in order to estimate the *regimes*. In general, without consideration of degrees of freedom per group, the higher L the better the approximation will be. However, there may be a value beyond which no 'significant improvement' is made, and it would be desirable to find this.

For example, if $p = 1$ and Z_1 has a uniform (a_1, a_2) distribution, $R^2(C)$ using L groups, say $R^2(C:L)$, would be given by $R^2(C:L) = 1 - (1/L^2)$. [See (20) and note the variance of a uniformly distributed random variable is simply 1/12 of the square of the range, i.e., $V[Z_1] = (a_2 - a_1)^2/12$;

in this case, *within each group*, the distribution would be uniform and the range equal to $(a_2-a_1)/L$ so, $D = (a_2-a_1)^2/(12L^2)$}. The values of $\mathcal{R}^2(C:L)$ for $L = 2,3,4,5,6$ and 7 are respectively .750, .889, .938, .960, .972 and .979. Hence beyond $L = 7$ little gain in $\mathcal{R}^2(C)$ would be obtained. In general, a procedure for determining the number of groups is to compute $\hat{\mathcal{R}}^2(C^*)$ (or $\mathcal{R}^2(C^*)$ if $p = 1$) for different values of L, and to choose that beyond which there is no substantial increase in $\hat{\mathcal{R}}^2(C^*)$ (or $\mathcal{R}^2(C^*)$ if $p = 1$). It is interesting to note that if $Z_1,\ldots,Z_p$ are all qualitative variables, so that Z_j can only take one of n_j values, then, by setting $L = \prod_{j=1}^{p} n_j$, we would have $\hat{\mathcal{R}}^2(C^*) = 1$ (or $\mathcal{R}^2(C^*) = 1$ if $p = 1$). In this case, C^* would be the classification of the individuals with each group consisting of individuals whose z_i are equal.

The determination of L may also be carried out within a Cluster Analysis framework. For instance, we could use Ward's (1963) clustering algorithm on the data $z_{hij}(C)/(V[Z_j])^{\frac{1}{2}}$; and note that if the individuals group 'appropriately' into L groups, then it is sensible to approximate $F(z_i)$ by a step-function of L pieces. Several indicators for this are found in the literature. For example, Beale (1969) suggests the use of

$$F_{[p(L_2-L_1),p(N-L_2)]} = \frac{b_{L_1} - b_{L_2}}{b_{L_2}} \left[\frac{N-L_1}{N-L_2} \left(\frac{L_2}{L_1} \right)^{2/p} - 1 \right]$$

where $b_L = \frac{N-L}{p}\, trace\{B(C)\}$. Using an F-Distribution, a significant result would mean that a subdivision into L_2 groups is significantly better than into a smaller number of groups L_1. Calinsky and Harabasz (1971) propose the use of $\lambda = [trace\{B(C)\}/(L-1)]/[trace\{W(C)\}/(N-L)]$, where $W(C) = \Sigma - B(C)$ is the matrix of the within groups sums of squares (see (26)). Here, if λ has its maximum value at L^*, we would set $L = L^*$. Yet another criterion is to use the L which maximizes $L^2 det\{W(C)\}$, as suggested by Marriot

(1971). All of these indicators require the computation of $B(C)$ (therefore $W(C)$ would be easily obtainable) and hence, in practice, several of these may be calculated before reaching a final decision on the number of groups to use.

5. SECOND STAGE: Estimation of Regimes and Macroparameter

The *second stage* of our procedure refers to the econometric estimation of the *regimes* $\beta(1),\ldots,\beta(L)$, and the *macroparameter* $\bar{\beta}$ (see (6) and (7)). In this section alternative estimators are presented. The results described are conditional on a given optimum classification defined by $I_1,\ldots,I_L$.

Two general approaches may be taken for the estimation of the *regimes*. The first uses information on the variables $Z_1,\ldots,Z_p$ and treats the model as one with systematic parameter variation. This approach is discussed in Subsection 5.1. The second approach ignores the information on $Z_1,\ldots,Z_p$ and estimates the *regimes* using random coefficient regression methods. This is treated in Subsection 5.2. Finally, in Subsection 5.3 the estimation of the *macroparameter* $\bar{\beta}$ is discussed.[8]

5.1 Systematic Parameter Variation Approach

It has been assumed that β_i is given by $\beta_i = F(z_i) + \varepsilon_i$, where $F(z_i) = (F_1(z_i),\ldots,F_K(z_i))'$ and $F_k(z_i)$ is an '*unknown smooth function*'. For derivation of the clustering criterion it was further assumed that $F_k(z_i)$ was approximated by a *step-function*. However, once the individuals have been grouped, it may seem appropriate to approximate $F_k(z_i)$ by a *linear function* within each group h, for $h = 1,2,\ldots,L$. One may write

$$\beta_i = \Gamma(h) \begin{bmatrix} 1 \\ z_i \end{bmatrix} + \varepsilon_{ih} \quad \text{for} \quad i \in I_h , \qquad (28)$$

and assume that $E[\varepsilon_{ih}] = 0$ for all i,h and that $E[\varepsilon_{ih}\varepsilon'_{jk}]$ is equal to Ω_h for $i = j$ and $h = k$ and 0 otherwise. The aim now is to estimate $\Gamma(h)$.

Substitution of (28) into the model $y_i = x'_i\beta_i$ gives the relation

$$y_i = ((1,z'_i) \otimes x'_i)Vec\{\Gamma(h)\} + x'_i\varepsilon_{ih} \qquad i \in I_h .$$

Let $\gamma(h) = Vec\{\Gamma(h)\}$, $y_h = (y_1,y_2,\ldots,y_{N_h})'$,[9] $\varepsilon_h = (\varepsilon'_{1h},\varepsilon'_{2h},\ldots,\varepsilon'_{N_h h})'$, $X_{dh} = diag\{x'_1,\ldots,x'_{N_h}\}$, $\varepsilon^*_h = X_{dh}\varepsilon_h$, and Z_h be the N_h by $(p+1)K$ matrix with i'th row equal to $((1,z'_i) \otimes x'_i)$. Assume the rank of Z_h is $(p+1)K$. Also let $X_h = (x_1,\ldots,x_{N_h})'$. Then the above relation may be written, in matrix form, as

$$y_h = Z_h\gamma(h) + \varepsilon^*_h . \tag{29}$$

Equation (29) is amenable to econometric estimation. For example, one may assume Ω_h to be *diagonal* and proceed à la Hildreth and Houck (1968); estimating $\gamma(h)$ by

$$\hat{\gamma}(h) = (Z'_h\hat{\Theta}^{-1}_h Z_h)^{-1}Z'_h\hat{\Theta}^{-1}_h y_h , \tag{30}$$

where $\hat{\Theta}_h$ is an estimator of the VCM of ε^*_h [e.g., one may use $[\hat{\Theta}_h]_{jk}$ equal $[\dot{X}_h\alpha_h]_j$, for $j = k$ and equal to zero for $j \neq k$, where $[\alpha_h]_j$ is the maximum between zero and $[(\dot{X}'_h\dot{M}'_h\dot{M}_h\dot{X}_h)^{-1}\dot{X}'_h\dot{M}'_h\dot{\eta}_h]_j$, and where $M_h = I_{N_h} - Z_h(Z'_hZ_h)^{-1}Z'_h$, $\eta_h = M_h y_h$ and $\dot{A} = [a^2_{ij}]$ if $A = [a_{ij}]$]. From $\hat{\gamma}(h)$ one obtains $\hat{\Gamma}(h)$; then one may predict β_i by $\hat{\Gamma}(h)[1\ z'_i]'$ and estimate $\beta(h)$ by

$$\hat{\beta}(h) = \hat{\Gamma}(h)\begin{bmatrix} 1 \\ \bar{z}_h \end{bmatrix} . \tag{31}$$

The assumption that Ω_h is *diagonal* may be avoided by following Swamy and Mehta (1975). They consider a prior distribution for $\gamma(h)$ with mean $\theta(h)$ and VCM ϕ_h, and suggest using the approximation to the minimum average risk estimator given by

$$\hat{\gamma}(h) = (Z_h' \hat{\Sigma}_h^{-1} Z_h + \phi_h^{-1})^{-1}[Z_h' \hat{\Sigma}_h^{-1} y_h + \phi_h^{-1}\theta(h)) \ , \qquad (32)$$

where $\hat{\Sigma}_h = X_{dh}(I_{N_h} \otimes \hat{\Omega}_h)X_{dh}'$ and where $\hat{\Omega}_h$ is an estimator of Ω_h (see Swamy and Mehta (1975, p.596)).

5.2 Random Coefficient Regression Approach

A second approach to the estimation of the *regimes* may be to consider the model as written out in (8) and (9), and proceed to estimate $\beta(h)$ and Δ_h as in a random coefficient regression model using the data corresponding to I_h. Although this approach neglects the information available in $z_1, z_2, \ldots, z_{N_h}$, it provides an alternative estimation procedure that would be particularly useful when the number of observations in a group is small. (In order to use (30) one requires that N_h be greater or equal to $(p+1)K$).

Define $v_h = (v_{1h}', \ldots, v_{N_h h}')'$ and $\xi_h = X_{dh} v_h$. Then, in matrix form, (8) and (9) may be written as

$$y_h = X_h \beta(h) + \xi_h \qquad h = 1,2,\ldots,L \ .$$

Assuming Δ_h to be diagonal one may follow Hildreth and Houck (1968) and estimate $\beta(h)$ by

$$\beta^+(h) = (X_h' \hat{\Lambda}_h^{-1} X_h)^{-1} X_h' \hat{\Lambda}_h^{-1} y_h \ , \qquad (33)$$

where $\hat{\Lambda}_h$ is an estimator of the variance-covariance matrix of ξ_h (see Hildreth and Houck (1968, p.589)).

Alternatively, one may avoid this assumption and follow Swamy and Mehta (1975) and use as estimator of $\beta(h)$

$$\tilde{\beta}(h) = (X_h' \tilde{\Sigma}_h^{-1} X_h + \psi_h^{-1})^{-1}(X_h' \tilde{\Sigma}_h^{-1} y_h + \psi_h^{-1} r(h)) , \qquad (34)$$

where $r(h)$ is the prior mean and ψ_h the prior VCM of $\beta(h)$, $\tilde{\Sigma}_h = X_{dh}(I_{N_h} \otimes \hat{\Delta}_h)X_{dh}'$ and $\hat{\Delta}_h$ is an estimator of Δ_h (see Swamy and Mehta (1975, p.596)).[10]

5.3 Estimation of the Macroparameter

So far, various estimators for the *regimes* have been presented. In practice, one may also be interested in estimating the *macroparameter* $\bar{\beta}$ (see (7)). For this one may consider the estimators of the *regimes* and use $\Sigma(N_h/N)\hat{\beta}(h)$ (see Subsection 5.1), or $\Sigma(N_h/N)\beta^+(h)$ or $\Sigma(N_h/N)\tilde{\beta}(h)$ (see Subsection 5.2). Alternatively, one may regard the $\beta(h)$ as a sample of independent identically distributed vector random variables with mean $\bar{\beta}$ and VCM Δ, and choose to estimate $\bar{\beta}$ and $\Delta, \Delta_1, \ldots, \Delta_L$. For this, one may follow Swamy and Mehta (1975, Section 3). They discuss the estimation of a random coefficient regression model from *panel data*, but their results are also applicable in a purely *cross-sectional* framework. This is now illustrated.

Firstly note that estimates of $\Delta_1, \ldots, \Delta_L$ may be obtained from estimation within I_h for $h = 1,2,\ldots,L$ (see comments below (34)). Now define $X = (X_1', X_2', \ldots, X_L')'$, $y = (y_1', y_2', \ldots, y_L')$, $b(h) = (X_h' \tilde{\Sigma}_h^{-1} X_h)^{-1} X_h' \tilde{\Sigma}_h^{-1} y_h$ and $S = \Sigma b(h)b'(h) - (1/L)(\Sigma b(h))(\Sigma b'(h))$. Swamy and Mehta (1975, p.600) suggest estimating Δ by

$$\hat{\Delta} = \frac{S}{L-1} - \frac{1}{L} \sum_{h=1}^{L} (X_h' \tilde{\Sigma}_h^{-1} X_h)^{-1}$$

and assuming that the prior mean of $\bar{\beta}$ is r and the prior VCM is ψ,

$$\hat{\bar{\beta}} = (X'\hat{\Sigma}^{-1}X + \psi^{-1})^{-1}(X'\hat{\Sigma}^{-1}y + \psi^{-1}r) \ ,$$

where $\hat{\Sigma} = diag\{X_1\hat{\Delta}X_1' + \tilde{\Sigma}_1, \ldots, X_L\hat{\Delta}X_L' + \tilde{\Sigma}_L\}$. It may be shown that $\hat{\bar{\beta}}$ is an approximation to the minimum average risk estimator of $\bar{\beta}$. Also, that under diffuse prior information (i.e. when setting ψ^{-1} equal to zero) this would reduce to a weighted sum of the estimators of the *regimes*, b(h), and given by

$$\hat{\bar{\beta}} = \sum_{h=1}^{L} \tau_h b(h)$$

where

$$\tau_h = \left[\sum_{j=1}^{L} [\hat{\Delta} + (X_j'\tilde{\Sigma}_j^{-1}X_j)^{-1}]^{-1} \right]^{-1} [\hat{\Delta} + (X_h'\tilde{\Sigma}_h^{-1}X_h)^{-1}]^{-1}.$$

6. NUMERICAL EXERCISE

In this section, a numerical exercise is presented for the comparison of three estimation procedures under alternative forms of *parameter variation.* For the study, $N = 100$, $K = 2$, $p = 1$ and $L = 5$. Variable X_{i1} is equal to one for all $i = 1,2,\ldots,N$ and the observations on X_2 are generated from a Normal (10,1). The disturbance terms $u_1,\ldots,u_N$ are generated from a Normal $(0,\sigma^2)$; and the observations on Z are generated from a Lognormal such that log (Z) is distributed as a Normal (3,1) (the subroutine used is described in Naylor et al. (1966)).

Four models are considered, defined by $y_i = x_i'\beta_i + u_i$, for $i = 1,\ldots,N$ and where β_i is non-random and given by the following expressions

$$(1) \qquad \beta_i = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} ,$$

$$(2) \qquad \beta_i = \begin{bmatrix} \theta_3 + \theta_4 z_i \\ \theta_5 + \theta_6 z_i \end{bmatrix} ,$$

$$(3) \qquad \beta_i = \begin{bmatrix} \theta_7 + \frac{1}{\theta_8} (z_i)^{\theta_9} \\ \theta_{10} + \frac{1}{\theta_{11}} (z_i)^{\theta_{12}} \end{bmatrix} , \quad \text{and}$$

$$(4) \qquad \beta_i = \begin{bmatrix} \theta_{13}/(\theta_{14} + \theta_{15}(z_i)^{\theta_{16}}) \\ \theta_{17}\exp(z_i/\theta_{18}) \end{bmatrix} ,$$

with

$$(\theta_1,\theta_2,\sigma) = (10,1,1) ,$$

$$(\theta_3,\theta_4,\theta_5,\theta_6,\sigma) = (60,-2,15,-1,70) ,$$

$$(\theta_7,\theta_8,\theta_9,\theta_{10},\theta_{11},\theta_{12},\sigma) = (1,250,2,2,7000,3,10) , \quad \text{and}$$

$$(\theta_{13},\theta_{14},\theta_{15},\theta_{16},\theta_{17},\theta_{18},\sigma) = (200,10,1,1,1,20,10) .$$

Each model is estimated using three procedures which are as follows.

(i) Estimating β_i by $\hat{\beta}_i = \hat{\Gamma}(1 \quad z_i)'$ where $Vec\{\hat{\Gamma}\} = (X_+'X_+)^{-1}X_+'y$ and X_+ is an N by 4 matrix with i'th row equal to $(1,x_{i2},z_i,x_{i2}z_i)$ and $y = (y_1,\ldots,y_N)'$. This procedure is equivalent to taking a linear approximation to $F_k(z_i)$ and in what remains is referred to as linear parameter variation - LPV.

(ii) The second procedure estimates β_i by $\hat{\beta}_i = \beta^+(h)$ for $i \in I_h$, where $\beta^+(h)$ is defined by (33) with $\hat{\Lambda}_h = I_{N_h}$, and I_h is determined by minimizing (19) using the cubic-root procedure of Singh (1975). It should be clear that this is a two stage estimation procedure (equivalent to OLS estimation within the optimum groups) and that it approximates $F_k(z_i)$ by a step-function. The procedure is referred to as 2S-OLS.

(iii) The third procedure estimates β_i by $\hat{\beta}_i = \hat{\Gamma}(h)(1 \quad z_i)'$ for $i \in I_h$, where $Vec\{\hat{\Gamma}(h)\}$ is given by (30) with $\hat{\Theta}_h = I_{N_h}$ and where I_h is the same as for (ii). This is also a two stage estimation procedure and approximates $F_k(z_i)$ by a piecewise linear function. The procedure is referred to as 2S-LPV.

As a goodness-of-fit measure we used R^2 adjusted for degrees of freedom, denoted by $\bar{R}^2$.[11] The values obtained for each of the four models, using the three estimators described, are reported in Table 1.

TABLE 1

Values of $\bar{R}^2$ for Four Models

MODEL	1	2	3	4
(i) LPV	.5537	.8364	.7979	.7499
(ii) 2S-OLS	.5402	.8001	.9123	.8451
(iii) 2S-LPV	.5538	.8363	.9711	.9465

The Table shows that for Model 1 (i.e. when parameter variation does not exist) the three procedures give approximately the same $\bar{R}^2$ (in fact, these are close to the one obtained using OLS, which is $\bar{R}^2 = .5607$).

Comparing LPV and 2S-OLS, it is observed that 2S-OLS performs better when parameter variation departs from linearity (e.g., for Model 3 - where parameter variation is quadratic for the intercept and cubic for the slope - the increase in $\bar{R}^2$ is approximately 14 per cent). Comparing 2S-OLS and 2S-LPV, one notes that 2S-OLS gives a lower $\bar{R}^2$ for all four models. This is reasonable, since piecewise linear functions approximate better than step-functions. Nevertheless, 2S-OLS should not be discarded given that it may be a useful estimation procedure when there are groups of small size (as mentioned before, 2S-LPV requires at least $(p+1)K$ observations per group in order for the econometric estimation to be possible). Other groupings of the observations were used to compute the estimator 2S-LPV and evaluate the effect of the optimum classification C^*. On average, the $\bar{R}^2$'s obtained were approximately 10 per cent lower than those computed with the optimum classification. Overall, these results indicate the preference of the *two stage procedures*, particularly 2S-LPV, and provide evidence that the increase in goodness-of-fit over LPV may be substantial (e.g., for Model 4 the increase in $\bar{R}^2$ is approximately 20 per cent when using 2S-LPV rather than LPV).

7. CONCLUDING REMARKS

In this paper, a two stage procedure for the estimation of systematic varying parameter models has been discussed. In the *first stage*, the individuals would be classified into groups by the use of a clustering criterion suggested. The *second stage* refers to the econometric estimation of the *regimes*, and several estimators for this were described in Section 5.

In our estimation procedure the clustering criterion is to maximize the "Overall Relative Explanatory Power", $\mathcal{R}^2(C)$, of a step-function approximation to the conditional expectation of β_i. We defined $\mathcal{R}^2(C)$ as

$$\mathcal{R}^2(C) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{R}_k^2(C) = \frac{1}{K} \sum_{k=1}^{K} \frac{R^2_{(2,k)}}{R^2_{(1,k)}} \quad ,$$

where $\mathcal{R}_k(C)$ is a correlation coefficient, and $R^2_{(2,k)}$ and $R^2_{(1,k)}$ are goodness-of-fit measures of particular regressions. When the VCM of ε_i in (3), i.e. Ω, is not diagonal, the equations defined by (3) are a system of *seemingly unrelated regressions*, and better measures of goodness-of-fit exist for such models (e.g., see Buse (1979)). An alternative definition of "Overall Relative Explanatory Power" is the ratio of Buse's goodness-of-fit measures for the systems (12) and (14). Unfortunately, this ratio depends on unknown quantities, such as Ω and, more importantly, for our problem this criterion is not numerically manageable. For these reasons we limited our discussion to the goodness-of-fit measure $\mathcal{R}^2(C)$.

Other approaches to the estimation problem may be considered. One may be to take the model as written in (8) and (9), and assume v_{ih} is normally distributed. We could then impose $N_h > K$, and maximize the likelihood function with respect to the classification and the parameters. Another may be to consider finding the classification that optimizes a function of the second order moments of some estimator of the *regimes*. An inconvenience with these approaches is that, unlike the two stage procedure that uses $\mathcal{R}^2(C)$, we may end up with a numerically unmanageable clustering criterion; and searching for the optimum classification by total enumeration of alternatives would be computationally inefficient due to the large number of these.

Throughout this paper, our discussion has been made assuming we have a cross-section of individuals. Of course our results are applicable when having observations on any set of entities, such as farms in a region or banks in a particular country. Also, our findings are applicable in

time-series studies. Here, we would need to modify slightly the algorithms used for the classification of the observations (as in McGee and Carleton (1970)), so the resulting groups contain subsequent observations in time.

Before concluding we will briefly mention some results of a Family Budget study which applied the two stage procedure suggested. The data was obtained from the 1975 Mexican Income-Expenditure Household Survey of the Mexican Ministry of Labor. We used the Extended Linear Expenditure System (ELES) derived by maximization of a Stone-Geary utility function, obtaining a system of equations for expenditures with current family income as the regressor.

Given that the study used cross-sectional data we postulated that the parameters of this system of equations were related to socioeconomic variables, such as income and family size and age and occupation of the head of the household [12]. The problem was how to group the 521 households for which data was available (only the Mexico City households were considered) so the parameters of the demand system were approximately the same.

The first step was to determine the number of grups into which the households were to be classified. For this, the R^2 (C^*) criterion was considered. However, in this application we had a system of m equations, so we took the average of the m individual R^2's as the criterion to optimize. We call this the "*Average Overall Relative Explanatory Power*" and denote it by R_+^2 (C^*). The values for $R_+^2(C^*)$ for L = 2,...,20 are reported in Table 2, together with the increase in $R_+^2(C^*)$ obtained when passing from L to L + 1. Although the choice of L is somewhat arbitrary, we decided to se L = 14 because it gives a value of R_+^2 (C^*) over .9, and because - beyond this - little increase in R_+^2 (C^*) is obtained. [For example, the increase in R_+^2 (C^*) from L = 13 to L = 14 would be .015; similarly, the increase from L = 14 to L = 15 would be .001.]

TABLE 2

Maximum 'Average Overall Relative Explanatory Power'

L	$R^2_+(C^*)$	$\Delta R^2_+(C^*)$	L	$R^2_+(C^*)$	$\Delta R^2_+(C^*)$
1	.000	-	11	.871	.010
2	.453	.453	12	.884	.013
3	.618	.165	13	.895	.011
4	.699	.081	14	.910	.015*
5	.740	.041	15	.911	.001
6	.780	.040	16	.912	.001
7	.827	.047	17	.913	.001
8	.843	.016	18	.916	.003
9	.857	.014	19	.919	.003
10	.861	.004	20	.919	.000

The classification C^* corresponding to $L = 14$ was taken, and some characteristics of the groups forming this were computed. The results are presented in Table 3. In the column relating to occupation we have written, for each of the 14 groups, the occupational category having most frequent occurrence, together with the corresponding frequency. For example, group 2 is formed by 24 households from which 23 have the head of the household unemployed (U). (We use W, E and T to denote, respectively, Worker, Entrepreneur and Technocrat). Other variables included in the Table are mean family income, mean family size, and mean average age of the head of the household. A striking feature of the clustering is the marked separation of households by occupational categories, more than by income classes (e.g., groups 1 and 3 have similar values for family income, size and age of head; and differ because group 1 is formed by Unemployed whereas group 3 is formed by Worker households). This indicates, apparently, that occupation exerts one of the main influences in the determination of consumption behaviour

TABLE 3

Characteristics of Groups from Cluster Analysis

Group	Occupation	Income	Family Size	Age	Number in Group
1	U(24/24)	715.	8.0	55.7	24
2	U(23/24)	1101.	3.3	65.9	24
3	W(41/42)	540.	9.8	49.0	42
4	W(95/95)	464.	7.1	38.3	95
5	W(68/69)	835.	3.6	31.8	69
6	W(47/47)	859.	4.4	54.6	47
7	E(13/28)	668.	10.4	45.9	28
8	E(41/41)	911.	5.4	42.2	41
9	T(62/65)	1154.	5.8	38.9	65
10	T(37/41)	1701.	3.1	31.9	41
11	T(21/26)	3534.	3.6	39.4	26
12	T(6/10)	5538.	3.1	49.3	10
13	T(5/5)	7196.	3.0	44.8	5
14	T(4/4)	9614.	4.0	41.0	4

We could,of course, estimate the ELES for each of these groups of households of 'homogeneous behaviour'. In some of these groups, however, we have few observations; for example, in groups 13 and 14 we have, respectively, 5 and 4 households (see last column in Table 3). Also, in some groups we have two or more occupational categories; for instance, group 11 has 21 Technocrats; 2 Entrepreneurs; 2 Workers and 1 Unemployed. Based on the results of the *Cluster Analysis*, we decided to redefine the groups to have

(i) sufficient number of observations for estimation; and

(ii) readily identifiable domains of study.

This required merging some groups, and group-reassignment of a few households. As pointed out previously, the most immediate split of the households is by occupational category.

Unemployed Households

Regarding the *Unemployed*, the natural further break-up was by family size, with splitting value 6 (The 'mostly Unemployed groups', i.e., groups 1 and 2 were such that group 1 had households of size basically greater or equal to 6 and group 2 less than 6. Also, there were no clear income nor age differentials between these groups). We defined large (L) households as those with family size greater or equal to 6, and small (S) households as those with family size less than 6. So, we divided Unemployed households into 2 groups: L and S; hereafter referred to as UL and US.

Worker Households

Regarding *Workers*, we had that groups 3 and 4 were *large* (L) size households; in group 3 most heads were over 45 years of age, and in group 4 most were under 45. We define households where the head is over 45 years of age as old (O) and those where the head is under 45 years of age as young (Y). [It is interesting to note that LPW (1977, p.122) also used this breaking point for age classification]. We also had that groups 5 and 6 were *small* (S) size Worker households; with group 5 being formed by young (Y) households and group 6 by old (O) households. [In general, groups 3 and 4 contained households with lower income than groups 5 and 6, reflecting that small households have higher incomes]. Because of these features, we decided to divide Worker households into 4 groups: LY, LO, SY and SO; hereafter referred to as WLY. WLO. WSY and WSO.

Entrepreneur Households

Regarding *Entrepreneurs*, we had that group 7 was formed by large (L) households and group 8 by mainly small (S) households. [Income and age differentials were not very significant among these groups, except for the fact that small size households tend to have slightly higher incomes]. So, we divided Entrepreneur households into 2 groups: L and S; hereafter referred to as EL and ES.

Technocrat Households

Finally, regarding *Technocrats*, we observed that groups 9 and 10 were basically 'low-income' (relative to other Technocrat) households, with incomes below 3000 pesos per capita per month. Households with incomes below this level were qualified by the symbol I1. Between groups 9 and 10 no clear age differentials are apparent; but group 9 was formed basically by large (L) size households and group 10 by small (S) households. [Again there was a tendency for smaller households to have higher incomes]. We then had group 11, consisting of 'middle-income' households, having incomes between 3000 and 5000 pesos. Households with income in this interval were qualified by the symbol I2 [It is interesting to note that, out of all the Unemployed, Worker or Entrepreneur households, only 11 had incomes within this interval, and none higher than 6230 pesos]. Finally, we had 'high-income' households in groups 12, 13 and 14; these had incomes over 5000 pesos and were qualified with the symbol I3. So, we formed 4 groups (note we previously had 6) of households of Technocrats, namely: I1L, I1S, I2 and I3; hereafter referred to as TI1L, TI1S, TI2 and TI3.

□

TABLE 4

Characteristics by Type of Households

Type of Household	Income	Family Size	Age	Number in Group
Unemployed				
L	761.30	7.9	55.4	23
S	1613.75	3.2	59.9	29
Worker				
L,Y	422.33	7.9	36.6	89
L,O	726.57	8.2	51.5	51
S,Y	874.41	3.8	31.9	68
S,O	849.89	3.9	56.9	33
Entrepreneur				
L	832.74	8.1	44.6	31
S	1159.93	4.0	41.8	30
Technocrat				
I1,L	1131.85	7.5	41.4	45
I1,S	1565.71	3.8	36.0	64
I2	3693.35	3.9	41.1	20
I3	6818.30	3.2	42.1	13

In all, we ended with 12 groups - or domains of study - containing different *types of households*. The average income, family size and age of the head of the household, together with the number of households in each group are given in Table 4. Average income for large households is lower than for small households (other attributes equal). Also, average incomes for Unemployed, Workers and E trepreneurs are smaller than for Technocrats. Large households (UL, WLY, WLO, EL and TI1L) have an average of approximately 8 members, and small households (US, WSY, WSO, ES and TI1S) have an average of approximately 4 members.

Having identified these 12 groups of households of homogeneous consumption behaviour (relative to the parameters of the demand system), we proceeded to the estimation of the ELES within each group. The complete description of the results is given in *Jarque (1982, Chapter 9)*.

APPENDIX

PROPOSITION 1: The explained variation of the regression of β_{ik} on $F_k(z_i)$ is given by $R^2_{(1,k)}$.

Proof: Equation (21) defines

$$\beta_{ik} = F_k(z_i) + \varepsilon_{ik} , \tag{A.1}$$

where

$$F_k(z_i) = \gamma_{ko} + \gamma_k' z_i .$$

We know $E[\beta_{ik}] = E_i E_{\varepsilon/i}[\beta_{ik}]$. From (A.1) we see $E_{\varepsilon/i}[\beta_{ik}] = \gamma_{ko} + \gamma_k' z_i$. It follows that

$$E[\beta_{ik}] = \gamma_{ko} + \gamma_k' \bar{z} , \tag{A.2}$$

where $\bar{z} = \sum_{i=1}^{N} z_i/N$. We also have

$$V[\beta_{ik}] = V_i E_{\varepsilon/i}[\beta_{ik}] + E_i V_{\varepsilon/i}[\beta_{ik}] .$$

But $E_{\varepsilon/i}[\beta_{ik}]$ is $\gamma_{ko} + \gamma_k' z_i$ and $V_{\varepsilon/i}[\beta_{ik}]$ is $V[\varepsilon_{ik}]$, reducing $V[\beta_{ik}]$ to

$$V[\beta_{ik}] = \gamma_k' \Sigma \gamma_k + V[\varepsilon_{ik}] ,$$

where

$$\Sigma = \sum_{i=1}^{N} (z_i - \bar{z})(z_i - \bar{z})'/N .$$

Therefore, it is seen that the explained variation of the regression of β_{ik} on $F_k(z_i)$ is simply $\gamma_k' \Sigma \gamma_k$ over $V[\beta_{ik}]$, which is equal to $R^2_{(1,k)}$. □

PROPOSITION 2: The explained variation of the regression of β_{ik} on $F_k^s(z_i)$ is given by $R^2_{(2,k)}$.

Proof: Our regression equation is now

$$\beta_{ik} = F_k^s(z_i) + v_{ik} , \tag{A.3}$$

where

$$F_k^s(z_i) = \sum_{h=1}^{L} \beta_k(h) D_{ih} .$$

By definition, $\beta(h) = \sum_{i \in I_h} F(z_i)/N_h$ (see (6)). We have $F_k(z_i) = \gamma_{ko} + \gamma_k' z_i$, so, $\beta_k(h) = \gamma_{ko} + \gamma_k' \bar{z}_h$, where $\bar{z}_h = \sum_{i \in I_h} z_i/N_h$. This means we can rewrite equation (A.3) in the form

$$\beta_{ik} = \sum_{h=1}^{L} (\gamma_{ko} + \gamma_k' \bar{z}_h) D_{ih} + v_{ik} . \tag{A.4}$$

Subtracting (A.2) from (A.4) and noting that $\sum_{h=1}^{L} D_{ih} = 1$, we obtain

$$\beta_{ik} - E[\beta_{ik}] = \gamma_k' \sum_{h=1}^{L} (\bar{z}_h - \bar{z}) D_{ih} + v_{ik} . \tag{A.5}$$

From (A.5) we can show that the variance of β_{ik} may be written as

$$V[\beta_{ik}] = \gamma_k' B \gamma_k + V[v_{ik}^2] , \tag{A.6}$$

where $B = \sum_{h=1}^{L} (N_h/N)(\bar{z}_h - \bar{z})(\bar{z}_h - \bar{z})'$. (To obtain this result note - in particular - that $D_{ih}^2 = D_{ih}$, since D_{ih} is either 1 or 0; $D_{ih}D_{jk} = 0$ except when $i = j$ and $h = k$; and $E[D_{ih}] = N_h/N$). From (A.6) the explained variation of the regression of β_{ik} on $F_k^s(z_i)$ is seen to be $\gamma_k' B \gamma_k$ over $V[\beta_{ik}]$, which is $R^2_{(2,k)}$. □

The helpful discussion of an earlier version of this paper with Deane Terrell and Adrian Pagan is gratefully acknowledged.

NOTES

[1] These Z-variables may contain expressions involving the regressors, e.g., if parameter variation is thought to be due to having misspecified the functional form in (1), despite inclusion of all relevant regressors.

[2] Other names for Cluster Analysis are Q-Analysis, Typology, Grouping, Clumping, Numerical Taxonomy and Unsupervised Pattern Recognition.

[3] The approach presented was motivated by the work of Aigner, Goldberger and Kalton (1975). They study the explanatory power of dummy variables in a regression equation where one independent variable is categorized. Here the dependent variable is the vector β_i and the more general case where p variables are to be 'categorized' is considered.

[4] Throughout this Subsection it is assumed that L is given. The determination of L is discussed in Subsection 4.2.

[5] In Section 7 we comment on the use of more general definitions.

[6] A more general form of $\mathfrak{R}^2(C)$ is $(1/K)\Sigma\nu_k\mathfrak{R}^2_k(C)$, where the ν_k are given weights. The use of this does not represent any additional problems in terms of computation of the optimum classification. A reasonable choice for ν_k is the ratio of the standard deviation of the predictions $\hat{\beta}_{ik}$ to the absolute value of their mean, with $\hat{\beta}_i = \hat{\Gamma}(1,z_i')'$ and where $\hat{\Gamma}$ is obtained by applying OLS to (5). For the $p = 1$ case the values of $\nu_1,\ldots,\nu_K$ do not matter since $\mathfrak{R}^2_k(C)$ is independent of k.

[7] Further considerations leading to a clustering procedure that avoids the estimation of Γ are given in Jarque (1980).

[8] The estimators we discuss in this section are not new, and are fully described in the sources cited. Our aim here is to illustrate their use in the present setting.

[9] In this section subindexes of y_i and x_i are in progressive order within each group (e.g., $y_1,\ldots,y_{N_h}$). Also note that whenever y has a subscript h (i.e., y_h) then it refers to an N_h by 1 vector.

[10] Another procedure is obtained by assuming prior exchangeability and following Lindley and Smith (1972). In addition, the analysis in this subsection may be extended to incorporate information about the z_i's as in Swamy and Tinsley (1980).

[11] $\bar{R}^2$ is given by $\bar{R}^2 = 1 - (\sum_{i=1}^{N} \hat{u}_i^2/(N-n))/\sigma_y^2$, where $\sigma_y^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2/(N-1)$, $\bar{Y} = \sum_{i=1}^{N} y_i/N$, $\hat{u}_i = y_i - x_i'\hat{\beta}_i$ and n is the number of parameters estimated. For (i) $n = 4$, for (ii) $n = 10$ and for (iii) $n = 20$. Of course, alternative goodness-of-fit measures could have been used.

[12] Other procedures for incorporating the effects of socioeconomic variables into demand systems are given in Pollak and Wales (1980) and Williams (1977).

REFERENCES

AIGNER, D.J., GOLDBERGER, A.S. and KALTON, G. (1975), "On the Explanatory Power of Dummy Variable Regressions", *International Economic Review*, 16, 503-510.

BALL, G.H. (1971), "Classification Analysis", *SRI Project* 5533, Stanford Research Institute.

BEALE, E.M.L. (1969), *Cluster Analysis* (London: Scientific Control Systems).

BOLSHEV, L.N. (1969), "Cluster Analysis", *Bulletin of the International Statistical Institute*, 43, 411-425.

BUSE, A. (1979), "Goodness of Fit in the Seemingly Unrelated Regressions Model: A Generalization", *Journal of Econometrics*, 10, 109-113.

CALINSKY, T. and HARABASZ, J. (1971), "A Dendrite Method for Cluster Analysis", unpublished manuscript.

CHENERY, H. and SYRQUIN, M. (1975), *Patterns of Development 1950-1970* (London: Oxford University Press).

CORMACK, R.M. (1971), "A Review of Classification", *Journal of the Royal Statistical Society (Series A)*, 134, 321-367.

EVERITT, B. (1974), *Cluster Analysis* (London: Heinemann).

GALLANT, A.R. and FULLER, W.A. (1973), "Fitting Segmented Polynomial Regression Models whose Join Points have to be Estimated", *Journal of the American Statistical Association*, 68, 144-147.

GOLDFELD, S.M. and QUANDT, R.E. (1973), "A Markov Model for Switching Regressions", *Journal of Econometrics*, 1, 3-16.

GOLDFELD, S.M. and QUANDT, R.E. (1976), eds., *Studies in Nonlinear Estimation* (Cambridge (Mass.): Ballinger).

HARTIGAN, J.A. (1975), *Clustering Algorithms* (New York: Wiley and Sons).

HILDRETH, C. and HOUCK, J.P. (1968), "Some Estimators for a Linear Model with Random Coefficients", *Journal of the American Statistical Association*, 63, 584-595.

HUDSON, D.J. (1966), "Fitting Segmented Curves whose Join Points have to be Estimated", *Journal of the American Statistical Association*, 61, 1097-1129.

JARQUE, C.M. (1980), "Estimation of Varying Parameter Models Using Cluster Analysis", *Working Papers in Economics and Econometrics*, 38, The Australian National University.

JARQUE, C.M. (1982), *Contributions to the Econometrics of Cross-Sections*, Ph.D. dissertation, The Australian National University.

KOOYMAN, M. (1976), *Dummy Variables in Econometrics* (Tilburg: Tilburg University Press).

LINDLEY, D.V. and SMITH, A.F.M. (1972), "Bayes Estimates for the Linear Model", *Journal of the Royal Statistical Association (Series B)*, 34, 1-18.

MCGEE, V.E. and CARLETON, W.T. (1970), "Piecewise Regression", *Journal of the American Statistical Association*, 65, 1109-1124.

MACQUEEN, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations", *Proceedings of the 5th Berkeley Symposium*, 1, 281-297.

MARRIOT, F.H.C. (1971), "Practical Problems in a Method of Cluster Analysis", *Biometrics*, 27, 501-514.

NAYLOR, T.H., BALINTFY, J.L., BURDICK, D.S. and CHU, K. (1966), *Computer Simulation Techniques* (New York: Wiley and Sons).

POIRIER, D.J. (1976), *The Econometrics of Structural Change* (New York: North-Holland).

POLLAK, R.A. and WALES, T.J. (1980), "Demographic Variables in Demand Analysis", paper presented at the *Fourth World Congress of the Econometric Society*, Aix-on-Provence.

SINGH, R. (1975), "On Optimum Stratification for Proportional Allocation", *Sankhya C*, 37, 109-115.

SPARKS, D.N. (1973), "Euclidean Cluster Analysis", *Journal of the Royal Statistical Society (Series C)*, 22, 126-130.

SWAMY, P.A.V.B. (1971), *Statistical Inference in Random Coefficient Regression Models* (New York: Springer-Verlag).

SWAMY, P.A.V.B. and MEHTA, J.S. (1975), "Bayesian and Non-Bayesian Analysis of Switching Regressions and Random Coefficient Regression Models", *Journal of the American Statistical Association*, 70, 593-602.

SWAMY, P.A.V.B. and TINSLEY, P.A. (1980), "Linear Prediction and Estimation Methods for Regression Models with Stationary Stochastic Coefficients", *Journal of Econometrics*, 12, 103-142.

WARD, J.H. (1963), "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association*, 58, 236-244.

WHITE, H. (1980), "A Heteroscedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity", *Econometrica*, 48, 817-838.

WILLIAMS, R.A. (1977), "Engel Curves and Demand Systems: Demographic Effects on Consumption Patterns in Australia", *Preliminary Working Paper* No. SP-07, IMPACT Project, Melbourne.

Serie Documentos de Trabajo 1983

No. I Bhaduri, Amit "Multimarket Classification of Unemployment"

No. II Ize, Alain y Salas, Javier "Prices and Output in the Mexican Economy: Empirical Testing of Alternative Hypotheses"

No. III Alberro, José Luis "Inventory Valuation, Realization Problems and Aggregate Demand"

No. IV Sachs, Jeffrey "Theoretical Issues in International Borrowing"

No. V Ize, Alain y Ortíz, Guillermo "Political Risk, Asset Substitution and Exchange Rate Dynamics"

No. VI Lustig, Nora "Un Análisis de Políticas de Consumo Alimentario en Equilibrio Parcial y Equilibrio General"

No. VII Seade, Jesús "Shifting Oligopolistic Equilibria: Profit-Raising Cost Increases and the Effects of Excise Tax"

No. VIII Jarque, Carlos M. "A Clustering Procedure for the Estimation of Econometric Models with Systematic Parameter Variation"

El Centro de Estudios Económicos de El Colegio de México, ha creado la serie "Documentos de Trabajo" para difundir investigaciones que contribuyen a la discusión de importantes problemas teóricos y empíricos aunque estén en versión preliminar. Con esta publicación se pretende estimular el análisis de las ideas aquí expuestas y la comunicación con sus autores. El contenido de los trabajos es responsabilidad exclusiva de los autores.

Editor: José Luis Alberro

Serie Documentos de Trabajo 1982

No. I Ize, Alain "Disequilibrium Theories, Imperfect Competition and Income Distribution: A Fix Price Analysis"

No. II Levy, Santiago "Un Modelo de Simulación de Precios para la Economía Mexicana"

No. III Persky, Joseph and Tam, Mo-Yin S. "On the Theory of Optimal Convergence"

No. IV Kehoe, Timothy J., Serra-Puche, Jaime y Solís, Leopoldo "A General Equilibrium Model of Domestic Commerce in Mexico"

No. V Guerrero, Víctor M. "Medición de los Efectos Inflacionarios Causados por Algunas Decisiones Gubernamentales: Teoría y Aplicaciones del Análisis de Intervención"

No. VI Gibson, Bill, Lustig, Nora and Taylor, Lance "Terms of Trade and Class Conflict in a Computable General Equilibrium Model for Mexico"

No. VII Dávila, Enrique "The Price System in Cantillon's Feudal Mercantile Model"

No. VIII Ize, Alain "A Dynamic Model of Financial Intermediation in a Semi-Industrialized Economy"

No. IX Seade, Jesús "On Utilitarianism and Horizontal Equity: When is the Equality of Incomes as such Desirable?"

No. X Cárdenas, Enrique "La Industrialización en México Durante la Gran Recesión: Política Pública y Respuesta Privada"