Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 2001, volum 25, núm. 2
Segona època

Entitats patrocinadores:

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Institut d'Estadística de Catalunya

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 2001, volum 25, núm. 2

SUMARI

EDITORIAL

Aquest segon número del volum 25, corresponent a l'any 2001, recull l'edició de set articles, publicats entre les quatre seccions temàtiques de la revista, a més d'una nova recensió a l'apartat dedicat al comentari de llibres i la relació actualitzada de novetats editorials de la Generalitat de Catalunya en matèria estadística.

Com ja s'ha fet en altres ocasions, *Qüestiió* promou la publicació de treballs originals que s'han donat a conèixer en el nostre país. En aquest cas, el present número enceta l'edició d'articles provinents de pòsters acceptats en el «Tercer Congrés Europeu de Matemàtiques» (Barcelona, 10-14 de juliol 2000) que va organitzar la Societat Catalana de Matemàtiques-IEC, sota els auspicis de la *European Mathematical Society*, en el qual hi van col·laborar quatre entitats patrocinadores de *Qüestiió*. Una vegada superada l'avaluació corresponent, s'inicia la sèrie de publicacions amb l'article d'H.M. Ramos i M.A. Sordo, en la secció d'Estadística, al qual seguiran altres originals en els propers números d'aquest volum i següents.

D'altra banda, els lectors apreciaran alguns canvis menors en les normes de presentació d'originals a la revista, orientades a l'agilització dels processos d'avaluació i autoedició d'articles en base a potenciar el tractament electrònic. Per últim, com és costum, s'ofereixen algunes dades il·lustratives de la difusió electrònica de *Qüestiió*, la qual continua oferint un ritme prou satisfactori: en el primer semestre de l'any 2001, els accessos gairebé han superat les 55.800 consultes al web de la revista que es van enregistrar en tot l'any 2000, a raó d'unes 300 peticions http de mitjana diària.

Comentari de les seccions «Estadística», «Investigació Operativa», «Estadística Oficial» i «Biometria»

Les seccions esmentades recullen quatre articles a la secció «Estadística», i un article a cadascuna de les altres seccions. En el primer cas, l'article *Some applications of the matrix Haffian in connection with differentiable matrix functions of a central Wishart variate*, de H. Neudecker, és una continuació d'altres articles del mateix autor —publicats en números anteriors de *Qüestiió*— sobre matrius Haffianes, que relaciona dues formes de derivació matricial, amb algunes aplicacions al tractament de les matrius de Wishart. El segon article, ja esmentat, d'H.M. Ramos i M.A. Sordo, *The proportional likelihood ratio order and applications*, introdueix i estudia un nou ordre estocàstic entre variables aleatòries, que pot ser utilitzat per caracteritzar certes variables amb

densitats log-còncaves i log-convexes, propietat que satisfan variables que descriuen la distribució de rendes. El tercer article, *Modelización de datos longitudinales con estructuras de covarianza no estacionarias: modelos de coeficientes aleatorios frente a modelos alternativos*, de V. Núñez-Anton i D.L. Zimmerman, examina les dades longitudinals amb estructura de covariància no estacionària, modelitzant les variàncies i correlacions que depenen d'altres factors a més del temps; els autors presenten i comparen dos exemples al respecte. En el darrer article de la secció, *El sesgo condicionado en el análisis de influencia: una revision*, de J.M. Muñoz, J.L. Moreno, T. Gómez i A. Enguix, es defineix el biaix condicionat a partir d'una fórmula d'Efron i Stein, i els seus autors ho apliquen a l'anàlisi d'influència del model lineal, a l'anàlisi de components principals i al mostreig de poblacions finites.

L'article de la secció «Investigació Operativa», *Modelización de un DSS para la gestión de productos perecederos*, de B. Díaz, J.A. Del Brío i B. González, proposa un sistema de suport a la presa de decisions aplicat a la gestió interhospitalària, el qual permet satisfer la demanda evitant la caducitat del producte i aconseguir, al mateix temps, una reducció notable dels productes que caduquen.

La secció «Estadística Oficial» publica un interessant treball de R. Vergés que, sota el títol *Tratamiento de datos territorializados de vivienda en el inventario de capital residencial*, explora les possibilitats d'una metodologia per al tractament analític (i prospectiu) de les components del parc d'habitatges, no obstant les bases de dades estadístiques deficitàries i incompletes al respecte.

Finalment, la secció «Biometria» presenta l'article de J. Orbe, *Análisis de duración mediante un modelo lineal generalizado semiparamétrico*, que estén la metodologia de Aitkin i Claiton sobre models de duració mitjançant models GLIM, proposant un model semiparamètric amb una component paramètrica que especifica la forma de la dependència, i una no paramètrica que també recull l'efecte de les variables explicatives, però sense suposar la forma funcional de dependència; l'article presenta una aplicació a dades de pacients amb SIDA.

Comentari d'altres seccions i apartats

A continuació, la «Secció docent i problemes» inclou la presentació de nous enunciats de problemes i la solució dels que s'han publicat en el número immediatament anterior d'aquest volum.

Seguidament, a la secció «Comentari de llibres» es fa una ressenya de C.M. Cuadras sobre la recent publicació de Joan M. Batista i Germà Coenders, *Modelos de ecuaciones*

estructurales (col·lecció Cuadernos de Estadística), en la qual es destaca l'esforç dels autors tant en el tractament monogràfic i actual de la modelització amb variables latents com de la potencialitat i limitacions de la seva aplicació en la investigació social.

El darrer apartat sobre «Ressenyes d'activitats institucionals», inclou la revisió actualitzada d'activitats de la Sociedad Española de Biometría, amb l'anunci dels cursos monogràfics organitzats en col·laboració amb altres entitats. En segon lloc, també s'actualitza l'habitual recensió del «Training for European Statisticians-TES Institute», amb la relació dels cursos del *Programme 2001* que s'imparteixen fins el novembre d'enguany, adreçats als membres dels instituts d'estadística oficial en l'àmbit comunitari. Finalment, s'anuncia el programa del workshop europeu que, sota el títol *Regional Data and Statistics in Europe* (Barcelona, 8-9 octubre 2001), organitzen l'Institut d'Estadística de Catalunya i el Centre Europeu de les Regions, amb el suport d'Eurostat-Comissió Europea i la presència de reconeguts especialistes en estadística oficial i l'administració regional. Per últim, les darreres pàgines es dediquen a novetats editorials en l'àmbit de l'estadística que ha publicat enguany la Generalitat de Catalunya.

Carles Cuadras, director executiu

Enric Ripoll, editor executiu

Estadística

QÜESTIIÓ, vol. 25, 2, p. 187-210, 2001

SOME APPLICATIONS OF THE MATRIX HAFFIAN IN CONNECTION WITH DIFFERENTIABLE MATRIX FUNCTIONS OF A CENTRAL WISHART VARIATE*

HEINZ NEUDECKER
Cesaro*

In this paper we revisit Haff's seminal work on the matrix Haffian as we proposed to call it. We review some results, and give new derivations. Use is made of the link between the matrix Haffian ∇F and the differential of the matrix function, dF.

Keywords: Kronecker product, commutation matrix, Hadamard product, matrix differentiation, matrix differentials, matrix partitioning

AMS Classification (MSC 2000): primary 62F0, secondary 62C99

*This research was supported by DGES.
*Oosterstraat, 13. 1741 GH Schagen. The Netherlands. E-mail: heinz@fee.uva.nl
-Received January 2001.
-Accepted April 2001.

1. INTRODUCTION

In the early eighties of last century Haff (1981, 1982) published seminal work on what I recently proposed to call the matrix Haffian. See Neudecker (2000*b*). Haff applied this matrix to various multivariate problems involving central Wishart variates. Relevant is a differentiable *square* matrix function $F(X)$, shortly F, which depends on a *symmetric* matrix X. Both matrices have the same dimension.

A strategic rôle is being played by a square matrix $\nabla = (d_{ij})$ of operators $d_{ij} := \frac{1}{2}(1+\delta_{ij})\frac{\partial}{\partial x_{ij}}$, where δ_{ij} is the Kronecker delta ($\delta_{ii} = 1, \delta_{ij} = 0$ when $i \neq j$). Haff used the symbol D, not ∇. The matrix ∇ applied to F yields the matrix Haffian ∇F. In parallel work on the kindred *scalar* Haffian I proposed to use the symbol ∇ (Neudecker, 2000*a*) in order to avoid confusion with the so-called duplication matrix which naturally cropped up in that context. Neudecker (2000*b*) presented a link between ∇F and dF, the differential of F.

Haff (1981) gave a fundamental identity based on the matrix Haffian involving a differentiable, not necessarily square, matrix function whose argument was a central Wishart variate. This Fundamental Identity (FI) was used to find expected values of occasionally complicated functions of a central Wishart variate. See also Haff (1982) for further results.

In the present paper we shall revisit Haff's seminal oeuvres, review some of his results, and give new derivations using the link between ∇F and dF.

We shall also consider other applications, drawing heavily on work by Legault-Giguère (1974), Giguère & Styan (1978) and Styan (1989).

2. THE FUNDAMENTAL IDENTITY

Haff (1981, Section 2, (4)) presents the following Fundamental Identity (FI) which holds under mild conditions on the input matrix, viz

$$\mathcal{E}F_1\Sigma^{-1}F_2 = 2\mathcal{E}F_1\nabla F_2 + 2\left(\mathcal{E}F_2'\nabla F_1'\right)' + (n-m-1)\mathcal{E}F_1S^{-1}F_2 \tag{1}$$

with $S \sim W_m(\Sigma, n)$, $n > m+1$ and $F_i := F_i(S)$ $(i = 1,2)$. As usual $\mathcal{E}$ is the expectation operator.

In Haff's presentation $F_1(F_2)$ is of dimension $p \times m$ $(m \times q)$. We shall have $p = q = m$, hence F_1, F_2, S and ∇ are all square of dimension m. This will do for our purposes.

3. THE LINK BETWEEN ∇F AND dF

In Neudecker (2000*b*) the following theorem was proved.

Theorem 1

For the differentiable matrix function $F(X)$ of symmetric X:

$$dF = P'(dX)Q \quad implies \quad \nabla F = \frac{1}{2}PQ + \frac{1}{2}(trP)Q,$$

where dF and dX are differentials of F and X.

In the sections to follow we shall apply Haff's FI and our Theorem 1 to a wide collection of matrix functions of a central Wishart variate. We shall therefore use S instead of X to denote the argument matrix. See Magnus and Neudecker (1999) on matrix differentials.

4. APPLICATIONS I

In this section we reconsider results given by Haff (1981). We shall occasionally use partitioned matrix Haffians. These were also developed by Haff (1981, Section 2). For a survey see the Appendix of this paper.

Theorem 2

$$\mathcal{E}S_{11\cdot 2} = (n - m_2)\Sigma_{11\cdot 2} \quad and \quad \mathcal{E}S_{22}^{-1}S_{21} = \Sigma_{22}^{-1}\Sigma_{21},$$

where $S_{11\cdot 2} := S_{11} - S_{12}S_{22}^{-1}S_{21}$, $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$, $m_1 \times m_1$ is the dimension of S_{11}, and $m_2 := m - m_1$.

$\Sigma_{11\cdot 2}$ is defined accordingly.

Proof

We take $F_1 = I_m$ and $F_2 = \begin{pmatrix} S_{11\cdot 2} & 0 \\ 0 & 0 \end{pmatrix}$. It is known that

$$S^{-1} = \begin{pmatrix} S_{11\cdot 2}^{-1} & -S_{11\cdot 2}^{-1}S_{12}S_{22}^{-1} \\ -S_{22}^{-1}S_{21}S_{11\cdot 2}^{-1} & S_{22\cdot 1}^{-1} \end{pmatrix}.$$

Further $S_{22\cdot 1}$ and Σ^{-1} are expressed analogously to $S_{11\cdot 2}$ and S^{-1}.

Haff's FI in partitioned form yields two equations, viz

(*i*) $\Sigma_{11\cdot 2}^{-1}\mathcal{E}S_{11\cdot 2} = (m_1+1)I_{m_1} + (n-m-1)I_{m_1}$

(*ii*) $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11\cdot 2}^{-1}\mathcal{E}S_{11\cdot 2} = (m_1+1)\mathcal{E}S_{22}^{-1}S_{21} + (n-m-1)\mathcal{E}S_{22}^{-1}S_{21}$, as $\nabla F_2 = \begin{pmatrix} \nabla_{11}S_{11\cdot 2} & 0 \\ \nabla_{21}S_{11\cdot 2} & 0 \end{pmatrix}$,

$$\nabla_{11}S_{11\cdot 2} = \frac{1}{2}(m_1+1)I_{m_1}$$

and

$$\nabla_{21}S_{11\cdot 2} = -\frac{1}{2}(m_1+1)S_{22}^{-1}S_{21}.$$

For details see Corollary 4 (1 & 4) of the Appendix. Solving the two equations yields the result.

□

Theorem 3

$$C\left\{(S_{11\cdot 2})_{ij}, (S_{11\cdot 2})_{kl}\right\} = (n-m_2)\left\{(\Sigma_{11\cdot 2})_{ik}(\Sigma_{11\cdot 2})_{jl} + (\Sigma_{11\cdot 2})_{jk}(\Sigma_{11\cdot 2})_{il}\right\},$$

where $(S_{11\cdot 2})_{ij}$ *is the* ij^{th} *element of* $S_{11\cdot 2}$. *Further* $C(\cdot)$ *denotes the covariance.*

Proof

Take $F_1 = I_m$ and $F_2 = \begin{pmatrix} S_{11\cdot 2}E_{jk}S_{11\cdot 2}E_{li} & 0 \\ 0 & 0 \end{pmatrix}$, with E_{jk} being the jk^{th} basis matrix of dimension $m_1 \times m_1$. Haff's FI in partitioned form yields two equations of which we need only one, viz

$$\begin{aligned} \Sigma_{11\cdot 2}^{-1}\mathcal{E}S_{11\cdot 2}E_{jk}S_{11\cdot 2}E_{li} &= 2\mathcal{E}\nabla_{11}S_{11\cdot 2}E_{jk}S_{11\cdot 2}E_{li} + \\ &+ (n-m-1)\mathcal{E}E_{jk}S_{11\cdot 2}E_{li}. \end{aligned}$$

From Corollary 5(1) of the Appendix emerges that

$$\begin{aligned} 2\nabla_{11}S_{11\cdot 2}E_{jk}S_{11\cdot 2}E_{li} &= (m_1+1)(S_{11\cdot 2})_{kl}E_{ji} + (S_{11\cdot 2})_{jl}E_{ki} + \\ &+ (S_{11\cdot 2})_{kj}E_{li}. \end{aligned}$$

Taking expectations and using Theorem 2 (first part) yields

$$\begin{aligned} \mathcal{E}S_{11\cdot 2}E_{jk}S_{11\cdot 2}E_{li} &= (m_1+1)(n-m_2)(\Sigma_{11\cdot 2})_{kl}\Sigma_{11\cdot 2}E_{ji} + \\ &+ (n-m_2)(\Sigma_{11\cdot 2})_{jl}\Sigma_{11\cdot 2}E_{ki} + (n-m_2)(\Sigma_{11\cdot 2})_{kj}\Sigma_{11\cdot 2}E_{li} + \\ &+ (n-m-1)(n-m_2)(\Sigma_{11\cdot 2})_{kl}\Sigma_{11\cdot 2}E_{ji}, \end{aligned}$$

and finally after taking the trace

$$\begin{aligned}\mathcal{E}(S_{11\cdot 2})_{ij}(S_{11\cdot 2})_{kl} &= (n-m_2)^2(\Sigma_{11\cdot 2})_{ij}(\Sigma_{11\cdot 2})_{kl}+ \\ &+ (n-m_2)\left\{(\Sigma_{11\cdot 2})_{ik}(\Sigma_{11\cdot 2})_{jl}+(\Sigma_{11\cdot 2})_{jk}(\Sigma_{11\cdot 2})_{il}\right\}\end{aligned}$$

from which the result follows.

□

Theorem 4

$$C\left(B'_{\cdot i},B'_{\cdot j}\right)=(n-m_2-1)^{-1}(\Sigma_{11\cdot 2})_{ij}\Sigma_{22}^{-1},$$

where $B':=S_{22}^{-1}S_{21}$ *and* $B'_{\cdot i}$ *is the* i^{th} *column of* B'. *Again* $C(\cdot)$ *denotes the covariance matrix.*

Proof

Write $B'_{\cdot i}=S_{22}^{-1}S_{21}e_i$. Then $C(B'_{\cdot i},B'_{\cdot j})=\mathcal{E}S_{22}^{-1}S_{21}E_{ij}S_{12}S_{22}^{-1}-\Sigma_{22}^{-1}\Sigma_{21}E_{ij}\Sigma_{12}\Sigma_{22}^{-1}$, by virtue of Theorem 2 (second part).

Take then

$$F_2=\begin{pmatrix}0 & S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1}\\ 0 & 0\end{pmatrix}.$$

Hence $\mathcal{E}S^{-1}F_2=\mathcal{E}\begin{pmatrix}0 & E_{ij}S_{12}S_{22}^{-1}\\ 0 & -S_{22}^{-1}S_{21}E_{ij}S_{12}S_{22}^{-1}\end{pmatrix}$.

With $F_1=I_m$ the FI yields

$$\begin{aligned}&\mathcal{E}\begin{pmatrix}0 & \Sigma_{11\cdot 2}^{-1}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1}\\ 0 & -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11\cdot 2}^{-1}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1}\end{pmatrix}\\ &=2\mathcal{E}\begin{pmatrix}0 & \nabla_{11}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1}\\ 0 & \nabla_{21}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1}\end{pmatrix}+\\ &+(n-m-1)\mathcal{E}\begin{pmatrix}0 & E_{ij}S_{12}S_{22}^{-1}\\ 0 & -S_{22}^{-1}S_{21}E_{ij}S_{12}S_{22}^{-1}\end{pmatrix}.\end{aligned}$$

We then get the following two equations:

$$\begin{aligned}(i)\qquad \mathcal{E}\Sigma_{11\cdot 2}^{-1}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1} &= 2\mathcal{E}\nabla_{11}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1}+\\ &+ (n-m-1)\mathcal{E}E_{ij}S_{12}S_{22}^{-1},\end{aligned}$$

$$(ii) \qquad \mathcal{E}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11\cdot 2}^{-1}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1} = -2\mathcal{E}\nabla_{21}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1} + \\ + (n-m-1)\mathcal{E}S_{22}^{-1}S_{21}E_{ij}S_{12}S_{22}^{-1}.$$

From (i) we derive

$$\mathcal{E}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1} = (n-m_2)\Sigma_{11\cdot 2}E_{ij}\Sigma_{12}\Sigma_{22}^{-1}$$

by Lemma 1 (1) of the Appendix and Theorem 2 (second part).

Insertion in (ii) leads to

$$(iii) \qquad (n-m_2)\Sigma_{22}^{-1}\Sigma_{21}E_{ij}\Sigma_{12}\Sigma_{22}^{-1} + \\ +2\mathcal{E}\nabla_{21}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1} = (n-m-1)\mathcal{E}S_{22}^{-1}S_{21}E_{ij}S_{12}S_{22}^{-1}.$$

We use the approach used earlier to find now

$$\nabla_{21}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1} = \frac{1}{2}(S_{11\cdot 2})_{ij}S_{22}^{-1} - \frac{1}{2}(m_1+1)S_{22}^{-1}S_{21}E_{ij}S_{12}S_{22}^{-1}.$$

In fact we applied Corollaries 4 (4) and 2(3) of the Appendix to split $\nabla_{21}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1}$ into two portions.

Hence $2\mathcal{E}\nabla_{21}S_{11\cdot 2}E_{ij}S_{12}S_{22}^{-1} = \mathcal{E}(S_{11\cdot 2})_{ij}S_{22}^{-1} -$

$$-(m_1+1)\mathcal{E}S_{22}^{-1}S_{21}E_{ij}S_{12}S_{22}^{-1} = (n-m_2)(n-m_2-1)^{-1}(\Sigma_{11\cdot 2})_{ij}\Sigma_{22}^{-1} -$$

$$-(m_1+1)\mathcal{E}S_{22}^{-1}S_{21}E_{ij}S_{12}S_{22}^{-1},$$

by Corollary 7 of the Appendix and Theorem 6 in Section 5.

Substitution in (iii) leads to

$$\mathcal{E}S_{22}^{-1}S_{21}E_{ij}S_{12}S_{22}^{-1} = (n-m_2-1)^{-1}(\Sigma_{11\cdot 2})_{ij}\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}E_{ij}\Sigma_{12}\Sigma_{22}^{-1},$$

from which the result follows immediately.

□

This completes Section 4.

5. APPLICATIONS II

In this section we shall consider results presented by various authors (including Haff), occasionally using *different* methods. We shall derive them by using matrix Haffians as advocated by us.

We shall start with some easy often well-known examples to exhibit the powerfulness of the method. They all involve $S \sim W_m(\Sigma, n)$, $n > m+1$.

Theorem 5

$$\mathcal{E}S = n\Sigma.$$

Proof

Take $F_1 = I_m, F_2 = S$ in the Fundamental Identity.

Clearly $dS = I_m(dS)I_m$, hence $\nabla S = \frac{1}{2}(m+1)I_m$ by Theorem 1. Then by the FI: $\Sigma^{-1}S = (m+1)I_m + (n-m-1)I_m = nI_m$, hence $\mathcal{E}S = n\Sigma$. We used $\nabla I_m = 0$.

□

Theorem 6

$$\mathcal{E}S^{-1} = (n-m-1)^{-1}\Sigma^{-1}.$$

Proof

Take $F_1 = F_2 = I_m$. This yields through the FI: $\Sigma^{-1} = (n-m-1)\mathcal{E}S^{-1}$, as $\nabla I_m = 0$.

□

Theorem 7

$$\mathcal{E}SAS = n^2\Sigma A\Sigma + n\Sigma A'\Sigma + n(\mathrm{tr}A\Sigma)\Sigma,$$

where A is a constant matrix.

Proof

Take $F_1 = I_m$ and $F_2 = SAS$. Hence by the FI: $\Sigma^{-1}\mathcal{E}SAS = (m+1)A\mathcal{E}S + A'\mathcal{E}S + \frac{1}{2}(\mathrm{tr}A\mathcal{E}S)I_m + (n-m-1)I_m + (n-m-1)A\mathcal{E}S = n^2A\Sigma + nA'\Sigma + \frac{1}{2}n(\mathrm{tr}A\Sigma)I_m$.

We applied $dSAS = I_m(dS)AS + SA(dS)I_m$, hence $\nabla SAS = \frac{1}{2}(m+1)AS + \frac{1}{2}A'S + \frac{1}{2}(\mathrm{tr}AS)I_m$. Also Theorem 5 was used.

□

Corollary 8

(1) $\mathcal{E}S^2 = n(n+1)\Sigma^2 + n(\mathrm{tr}\Sigma)\Sigma$.

(2) $\mathcal{E}(S \otimes S) = n^2(\Sigma \otimes \Sigma) + nK_{mm}(\Sigma \otimes \Sigma) + n(\mathrm{vec}\Sigma)(\mathrm{vec}\Sigma)'$.

(3) $\mathcal{E}(S \odot S) = n(n+1)\Sigma \odot \Sigma + \Sigma_d 1_m 1_m' \Sigma_d$.

Proof

(1) is obvious. (2) follows by vectorization, viz

$$\begin{aligned}\mathcal{E}(S \otimes S)\mathrm{vec}A &= n^2(\Sigma \otimes \Sigma)\mathrm{vec}A + n(\Sigma \otimes \Sigma)\mathrm{vec}A' + n(\mathrm{vec}\Sigma)(\mathrm{vec}\Sigma)'\mathrm{vec}A \\ &= n^2(\Sigma \otimes \Sigma)\mathrm{vec}A + n(\Sigma \otimes \Sigma)K_{mm}\mathrm{vec}A + n(\mathrm{vec}\Sigma)(\mathrm{vec}\Sigma)'\mathrm{vec}A \\ &= n^2(\Sigma \otimes \Sigma)\mathrm{vec}A + nK_{mm}(\Sigma \otimes \Sigma)\mathrm{vec}A + n(\mathrm{vec}\Sigma)(\mathrm{vec}\Sigma)'\mathrm{vec}A,\end{aligned}$$

where K_{mm} is a commutation matrix.

This equality holds for *any* A. We prove (3) by using the relation $S \odot S = W_m'(S \otimes S)W_m$, and the equalities $K_{mm}W_m = W_m$ and $W_m'\mathrm{vec}\Sigma = \Sigma_d 1_m$, where Σ_d is a diagonal matrix displaying the diagonal of Σ and 1_m is a column vector consisting of m ones.

For these and other properties of the Hadamard product see, e.g. Neudecker, Liu and Polasek (1995).

Theorem 9

$$\mathcal{E}SAS^{-1} = n(n-m-1)^{-1}\Sigma A\Sigma^{-1} - (n-m-1)^{-1}A' - (n-m-1)^{-1}(\mathrm{tr}A)I_m.$$

Proof

Take $F_1 = SA$ and $F_2 = I_m$. Hence by the FI $\mathcal{E}SA\Sigma^{-1} = 2(\mathcal{E}\nabla A'S)' + (n-m-1)\mathcal{E}SAS^{-1}$, which yields $n\Sigma A\Sigma^{-1} = A' + (\mathrm{tr}A)I_m + (n-m-1)\mathcal{E}SAS^{-1}$. We used $dA'S = A'(dS)I_m$ hence $\nabla A'S = \frac{1}{2}A + \frac{1}{2}(\mathrm{tr}A)I_m$.

□

Corollary 10

$$\mathcal{E}S^{-1}AS = n(n-m-1)^{-1}\Sigma^{-1}A\Sigma - (n-m-1)^{-1}A' - (n-m-1)^{-1}(\mathrm{tr}A)I_m.$$

Proof

Transpose the result of Theorem 9 and replace A' by A (A by A').

□

Corollary 11

(1) $\mathcal{E}(S\otimes S^{-1}) = n(n-m-1)^{-1}\Sigma\otimes\Sigma^{-1} - (n-m-1)^{-1}K_{mm} - (n-m-1)^{-1}(\mathrm{vec}\, I_m)(\mathrm{vec}\, I_m)'$.

(2) $\mathcal{E}(S\odot S^{-1}) = n(n-m-1)^{-1}\Sigma\odot\Sigma^{-1} - (n-m-1)^{-1}I_m - (n-m-1)^{-1}1_m 1_m'$.

Proof

As before. Use $W_m'W_m = I_m$.

□

Theorem 12

(1) $\mathcal{E}s_{ij}S = n^2\sigma_{ij}\Sigma + n\Sigma(E_{ij}+E_{ji})\Sigma$

(2) $\mathcal{E}s_{ij}S^{-1} = n(n-m-1)^{-1}\sigma_{ij}\Sigma^{-1} - (n-m-1)^{-1}(E_{ij}+E_{ji})$

(3) $\mathcal{E}s^{ij}S = n(n-m-1)^{-1}\sigma^{ij}\Sigma - (n-m-1)^{-1}(E_{ij}+E_{ji})$

where $s^{ij} = \left(S^{-1}\right)_{ij}$.

Proof

(1) Premultiply in Corollary 8 (2) the expression $\mathcal{E}(S\otimes S)$ by $e_i'\otimes I_m$ and postmultiply by $e_j\otimes I_m$. Use $(e_i'\otimes I_m)K_{mm} = I_m\otimes e_i'$, $(e_i'\otimes I_m)\mathrm{vec}\Sigma = \Sigma e_i$ and $\Sigma e_i\otimes e_j'\Sigma = \Sigma E_{ij}\Sigma$.

(2) Subject Corollary 11 (1) to the same treatment.

(3) Follows from (2) immediately.

□

Corollary 13

(1) $\mathcal{E}(\mathrm{tr}AS)S = n^2(\mathrm{tr}A\Sigma)\Sigma + n\Sigma(A+A')\Sigma$

(2) $\mathcal{E}(\mathrm{tr}AS)S^{-1} = n(n-m-1)^{-1}(\mathrm{tr}A\Sigma)\Sigma^{-1} - (n-m-1)^{-1}(A+A')$

(3) $\mathcal{E}(\mathrm{tr}AS^{-1})S = n(n-m-1)^{-1}(\mathrm{tr}A\Sigma^{-1})\Sigma - (n-m-1)^{-1}(A+A')$.

Proof

Use $\mathrm{tr}AS = \sum_{ij} a_{ij} s_{ij}, \sum_{ij} a_{ij} E_{ij} = A.$

□

Finding $\mathcal{E}(\mathrm{tr}AS^{-1})S^{-1}$ is not so easy. We need this for getting $\mathcal{E}S^{-1}AS^{-1}$.

We shall accomplish this in stages.

Theorem 14

For $\Sigma = I_m$:

$$\mathcal{E}(\mathrm{tr}S^{-1})S^{-1} = (n-m)^{-1}(n-m-1)^{-1}(n-m-3)^{-1}\{m(n-m-2)+2\}I_m.$$

Proof

We apply the FI with $F_1 = I_m$ and $F_2 = AS^{-1}$.

We then get by employing Theorem 1:

$$(n-m-1)^{-1}A = -\mathcal{E}S^{-1}A'S^{-1} - \mathcal{E}(\mathrm{tr}AS^{-1})S^{-1} + (n-m-1)\mathcal{E}S^{-1}AS^{-1}.$$

Expected values of the expressions $(s^{ii})^2$, $s^{ii}s^{ij}$, $s^{ii}s^{jj}$, $s^{ii}s^{jk}$, $(s^{ij})^2$, $s^{ij}s^{ik}$ and $s^{ij}s^{kl}$ have to be determined, where i, j, k and l are distinct.

This will be done by choosing appropriate values of A.

(*i*) $A = E_{ii}$ yields the equation

$$(n-m-1)^{-1}E_{ii} = -\mathcal{E}S^{-1}E_{ii}S^{-1} - \mathcal{E}s^{ii}S^{-1} + (n-m-1)\mathcal{E}S^{-1}E_{ii}S^{-1}.$$

Pre(post)multiplication by $e_i'(e_i), e_i'(e_j), e_j'(e_j)$ and $e_k'(e_l)$ yields

$$\mathcal{E}(s^{ii})^2 = (n-m-1)^{-1}(n-m-3)^{-1}$$

$$\mathcal{E}s^{ii}s^{ij} = 0$$

(1) $$\mathcal{E}(s^{ij})^2 = (n-m-2)^{-1}\mathcal{E}s^{ii}s^{jj}$$

(2) $$\mathcal{E}s^{ik}s^{il} = (n-m-2)^{-1}\mathcal{E}s^{ii}s^{kl}$$

(*ii*) $A = E_{ij}$ leads to the equation

$$(n-m-1)^{-1}E_{ij} = -\mathcal{E}S^{-1}E_{ji}S^{-1} - \mathcal{E}s^{ij}S^{-1} + (n-m-1)\mathcal{E}S^{-1}E_{ij}S^{-1}.$$

Pre(post)multiplication by $e_i'(e_j)$ yields

$$\mathcal{E}(s^{ij})^2 = \frac{1}{2}(n-m-1)\mathcal{E}s^{ii}s^{jj} - \frac{1}{2}(n-m-1)^{-1},$$

which in combination with (1) gives

$$\begin{aligned}\mathcal{E}(s^{ij})^2 &= (n-m)^{-1}(n-m-1)^{-1}(n-m-3)^{-1},\\ \mathcal{E}s^{ii}s^{jj} &= (n-m)^{-1}(n-m-1)^{-1}(n-m-2)(n-m-3)^{-1}.\end{aligned}$$

Pre(post)multiplication by $e_i'(e_k)$ yields

$$\mathcal{E}s^{ij}s^{ik} = \frac{1}{2}(n-m-1)\mathcal{E}s^{ii}s^{jk}$$

which in combination with (2) gives

$$\mathcal{E}s^{ii}s^{kl} = \mathcal{E}s^{ik}s^{il} = 0.$$

Finally, pre(post)multiplication by $e_k'(e_l)$ leads to

$$\mathcal{E}s^{jk}s^{il} + \mathcal{E}s^{ij}s^{kl} = (n-m-1)\mathcal{E}s^{ik}s^{jl}$$

which implies

$$\mathcal{E}s^{ij}s^{kl} = 0,$$

as all these terms are identical.

We conclude that

$$\begin{aligned}\mathcal{E}s^{ii}S^{-1} &= d_1 I_m + 2d_2 E_{ii}\\ \mathcal{E}s^{ij}S^{-1} &= d_2(E_{ij}+E_{ji})\end{aligned}$$

with

$$\begin{aligned}d_1 &:= (n-m)^{-1}(n-m-1)^{-1}(n-m-2)(n-m-3)^{-1}\\ d_2 &:= (n-m)^{-1}(n-m-1)^{-1}(n-m-3)^{-1}.\end{aligned}$$

As $\sum_{i=1}^{m}(d_1 I_m + 2d_2 E_{ii}) = (md_1 + 2d_2)I_m$, the theorem has been proved.

□

Theorem 15

For $\Sigma = I_m$:

$$\mathcal{E}(\mathrm{tr}AS^{-1})S^{-1} = (n-m)^{-1}(n-m-1)^{-1}(n-m-3)^{-1}\left[A+A'+(n-m-2)(\mathrm{tr}A)I_m\right].$$

Proof

Write $\mathrm{tr}AS^{-1} = \sum_i a_{ii}s^{ii} + \sum_{i\neq j} a_{ij}s^{ij}$.

$$\begin{aligned}
\text{Hence } \mathcal{E}(\mathrm{tr}AS^{-1})S^{-1} &= \sum_i a_{ii}\mathcal{E}s^{ii}S^{-1} + \sum_{i\neq j} a_{ij}\mathcal{E}s^{ij}S^{-1} \\
&= \sum_i a_{ii}(d_1 I_m + 2d_2 E_{ii}) + \sum_{i\neq j} a_{ij}d_2(E_{ij}+E_{ji}) \\
&= d_1(\mathrm{tr}A)I_m + 2d_2\sum_i a_{ii}E_{ii} + d_2\sum_{i\neq j} a_{ij}E_{ij} + d_2\sum_{i\neq j} a_{ij}E_{ji} \\
&= d_1(\mathrm{tr}A)I_m + d_2\sum_{ij} a_{ij}E_{ij} + d_2\sum_{ij} a_{ij}E_{ji} \\
&= d_1(\mathrm{tr}A)I_m + d_2(A+A').
\end{aligned}$$

□

Having found $\mathcal{E}(\mathrm{tr}AS^{-1})S^{-1}$ with $\Sigma = I_m$ we can finally determine $\mathcal{E}(\mathrm{tr}AS^{-1})S^{-1}$ for scale parameter $\Sigma \neq I_m$.

Theorem 16

When $S \sim W_m(\Sigma, n)$ *then*

$$\begin{aligned}
\mathcal{E}(\mathrm{tr}AS^{-1})S^{-1} = (n-m)^{-1}(n-m-1)^{-1}(n-m-3)^{-1}\cdot \\
\cdot\left[\Sigma^{-1}(A+A')\Sigma^{-1} + (n-m-2)(\mathrm{tr}A\Sigma^{-1})\Sigma^{-1}\right].
\end{aligned}$$

Proof

When $S \sim W_m(\Sigma, n)$ then $\tilde{S} \equiv \Sigma^{-\frac{1}{2}}S\Sigma^{-\frac{1}{2}} \sim W_m(I_m, n)$.

$$\begin{aligned}
\text{Hence } \mathcal{E}(\mathrm{tr}AS^{-1})S^{-1} &= \mathcal{E}(\mathrm{tr}\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}\tilde{S}^{-1})\Sigma^{-\frac{1}{2}}\tilde{S}^{-1}\Sigma^{-\frac{1}{2}} \\
&= d_2\Sigma^{-\frac{1}{2}}\left[\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}} + \Sigma^{-\frac{1}{2}}A'\Sigma^{-\frac{1}{2}} + (n-m-2)(\mathrm{tr}\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}})I_m\right]\Sigma^{-\frac{1}{2}} \\
&= d_2\left[\Sigma^{-1}(A+A')\Sigma^{-1} + (n-m-2)(\mathrm{tr}A\Sigma^{-1})\Sigma^{-1}\right], \quad \text{by Theorem 15.}
\end{aligned}$$

□

Having obtained this result we now present

Theorem 17

$$\mathcal{E}(S^{-1}AS^{-1}) = d_1\Sigma^{-1}A\Sigma^{-1} + d_2\left[\Sigma^{-1}A'\Sigma^{-1} + (\mathrm{tr}A\Sigma^{-1})\Sigma^{-1}\right],$$

where

$$d_1 := (n-m)^{-1}(n-m-1)^{-1}(n-m-2)(n-m-3)^{-1}$$
$$d_2 := (n-m)^{-1}(n-m-1)^{-1}(n-m-3)^{-1}.$$

Proof

Take $F_1 = I_m$ and $F_2 = AS^{-1}$. We get $dF_2 = -AS^{-1}(dS)S^{-1}$ which implies

$$\nabla F_2 = -\frac{1}{2}S^{-1}A'S^{-1} - \frac{1}{2}(\mathrm{tr}AS^{-1})S^{-1}.$$

Applying the FI we get

$$\Sigma^{-1}\mathcal{E}AS^{-1} = -\mathcal{E}S^{-1}A'S^{-1} - \mathcal{E}(\mathrm{tr}AS^{-1})S^{-1} + (n-m-1)\mathcal{E}S^{-1}AS^{-1}.$$

Using Theorems 6 and 16 we arrive at

$$(n-m-1)^{-1}\Sigma^{-1}A\Sigma^{-1} + d_2\Sigma^{-1}(A+A')\Sigma^{-1} + d_1(\mathrm{tr}A\Sigma^{-1})\Sigma^{-1}$$
$$= (n-m-1)\mathcal{E}S^{-1}AS^{-1} - \mathcal{E}S^{-1}A'S^{-1}.$$

Hence by transposition:

$$(n-m-1)^{-1}\Sigma^{-1}A'\Sigma^{-1} + d_2\Sigma^{-1}(A+A')\Sigma^{-1} + d_1(\mathrm{tr}A\Sigma^{-1})\Sigma^{-1}$$
$$= (n-m-1)\mathcal{E}S^{-1}A'S^{-1} - \mathcal{E}S^{-1}AS^{-1}.$$

The first equation we rewrite as

$$\begin{aligned}(n-m-1)\mathcal{E}S^{-1}AS^{-1} &= (n-m-1)^{-1}\Sigma^{-1}A\Sigma^{-1} + d_2\Sigma^{-1}(A+A')\Sigma^{-1} + \\ &+ d_1(\mathrm{tr}A\Sigma^{-1})\Sigma^{-1} + \mathcal{E}S^{-1}A'S^{-1} = (n-m-1)^{-1}\Sigma^{-1}A\Sigma^{-1} + \\ &+ d_2\Sigma^{-1}(A+A')\Sigma^{-1} + d_1(\mathrm{tr}A\Sigma^{-1})\Sigma^{-1} + (n-m-1)^{-2}\Sigma^{-1}A'\Sigma^{-1} + \\ &+ d_2(n-m-1)^{-1}\Sigma^{-1}(A+A')\Sigma^{-1} + d_1(n-m-1)^{-1}(\mathrm{tr}A\Sigma^{-1})\Sigma^{-1} + \\ &+ (n-m-1)^{-1}\mathcal{E}S^{-1}AS^{-1}.\end{aligned}$$

Hence

$$\begin{aligned}(n-m)(n-m-1)^{-1}(n-m-2)\mathcal{E}S^{-1}AS^{-1} &= d_1(n-m)(n-m-1)^{-1}(\mathrm{tr}A\Sigma^{-1})\Sigma^{-1} + \\ &+ d_1(n-m)(n-m-1)^{-1}(n-m-2)\Sigma^{-1}A\Sigma^{-1} + \\ &+ d_2(n-m)(n-m-1)^{-1}(n-m-2)\Sigma^{-1}A'\Sigma^{-1},\end{aligned}$$

which proves the theorem as $d_1 = (n-m-2)d_2$.

□

Corollary 18

$$\begin{aligned}
&(1)\quad \mathcal{E}(S^{-1}\otimes S^{-1}) &&= d_1\Sigma^{-1}\otimes\Sigma^{-1} + d_2K_{mm}(\Sigma^{-1}\otimes\Sigma^{-1}) + \\
&&&+\ d_2(\mathrm{vec}\,\Sigma^{-1})(\mathrm{vec}\,\Sigma^{-1})' \\
&(2)\quad \mathcal{E}(S^{-1}\odot S^{-1}) &&= (d_1+d_2)\Sigma^{-1}\odot\Sigma^{-1} + d_2(\Sigma^{-1})_d 1_m 1_m'(\Sigma^{-1})_d. \\
&(3)\quad \mathcal{E}S^{-2} &&= (d_1+d_2)\Sigma^{-2} + d_2(\mathrm{tr}\,\Sigma^{-1})\Sigma^{-1},
\end{aligned}$$

with $d_1 := (n-m)^{-1}(n-m-1)^{-1}(n-m-2)(n-m-3)^{-1}$

$d_2 := (n-m)^{-1}(n-m-1)^{-1}(n-m-3)^{-1}$, hence

$d_1 + d_2 := (n-m)^{-1}(n-m-3)^{-1}$.

Proof

As before.

□

Theorem 19

$$\begin{aligned}
\mathcal{E}SASBS &= n[\Sigma(nA+A') + (\mathrm{tr}A\Sigma)I_m]\Sigma[(nB+B')\Sigma + (\mathrm{tr}B\Sigma)I_m] + \\
&+\ n\Sigma B\Sigma(A+A')\Sigma + n\Sigma B'\Sigma(nA'+A)\Sigma + \\
&+\ n\{\mathrm{tr}A\Sigma(nB+B')\Sigma\}\Sigma
\end{aligned}$$

Proof

Take $F_1 = I_m$ and $F_2 = SASBS$. The FI yields the equality

$$\Sigma^{-1}\mathcal{E}SASBS = 2\mathcal{E}\nabla SASBS + (n-m-1)\mathcal{E}ASBS.$$

It is easy to see that

$$\begin{aligned}
2\nabla SASBS + (n-m-1)ASBS &= nASBS + A'SBS + \\
&+\ B'SA'S + (\mathrm{tr}AS)BS + (\mathrm{tr}ASBS)I_m.
\end{aligned}$$

Its expected value is equal to

$$n^3A\Sigma B\Sigma + n^2A\Sigma B'\Sigma + n^2(\mathrm{tr}B\Sigma)A\Sigma + n^2A'\Sigma B\Sigma +$$

$$+nA'\Sigma B'\Sigma + n(\mathrm{tr}B\Sigma)A'\Sigma + n^2B'\Sigma A'\Sigma + nB'\Sigma A\Sigma +$$

$$+n(\mathrm{tr}A\Sigma)B'\Sigma + n^2(\mathrm{tr}A\Sigma)B\Sigma + nB\Sigma(A+A')\Sigma +$$

$$+n[n\mathrm{tr}A\Sigma B\Sigma + \mathrm{tr}A\Sigma B'\Sigma + (\mathrm{tr}A\Sigma)(\mathrm{tr}B\Sigma)]I_m.$$

We used Theorems 1 and 7, and Corollary 13 (1). Premultiplication by Σ and some rearranging yields the result.

□

Corollary 20

(1) $\mathcal{E}SAS^2 = n[\Sigma(nA+A') + (\mathrm{tr}A\Sigma)I_m]\Sigma[(n+1)\Sigma + (\mathrm{tr}\Sigma)I_m] +$
$+n(n+1)(\mathrm{tr}A\Sigma^2)\Sigma + n\Sigma^2[(n+1)A' + 2A]\Sigma.$

(2) $\mathcal{E}S^2AS = n[(n+1)\Sigma + (\mathrm{tr}\Sigma)I_m]\Sigma[(nA+A')\Sigma + (\mathrm{tr}A\Sigma)I_m] +$
$+n(n+1)(\mathrm{tr}A\Sigma^2)\Sigma + n\Sigma[(n+1)A' + 2A]\Sigma^2$

(3) $\mathcal{E}S^3 = n(n^2+3n+4)\Sigma^3 + 2n(n+1)(\mathrm{tr}\Sigma)\Sigma^2 +$
$+n\left[(\mathrm{tr}\Sigma)^2 + (n+1)\mathrm{tr}\Sigma^2\right]\Sigma.$

(4) $\mathcal{E}(S\otimes S^2) = n^2(n+1)\Sigma\otimes\Sigma^2 + n^2(\mathrm{tr}\Sigma)\Sigma\otimes\Sigma + n(n+1)(\mathrm{vec}\Sigma)(\mathrm{vec}\Sigma^2)' +$
$+n(n+1)\left(\Sigma\otimes\Sigma^2 + \Sigma^2\otimes\Sigma\right)K_{mm} + n(\mathrm{tr}\Sigma)(\Sigma\otimes\Sigma)K_{mm} + n(n+1)(\mathrm{vec}\Sigma^2)(\mathrm{vec}\Sigma)' +$
$+n(\mathrm{tr}\Sigma)(\mathrm{vec}\Sigma)(\mathrm{vec}\Sigma)' + 2n\Sigma^2\otimes\Sigma.$

(5) $\mathcal{E}(S\odot S^2) = n(n^2+3n+4)\Sigma\odot\Sigma^2 + n(n+1)(\mathrm{tr}\Sigma)\Sigma\odot\Sigma +$
$+n(\mathrm{tr}\Sigma)\Sigma_d 1_m 1_m'\Sigma_d + n(n+1)\Sigma_d 1_m 1_m'\Sigma_d^2 + n(n+1)\Sigma_d^2 1_m 1_m'\Sigma_d.$

Proof

(1) Replace B by I_m in Theorem 19.

(2) Replace A by I_m and B by A in Theorem 19 *or* transpose Corollary 20 (1) and interchange A and A' in the result.

(3) Replace A by I_m in Corollary 20 (1) or (2).

(4) Vectorize Corollary 20 (2) and omit $\mathrm{vec}A$. This goes as follows. Vectorization of the LHS expression leads to $\mathcal{E}(S\otimes S^2)\mathrm{vec}A$.

Vectorization of the RHS expressions yields

$$\{nI_m \otimes [(n+1)\Sigma^2 + (\mathrm{tr}\Sigma)\Sigma]\}\{(\Sigma \otimes I_m)(nI_{m^2} + K_{mm}) +$$

$$+n(\mathrm{tr}\Sigma)(\mathrm{vec}\Sigma)(\mathrm{vec}\Sigma)'\}\,\mathrm{vec}A + n(n+1)(\mathrm{vec}\Sigma)(\mathrm{vec}\Sigma^2)'\mathrm{vec}A +$$

$$+n(n+1)(\mathrm{vec}\Sigma^2)(\mathrm{vec}\Sigma)'\mathrm{vec}A +$$

$$+n(\Sigma^2 \otimes \Sigma)\{(n+1)K_{mm} + 2I_{m^2}\}\,\mathrm{vec}A.$$

We then cancel vec$A$.

(5) Follows from (4) immediately, see e.g. Corollary 8 (3).

□

Theorem 21

(1) $\mathcal{E}s_{ij}S^2 = n^2(n+1)\sigma_{ij}\Sigma^2 + n^2(\mathrm{tr}\Sigma)\sigma_{ij}\Sigma + n(n+1)\Sigma E_{ij}\Sigma^2 +$
$+n(\mathrm{tr}\Sigma)\Sigma E_{ij}\Sigma + n(n+1)\left(\Sigma^2 E_{ji}\Sigma + \Sigma E_{ji}\Sigma^2\right) +$
$+n(\mathrm{tr}\Sigma)\Sigma E_{ji}\Sigma + n(n+1)\Sigma^2 E_{ij}\Sigma + 2n\left(\Sigma^2\right)_{ij}\Sigma$

(2) $\mathcal{E}\left(S^2\right)_{ij}S = n^2(n+1)\left(\Sigma^2\right)_{ij}\Sigma + n^2(\mathrm{tr}\Sigma)\sigma_{ij}\Sigma + n(n+1)\Sigma E_{ij}\Sigma^2 +$
$+n(n+1)(\Sigma E_{ji}\Sigma^2 + \Sigma^2 E_{ji}\Sigma) + n(\mathrm{tr}\Sigma)\Sigma E_{ji}\Sigma + n(n+1)\Sigma^2 E_{ij}\Sigma +$
$+n(\mathrm{tr}\Sigma)\Sigma E_{ij}\Sigma + 2n\sigma_{ij}\Sigma^2$

Proof

(1) Premultiply in Corollary 20 (4) the expression $\mathcal{E}(S \otimes S^2)$ by $e_i' \otimes I_m$ and postmultiply by $e_j \otimes I_m$. Use $K_{mm}(e_i \otimes I_m) = I_m \otimes e_i$ and $a' \otimes b = ba'$.

(2) Pre(post)multiply in Corollary 20 (4) the expression $\mathcal{E}(S \otimes S^2)$ by $I_m \otimes e_i'$ $(I_m \otimes e_j)$. Use $a \otimes b' = ab'$.

□

Corollary 22

(1) $\mathcal{E}(\mathrm{tr}AS)S^2 = n^2(n+1)(\mathrm{tr}A\Sigma)\Sigma^2 + n^2(\mathrm{tr}A\Sigma)(\mathrm{tr}\Sigma)\Sigma +$
$+n(n+1)\Sigma A\Sigma^2 + n(\mathrm{tr}\Sigma)\Sigma A\Sigma + n(n+1)\left(\Sigma^2 A'\Sigma + \Sigma A'\Sigma^2\right) + n(\mathrm{tr}\Sigma)\Sigma A'\Sigma +$
$+n(n+1)\Sigma^2 A\Sigma + 2n(\mathrm{tr}A\Sigma^2)\Sigma$

(2) $\mathcal{E}(\mathrm{tr}AS^2)S = n^2(n+1)(\mathrm{tr}A\Sigma^2)\Sigma + n^2(\mathrm{tr}A\Sigma)(\mathrm{tr}\Sigma)\Sigma +$
$+n(n+1)\Sigma A\Sigma^2 + n(n+1)\left(\Sigma A'\Sigma^2 + \Sigma^2 A'\Sigma\right) + n(\mathrm{tr}\Sigma)\Sigma A'\Sigma +$
$+n(n+1)\Sigma^2 A\Sigma + n(\mathrm{tr}\Sigma)\Sigma A\Sigma + 2n(\mathrm{tr}A\Sigma)\Sigma^2$

Proof

Use $\mathrm{tr}AS = \sum_{ij} a_{ij}s_{ij}$ and $\sum_{ij} a_{ij}E_{ij} = A$.

□

This has brought us to the end of the article. We want to mention that Theorem 6 and Corollary 18 (3) have been given by Haff (1982). Legault-Giguère (1974) derived Theorems 5, 6, 7, 9, 15, 17 and Corollary 18 (3) in a completely different way.

For Theorems 5, 6, 7, 17 (for $\Sigma = I$) see also Giguère and Styan (1978).

Corollary 10 and Theorems 9 and 17 can also be found in Styan (1989).

For completely different proofs of Theorem 7 see Ghazal and Neudecker (2000) and Neudecker (2000*c*).

Corollaries 8(1) and 20(3) have been established by de Waal and Nel (1973) using a different method.

6. REFERENCES

Ghazal, G. A. and Neudecker, H. (2000). «On second-order and fourth-order moments of jointly distributed random matrices: a survey». *Linear Algebra Appl.*, 321, 61-93.

Giguère, M. A. and Styan, G. P. H. (1978). «Multivariate normal estimation with missing data on several variates». Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, and of the Eighth European Meeting of Statisticians. *Academia, Prague and D. Reidel, Dordrecht*, vol. B, 129-139.

Haff, L. R. (1981). «Further identities for the Wishart distribution with applications in regression». *Canad. J. Statist.*, 9, 215-224.

Haff, L. R. (1982). «Identities for the inverse Wishart distribution with computational results in linear and quadratic discrimination». *Sankhyã*, B, 44, 245-258.

Legault-Giguère, M. A. (1974). «Multivariate normal estimation with missing data». MSc Thesis, McGill University, Montréal, Québec, Canada.

Magnus, J. R. and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, revised edition. John Wiley, Chichester, England.

Neudecker, H. (2000*a*). «A note on the scalar Haffian». *Qüestiió*, 24, 243-249.
Neudecker, H. (2000*b*). «A note on the matrix Haffian». *Qüestiió*, 24, 419-424.
Neudecker, H. (2000*c*). «On expected values of fourth-degree matrix products of a multinormal matrix variate». *New Trends in Probability and Statistics. Proceedings of the 6th Tartu Conference on Multivariate Statistics*, TEV Vilnius/Utrecht.
Neudecker, H., Liu, S. and Polasek, W. (1995). «The Hadamard product and some of its applications in statistics». *Statistics*, 26, 365-373.
Styan, G. P. H. (1989). «Three useful expressions for expectations involving a Wishart matrix and its inverse». *Statistical Data Analysis and Inference*. Elsevier Science Publishers, Amsterdam, 203-296.
de Waal, D. J. and Nel, D. G. (1973). «On some expectations with respect to Wishart matrices». *S. Afr. Statist. J.*, 7, 61-68.

APPENDIX

Partitioned matrix Haffians

Occasionally we meet with lower-dimensional (not necessarily square) matrix functions of a symmetric matrix X.

Examples are $X_{11}^{-1}, X_{11\cdot 2} := X_{11} - X_{12}X_{22}^{-1}X_{21}, X_{22}^{-1}X_{21}$ and $X_{11\cdot 2}\, E_{jk}\, X_{11\cdot 2}E_{li}$, where E_{jk} is the jk^{th} unit matrix of appropriate dimension, and

$$X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}.$$

The submatrices X_{11} and X_{12} are usually of dimension $m_1 \times m_1$ and $m_2 \times m_2$ respectively with $m_1 + m_2 = m$.

The application of the Fundamental Identity and of Theorem 1 is then not clear-cut. It is obvious that X_{11}^{-1} depends on X_{11}, $X_{11\cdot 2}$ depends on X_{11}, X_{12} and X_{22} (with $X_{12} = X_{21}'$) etc.

We can immediately find $\nabla_{11}X_{11}^{-1}, \nabla_{11}X_{11\cdot 2}$ and $\nabla_{11}X_{11\cdot 2}E_{jk}X_{11\cdot 2}E_{li}$ (when E_{li} is square), because operator and operand have equal dimensions in all these cases, viz. $m_1 \times m_1$.

Finding e.g. $\nabla_{12}PX_{11\cdot 2}Q$, $\nabla_{22}PX_{11\cdot 2}Q$ and $\nabla_{21}PX_{11\cdot 2}Q$ (where the generic constant matrices P and Q have such dimensions that operators and operands fit and the products are square) is not trivial.

The application of the FI and of Theorem 1 will be greatly facilitated by partitioning of the operator ∇, viz as

$$\nabla = \begin{pmatrix} \nabla_{11} & \nabla_{12} \\ \nabla_{21} & \nabla_{22} \end{pmatrix}.$$

As ∇ is symmetric, the off-diagonal block matrices ∇_{12} and ∇_{21} satisfy $\nabla_{21} = \nabla_{12}'$. The symmetry of ∇ follows from the circumstance that the ij^{th} scalar element of ∇ is

$$\frac{1}{2}(1+\delta_{ij})\frac{\partial}{\partial x_{ij}} \qquad (i,j = 1,\ldots,m)$$

Haff (1981, Lemma 3) presented a collection of useful results on partitioned Haffians. We shall summarize these, in a streamlined and sometimes generalized form. The proofs will be very similar to those of Haff's Lemma 3.

Lemma 1

1. $\nabla_{11} P' X_{11} Q = \frac{1}{2} PQ + \frac{1}{2}(\mathrm{tr}\, P) Q$
2. $\nabla_{12} P' X_{12} Q = \frac{1}{2} PQ$
3. $\nabla_{12} P' X_{21} Q = \frac{1}{2}(\mathrm{tr}\, P) Q,$

where P and Q are generic constant matrices.

Proof

1. Apply Theorem 1 with X and F replaced by X_{11} and $P' X_{11} Q$ respectively.
2. Take
$$F = \begin{pmatrix} 0 & 0 \\ P' & 0 \end{pmatrix} X \begin{pmatrix} 0 & 0 \\ Q & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ P' X_{12} Q & 0 \end{pmatrix}.$$
Clearly $\nabla F = \frac{1}{2} \begin{pmatrix} PQ & 0 \\ 0 & 0 \end{pmatrix}$.
As $\nabla F = \begin{pmatrix} \nabla_{12} P' X_{12} Q & 0 \\ 0 & 0 \end{pmatrix}$, the result follows.
3. Take
$$F = \begin{pmatrix} 0 & 0 \\ 0 & P' \end{pmatrix} X \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ P' X_{21} Q & 0 \end{pmatrix}.$$
Then $\nabla F = \frac{1}{2}(\mathrm{tr}\, P) \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix}$ and
$$\nabla_{12} P' X_{21} Q = \frac{1}{2}(\mathrm{tr}\, P) Q.$$
□

Corollary 2

1. $\nabla_{22} P' X_{22} Q = \frac{1}{2} PQ + \frac{1}{2}(\mathrm{tr}\, P) Q$
2. $\nabla_{21} P' X_{21} Q = \frac{1}{2} PQ$
3. $\nabla_{21} P' X_{12} Q = \frac{1}{2}(\mathrm{tr}\, P) Q.$

Proof

2. Take $F = \begin{pmatrix} 0 & P' \\ 0 & 0 \end{pmatrix} X \begin{pmatrix} 0 & Q \\ 0 & 0 \end{pmatrix}$.
3. Take $F = \begin{pmatrix} P' & 0 \\ 0 & 0 \end{pmatrix} X \begin{pmatrix} 0 & 0 \\ 0 & Q \end{pmatrix}$.

□

We shall now consider some special results.

Corollary 3

1. $\nabla_{11} P' X_{11}^{-1} Q = -\frac{1}{2} X_{11}^{-1} P X_{11}^{-1} Q - \frac{1}{2} (\mathrm{tr}\, P X_{11}^{-1}) X_{11}^{-1} Q$

2. $\nabla_{12} P' X_{12} X_{22}^{-1} Q = \frac{1}{2} P X_{22}^{-1} Q$

3. $\nabla_{12} P' X_{21} X_{11}^{-1} Q = \frac{1}{2} (\mathrm{tr}\, P) X_{11}^{-1} Q$

4. $\nabla_{11} P' X_{21} X_{11}^{-1} Q = -\frac{1}{2} X_{11}^{-1} X_{12} P X_{11}^{-1} Q - \frac{1}{2} (\mathrm{tr}\, P X_{11}^{-1} X_{12}) X_{11}^{-1} Q$

5. $\nabla_{12} P' X_{22}^{-1} X_{21} Q = \frac{1}{2} (\mathrm{tr}\, P X_{22}^{-1}) Q$

6. $\nabla_{12} P' X_{11}^{-1} X_{12} Q = \frac{1}{2} X_{11}^{-1} P Q$

7. $\nabla_{22} P' X_{22}^{-1} X_{21} Q = -\frac{1}{2} X_{22}^{-1} P X_{22}^{-1} X_{21} Q - \frac{1}{2} (\mathrm{tr}\, P X_{22}^{-1}) X_{22}^{-1} X_{21} Q$

Proof

1. Consider $dP' X_{11}^{-1} Q = P' (dX_{11}^{-1}) Q = -P' X_{11}^{-1} (dX_{11}) X_{11}^{-1} Q$.
   Replace then P' by $-P' X_{11}^{-1}$ and Q by $X_{11}^{-1} Q$ in Lemma 1 (1).
2. Replace Q by $X_{22}^{-1} Q$ in Lemma 1 (2).
3. Replace Q by $X_{11}^{-1} Q$ in Lemma 1 (3).
4. Replace P' by $P' X_{21}$ in 1 of this corollary.
5. Replace P' by $P' X_{22}^{-1}$ in Lemma 1 (3).
6. Replace P' by $P' X_{11}^{-1}$ in Lemma 1 (2).

7. Replace ∇_{11} by ∇_{22}, X_{11}^{-1} by X_{22}^{-1} and Q by $X_{21}Q$ in 1 of this corollary.

□

Note. Haff's Lemma 3 (*e*) is a special case of 5 in this corollary.

Corollary 4

1. $\nabla_{11}P'X_{11\cdot 2}Q = \frac{1}{2}PQ + \frac{1}{2}(\mathrm{tr}\,P)Q$
2. $\nabla_{12}P'X_{11\cdot 2}Q = -\frac{1}{2}PX_{22}^{-1}X_{21}Q - \frac{1}{2}(\mathrm{tr}\,PX_{22}^{-1}X_{21})Q$
3. $\nabla_{22}P'X_{11\cdot 2}Q = \frac{1}{2}X_{22}^{-1}X_{21}PX_{22}^{-1}X_{21}Q + \frac{1}{2}(\mathrm{tr}\,PX_{22}^{-1}X_{21})X_{22}^{-1}X_{21}Q$
4. $\nabla_{21}P'X_{11\cdot 2}Q = -\frac{1}{2}X_{22}^{-1}X_{21}PQ - \frac{1}{2}(\mathrm{tr}\,P)X_{22}^{-1}X_{21}Q$

Proof

1. As only X_{11} varies this result equals that of Lemma 1 (1).
2. This follows from Lemma 1 (2 & 3 combined).

   The reason is that now $dX_{11\cdot 2} = -(dX_{12})X_{22}^{-1}X_{21} - X_{12}X_{22}^{-1}dX_{21}$. Hence we replace Q by $-X_{22}^{-1}X_{21}Q$ in 2 and P' by $P'X_{12}X_{22}^{-1}$ in 3 and add the resulting two expressions together.
3. This follows from Corollary 2 (1). Now

$$dX_{11\cdot 2} = -X_{12}(dX_{22}^{-1})X_{21} = X_{12}X_{22}^{-1}(dX_{22})X_{22}^{-1}X_{21}.$$

   Hence we replace P' by $P'X_{12}X_{22}^{-1}$ and Q by $X_{22}^{-1}X_{21}Q$ in 1.
4. This follows from Lemma 1 (4 & 5 combined).

   The reason is that $dX_{11\cdot 2} = -(dX_{12})X_{22}^{-1}X_{21} - X_{12}X_{22}^{-1}dX_{21}$. Hence we substitute $-X_{22}^{-1}X_{21}Q$ for Q in 4 and $-P'X_{12}X_{22}^{-1}$ for P' in 5 and add the resulting two expressions together.

□

Corollary 5

1. $\nabla_{11} P' X_{11\cdot 2} Q X_{11\cdot 2} R = \frac{1}{2} P Q X_{11\cdot 2} R + \frac{1}{2} (\mathrm{tr}\, P) Q X_{11\cdot 2} R + \frac{1}{2} Q' X_{11\cdot 2} P R + \frac{1}{2} (\mathrm{tr}\, P' X_{11\cdot 2} Q) R$
2. $\nabla_{12} P' X_{11\cdot 2} Q X_{11\cdot 2} R = -\frac{1}{2} P X_{22}^{-1} X_{21} Q X_{11\cdot 2} R - \frac{1}{2} Q' X_{11\cdot 2} P X_{22}^{-1} X_{21} R -$
   $-\frac{1}{2} (\mathrm{tr}\, P X_{22}^{-1} X_{21} Q' X_{11\cdot 2}) R - \frac{1}{2} (\mathrm{tr}\, P X_{22}^{-1} X_{21}) Q X_{11\cdot 2} R$
3. $\nabla_{22} P' X_{11\cdot 2} Q X_{11\cdot 2} R = \frac{1}{2} X_{22}^{-1} X_{21} P X_{22}^{-1} X_{21} Q X_{11\cdot 2} R + \frac{1}{2} (\mathrm{tr}\, P X_{22}^{-1} X_{21}) X_{22}^{-1} X_{21} Q X_{11\cdot 2} R +$
   $+\frac{1}{2} X_{22}^{-1} X_{21} Q' X_{11\cdot 2} P X_{22}^{-1} X_{21} R + \frac{1}{2} (\mathrm{tr}\, P X_{22}^{-1} X_{21} Q' X_{11\cdot 2}) X_{22}^{-1} X_{21} R$
4. $\nabla_{21} P' X_{11\cdot 2} Q X_{11\cdot 2} R = -\frac{1}{2} X_{22}^{-1} X_{21} P Q X_{11\cdot 2} R - \frac{1}{2} (\mathrm{tr}\, P) X_{22}^{-1} X_{21} Q X_{11\cdot 2} R -$
   $-\frac{1}{2} X_{22}^{-1} X_{21} Q' X_{11\cdot 2} P X_{11\cdot 2} R - \frac{1}{2} (\mathrm{tr}\, Q' X_{11\cdot 2} P) X_{22}^{-1} X_{21} X_{11\cdot 2} R$

Proof

1. Using Theorem 1 we conclude from

$$dP' X_{11\cdot 2} Q X_{11\cdot 2} R = P' (dX_{11}) Q X_{11\cdot 2} R + P' X_{11\cdot 2} Q (dX_{11}) R$$

   that the identity holds.
2. This is proved in the same way as Corollary 4 (2). The expression $P' X_{11\cdot 2} Q X_{11\cdot 2} R$ is split into $P' X_{11\cdot 2} (Q X_{11\cdot 2} R)$ and $(P' X_{11\cdot 2} Q) X_{11\cdot 2} R$. We then make the following substitutions in Corollary 4 (2): *(i)* P remains P, Q becomes $Q X_{11\cdot 2} R$ and *(ii)* P becomes $Q' X_{11\cdot 2} P$, Q becomes R.

   This yields the result.
3. This is proved in the same way as Corollary 4 (3). We make the same substitutions as previously.
4. The proof is similar to that of Corollary 4 (4).

   The same substitutions are used as above.

□

Lemma 6

$$\mathcal{E} \left(S_{22}^{-1}\right)_{ij} S_{11\cdot 2} = (n - m_2) \mathcal{E} \left(S_{22}^{-1}\right)_{ij} \Sigma_{11\cdot 2}.$$

Proof

Take

$$F_1 = I_{m_1} \quad \text{and} \quad F_2 = \begin{pmatrix} \left(S_{22}^{-1}\right)_{ij} S_{11\cdot 2} & 0 \\ 0 & 0 \end{pmatrix}.$$

Then

$$\Sigma^{-1} F_2 = \begin{pmatrix} \left(S_{22}^{-1}\right)_{ij} \Sigma_{11\cdot 2}^{-1} S_{11\cdot 2} & 0 \\ -\left(S_{22}^{-1}\right)_{ij} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11\cdot 2}^{-1} S_{11\cdot 2} & 0 \end{pmatrix},$$

$$\nabla F_2 = \begin{pmatrix} \nabla_{11} \left(S_{22}^{-1}\right)_{ij} S_{11\cdot 2} & 0 \\ \nabla_{21} \left(S_{22}^{-1}\right)_{ij} S_{11\cdot 2} & 0 \end{pmatrix}$$

and

$$S^{-1} F_2 = \begin{pmatrix} \left(S_{22}^{-1}\right)_{ij} I_{m_1} & 0 \\ -\left(S_{22}^{-1}\right)_{ij} S_{22}^{-1} S_{21} S_{11\cdot 2} & 0 \end{pmatrix}.$$

Hence by the FI we get

$$\begin{aligned} \mathcal{E}\left(S_{22}^{-1}\right)_{ij} \Sigma_{11\cdot 2}^{-1} S_{11\cdot 2} &= (m_1+1)\mathcal{E}\left(S_{22}^{-1}\right)_{ij} I_{m_1} + (n-m-1)\left(\mathcal{E} S_{22}^{-1}\right)_{ij} I_{m_1} \\ &= (n-m_2)\mathcal{E}\left(S_{22}^{-1}\right)_{ij} I_{m_1}, \end{aligned}$$

by virtue of Lemma 1 (1). This yields

$$\mathcal{E}\left(S_{22}^{-1}\right)_{ij} S_{11\cdot 2} = (n-m_2)\mathcal{E}\left(S_{22}^{-1}\right)_{ij} \Sigma_{11\cdot 2}.$$

□

Corollary 7

$$\mathcal{E}\left(S_{11\cdot 2} \otimes S_{22}^{-1}\right) = (n-m_2)\Sigma_{11\cdot 2} \otimes \mathcal{E} S_{22}^{-1}.$$

Proof

Immediate from Lemma 6.

□

QÜESTIIÓ, vol. 25, 2, p. 211-223, 2001

THE PROPORTIONAL LIKELIHOOD RATIO ORDER AND APPLICATIONS

H. M. RAMOS ROMERO
M. A. SORDO DÍAZ
Universidad de Cádiz*

In this paper, we introduce a new stochastic order between continuous non-negative random variables called the PLR (proportional likelihood ratio) order, which is closely related to the usual likelihood ratio order. The PLR order can be used to characterize random variables whose logarithms have log-concave (log-convex) densities. Many income random variables satisfy this property and they are said to have the IPLR (increasing proportional likelihood ratio) property (DPLR property). As an application, we show that the IPLR and DPLR properties are sufficient conditions for the Lorenz ordering of truncated distributions.

Keywords: Likelihood ratio order, proportional likelihood ratio order, ILR, DLR, IPLR, DPLR, log-concave density function, Lorenz order, truncated distributions

AMS Classification (MSC 2000): 60E15, 60K10

* Departamento de Estadística e I. O. Universidad de Cádiz. Duque de Nájera, 8. 11002 Cádiz. Spain. Corresponding author: Tel.: 34-956015406; fax: 34-956-015385; e-mail: hector.ramos@uca.es

– Received December 2000.
– Accepted March 2001.

1. INTRODUCTION

In this paper, we introduce the PLR (proportional likelihood ratio) order as a new stochastic order among continuous random variables, and two classes of probability distributions, the IPLR and DPLR classes, based on this order. Throughout this paper, the term *increasing* means *non-decreasing but not identically equal to a constant* and *decreasing* has an analogous meaning.

The PLR order is related to the likelihood ratio order, which is defined as follows (see, e.g., Ross, 1983).

Definition 1. *Let X and Y be continuous random variables with densities f and g, respectively, such that*

$$\frac{f(x)}{g(x)} \text{ decreases over the union of the supports of } X \text{ and } Y$$

(here $a/0$ is taken to be equal to ∞ whenever $a > 0$). Then X is said to be smaller than Y in the likelihood ratio order (denoted by $X \leq_{lr} Y$).

Many properties of the likelihood ratio order are listed in Section 1.C of Shaked and Shanthikumar (1994). The order $\leq_{lr}$ can be used to characterize random variables whose logarithms have log-concave (log-convex) densities (see Shaked and Shanthikumar,1994, Theorem 1.C.22). It can be shown that

$$X + t \leq_{lr} X + t' \text{ for all } t \leq t' \iff \log f(x) \text{ is concave.} \tag{1}$$

The characterization (1) shows that log-concavity of densities can be interpreted as an ageing notion in reliability theory. In this sense, we have the following definition (see Ross, 1983).

Definition 2. *The continuous random variable X having density f is said to have the ILR (increasing likelihood ratio) property if* $\log f(x)$ *is concave and is said to have the DLR (decreasing likelihood ratio) property if* $\log f(x)$ *is convex.*

In Section 3 we introduce the IPLR (increasing proportional likelihood ratio) and DPLR (decreasing proportional likelihood ratio) classes. The basic properties of these classes are proven. In particular we show that the IPLR (DPLR) property can be used to characterize non-negative random variables whose logarithms have log-concave (log-convex) densities. A purpose of this Section is to find conditions under which a continuous random variable X is said to have the IPLR (DPLR) property.

In Section 4 we apply the IPLR and DPLR properties to comparisons of truncated random variables according to the Lorenz order. The Lorenz order is closely connected to the so-called Lorenz curve, defined as follows. Suppose $F(x)$ is the distribution function of a non-negative random variable X with finite mean μ. Let F^{-1} denote the inverse of F defined by

$$F^{-1}(p) = \inf\{x : F_X(x) \geq p\}, \quad p \in [0,1],$$

then the Lorenz curve corresponding to X can be defined (Gastwirth, 1971) as:

$$L_X(p) = \frac{1}{\mu}\int_0^p F^{-1}(t)dt \qquad 0 \leq p \leq 1. \tag{2}$$

The Lorenz curve is used in economics to measure the inequality of incomes. If X represents annual income, $L_X(p)$ is the proportion of total income that accrues to individuals having the $100p\%$ lowest incomes. The Lorenz curve provides the next partial ordering between random variables with finite means (see Arnold, 1987).

Definition 3. *We say that* $X \leq_L Y \Leftrightarrow L_X(p) \geq L_Y(p)$ *for every* $0 \leq p \leq 1$.

Let $S(f)$ be the number of sign changes of the function $f(x)$. The next theorem, from Arnold (1987), will be used in Section 4 to show that IPLR and DPLR properties are sufficient conditions to obtain orderings of the truncated distribution by the Lorenz order.

Theorem 1. *Let X and Y be two non-negative random variables with finite means μ_X and μ_Y, respectively, and let F and G be the corresponding distribution functions. If $S(F(x\mu_X) - G(x\mu_Y)) = 1$ and the sign sequence is $-,+$, then $X \leq_L Y$.*

2. THE PROPORTIONAL LIKELIHOOD RATIO ORDER

A new order closely related to the likelihood ratio order will next be described.

Definition 4. *Let X and Y be non-negative and absolutely continuous random variables with supports* $\mathrm{supp}(X)$ *and* $\mathrm{supp}(Y)$, *respectively. Denote the density functions of X and Y by f and g, respectively. Suppose that*

$$\frac{g(\lambda x)}{f(x)} \text{ } \textit{increases} \text{ in } x \text{ for any positive constant } \lambda < 1 \tag{3}$$

over the union of the supports of X and Y (here $a/0$ is taken to be equal to ∞ whenever $a > 0$). Then, we say that X is smaller than Y in the proportional likelihood ratio order (denoted as $X \leq_{plr} Y$).

Example 1. *Let X_i be, $i = 1,2$, the exponential random variable with parameter α_i. Its probability density function is $f_i(x) = \alpha_i \exp\{-\alpha_i x\}$, $x > 0$. Then, it is easy to see that $X_2 \leq_{plr} X_1$ whenever $\alpha_1 < \alpha_2$.*

A consequence of definition 4 is shown next. Suppose that X and Y are random variables whose supports are intervals with non-empty intersection, and let $l_X = \inf\{x : x \in \text{supp}(X)\}$ and $u_X = \sup\{x : x \in \text{supp}(X)\}$. Define l_Y and u_Y similarly.

Theorem 2. *If $X \leq_{plr} Y$, then $l_X \leq l_Y$ and $u_X \leq u_Y$.*

Proof. Suppose that $l_X > l_Y$. Let t_1, t_2 be such that $l_Y < t_1 < l_X < t_2 < \min\{u_X, u_Y\}$ and let $\lambda \in (0,1)$ such that $l_Y < \lambda t_1 < l_X < \lambda t_2 < \min\{u_X, u_Y\}$. Then $g(\lambda t_1)/f(t_1) = \infty > g(\lambda t_2)/f(t_2)$, in contradiction to (3). Therefore we must have $l_X \leq l_Y$. Similarly, it can be shown that $u_X \leq u_Y$. □

If X and Y are two random variables with respective supports (l_X, u_X) and (l_Y, u_Y) such that $l_X \leq l_Y$ and $u_X \leq u_Y$, it should be noted here that in (3) it is sufficient to consider only f and g such that $g(\lambda x)/f(x)$ increases in x over

$$\Lambda(\lambda) = \{x \in \text{supp}(X) \text{ such that } \lambda x \in \text{supp}(Y)\},$$

for all $\lambda \in (0,1)$ rather than over the union of the supports of X and Y.

The next result shows that the *plr* order has the property of ordering by size.

Theorem 3. *Let X and Y be non-negative and absolutely continuous random variables. If $X \leq_{plr} Y$, then $\mu_X \leq \mu_Y$.*

Proof. Let f and g be the density functions of X and Y, respectively, and for each $\lambda \in (0,1)$ let $g_{\frac{Y}{\lambda}}$ denote the density function of the random variable $\frac{Y}{\lambda}$. Suppose, by contradiction, that $\mu_X > \mu_Y$. Since

$$g(\lambda x) = \frac{1}{\lambda} g_{\frac{Y}{\lambda}}(x)$$

it follows from the assumptions that

(4) $g_{\frac{Y}{\lambda}}(x)/f(x)$ is increasing in x for all λ in $(0,1)$.

Hence

$$S\left(g_{\frac{Y}{\lambda}} - f\right) = 1 \text{ for each } \lambda \in (0,1),$$

that is, X and $\frac{Y}{\lambda}$ are stochastically ordered for each λ in $(0,1)$. In particular, by taking

$$\lambda = \frac{\mu_Y}{\mu_X} < 1$$

it follows that the random variables X and $\frac{\mu_X}{\mu_Y}Y$ are stochastically ordered. Since X and $\frac{\mu_X}{\mu_Y}Y$ have the same mean, ordinary stochastic order is only possible if they have the same distribution. This contradicts (4) and hence $\mu_X \leq \mu_Y$ holds. □

The following result characterizes the proportional likelihood ratio order by means of the order $\leq_{lr}$.

Theorem 4. *The two absolutely continuous random variables X and Y satisfy $X \leq_{plr} Y$ if and only if $X \leq_{lr} aY$ for all $a > 1$.*

Proof. Note that

$$\frac{g_{aY}(x)}{f(x)} = \frac{g(x/a)}{af(x)} = \lambda \frac{g(\lambda x)}{f(x)}, \quad \lambda = \frac{1}{a} < 1$$

and the result holds. □

Theorem 5. *Let X and Y be non-negative and absolutely continuous random variables. Suppose that Y has a log-concave density function. Then*

(5) $$X \leq_{lr} Y \Longrightarrow X \leq_{plr} Y.$$

Proof. It is well known that if a non-negative random variable Y has a log-concave density and $a > 1$, then $Y \leq_{lr} aY$ (see, for example, Section 1.C. in Shaked and Shanthikumar, 1994). Since $X \leq_{lr} Y$ by assumption and the relation $\leq_{lr}$ is a transitive order, it follows that $X \leq_{lr} aY$. From Theorem 4 we obtain (5). □

3. INCREASING AND DECREASING PROPORTIONAL LIKELIHOOD RATIO

Definition 5. *Let X be a continuous non-negative random variable with density f. It will be said that X is increasing proportional likelihood ratio (IPLR) if*

$$\frac{f(\lambda x)}{f(x)} \text{ is increasing in } x \text{ for any positive constant } \lambda < 1 \tag{6}$$

It will be said that X is decreasing proportional likelihood ratio (DPLR) if

$$\frac{f(\lambda x)}{f(x)} \text{ is decreasing in } x \text{ for any positive constant } \lambda < 1.$$

(By convention, $\frac{a}{0} = +\infty$ whenever $a > 0$).

The study of the increase of $f(\lambda x)/f(x)$ can be restricted to the case of both arguments are in the support of X, as the next result shows. The proof is easy and is therefore omitted.

Theorem 6. *Let X be a continuous non-negative random variable with density f and suppose that the support of X is an interval. Then, X is IPLR if and only if*

$$\frac{f(\lambda x)}{f(x)} \text{ is increasing in } x \text{ over } \Lambda(\lambda) \text{ for all } \lambda \in (0,1)$$

where

$$\Lambda(\lambda) = \{x \in \text{supp}(X) \text{ such that } \lambda x \in \text{supp}(X)\}.$$

From Theorem 4 we have the following characterization of IPLR random variables in terms of the $\leq_{plr}$ and $\leq_{lr}$ orders. □

Theorem 7. *Let X be a non-negative and absolutely continuous random variable. The following conditions are equivalent:*

a) X is IPLR.
b) $X \leq_{lr} aX, \forall a > 1$.
c) $X \leq_{plr} X$.

Now, combining Theorem 7 with the argument used in the proof of Theorem 5, we obtain the following sufficient condition for the property IPLR.

Theorem 8. *Let X be a non-negative and absolutely continuous random variable with a log-concave density function. Then, X is IPLR.*

In other words, using Definition 2 we have that

$$X \text{ is ILR} \Longrightarrow X \text{ is IPLR.}$$

The class of IPLR (DPLR) random variables can be used to characterize random variables, whose logarithms have log-concave (log-convex) densities. This is shown in the following results.

Theorem 9. *Let X be an absolutely continuous non-negative random variable with density f. Then, X is IPLR (DPLR) if and only if $f(e^x)$ is log-concave (log-convex).*

Proof. We will prove the result for the IPLR case; the DPLR case can be proven in a similar way. Denote $g(x) = f(e^x)$ and suppose that $g(x)$ is log-concave, that is,

$$g(ax+(1-a)y) \geq g(x)^a g(y)^{1-a}, \ 0 \leq a \leq 1$$

for all x and y in the domain of g or, equivalently,

$$\frac{g(y_2-x_2)\,g(y_1-x_1)}{g(y_2-x_1)\,g(y_1-x_2)} \geq 1, \quad \forall x_1 < x_2, \forall y_1 < y_2. \tag{7}$$

Let $\lambda < 1$ and select t_1 and t_2 such that $0 \leq t_1 < t_2$. By taking $y_1 = \log t_1$, $y_2 = \log t_2$, $x_1 = 0$, $x_2 = -\log \lambda$ in (7), one obtains

$$\frac{f(\lambda t_1)}{f(t_1)} \leq \frac{f(\lambda t_2)}{f(t_2)},$$

that is, X is IPLR. Conversely, assume that X is IPLR and let $a > b$. Since

$$\frac{f(at)}{f(bt)} = \frac{f\left(\frac{a}{b}bt\right)}{f(bt)} = \frac{f(\lambda t')}{f(t')}, \ \lambda = a/b < 1, t' = bt,$$

the ratio $f(at)/f(bt)$ increases in t, that is,

$$\frac{f(at_1)}{f(bt_1)} \leq \frac{f(at_2)}{f(bt_2)}, \ \forall t_1 < t_2, \forall a < b. \tag{8}$$

Now, let $x_1 < x_2$ and $y_1 < y_2$. By taking $t_1 = e^{-x_2}$, $t_2 = e^{x_1}$, $a = e^{y_1}$, $b = e^{y_2}$ in (8) we obtain (7) and the result holds. □

The next result follows from Theorem 9. The proof is obvious and it is omitted.

Corollary 1. *Let X be an absolutely continuous non-negative random variable with density f. Then, X is IPLR (DPLR) if and only if* $\log X$ *has a log-concave (log-convex) density. Equivalently, using Definition 2, we can say that*

$$X \text{ is IPLR (DPLR)} \Longleftrightarrow \log X \text{ is ILR (DLR).}$$

Let X and Y be two absolutely continuous non-negative random variables. If $X \leq_{lr} Y$, then it is not necessarily true that $X \leq_{plr} Y$. However, if one of these random variables is IPLR, then the relationship is verified (when X and Y have the same support $\mathbb{R}^+$). Since

$$\frac{f(\lambda x)}{g(x)} = \frac{f(\lambda x)}{f(x)} \frac{f(x)}{g(x)} = \frac{f(\lambda x)}{g(\lambda x)} \frac{g(\lambda x)}{g(x)}, \ \forall x, \lambda x \in \mathbb{R}^+,$$

it is easy to prove the next result.

Theorem 10. *Let X and Y be two non-negative and absolutely continuous random variables having the same support $\mathbb{R}^+$. If $X \leq_{lr} Y$ and X or Y is IPLR, then $X \leq_{plr} Y$.*

The next result yields random variables with the IPLR property by means of a simple factorization of the density function.

Theorem 11. *Let X be a continuous non-negative random variable with finite mean, the support of which is an interval. If f, the density function of X, satisfies that*

$$f(\lambda x) = A(\lambda) \cdot B(x) \cdot \exp\{C(\lambda) \cdot D(x)\}, \quad \forall \lambda \tag{9}$$

whenever $x, \lambda x \in \operatorname{supp}(X)$, where:
$A(\lambda)$ and $C(\lambda)$ are independent of x,
$B(x)$ and $D(x)$ are independent of λ,
$C(\lambda)$ decreases in λ,
$D(x)$ increases in x,
then X is IPLR.

Proof. Let λ be a positive constant, $\lambda < 1$. Consider the ratio

$$h(x, \lambda) = \frac{f(\lambda x)}{f(x)} = \frac{A(\lambda)}{A(1)} \cdot \exp\{[C(\lambda) - C(1)] D(x)\} > 0$$

then

$$\frac{\partial}{\partial x}h(x,\lambda) = [C(\lambda) - C(1)] \cdot D'(x) \cdot h(x,\lambda) > 0$$

by assumption. It follows that $h(x,\lambda)$ is increasing, that is, X is IPLR. □

Remark 1. *Note that if $C(\lambda)$ increases in λ, then X is DPLR.*

In many distributions, (9) is very easy to verify. As an example, consider the three-parameter Amoroso distribution, with density function

$$f(x) = \frac{a^p}{|s|\Gamma(p)} x^{\frac{p}{s}-1} \exp\left\{-ax^{\frac{1}{s}}\right\},\ x > 0,\ p > 0,\ a > 0,\ \frac{p}{s} \neq 0$$

Then

$$f(\lambda x) = \frac{a^p}{|s|\Gamma(p)} \cdot \lambda^{\frac{p}{s}-1} \cdot x^{\frac{p}{s}-1} \cdot \exp\left\{-a\lambda^{\frac{1}{s}}x^{\frac{1}{s}}\right\}.$$

By taking

$$A(\lambda) = \frac{a^p}{|s|\Gamma(p)}\lambda^{\frac{p}{s}-1},\ B(x) = x^{\frac{p}{s}-1},$$

$$C(\lambda) = \begin{cases} -a\lambda^{\frac{1}{s}} & if\ s > 0 \\ a\lambda^{\frac{1}{s}} & if\ s < 0 \end{cases}$$

$$D(x) = \begin{cases} x^{\frac{1}{s}} & if\ s > 0 \\ -x^{\frac{1}{s}} & if\ s < 0 \end{cases}$$

it follows that $f(x)$ satisfies the property (9).

The Amoroso family includes the standard Gamma ($\lambda = 1, s = 1$), March ($s = 1$), Vinci ($s = -1$), Weibull ($p = 1$), Exponential ($p = 1, s = 1$), Rayleigh $\left(p = 1, s = \frac{1}{2}\right)$, Chi-Square $\left(\lambda = \frac{1}{2}, p = \frac{n}{2}\right)$, Half-Normal $\left(\lambda = \frac{1}{2\sigma^2},\ p = \frac{1}{2}, s = \frac{1}{2}\right)$, and Maxwell distributions $\left(p = \frac{3}{2}, s = \frac{1}{2}\right)$.

Similarly, the Dagum type I, Singh-Maddala, Generalized Beta of second kind, Three-Parameter Generalized Gamma, Log-Gomperz and Lognormal distributions have the property of IPLR. On the other hand, the random variable X having density function

$$f(x) = e^x,\ 0 < x < \log 2$$

is an example of DPLR distribution.

4. APPLICATIONS

Several authors have studied the effects of truncation of the random variable upon the Lorenz curve. Bhattacharya (1963) showed that under certain conditions on the support of the distribution, the Lorenz curve of a left truncated income distribution is independent of the point of truncation if, and only if, the incomes follow the Pareto law. For the right truncation case, Moothathu (1991) showed that the Lorenz curve is independent of the point of truncation if, and only if, incomes follow a power distribution. For random variables with absolutely continuous distributions, Ord *et al.* (1983) obtained an ordering in the Lorenz sense of the left truncated random variables, in terms of the mean residual life. Also for absolutely continuous random variables, Belzunce *et al.* (1995) gave some conditions in terms of the proportional failure rate and the elasticity of the random variable to obtain orderings of the truncated random variables by the Lorenz order.

Consider a continuous non-negative random variable X with distribution function F and survival function $\overline{F} = 1 - F$. The left truncated random variable of X in t is

$$X_{(t,\infty)} = \{X \mid X > t\}, \quad t \in \operatorname{supp}(X),$$

and the corresponding survival function is given by

$$\overline{F}_{(t,\infty)}(x) = \begin{cases} 1 & x < t \\ \frac{\overline{F}(x)}{\overline{F}(t)} & x \geq t. \end{cases}$$

The right truncated random variable of X in t is

$$X_{(-\infty,t)} = \{X \mid X < t\}, \quad t \in \operatorname{supp}(X),$$

whose survival function is given by

$$\overline{F}_{(-\infty,t)}(x) = \begin{cases} \frac{F(t)-F(x)}{F(t)} & x \leq t \\ 0 & x > t. \end{cases}$$

Before obtaining the main results of this section, we need to state the following definition.

Definition 6. *We say that X is an increasing failure rate (IFR) random variable if $\overline{F}$ is log-concave and we say that it is a decreasing failure rate (DFR) random variable if $\overline{F}$ is log-convex on its support.*

The IFR or DFR random variables are of interest in reliability theory. It can be shown (see, for example, Bryson and Siddiqui, 1969; Barlow and Proschan, 1975; Ross, 1983) that X is IFR (DFR) if and only if

(10) $\dfrac{\overline{F}(x+c)}{\overline{F}(x)}$ is decreasing (increasing) in $x > 0$ for all $c \geq 0$.

If we let $\widetilde{F}(x) = \overline{F}(e^x)$, it follows from (10) that the random variable $\log X$ is IFR (DFR) if and only if $\widetilde{F}(x+c)/\widetilde{F}(x)$ is decreasing (increasing) in x for all $c \geq 0$. Substituting $e^x = t$ it is seen that $\log X$ is IFR (DFR) if and only if

(11) $\dfrac{\overline{F}(cx)}{\overline{F}(x)}$ is decreasing (increasing) in $x \in \text{supp}(X)$ for all $c > 1$.

Theorem 12. *Let X be a non-negative continuous random variable. If the random variable $\log X$ is IFR (DFR) then $X_{(b,\infty)} \leq_L X_{(a,\infty)}$ $(\geq_L)$ for all $a < b$, $a, b \in \text{supp}(X)$.*

Proof. We give the proof for the IFR case; the proof for the DFR case is similar.

Denote by $\mu_{(a,\infty)}$ the mean of $X_{(a,\infty)}$. It is easy to see that $\mu_{(a,\infty)} < \mu_{(b,\infty)}$ for all $a < b$ and since

$$\frac{\overline{F}\left(x\mu_{(b,\infty)}\right)}{\overline{F}\left(x\mu_{(a,\infty)}\right)} = \frac{\overline{F}\left[x\mu_{(a,\infty)}\left(\frac{\mu_{(b,\infty)}}{\mu_{(a,\infty)}}\right)\right]}{\overline{F}\left(x\mu_{(a,\infty)}\right)} =$$

$$= \frac{\overline{F}(yt)}{\overline{F}(y)},\ t = \frac{\mu_{(b,\infty)}}{\mu_{(a,\infty)}},\ y = x\mu_{(a,\infty)},$$

it follows that if $\log X$ is IFR then $\overline{F}\left(x\mu_{(b,\infty)}\right)/\overline{F}\left(x\mu_{(a,\infty)}\right)$ decreases in x for all $a < b$. Now consider the ratio

(12) $$r_{ab}(x) = \frac{\overline{F}_{(a,\infty)}\left(x\mu_{(a,\infty)}\right)}{\overline{F}_{(b,\infty)}\left(x\mu_{(b,\infty)}\right)},\ x > 0,\ a < b,\ (a, b \in \text{supp}(X))$$

where

$$\overline{F}_{(t,\infty)}(cx) = \begin{cases} 1 & \text{if} \quad x < t/c \\ \overline{F}(cx)/\overline{F}(t) & \text{if} \quad x \geq t/c \end{cases}$$

(in (12), $k/0$ is taken to be equal to ∞ whenever $k > 0$). If $\log X$ is IFR then the ratio $\overline{F}\left(x\mu_{(a,\infty)}\right)/\overline{F}\left(x\mu_{(b,\infty)}\right)$ is increasing in x and we have that $r_{ab}(x) \geq 1$ for every x whenever $b/\mu_{(b,\infty)} \leq a/\mu_{(a,\infty)}$, and $S(r_{ab}(x) - 1) \leq 1$ with the sign sequence being $-, +$, when equality holds whenever $b/\mu_{(b,\infty)} > a/\mu_{(a,\infty)}$. Therefore, $S(r_{ab}(x) - 1) \leq 1$ with sign sequence $-, +$ in the case of equality. On the other hand,

$$S(r_{ab}(x) - 1) = S\left(\overline{F}_{(a,\infty)}\left(x\mu_{(a,\infty)}\right) - \overline{F}_{(b,\infty)}\left(x\mu_{(b,\infty)}\right)\right)$$

$$= S\left(F_{(b,\infty)}\left(x\mu_{(b,\infty)}\right) - F_{(a,\infty)}\left(x\mu_{(a,\infty)}\right)\right) \text{ for all } a < b$$

and from Theorem 1 it follows that $X_{(b,\infty)} \leq_L X_{(a,\infty)}$. □

Theorem 13. *Let X be a non-negative continuous random variable. If $F(e^x)$ is log-concave (log-convex) then $X_{(0,a)} \leq_L X_{(0,b)}$ $(\geq_L)$ for all $a < b$, $a, b \in \text{supp}(X)$.*

Proof. Since $F(e^x)$ is log-concave (log-convex) if and only if

$$\frac{F(cx)}{F(x)} \text{ is decreasing (increasing) in} x > 0 \text{ for all } c \geq 0,$$

the result can be proven in the same way as the proof of Theorem 12. □

The previous results will allow us to show that the IPLR and DPLR properties (satisfied for many income distributions, as can be seen from Theorem 11) are sufficient conditions for the ordering of truncated distributions.

Corollary 2. *Let X be a non-negative and absolutely continuous random variable. If X is IPLR, then $X_{(a,\infty)} \geq_L X_{(b,\infty)}$ and $X_{(0,a)} \leq_L X_{(0,b)}$ for all $a < b$, $a, b \in \text{supp}(X)$.*

Proof. If X is IPLR, it follows from Theorem 9 that $\log X$ has a log-concave density. It is well known (see, e.g., Prekopa, 1973) that the hypothesis of logconcavity of the density function implies the logconcavity of the distribution function and the survival function. Now, the result follows by applying Theorems 12 and 13. □

The following corollary can be proven in an analogous way to the previous.

Corollary 3. *Let X be a non-negative and absolutely continuous random variable. If X is DPLR, then $X_{(a,\infty)} \leq_L X_{(b,\infty)}$ and $X_{(0,a)} \geq_L X_{(0,b)}$ for all $a < b$, $a, b \in \text{supp}(X)$.*

REFERENCES

Arnold, B. C. (1987). «Majorization and the Lorenz order: A brief introduction». *Lecture Notes in Statistics*, 43, Springer-Verlag.

Barlow, R. E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart and Winston, New York.

Belzunce, F., Candel, J. and Ruiz, J. M. (1995). «Ordering of truncated distributions through concentration curves». *Sankhyã*, 57, Series A, 375-383.

Bhattacharya, N. (1963). «A property of the Pareto distribution». *Sankhyã*, 25, Series B, 195-196.

Bryson, M. C. and Siddiqui, M. M. (1969). «Some criteria for ageing». *Journal of the American Statistical Association*, 64, 1472-1483.

Gastwirth, J. L. (1971). «A general definition of the Lorenz curve>». *Econometrica*, 39, 1037-1039.

Moothathu, T. S. K. (1986). «A characterization of power distribution through a property of the Lorenz curve». *Sankhyã*, 48, Series B, 262-265.

Ord, J. K., Patil, G. P. and Taillie, C. (1983). «Truncated distributions and measures of income inequality». *Sankhyã*, 45, Series B, 413-430.

Prekopa, A. (1973). «On logarithmic concave measures and functions». *Acta Math. Acad. Sci. Hungar*, 34, 335-343.

Ross, S. M. (1983). *Stochastic Processes*. Wiley, New York.

Shaked, M. and Shanthikumar, J. G. (1994). *Stochastic Orders and their Applications*. Academic Press, New York.

QÜESTIIÓ, vol. 25, 2, p. 225-262, 2001

MODELIZACIÓN DE DATOS LONGITUDINALES CON ESTRUCTURAS DE COVARIANZA NO ESTACIONARIAS: MODELOS DE COEFICIENTES ALEATORIOS FRENTE A MODELOS ALTERNATIVOS

VICENTE NÚÑEZ-ANTÓN*
DALE L. ZIMMERMAN**

Un tema que ha suscitado el interés de los investigadores en datos longitudinales durante las dos últimas décadas, ha sido el desarrollo y uso de modelos paramétricos explícitos para la estructura de covarianza de los datos. Sin embargo, el análisis de estructuras de covarianza no estacionarias en el contexto de datos longitudinales no se ha realizado de forma detallada principalmente debido a que las distintas aplicaciones no hacían necesario su uso. Muchos son los modelos propuestos recientemente, pero la mayoría son estacionarios de segundo orden. Algunos de éstos, sin embargo, son no estacionarios y suficientemente flexibles, de tal forma que es posible modelizar varianzas no constantes y/o correlaciones que no sean sólo función del tiempo que separa a dos observaciones dadas. Estudiaremos algunas de estas propuestas y las compararemos con los modelos de coeficientes aleatorios, evaluando sus ventajas y desventajas e indicando cuándo su uso no es apropiado o útil. Presentaremos dos ejemplos para ilustrar el ajuste de estos modelos y los compararemos entre sí, mostrando de esta forma cómo pueden modelizarse datos longitudinales de forma efectiva y simple. En estos ejemplos, los distintos modelos alternativos, especialmente los modelos antedependientes, fueron superiores a los modelos de coeficientes aleatorios.

Modelling longitudinal data with nonstationary covariance structures: random coefficients models versus alternative models

Palabras clave: Antedependencia, modelos Arima, AIC, BIC, estructuras de covarianza, máxima verosilimitud residual, modelos mixtos

Clasificación AMS (MSC 2000): 62J05, 62F10, 62P10

* Departamento de Econometría y Estadística (E.A. III). Universidad del País Vasco. Avenida Lehendakari Aguirre, 83. 48015 Bilbao. E-mail: vn@alcib.bs.ehu.es.

** Department of Statistics and Actuarial Science. The University of Iowa. Iowa City, Iowa 52242. Estados Unidos.

- Recibido en abril de 2000.
- Aceptado en enero de 2001.

1. INTRODUCCIÓN

Un tema importante en el análisis de modelos de regresión y, especialmente, en su utilización para el análisis de datos longitudinales, ha sido el desarrollo paralelo de los distintos modelos paramétricos para la matriz de varianzas y covarianzas de los datos y de los modelos de coeficientes aleatorios. Recordemos que en el análisis de datos longitudinales, se realizan mediciones a lo largo del tiempo en cada una de las unidades experimentales (normalmente asignadas a diferentes grupos o tratamientos). A este respecto, podemos consultar como referencias bibliográficas, por ejemplo, a Laird y Ware (1982), Crowder y Hand (1990, cap. 5 y 6), Jones (1993, cap. 2 y 3), Diggle, Liang y Zeger (1994, cap. 4 y 5), Wolfinger (1996), y Verbeke y Molenberghs (1997, cap. 3). Sin embargo, uno de los aspectos a resaltar debe ser que los modelos de coeficientes aleatorios se han considerado típicamente distintos de los modelos que especifican de forma paramétrica la matriz de varianzas y covarianzas de los datos. Esto es principalmente debido a que en los modelos de coeficientes aleatorios se ve el origen de la estructura de covarianzas como regresiones que varían entre los individuos o entes en el estudio, en lugar de verla como una consideración relativa a similitud de la estructura intra-individuos. En cualquier caso, en general, los modelos de coeficientes aleatorios pueden usarse y ser de mucha utilidad en el contexto de datos longitudinales.

En cuanto a los modelos que utilizan estructuras paramétricas para la matriz de varianzas y covarianzas de los datos, podemos mencionar que la modelización paramétrica de la estructura de covarianzas tiene algunas ventajas: (*i*) permite obtener de una forma más eficiente los estimadores de los parámetros usados en la modelización de la estructura de medias en los datos; (*ii*) permite obtener estimadores más adecuados para los errores estándar de los estimadores de los parámetros usados en la modelización de la estructura de medias; (*iii*) en muchos casos, permite solucionar de forma efectiva los problemas relativos a datos faltantes o datos perdidos o a datos en los que los tiempos de medición no sean los mismos para todos los individuos; y (*iv*) puede utilizarse aún cuando el número de ocasiones en que se realizan mediciones en los individuos es grande en comparación con el número de individuos.

Uno de los modelos paramétricos más utilizados para la estructura de varianzas y covarianzas de los datos es el modelo de correlación en serie. Es decir, el modelo en el que las correlaciones muestrales para un individuo determinado decrecen a medida que el tiempo de separación entre dichas observaciones aumenta. Entre estos modelos el más utilizado es el modelo estacionario autorregresivo (modelo AR) y otras versiones no muy parametrizadas de modelos estacionarios de segundo orden (véase, por ejemplo, Jennrich y Schluchter, 1986; Diggle, 1988; Jones y Boadi-Boateng, 1991; y Muñoz y otros, 1992). En estos modelos, las varianzas son constantes en el tiempo y las correlaciones entre mediciones equidistantes en el tiempo son las mismas. Cuando éste no sea el caso, el uso de los modelos estacionarios no es aconsejable. Es decir, se deben considerar otros modelos capaces de modelizar esta no estacionariedad.

Si la no estacionariedad se da en las varianzas, se puede optar por transformar los datos para estabilizar las varianzas o por utilizar modelos que permitan tener varianzas heterogéneas (véase, por ejemplo, Wolfinger, 1996). Sin embargo, estas alternativas pueden llegar a no resolver el problema si los datos, además, presentan una no estacionariedad en correlación. A pesar de esto, existen modelos alternativos, aplicables a datos longitudinales, en los que se permite que exista no estacionariedad en varianza y en correlación. Posiblemente el más obvio de ellos es el modelo clásico multivariante completamente no estructurado. En muchos casos, sin embargo, la no estacionariedad de los datos tiene una estructura que puede modelizarse usando un modelo con pocos parámetros e ignorar esto puede, obviamente, eliminar todas las ventajas que hemos mencionado anteriormente sobre la modelización paramétrica de la estructura de covarianzas. Una familia de modelos paramétricos que puede permitir correlaciones no estacionarias es una generalización de los modelos AR, conocida como la familia de los modelos antedependientes: no estructurados (Gabriel, 1962; Kenward, 1987) y estructurados (Núñez Antón y Woodworth, 1994; Núñez Antón, 1997; Zimmerman y Núñez Antón, 1997). Otra generalización no estacionaria de los modelos AR, son los modelos autorregresivos integrados de medias móviles (modelos ARIMA) (véase, por ejemplo, Diggle, 1990). Una posibilidad final, de la que ya hemos hablado brevemente, son los modelos de coeficientes aleatorios. Nuevamente indicamos que a pesar de que estos modelos son típicamente modelos para regresiones de una variable de respuesta en el tiempo (o, posiblemente, en otras covariables) que varía de individuo a individuo, un modelo de coeficientes aleatorios puede usarse para modelizar ciertos tipos de no estacionariedad (véase, por ejemplo, Diggle, Liang y Zeger, 1994).

Cada uno de los modelos descritos anteriormente, además de los modelos estacionarios, han sido propuestos para su uso en contextos de datos longitudinales por al menos un autor, pero nunca han sido comparados ni evaluados en conjunto, menos aún ante la presencia de no estacionariedad en varianza y correlación. La práctica general suele ser que, ante un ajuste erróneo de los modelos, presumiblemente debido a la presencia de no estacionariedad, un modelo de coeficientes aleatorios sería el adecuado y el ajustado (véase, por ejemplo, Jones, 1990). Así, ante la falta de esta comparación entre los distintos modelos en datos reales, intentaremos comparar tanto los modelos estacionarios como los no estacionarios con el más utilizado de estos últimos, el de coeficientes aleatorios, en situaciones de no estacionariedad.

En este trabajo, consideraremos las ventajas y limitaciones de cada uno de los modelos, enfatizando los casos en que sus usos sean inadecuados. Para ello, presentaremos dos ejemplos que ilustrarán el ajuste y la comparación entre los modelos alternativos y el de efectos aleatorios. Los ejemplos demostrarán que es posible modelizar datos longitudinales no estacionarios de forma efectiva y, en algunos casos, utilizando modelos paramétricos estructurados. En la Sección 2 describimos los dos estudios longitudinales que analizaremos posteriormente en la Sección 5. Las hipótesis y notación que utilizaremos se mencionan en la Sección 3. En la Sección 4 realizamos una breve descripción

de los distintos modelos paramétricos para la estructura de covarianzas intra-individuos y de los aspectos computacionales para el ajuste de los mismos. Finalmente, en la Sección 6, evaluamos las distintas propuestas del trabajo y establecemos las conclusiones del mismo.

2. DATOS

Utilizaremos datos provenientes de dos estudios longitudinales para motivar la consideración de modelos no estacionarios distintos al de coeficientes aleatorios. Estos datos también los usaremos, en la Sección 5, para ilustrar el ajuste y comparación de los distintos modelos entre ellos y con el de coeficientes aleatorios.

Los datos que denominaremos *race data*, que amablemente nos ha dejado utilizar Ian Jollife de la Universidad de Kent, corresponden a cada uno de los tiempos parciales (en minutos) para cada uno de los 80 competidores en cada una de las secciones de 10 kilómetros de una carrera con un total de 100 kilómetros, que se llevó a cabo en el Reino Unido en 1984. Además de los tiempos parciales, los datos contienen la edad de 76 de los 80 competidores.

Tabla 1. Medias muestrales (en minutos), varianzas y correlaciones muestrales correspondientes a los datos del estudio longitudinal *race data*.

Sección (t)	1	2	3	4	5	6	7	8	9	10
Corr.:										
	1.0									
	.95	1.0								
	.84	.89	1.0							
	.78	.82	.92	1.0						
	.60	.63	.75	.88	1.0					
	.60	.62	.72	.84	.94	1.0				
	.52	.54	.60	.69	.75	.84	1.0			
	.45	.48	.61	.69	.78	.84	.78	1.0		
	.51	.51	.56	.65	.73	.77	.69	.75	1.0	
	.38	.40	.44	.49	.52	.64	.72	.65	.77	1.0
Medias:	47.8	50.9	49.6	53.2	54.7	60.1	62.4	69.3	68.7	67.4
Var.:	26.9	34.8	49.0	58.9	91.4	149.9	107.9	152.2	145.0	167.2

Los datos que denominaremos *cattle data* corresponden a un experimento descrito por Kenward (1987). Un grupo de vacas fue sometido a uno de los dos posibles tratamientos para parásitos intestinales, a los que denominaremos A y B. Las vacas fueron pesadas 11 veces en un periodo de tiempo de 133 días. Treinta vacas recibieron cada uno de los tratamientos. Las primeras 10 mediciones se realizaron cada dos semanas, mientras que la última medición se realizó una semana después de la décima.

Tabla 2. Medias (en Kg.), varianzas y correlaciones muestrales correspondiente a los datos del estudio longitudinal *cattle data*, tratamiento A.

Tiempo (en días)	0	14	28	42	56	70	84	98	112	126	133
Corr.:											
	1.0										
	.82	1.0									
	.76	.91	1.0								
	.66	.84	.93	1.0							
	.64	.80	.88	.94	1.0						
	.59	.74	.85	.91	.94	1.0					
	.52	.63	.75	.83	.87	.93	1.0				
	.53	.67	.77	.84	.89	.94	.93	1.0			
	.52	.60	.71	.77	.84	.90	.93	.97	1.0		
	.48	.58	.70	.73	.80	.87	.88	.94	.96	1.0	
	.48	.55	.68	.71	.77	.83	.86	.92	.96	.98	1.0
Medias:	226.2	230.3	246.9	265.6	281.2	294.9	304.7	312.9	315.1	324.1	325.5
Var.:	105.6	155.1	165.2	184.9	243.0	283.8	306.6	340.7	389.2	470.1	444.6

Las Tablas 1, 2 y 3 muestran las medias, varianzas y correlaciones muestrales correspondientes a cada uno de los dos conjuntos de datos, mientras que las Figuras 1 y 2 muestran la evolución de la media muestral en el tiempo para cada uno de los estudios longitudinales. En el caso del *cattle data*, el contraste de homogeneidad indicó que no era razonable utilizar la misma matriz de varianzas y covarianzas para los dos grupos de tratamientos. Las medias muestrales correspondientes al *race data* (ver Tabla 1 y Figura 1) indican que los tiempos tienden a incrementarse hasta los primeros 80 kilómetros, pero luego decrecen de forma leve en los últimos 20 kilómetros. Esto indica que un modelo con estructura lineal en t no sería el adecuado y que habría que intentar ajustar un modelo cuadrático o cúbico en t. En el sentido práctico, estos resultados indican que, a medida que la carrera avanza, los corredores se cansan más y, por tanto, tardarán cada vez mas en cubrir cada tramo de diez kilómetros.

Tabla 3. Medias (en Kg.), varianzas y correlaciones muestrales correspondiente a los datos del estudio longitudinal *cattle data*, tratamiento B.

Tiempo (en días)	0	14	28	42	56	70	84	98	112	126	133
Corr.:											
	1.0										
	.86	1.0									
	.83	.94	1.0								
	.68	.89	.93	1.0							
	.67	.84	.88	.95	1.0						
	.66	.84	.87	.95	.98	1.0					
	.61	.78	.82	.91	.94	.97	1.0				
	.63	.81	.84	.92	.92	.95	.95	1.0			
	.63	.80	.79	.90	.93	.95	.93	.96	1.0		
	.48	.65	.67	.78	.78	.82	.76	.78	.83	1.0	
	.44	.57	.62	.73	.68	.74	.71	.71	.75	.92	1.0
Medias:	224.6	227.9	243.5	262.5	276.4	290.1	299.2	317.7	319.7	326.9	320.5
Var.:	105.3	108.4	147.1	198.5	217.7	250.4	248.2	234.1	287.0	404.7	598.6

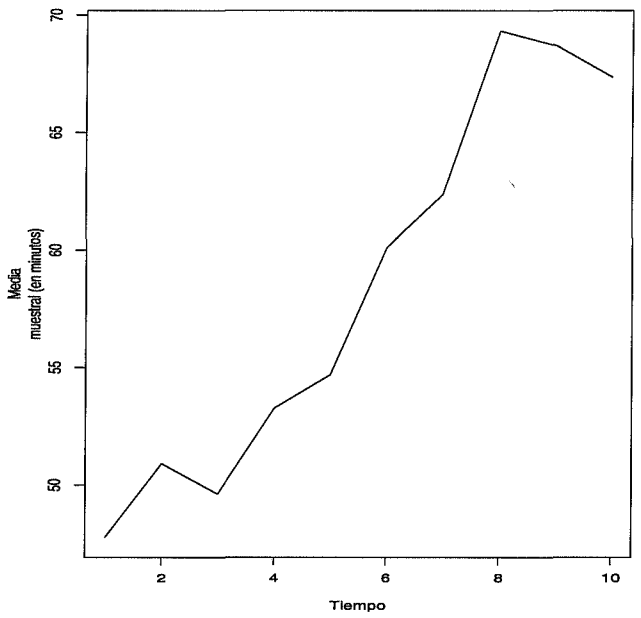


Figura 1. Medias muestrales para el estudio longitudinal *race data* en función de la sección de la carrera (tiempo).

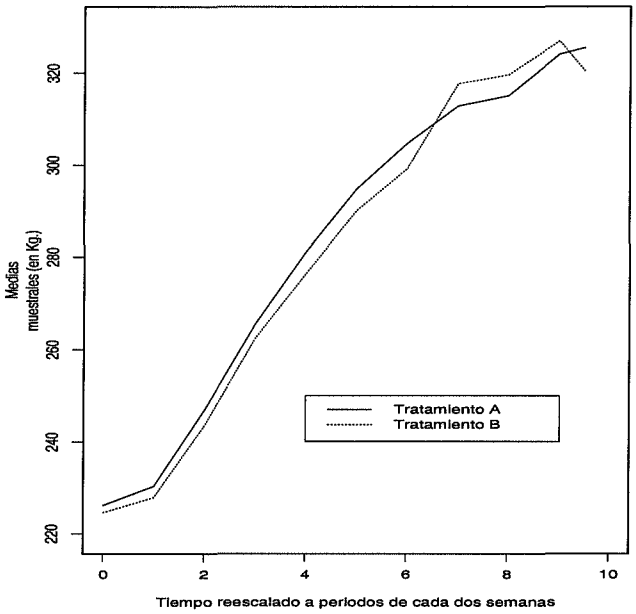


Figura 2. Medias muestrales para el estudio longitudinal *cattle data*, grupos A y B, en función del tiempo reescalado a periodos de cada dos semanas (tiempo=t/14).

En el caso del estudio *cattle data* (ver Tablas 2 y 3 y Figura 2), se observa que las medias correspondientes a ambos tratamientos tienden a incrementarse con el tiempo, aunque no linealmente. Las medias del grupo correspondiente al tratamiento A son ligeramente superiores hasta aproximadamente la séptima medición ($t = 6$, que corresponde a una medición efectuada a las 12 semanas o a los 84 días, en la Figura 2). A partir de ese instante, las medias muestrales del tratamiento B superan a las del tratamiento A, aunque al final del estudio, a partir de la medición efectuada a las 18 semanas, la situación vuelve a invertirse. Nuevamente, estas tendencias indican que las vacas tienden a ganar peso a medida que el tratamiento al que se les somete empieza a hacer efecto o se empieza a observar sus resultados.

Estas matrices muestran algunas características interesantes, comunes a muchos datos en el coxtexto de datos longitudinales:

1. Las varianzas no son homogéneas y tienden a incrementarse con el tiempo (es decir, en el caso del *race data* a medida que la carrera avanza). Este incremento es bastante monótono para el *cattle data* y no tanto para el caso del *race data*. Una forma de interpretar esta característica para el caso del *race data* es pensar que a medida que la carrera avanza, los corredores tenderán a diferenciarse cada vez más unos de

otros en cuanto al tiempo necesario para completar cada tramo de la misma, lo que incrementará la varianza a medida que la carrera avanza. En el caso del *cattle data* un razonamiento similar, pero aplicado a los pesos de las vacas y su variabilidad a medida que el estudio avanza, permite comprender este incremento de varianza con el tiempo.

2. Las correlaciones son todas positivas.
3. Existe correlación en serie dado que las correlaciones dentro de una columna específica tienden a decrecer hacia cero (a menos que estén cercanas a cero inicialmente).
4. Las correlaciones entre observaciones separadas por la misma distancia (es decir, subdiagonales o superdiagonales de la matriz) no son constantes. Por el contrario, en el caso del tratamiento A del *cattle data* tienden a incrementarse al principio del estudio antes de nivelarse o, en algunos casos (por ejemplo, tratamiento B para el *cattle data*) decrecer levemente al final del estudio. Por otro lado, en el caso del *race data*, son menores al final del estudio que al principio del mismo. Una interpretación práctica de este hecho en el caso del *race data* podría ser que los tiempos utilizados en una sección de la carrera serán predictores más fiables de los tiempos en las secciones siguientes al principio de la carrera que al final de la misma.

Para estabilizar las varianzas, transformamos las respuestas para cada uno de los datos utilizando diversas alternativas, sin conseguir el resultado esperado. Esto era previsible dado que en estos casos las varianzas no parecían ser funciones suaves de la media. Aún en los casos en que la transformación estabilizaba las varianzas, la no estacionariedad de las correlaciones permanecía presente en los mismos.

3. EL MODELO GENERAL: HIPÓTESIS Y NOTACIÓN

Supondremos que estamos en el contexto de datos longitudinales. Es decir, se realizan mediciones a lo largo del tiempo en cada uno de los m individuos o entes bajo estudio. Sea $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ el vector de las n_i mediciones que se han realizado en el i-ésimo individuo y sea $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})'$ el vector que contiene los tiempos correspondientes en los que se han realizado dichas mediciones. Además suponemos que observamos un vector p-variante de covariables, $\mathbf{x}_{ij}$, asociado con y_{ij}. Así utilizando una notación matricial más compacta, sean $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_m')'$, $\mathbf{t} = (\mathbf{t}_1', \ldots, \mathbf{t}_m')'$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})'$, $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_m')'$, y $N = \sum_{i=1}^{m} n_i$.

Denominaremos diseño de medición al conjunto de los tiempos en que las mediciones se han realizado. No imponemos ninguna restricción en el diseño de medición. En general, los tiempos en los que se realizan las mediciones para un mismo individuo pueden estar espaciados de forma irregular y pueden ser distintos para los diferentes individuos. Si los tiempos de medición son los mismos para todos los individuos, entonces tendre-

mos un diseño de medición rectangular, como es el caso en los dos ejemplos descritos en la Sección 2. Analizaremos, posteriormente, la influencia de tener o no tener este tipo de diseño en la modelización de la estructura de covarianzas.

Las hipótesis generales que tendremos presentes en nuestra modelización paramétrica de estructuras de covarianzas son:

1. Las mediciones realizadas en individuos distintos son independientes entre sí, aunque las mediciones realizadas en un mismo individuo podrán ser dependientes.
2. La respuesta media (para los individuos) es una combinación lineal de funciones conocidas de las covariables y/o del tiempo.
3. Las respuestas están normalmente distribuidas (posiblemente después de hacer alguna transformación).
4. La naturaleza de la variación de segundo orden intra-individuos es la misma para todos los individuos.
5. Si hay datos perdidos, éstos se asumen ignorables (Laird, 1988).

Estas hipótesis nos dan el modelo $\mathbf{y} \sim MVN\,(\mathbf{X}\beta, \Sigma)$, donde β es un vector p-dimensional de parámetros fijos, desconocidos y, típicamente, no restringidos y Σ es la matriz de covarianzas desconocida de $N \times N$, diagonal en bloques, con elementos no nulos en los bloques respectivos dados por $\mathbf{V}_i = \text{var}(\mathbf{y}_i)$. En el caso de diseño de medición rectangular, las $\mathbf{V}_i$ son todas iguales.

A continuación, para tener una modelización paramétrica de la estructura de covarianzas, tendremos que establecer un modelo para Σ:

$$\mathbf{y} \sim MVN\,(\mathbf{X}\beta, \Sigma(\mathbf{t}, \theta)), \tag{1}$$

donde θ es un vector q-dimensional de parámetros desconocidos, restringido a un espacio de parámetros Θ que es o el conjunto de todos los vectores θ para los que Σ es definida positiva o algún subconjunto de este conjunto. Los elementos de Σ pueden ser funciones de $\mathbf{t}$, pero no de $\mathbf{X}$. Además, Σ es definida positiva si y sólo si $\mathbf{V}_i$ es definida positiva para cada i. La estimación de los parámetros β y θ en este modelo se realiza utilizando el método de máxima verosimilitud o el de máxima verosimilitud residual (MVR), que a menudo debe implementarse utilizando subrutinas de optimización numérica. Los detalles adicionales y más específicos sobre cómo se ajustan los distintos modelos y la comparación entre los mismos se mencionarán en la Sección 5.

A continuación describiremos la mayoría de los modelos que se utilizan frecuentemente para la modelización de la estructura de covarianzas en el contexto de datos longitudinales. Estos modelos serán posteriormente ajustados y comparados con el modelo de coeficientes aleatorios. En cada caso, mencionaremos sus propiedades principales, ventajas y limitaciones. Dado que los modelos se utilizan para la estructura de covarianzas

intra-individuos, eliminaremos, al menos inicialmente, el subíndice i (denotando al individuo) en y_{ij}, n_i, x_{ij}, y $\mathbf{V}_i$. Además, $\{v_{ju}\}$ denotará a los elementos de $\mathbf{V}$ y $\{\rho_{ju}\}$ a los elementos correspondiente de la matriz de correlaciones.

4. MODELOS PARAMÉTRICOS PARA LA ESTRUCTURA DE COVARIANZAS INTRA-INDIVIDUOS

En esta sección describimos los modelos más utilizados para la estructura de covarianzas intra-individuos: (*i*) de coeficientes aleatorios (CA), (*ii*) de simetría compuesta (SC), (*iii*) autorregresivo de primer orden (AR(1)), (*iv*) Huynh-Feldt (HF), (*v*) Toeplitz (TOEP y TOEPH), (*vi*) no estructurado (NE), (*vii*) antedependientes no estructurados (AD), (*viii*) antedependientes estructurados (ADE) y (*ix*) ARIMA. En la descripción de cada uno de estos modelos, ilustraremos la estructura del modelo para el caso en que todos los individuos tengan $n_i = n = 4$ $(i = 1,\ldots,m)$ observaciones igualmente espaciadas en los tiempos $t = 1,2,3,4$. A continuación realizamos una descripción de los modelos y luego, al final de esta sección, mencionaremos algunos aspectos computacionales básicos que indican la forma de ajustar estos modelos.

4.1. Modelo de Coeficientes Aleatorios (CA)

Los modelos de coeficientes aleatorios fueron introducidos por Rao (1959) y su popularidad ha ido en aumento debido a su interpretabilidad intuitiva y su relación cercana con las técnicas Bayesianas (véase, por ejemplo, Laird y Ware, 1982; Reinsel, 1982; o Rutter y Elashoff, 1994).

El modelo general de coeficientes aleatorios viene dado por la ecuación

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i \qquad (i = 1,\ldots,m),$$

donde las $\mathbf{Z}_i$'s representan matrices conocidas, cuya estructura describiremos en breve, los $\mathbf{u}_i$ son vectores de coeficientes aleatorios distribuidos, independientes entre sí, como $MVN(\mathbf{0},\mathbf{G}_i)$, las $\mathbf{G}_i$'s son matrices definidas positivas y, aparte de esto, matrices no restringidas, y los $\mathbf{e}_i$'s están distribuidos, independientes de los $\mathbf{u}_i$'s y entre sí, como $MVN(\mathbf{0},\sigma^2\mathbf{I}_{n_i})$. En general, las $\mathbf{G}_i$ son las mismas para todos los individuos y, por tanto, la matriz de covarianzas de $\mathbf{y}_i$ será $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \sigma^2\mathbf{I}_{n_i}$. Entre los casos especiales tenemos el modelo de coeficientes aleatorios lineales (CAL) y el modelo de coeficientes aleatorios cuadráticos (CAC). En el caso cuadrático, $\mathbf{Z}_i = [\mathbf{1}_{n_i}, \mathbf{t}_i, (t_{i1}^2, t_{i2}^2, \ldots, t_{in_i}^2)']$, y

$$\mathbf{G} = \begin{pmatrix} \sigma_{00} & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_{11} & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_{22} \end{pmatrix}$$

En el caso lineal, $\mathbf{Z}_i = [\mathbf{1}_{n_i}, \mathbf{t}_i]$. Otro caso especial que se obtiene si $\mathbf{Z}_i = \mathbf{1}_{n_i}$, es equivalente al comúnmente utilizado modelo de simetría compuesta.

Los modelos de coeficientes aleatorios se han considerado típicamente distintos de los modelos que especifican de forma paramétrica la matriz de varianzas y covarianzas de los datos. Esto es principalmente debido a que en los modelos de coeficientes aleatorios se ve el origen de la estructura de covarianzas como regresiones que varían entre los individuos o entes en el estudio, en lugar de verla como una consideración relativa a similitud de la estructura intra-individuos. En cualquier caso, los modelos de coeficientes aleatorios pueden, en general, usarse y ser de mucha utilidad en el contexto de datos longitudinales para modelizar estructuras que presentan varianzas no constantes y correlaciones no estacionarias, con la excepción del modelo de simetría compuesta. Este es un hecho muy poco conocido y, obviamente, poco utilizado. Consideremos, por ejemplo, la estructura CAL para un individuo en el que se han realizado mediciones en tiempos igualmente espaciados, digamos $t_1 = 1, \ldots, t_n = n$. Para este modelo, tendremos que $\text{var}(y_{ij}) = \sigma^2 + \sigma_{00} + 2\sigma_{01}j + \sigma_{11}j^2$ y

$$\text{corr}(y_{ij}, y_{ik}) = \frac{\sigma_{00} + \sigma_{01}(j+k) + \sigma_{11}jk}{\sqrt{\sigma^2 + \sigma_{00} + 2\sigma_{01}j + \sigma_{11}j^2}\sqrt{\sigma^2 + \sigma_{00} + 2\sigma_{01}k + \sigma_{11}k^2}}$$

Como puede observarse, éste es un modelo muy flexible que permite modelizar diversos tipos de comportamientos en las varianzas y correlaciones, entre los que se incluyen varianzas que crecen o decrecen, así como correlaciones que pueden ser negativas o positivas. Sin embargo, no permite que la varianza sea una función cóncava hacia abajo del tiempo o que la varianza sea constante si las correlaciones entre observaciones igualmente espaciadas no lo son. Además, el número de parámetros en estos modelos no está relacionado con el número de ocasiones en las que se realizan mediciones.

4.2. Modelos de Simetría Compuesta (SC) y Autorregresivos de Primer Orden (AR(1))

Estos dos modelos paramétricos para la estructura de covarianzas se consideran modelos homogéneos. Es decir, la varianza a lo largo de la diagonal principal de esta matriz permanece constante. Sin embargo, difieren en el trato de la covarianza. Para el modelo SC, ésta permanece constante y para el modelo AR(1), ésta decrece exponencialmente. Las matrices de varianzas y covarianzas para los modelos SC y AR(1) son

$$v\begin{bmatrix} 1 & \rho & \rho & \rho \\ & 1 & \rho & \rho \\ & & 1 & \rho \\ & & & 1 \end{bmatrix} \quad \text{y} \quad v\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{bmatrix},$$

respectivamente, en donde $v = \text{Var}(y_{ij})$. Como puede verse, estos modelos tienen dos parámetros cada uno. Es decir, son modelos muy restringidos que deben utilizarse sólo

en los casos en que se tenga correlaciones constantes o que decrezcan exponencialmente, además de varianzas constantes. Sin embargo, es posible obtener una generalización directa de los modelos SC y AR(1) permitiendo que las varianzas en la diagonal principal de las matrices de covarianzas sean distintas. Denominaremos a estos modelos, que son una extensión heterogénea de los anteriores, SCH y ARH(1), respectivamente. Las matrices de varianzas y covarianzas para estos modelos son, por tanto,

$$\begin{bmatrix} v_{11} & \sqrt{v_{11}v_{22}}\rho & \sqrt{v_{11}v_{33}}\rho & \sqrt{v_{11}v_{44}}\rho \\ & v_{22} & \sqrt{v_{22}v_{33}}\rho & \sqrt{v_{22}v_{44}}\rho \\ & & v_{33} & \sqrt{v_{33}v_{44}}\rho \\ & & & v_{44} \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} v_{11} & \sqrt{v_{11}v_{22}}\rho & \sqrt{v_{11}v_{33}}\rho^2 & \sqrt{v_{11}v_{44}}\rho^3 \\ & v_{22} & \sqrt{v_{22}v_{33}}\rho & \sqrt{v_{22}v_{44}}\rho^2 \\ & & v_{33} & \sqrt{v_{33}v_{44}}\rho \\ & & & v_{44} \end{bmatrix},$$

respectivamente. Como puede verse, estos modelos tienen cinco parámetros cada uno. Son modelos menos restringidos que sus versiones homogéneas en varianza, pero siguen siendo válidos sólo para los casos de correlaciones constantes o exponencialmente decrecientes.

4.3. Modelos Huynh-Feldt (HF) y Toeplitz (TOEP y TOEPH)

Los contrastes de esfericidad de Huynh y Feldt (1970) han sido muy relevantes para decidir si optar por un análisis univariante o multivariante para datos de medidas repetidas o longitudinales. Sin embargo, pocas veces hemos visto que la estructura de covarianzas Huynh-Feldt haya sido ajustada directamente a los datos principalmente debido a que los análisis estándar utilizan el hecho de que un conjunto de contrastes ortogonales de datos con esta estructura de covarianzas tienen de por si una estructura de covarianzas diagonal. El modelo HF es similar al modelo SCH, tanto en el número de parámetros como en el hecho de poseer una heterogeneidad no estructurada para las varianzas en la diagonal principal. Sin embargo, la estructura HF construye los elementos fuera de la diagonal principal utilizando medias aritméticas en lugar de medias geométricas. Así, la estructura de varianzas y covarianzas Huynh-Feldt (HF), que tiene 5 parámetros, puede expresarse como:

$$\begin{bmatrix} v_{11} & (v_{11}+v_{22})/2-\lambda & (v_{11}+v_{33})/2-\lambda & (v_{11}+v_{44})/2-\lambda \\ & v_{22} & (v_{22}+v_{33})/2-\lambda & (v_{22}+v_{44})/2-\lambda \\ & & v_{33} & (v_{33}+v_{44})/2-\lambda \\ & & & v_{44} \end{bmatrix}$$

Por otro lado, Los modelos TOEP y TOEPH generalizan, respectivamente, a los modelos AR(1) y ARH(1). En estos modelos no se asume que las correlaciones decrezcan de forma exponencial y, por el contrario, se las deja variar de forma no estructurada. Sin embargo, seguimos asumiendo que las correlaciones entre observaciones igualmente espaciadas son las mismas. Estos modelos tienen 4 y 7 parámetros y sus matrices de varianzas y covarianzas son

$$v\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ & 1 & \rho_1 & \rho_2 \\ & & 1 & \rho_1 \\ & & & 1 \end{bmatrix} \quad y \quad \begin{bmatrix} v_{11} & \sqrt{v_{11}v_{22}}\rho_1 & \sqrt{v_{11}v_{33}}\rho_2 & \sqrt{v_{11}v_{44}}\rho_3 \\ & v_{22} & \sqrt{v_{22}v_{33}}\rho_1 & \sqrt{v_{22}v_{44}}\rho_2 \\ & & v_{33} & \sqrt{v_{33}v_{44}}\rho_1 \\ & & & v_{44} \end{bmatrix},$$

respectivamente, en donde $v = \text{Var}(y_{ij})$.

4.4. Modelo No Estructurado (NE)

El modelo no estructurado de covarianzas es el caso extremo de modelización paramétrica de estructuras de covarianza en el que θ tiene $n(n+1)/2$ varianzas y covarianzas, o equivalentemente $n(n+1)/2$ varianzas y correlaciones. Dado que es un modelo completamente general para la estructura de covarianzas intra-individuo, no requiere que las observaciones se encuentren igualmente espaciadas entre sí. Para este modelo, el espacio de parámetros $\Theta = \{\theta\text{: } \mathbf{V} \text{ es definida positiva}\}$ no puede ser expresado como restricciones de desigualdades lineales en cada uno de los parámetros, lo que origina que estas restricciones sean muy difíciles de cumplir. El modelo NE tiene 10 parámetros y puede expresarse como:

$$\begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} \\ & v_{22} & v_{23} & v_{24} \\ & & v_{33} & v_{34} \\ & & & v_{44} \end{bmatrix}$$

4.5. Modelos Antedependientes no Estructurados (AD)

Las observaciones normales multivariantes $y_1, \ldots, y_n$ son antedependientes de orden s [AD(s)] si y_j e y_{j+k+1}, condicionadas a las observaciones intermedias $y_{j+1}, \ldots, y_{j+k}$, son independientes, para todo $j = 1, \ldots, n-k-1$ y todo $k \geq s$. La definición original de antedependencia fue propuesta por Gabriel (1962), aunque en la misma nunca se hizo referencia alguna a una estructura paramétrica. Una definición equivalente para el modelo AD(s), pero especificada de forma paramétrica, puede expresarse con las ecuaciones:

$$\begin{aligned} y_1 &= \mathbf{x}_1'\beta + \varepsilon_1, \\ y_j &= \mathbf{x}_j'\beta + \sum_{k=1}^{s^*} \phi_{jk}(y_{j-k} - \mathbf{x}_{j-k}'\beta) + \varepsilon_j \quad (j = 2, \ldots, n) \end{aligned} \tag{2}$$

donde $s^* = \min(s, j-1)$, los ε_j's son variables aleatorias normales independientes con media cero y varianzas, $\sigma_j^2 > 0$, que puede depender del tiempo. Los ϕ_{jk}'s son parámetros no estructurados. Es importante indicar que σ_j^2 y v_{jj} no son las mismas cantidades. De la especificación paramétrica en (2), podemos observar cómo los modelos AD generalizan a los modelos estacionarios AR: tal y como ocurre en los modelos AR, los modelos AD permiten la existencia de correlación en serie en las observaciones realizadas en cada individuo pero, a diferencia de los modelos AR, los modelos AD no exigen que las varianzas sean constantes o que las correlaciones entre observaciones igualmente espaciadas en el tiempo sean las mismas. En resumen, podríamos decir que el modelo AR(s) es un caso especial del modelo AD(s) en el que: (a) $\phi_{jk} \equiv \phi_k$ para $j = s+1, \ldots, n$ y $k = 1, \ldots, s$; (b) las s raíces de la ecuación característica del modelo AR(s), $1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_s x^s = 0$, todas son mayores que uno en valor absoluto; (c) $\sigma_{s+1}^2 = \sigma_{s+2}^2 = \cdots = \sigma_n^2 > 0$; y (d) los «valores iniciales» $\{\phi_{jk}\colon\ j = 2, \ldots, s;\ k = 1, \ldots, j-1\}$ y $\sigma_1^2, \sigma_2^2, \ldots, \sigma_s^2$ se eligen de una forma adecuada. Además, mencionaremos que el modelo AD($n-1$) es equivalente al modelo NE y que el modelo AD(0) es equivalente al modelo de independencia heterogéneo (es decir, $\mathbf{V} = \text{diag}(v_{11}, \ldots, v_{nn})$). Por ejemplo, podemos escribir la matriz de varianzas y covarianzas para el modelo AD(1), que tiene 7 parámetros.

$$\begin{bmatrix} v_{11} & \sqrt{v_{11}v_{22}}\rho_1 & \sqrt{v_{11}v_{33}}\rho_1\rho_2 & \sqrt{v_{11}v_{44}}\rho_1\rho_2\rho_3 \\ & v_{22} & \sqrt{v_{22}v_{33}}\rho_2 & \sqrt{v_{22}v_{44}}\rho_2\rho_3 \\ & & v_{33} & \sqrt{v_{33}v_{44}}\rho_3 \\ & & & v_{44} \end{bmatrix}$$

Nos referiremos al modelo (2) como un modelo AD no estructurado de orden s [AD(s)], donde al decir «no estructurado» nos referimos al hecho de que los parámetros ϕ_{jk} y σ_j^2 no se pueden expresar como funciones de un número menor de parámetros. Un modelo de covarianzas AD(s) tiene $(s+1)(2n-s)/2$ parámetros. Gabriel (1962) además propuso un contraste de razón de verosimilitudes que permite determinar el orden adecuado de antedependencia para un conjunto de datos específicos.

La ecuación (2) es una especificación *autorregresiva* de un modelo AD(s), es decir una especificación paramétrica en función de los coeficientes autorregresivos $\{\phi_{jk}\colon\ j = 2\ldots, n;\ k = 1, \ldots, s^*\}$ y la varianzas específicas $\{\sigma_j^2\colon\ j = 1, \ldots, n\}$. Una segunda forma equivalente de especificar (o reparametrizar) el modelo AD(s) se denomina especificación *de covarianza*, la que es una especificación paramétrica en función de las varianzas y covarianzas (o correlaciones) de las observaciones. Es fácilmente demostrable que las varianzas y las covarianzas en las primeras s subdiagonales (or superdiagonales) de la matriz de covarianzas de un modelo AD(s) permanecen no estructuradas, mientras que las restantes covarianzas se encuentran completamente determinadas por estos valores. Por ejemplo, en el caso de un modelo de primer orden, la matriz $\mathbf{V}$ puede escribirse de la siguiente forma:

$$
(3)\qquad
\begin{bmatrix}
v_{11} & & & & \\
\sqrt{v_{11}v_{22}}\rho_1 & v_{22} & & & \\
\sqrt{v_{11}v_{33}}\rho_1\rho_2 & \sqrt{v_{22}v_{33}}\rho_2 & v_{33} & & \\
\sqrt{v_{11}v_{44}}\rho_1\rho_2\rho_3 & \sqrt{v_{22}v_{44}}\rho_2\rho_3 & \sqrt{v_{33}v_{44}}\rho_3 & \ddots & \\
\vdots & \vdots & \vdots & \ddots & v_{n-1,n-1} \\
\sqrt{v_{11}v_{nn}}\prod_{j=1}^{n-1}\rho_j & \sqrt{v_{22}v_{nn}}\prod_{j=2}^{n-1}\rho_j & \cdots & \cdots & \sqrt{v_{n-1,n-1}v_{nn}}\rho_{n-1} \quad v_{nn}
\end{bmatrix}
$$

donde $\rho_j \equiv \rho_{j,j+1}$. En el caso en que $s > 1$, las correlaciones en las subdiagonales $s+1,\ldots,n-1$ de $\mathbf{V}$ dependerán funcionalmente de las correlaciones en las primeras s subdiagonales, pero esta dependencia es mucho más complicada que la estructura multiplicativa evidente en (3).

Una tercera especificación posible para un modelo AD(s) se denomina la especificación de *concentración*. En este caso, se especifican los elementos en la diagonal principal y las primeras s subdiagonales (y superdiagonales) de la matriz que, comúnmente, se denomina matriz de concentración $\mathbf{V}^{-1}$. Los elementos restantes de $\mathbf{V}^{-1}$ son cero (Gabriel, 1962, Núñez-Antón y otros, 1995, Zimmerman y otros, 1998). Se pueden encontrar más detalles sobre los tres tipos de especificación de estos modelos en Zimmerman y Núñez-Antón (1997).

En general no es posible establecer el espacio de parámetros en las especificaciones de covarianza y de concentración en función de restricciones simples sobre cada uno de los parámetros. Una excepción es el caso de especificación de covarianza para el modelo AD(1) en (3), para el que $\Theta = \{\theta\colon -1 < \rho_j < 1 \text{ para } j = 1,\ldots,n-1;\ v_{jj} > 0 \text{ para } j = 1,\ldots,n\}$. Es decir, con la excepción del modelo de primer orden, la especificación autorregresiva, que no impone restricciones en los coeficientes autorregresivos, es la más conveniente desde el punto de vista de imponer restricciones en los parámetros. De hecho, la especificación autorregresiva puede incluso evitar imponer todo tipo de restricciones en los parámetros si se reparametriza las varianzas específicas de forma que $\gamma_j = \log\sigma_j^2$. Cualquier especificación del modelo AD(s) se puede utilizar sin importar el espaciamiento temporal entre las mediciones.

4.6. Modelos Antedependientes Estructurados (ADE)

Aunque sabemos que los modelos AD tienen menos parámetros que el modelo NE, aún pueden tener demasiados parámetros en situaciones reales. Por ejemplo, si $n = 10$ y tenemos un diseño de medición rectangular, el modelo NE tiene 45 parámetros, mientras que los modelos AD(1) y AD(2) tienen 19 y 27 parámetros, respectivamente. De hecho, el modelo NE tiene $O(n^2)$ parámetros, mientras que los modelos AD tienen $O(n)$ parámetros y los modelos estacionarios más utilizados sólo tienen $O(1)$ parámetros.

Esta situación originó que Zimmerman y Núñez-Antón (1997) propusiesen versiones estructuradas y menos generales de los modelos AD, a las que denominaron modelos antedependientes estructurados (ADE). En estos modelos, los coeficientes autorregresivos, las correlaciones, o las correlaciones parciales (dependiendo de si la especificación del modelo es autorregresiva, de covarianza, o de concentración) siguen una ley de potencias de Box-Cox, y las varianzas específicas, varianzas, o varianzas parciales (nuevamente dependiendo de la especificación utilizada) son funciones polinomiales o por tramos (es decir, que toma valores distintos en distintos tramos de la escala temporal) del tiempo de medición. Como ejemplo, escribiremos de forma detallada las especificaciones estructuradas autorregresiva y de covarianza. En ambos casos, f es una función dada por (Núñez-Antón y Woodworth, 1994):

$$f(t;\lambda) = \begin{cases} (t^{\lambda}-1)/\lambda & \text{si } \lambda \neq 0 \\ \log t & \text{si } \lambda = 0 \end{cases} \tag{4}$$

y g, para que en la práctica tenga una especificación útil, es una función de pocos parámetros (e.g., una función polinomial de bajo orden).

Especificación autorregresiva (ADE-EA):

$$\begin{aligned} \phi_{jk} &= \phi_k^{f(t_j;\lambda_k)-f(t_{j-k};\lambda_k)} \quad (j=s+1,\ldots,n;\ k=1,\ldots,s), \\ \sigma_j^2 &= \sigma^2 g(t_j;\psi) \quad (j=s+1,\ldots,n). \end{aligned} \tag{5}$$

Especificación de covarianza (ADE-EC):

$$\begin{aligned} \rho_{j,j-k} &= \rho_k^{f(t_j;\lambda_k)-f(t_{j-k};\lambda_k)} \quad (j=k+1,\ldots,n;\ k=1,\ldots,s), \\ v_{jj} &= \sigma^2 g(t_j;\psi) \quad (j=1,\ldots,n). \end{aligned} \tag{6}$$

Indicaremos que en el caso del modelo ADE-EA, $\{\phi_{jk} : j=2,\ldots,s;\ k=1,\ldots,j-1\}$ y $\sigma_1^2,\ldots,\sigma_s^2$ se dejan sin estructurar (es decir, como parámetros a estimar en el modelo). Además, la forma funcional de Box-Cox que tiene f especifica la estructura del modelo de tal forma que, los coeficientes autorregresivos de orden k (es decir, ϕ_k en el modelo ADE-EA) o las correlaciones en la k-ésima diagonal (en el modelo ADE-EC) son monótonas creciente si $\lambda_k < 1$, monótonas decreciente si $\lambda_k > 1$, o constantes si $\lambda_k = 1$ $(k=1,\ldots,s)$. Explicado de una forma más simple, el efecto de f es el de transformar la escala temporal de una forma no lineal para poder lograr que los coeficientes autorregresivos (en el modelo ADE-EA) o las correlaciones entre mediciones equidistantes en el tiempo (en el modelo ADE-EC), en la escala transformada, sean constantes.

Un modelo ADE tiene bastante menos parámetros que un modelo AD del mismo orden. Por ejemplo, si g es cuadrática, entonces, los modelos ADE-EA(1) y ADE-EA(2) tienen

6 y 10 parámetros, respectivamente, y los modelos ADE-EC(1) y ADE-EC(2) tienen 5 y 7 parámetros, respectivamente. Además, Cualquier especificación del modelo ADE(s) se puede utilizar sin importar el espaciamiento temporal entre las mediciones. La estimación de los parámetros en estos modelos precisa del uso de optimización numérica. Sin embargo, a diferencia de los modelos AD, los parámetros en estos modelos tienen restricciones, aún en el caso de la especificación autorregresiva. En particular, las restricciones asociadas al modelo (5) son $\phi_k > 0$, $\sigma^2 > 0$, y $\{\psi\colon\ g(t_j;\psi) > 0\}$. Las restricciones para el modelo (6) son similares pero sustituyendo $\phi_k > 0$ por $\rho_k > 0$.

4.7. Modelos ARIMA

Un modelo ARIMA(s,d,q) generaliza un modelo autorregresivo de medias móviles (ARMA), ya que especifica que las diferencias de orden d entre mediciones consecutivas, en lugar de las propias mediciones, siguen un modelo estacionario ARMA(s,q). Un caso particular es el modelo ARIMA(0,1,0) o modelo de paseo aleatorio

$$y_j - \mathbf{x}_j'\beta = \sum_{t=1}^{j} a_t \qquad (j = 1\ldots,n) \tag{7}$$

donde $a_1,\ldots,a_n$ son variables aleatorias independientes y con una distribución $N(0,v_a)$. Para este proceso, tenemos que $\text{var}(y_j) = jv_a$, $\text{cov}(y_j,y_u) = jv_a$ para $1 \leq j \leq u \leq n$ y $\text{corr}(y_j,y_u) = \sqrt{j/u}$ para $1 \leq j \leq u \leq n$. Por tanto, las varianzas crecen (linealmente) con el tiempo y las correlaciones entre observaciones igualmente espaciadas también crecen (no linealmente) con el tiempo. Este comportamiento es típico en los modelos ARIMA (véase Cryer, 1986, cap. 5). Existen otros dos casos dignos de mencionar en los modelos ARIMA: el modelo ARIMA(0,1,1) [o IMA(1,1)]

$$y_j - \mathbf{x}_j'\beta = y_{j-1} - \mathbf{x}_{j-1}'\beta + a_t - \gamma a_{t-1}$$

y el modelo ARIMA(1,1,0) [o ARI(1,1)]

$$y_j - \mathbf{x}_j'\beta - (y_{j-1} - \mathbf{x}_{j-1}'\beta) = \phi[y_{j-1} - \mathbf{x}_{j-1}'\beta - (y_{j-2} - \mathbf{x}_{j-2}'\beta)] + a_t.$$

Una desventaja de este tipo de modelos es que, para que puedan utilizarse en un contexto de datos longitudinales, el diseño de medición debe ser rectangular y los tiempos de medición equiespaciados entre sí. Sin embargo, existen modelos aplicables al caso de tiempos continuos y no discretos, lo que permitiría que estas restricciones se puedan relajar (véase, por ejemplo, Bergstrom, 1985). En este trabajo, consideraremos sólo uno de estos casos, el proceso de Wiener (WI), que es el análogo en tiempo continuo del modelo de paseo aleatorio. La función de covarianza para un proceso de Wiener está dada por $\text{cov}(y_j,y_k) = v\min(t_j,t_k)$, que coincide con la función de covarianza en (7), para el caso de datos equiespaciados en el tiempo, tomando $v = v_a$ y $t_j = j$. Así

tendremos que, para el modelo WI, la estructura de covarianzas intra-individuo, que tiene un solo parámetro ($v > 0$), se puede escribir como:

$$\mathbf{V} = v \begin{pmatrix} t_1 & t_1 & t_1 & \cdots & t_1 \\ t_1 & t_2 & t_2 & \cdots & t_2 \\ t_1 & t_2 & t_3 & \cdots & t_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1 & t_2 & t_3 & \cdots & t_n \end{pmatrix} \equiv v\mathbf{H}. \tag{8}$$

4.8. Aspectos Computacionales del Ajuste

La primera desventaja de los modelos que utilizan estructuras paramétricas para la matriz de covarianzas intra-individuos ha sido, hasta hace poco, la ausencia de paquetes estadísticos que permitiesen ajustar los mismos. Recientemente, el paquete estadístico SAS, con su PROC MIXED (SAS Institute Inc., 1996) ha logrado que muchos de estos modelos se puedan ajustar y comparar a otros modelos alternativos. Sin embargo, existen algunos modelos que deben ajustarse utilizando otros métodos.

Los modelos SC, AR(1), SCH, ARH(1), HF, TOEP, TOEPH, CA, NE, AD y ADE permiten tener cualquier diseño de medición y el hecho de tener observaciones igualmente espaciadas o diseño de medición rectangular no tiene mayores ventajas, aparte de las que garanticen su ajuste. Así, por ejemplo, para el ajuste del modelo NE, es necesario que el diseño no se aleje mucho de la rectangularidad. Si el diseño de medición es rectangular, y m es suficientemente grande, existen expresiones explícitas para los estimadores MV y MVR de $\mathbf{V}$. Sin embargo, si el diseño de medición no es rectangular, no existen expresiones explícitas para los estimadores MV o MVR. En dicho caso, y dependiendo del alejamiento del diseño rectangular, puede ser difícil o imposible maximizar la verosimilitud. En particular, la función de verosimilitud puede ser muy plana (debido al gran número de parámetros), lo que puede causar, y de hecho causa, problemas de convergencia. Los modelos ARIMA, por otro lado, sólo pueden utilizarse en un contexto de datos longitudinales en el caso de diseño rectangular.

En el caso de los modelos AD, la rectangularidad en el diseño simplifica de manera sustancial el proceso de estimación. Si el diseño de medición es rectangular y m es suficientemente grande, existen expresiones para los estimadores MV o MVR de $\mathbf{V}$ en el caso de un modelo de primer orden (véase Byrne y Arnold, 1983). En el caso de un diseño rectangular y modelos de órdenes superiores a uno, no existen expresiones simples para obtener los estimadores MV y MVR pero se pueden obtener a partir de los elementos de $\mathbf{S}$, utilizando un proceso recursivo, que no requiere del uso de técnicas de optimización numéricas (Johnson, 1989). Sin embargo, en los casos de diseños no rectangulares, debemos utilizar optimización numérica. En estos casos, el modelo AD(1) con especificación de covarianza tiene ventajas computacionales sobre las otras

especificaciones dado que existen fórmulas explícitas para la inversa y el determinante de la matriz de covarianzas de los modelos AD(1) y ADE(1) (Núñez-Antón y otros, 1995, Núñez-Antón, 1997 y Zimmerman y otros, 1998).

Los modelos de coeficientes aleatorios, de simetría compuesta, de simetría compuesta heterogéneo, autorregresivos de primer orden, autorregresivos de primer orden heterogéneo, Huynh-Feldt, Toeplitz, no estructurado, antedependiente no estructurado de orden uno y ciertos modelos ARIMA se pueden ajustar utilizando PROC MIXED, aunque en los modelos CA suelen existir problemas de convergencia en el algoritmo. Sin embargo, mencionaremos que el ajuste de los modelos AR(1), ARH(1), TOEP y TOEPH en SAS sólo se puede realizar si los tiempos en los que se realizan las mediciones son equidistantes y, además, van desde $t_1 = 1$ hasta $t_n = n$.

Es posible ajustar el modelo AD(1) utilizando PROC MIXED, para lo cual es necesario utilizar la opción ANTE(1). La especificación utilizada es la de covarianza, con las restricciones que aseguran que la matriz sea definida positiva. Para que se dé la convergencia, se permite que haya ciertos tipos de no rectangularidad en el diseño de medición. Los modelos AD de órdenes superiores no pueden ajustarse utilizando PROC MIXED. Sin embargo, SAS no puede utilizarse para ajustar los modelos ADE y programas específicos que permitan estos ajustes deben ser utilizados.

Algunos modelos ARIMA pueden ajustarse en PROC MIXED, utilizando los comandos adecuados para que sean aplicados a los datos diferenciados. Por ejemplo, los modelos ARI(1,1) e IMA(1,1) pueden ajustarse utilizando los comandos para ajustar los modelos AR(1) o TOEP(2) en PROC MIXED, respectivamente, a las diferencias de primer orden de las observaciones originales. El modelo WI se puede ajustar directamente a los datos no diferenciados utilizando el comando LIN y especificando $\mathbf{H}$ en la ecuación (8). De forma alternativa, el modelo WI se puede ajustar usando mínimos cuadrados ordinarios para las diferencias de primer orden. Sin embargo, esto no es recomendable debido a que reduce el número de observaciones y, además, elimina cualquier variable explicativa que sea constante dentro de un mismo individuo, como, por ejemplo, los efectos de uno o varios tratamientos.

5. EJEMPLOS: AJUSTE DE MODELOS

5.1. Aspectos Generales del Ajuste

Para ilustrar el ajuste de los modelos introducidos en la Sección 4 y, además, para poder compararlos entre sí y con el modelo de coeficientes aleatorios, en esta sección ajustaremos algunos de estos modelos a los datos introducidos en la Sección 2. A continuación, mencionaremos algunos aspectos generales que utilizaremos en el ajuste de los modelos paramétricos de la estructura de covarianza en el contexto de datos longitudinales:

1. Dado que nuestro principal interés en este trabajo es la modelización de la estructura de covarianzas, utilizaremos un modelo tan saturado como nos sea posible para la estructura de la respuesta media. Puesto que para el *cattle data* tenemos dos posibles tratamientos, utilizaremos el siguiente modelo para la estructura de la respuesta media:

$$E(y_{ij}) = \begin{cases} \mu_{Aj} & \text{si el individuo } i \text{ recibe el tratamiento A} \\ \mu_{Bj} & \text{si el individuo } i \text{ recibe el tratamiento B,} \end{cases} \tag{9}$$

para $i = 1,\ldots,60$ y $j = 1,\ldots,7,8,9,10,11$. Los tiempos han sido reescalados dividiéndolos por 14, que son los días que hay en un periodo de dos semanas. Es decir, los tiempos que utilizaremos en estos datos son $t = 1,\ldots,7,8,9,9.5$. Mencionaremos que, en base a lo que vemos en la Figura 2, es posible que un modelo cuadrático en el tiempo sea una buena alternativa para modelizar la respuesta media en estos datos. Sin enbargo, análisis previos de estos datos que también centraron su interés en la modelización de la estructura de covarianzas (véase, por ejemplo, Kenward, 1987 y Zimmerman y Núñez-Antón, 1997), también utilizaron un modelo similar al de la ecuación (9).

En el caso del *race data*, utilizaremos un modelo parecido al anterior pero considerando que en este caso no tenemos ningún tratamiento y simplemente estamos midiendo la variable de respuesta para cada uno de los tiempos o tramos de la carrera. Así, el modelo utilizado será

$$E(y_{ij}) = \mu_j, \tag{10}$$

para $i = 1,\ldots,80$ y $j = 1,\ldots,10$. Los tiempos que utilizaremos representarán a cada una de las diez secciones (cronológicamente hablando) de la carrera. Es decir, $t = 1,\ldots,10$. Estos datos han sido analizados previamente de forma exploratoria y gráfica por Everitt (1994*a*, 1994*b*). Además, Zimmerman y otros (1998) ajustaron inicialmente un modelo paramétrico cúbico en el tiempo para la estructura de medias y un modelo no estacionario ADE para la estructura de covarianzas intra-individuos. En ninguno de estos análisis previos el objetivo principal de los mismos era el de comparar diferentes estructuras de covarianzas, por lo que, dado que este es el objetivo de este trabajo, hemos decidido mantener el modelo (10) para la estructura de medias.

En cualquier caso, creemos que al utilizar una estructura de medias saturada, y concentrarnos en la modelización de la estructura de covarianzas, no estamos realizando un análisis todo lo riguroso que querríamos. Pensamos que esta propuesta debe complementarse con una posible reducción en la estructura de medias de forma que se

reduzca el número de parámetros presentes en la misma. Por ejemplo, podría plantearse el uso de un modelo paramétrico para esta estructura en cualquiera de los dos conjuntos de datos o, como alternativa, podría contrastarse hipótesis sobre el modelo propuesto que permitan su reducción.

2. Estimaremos los parámetros utilizando el método de máxima verosimilitud residual o restringida (MVR). En este contexto, menos dos veces el logaritmo de la verosimilitud residual para los datos será igual a:

$$-2l_R(\theta) = \sum_{i=1}^{m} \log|\mathbf{V}_i(\theta)| + \sum_{i=1}^{m} [\mathbf{y}_i - \mathbf{X}_i\hat{\beta}(\theta)]'[\mathbf{V}_i(\theta)]^{-1}[\mathbf{y}_i - \mathbf{X}_i\hat{\beta}(\theta)] + \log|\sum_{i=1}^{m} \mathbf{X}_i'[\mathbf{V}_i(\theta)]^{-1}\mathbf{X}_i| + (N-p)\log 2\pi$$

donde $\hat{\beta}(\theta) = (\sum_{i=1}^{m} \mathbf{X}_i'[\mathbf{V}_i(\theta)]^{-1}\mathbf{X}_i)^{-1} \sum_{i=1}^{m} \mathbf{X}_i'[\mathbf{V}_i(\theta)^{-1}]\mathbf{y}_i$.

3. Dado que en la Sección 2 hemos indicado que existe una no estacionariedad en varianza y en correlación, ajustaremos los distintos modelos no estacionarios mencionados en la Sección 4 (es decir, modelo NE, modelo WI, uno o más modelos AD, y uno o más modelos ADE) y los compararemos con los modelos de coeficientes aleatorios lineales (CAL) y cuadráticos (CAC). Usaremos PROC MIXED para ajustar los modelos NE, WI, CAL y CAC. Los modelos ADE y AD de órdenes mayores los ajustaremos utilizando programas en FORTRAN escritos por los autores y usados conjuntamente con las subrutinas IMSL (IMSL, Inc., 1991*a*, 1991*b*). Estos programas, en su proceso de optimización de la función de verosimilitud residual $l_R(\theta)$, utilizan el algoritmo simplex de Nelder y Mead (Nelder y Mead, 1965). Todos los modelos AD y ADE ajustados a los datos tienen una especificación autorregresiva. Además, para poder comparar todos estos modelos entre sí y con los modelos de coeficientes aleatorios, hemos ajustado los modelos Toeplitz (TOEP) y su extensión heterogénea (TOEPH) (sólo para el caso de race data), Huynh-Feldt (HF), además de los modelos estacionarios de simetría compuesta (SC) y autorregresivo de primer orden (AR(1)), y sus extensiones heterogéneas (SCH y ARH(1)).

   Los modelos TOEP y TOEPH, HF, SC y SCH se ajustaron utilizando PROC MIXED. Los modelos AR(1) y ARH(1), casos especiales de los modelos ADE, además del modelo WI, se ajustaron (sólo en el caso del *cattle data* por existir observaciones no equiespaciadas entre sí) utilizando versiones adaptadas de nuestros programas. Los modelos AD, especificación autorregresiva, se ajustaron siguendo las pautas que a este efecto se mencionan en Macchiavelli y Arnold (1994) y Zimmerman y Núñez-Antón (1997).

4. Los ajustes de los modelos paramétricos utilizados para la estructura de covarianzas intra-individuo se compararon a través de dos criterios comúnmente propuestos para este fin. Ambos se encuentran especificados de tal forma que un valor mayor de estos criterios indica un mejor modelo de covarianzas. Los criterios son: el criterio de

información de Akaike AIC $= l_R(\hat{\theta}) - q$, y el criterio de información Bayesiana de Schwarz BIC $= l_R(\hat{\theta}) - \frac{q}{2}\log(N-p)$. En estas ecuaciones, $\hat{\theta}$ es el estimador MVR de θ, y q es la dimensión de θ. Como ya es sabido, el criterio BIC tiene una penalización mayor por los parámetros adicionales en el modelo, por lo que tenderá a favorecer modelos con menos parámetros que los seleccionados utilizando AIC. Otro criterio de comparación de estos ajustes es el contraste de razón de vesosimilitudes residuales (CRVR). Este contraste se realiza haciendo la diferencia entre los valores de $-2l_R$ para dos modelos anidados y comparando este valor con el percentil correspondiente de una distribución χ^2 con grados de libertad iguales a la diferencia en el número de parámetros en las dos estructuras de covarianza que se estén contrastando.

5.2. Análisis del Race Data

Los primeros trabajos que realizan un análisis de tipo gráfico y exploratorio para estos datos son Everitt (1994*a* y 1994*b*). Zimmerman y otros (1998) ajustaron un modelo de pocos parámetros para describir la relación entre los tiempos que cada corredor utilizaba en recorrer cada sección de 10 kilómetros de la carrera y el número de la sección correspondiente ($t = 1,\ldots,10$) y la edad del corredor. El modelo final que se utilizó para la estructura de medias era un modelo lineal en el tiempo y que eliminó la variable edad. Su modelo para la estructura de covarianzas intra-individuos fue un modelo ADE(1) con especificación de covarianza (es decir, ADE-EC(1)), y una varianza que cambiaba de forma lineal con el tiempo (véase el modelo (6)). Este análisis previo, aunque muy limitado, nos puede dar alguna pista sobre los modelos que podrían ser adecuados para estos datos. Hay una clara presencia de no estacionariedad en varianza y en correlación, por lo que los modelos no estacionarios, entre los que se encuentra el de coeficientes aleatorios, serán seguramente las mejores opciones a ajustar.

El contraste de verosimilitudes de Gabriel (1962) para determinar el orden adecuado de antedependencia sugiere que un modelo de orden dos será suficiente para estos datos. Es decir, un modelo AD(2) será mejor que modelos de órdenes inferiores y tan bueno como modelos de órdenes superiores (es decir, mayores que dos). Así, tendremos que al menos ajustar modelos estructurados y no estructurados de orden dos y uno. Sin embargo, dado que el contraste de Gabriel es un contraste aproximado, también ajustaremos modelos no estructurados de órdenes tres y cuatro.

Para las varianzas específicas, en la especificación autorregresiva estructurada, utilizaremos modelos lineales (ADEL) y cuadráticos (ADEQ) en el tiempo (véase el modelo (5)). Es decir, para el caso del modelo ADEQ(1), tendremos que:

$$\begin{aligned} \phi_j &= \phi_1^{f(t_j;\lambda_1)-f(t_{j-k};\lambda_1)} \quad (j=2,\ldots,n), \\ \sigma_j^2 &= \sigma^2\left(1+\psi_1 t_j+\psi_2 t_j^2\right) \quad (j=2,\ldots,n). \end{aligned}$$

En este caso, tendremos que el modelo tiene 6 parámetros, de tal forma que $\theta = (\phi_1, \lambda_1, \sigma^2, \sigma_1^2, \psi_1, \psi_2)$. Así, en el caso del modelo ADEL(1), $\psi_2 = 0$ y, por tanto, el modelo tendrá 5 parámetros.

Tabla 4. Resultados del análisis de las distintas propuestas de modelos para la estructura de covarianzas intra-individuo para el *race data*.

Estructura	q	AIC	BIC	$-2l_R$	MC	ν^*	χ^2	$Pr > \chi^2$
SC	2	−2673.6	−2678.2	5343.1				
AR(1)	2	−2550.6	−2555.3	5097.2				
AD(0)	10	−2897.8	−2921.1	5775.5				
WI	1	−2532.5	−2534.9	5063.1				
SCH	11	−2558.6	−2584.3	5095.2	SC	9	247.9	0.00
CAL	4	−2566.6	−2575.9	5125.1	SC	2	218.0	0.00
CAC	7	−2525.0	−2541.4	5036.1	CAL	3	89.0	0.00
HF	11	−2634.7	−2660.4	5247.5	SC	9	95.6	0.00
ARH(1)	11	−2395.7	−2421.4	4769.4	AR(1)	9	327.8	0.00
TOEP	10	−2539.3	−2562.7	5058.6	AR(1)	8	38.6	0.00
TOEPH	19	−2396.9	−2441.3	4755.8	ARH(1)	9	13.6	0.09
ADEL(1)	5	−2504.5	−2516.2	4999.0	AD(1)	14	283.0	0.00
ADEQ(1)	6	−2443.3	−2457.3	4874.6	AD(1)	13	158.6	0.00
ADEL(2)	9	−2469.5	−2490.5	4920.9	ADEL(1)	4	78.1	0.00
ADEQ(2)	10	−2422.0	−2445.3	4823.9	ADEL(2)	1	97.0	0.00
ADEL(5)	27	−2445.3	−2508.4	4836.3	ADEL(2)	18	84.6	0.00
ADEQ(5)	28	−2421.0	−2486.4	4786.0	ADEL(5)	1	50.3	0.00
AD(1)	19	−2377.0	−2421.4	4716.0	ARH(1)	8	53.4	0.00
AD(2)	27	−2361.1	−2424.2	4668.2	AD(1)	8	47.8	0.00
AD(3)	34	−2357.3	−2436.7	4646.6	AD(2)	7	21.6	0.00
AD(4)	40	−2361.0	−2454.4	4642.0	AD(3)	6	4.6	0.60
NE	55	−2365.2	−2493.7	4620.4	AD(3)	21	26.2	0.20

En la Tabla 4 y posteriores, q es el número de parámetros en la estructura de covarianzas, MC es el modelo anidado con el que se compara el modelo en cuestión, ν^* son los grados de libertad para el contraste de verosimilitudes residuales, y la última columna contiene el valor de probabilidad que permitirá o no rechazar el modelo en la hipótesis nula. Para el cálculo de BIC, $N = 800$ y $p = 10$.

Si la selección del modelo se basa en el criterio AIC, el mejor modelo ajustado es el AD(3), con $q = 34$ parámetros. Los modelos AD(2) y AD(4), que tienen $q = 27$ y $q = 40$ parámetros, respectivamente, son también modelos con un buen ajuste. Si, por el contrario, usamos el criterio BIC, los mejores modelos ajustados son ARH(1) y AD(1), que tienen, respectivamente $q = 11$ y $q = 19$ parámetros. El contraste CRVR con 19-11=8

Tabla 5. Resultados del análisis de las distintas propuestas de modelos para la estructura de covarianzas intra-individuo para el *cattle data*-tratamiento A.

Estructura	q	AIC	BIC	$-2l_R$	MC	ν^*	χ^2	P
SC	2	−1192.2	−1196.0	2380.4				
AR(1)	2	−1052.9	−1056.7	2101.8				
AD(0)	11	−1366.0	−1386.7	2710.0				
WI	1	−1053.7	−1055.6	2105.4				
SCH	12	−1172.3	−1194.9	2320.6	SC	10	59.8	0.00
HF	12	−1190.0	−1212.6	2355.9	SC	10	24.5	0.01
CAL	4	−1080.8	−1088.3	2153.6	SC	2	226.8	0.00
CAC	7	−1051.2	−1064.4	2088.3	CAL	3	65.3	0.00
ARH(1)	12	−1057.0	−1079.6	2089.9	AR(1)	10	11.8	0.30
ADE(1)	4	−1048.9	−1056.4	2089.7	AR(1)	2	12.0	0.00
ADE(2)	8	−1051.2	−1066.2	2086.4	ADE(1)	4	3.4	0.49
AD(1)	21	−1055.9	−1095.4	2069.8	ARH(1)	9	20.2	0.02
					ADE(1)	17	19.9	0.28
AD(2)	30	−1055.6	−1112.1	2051.2	AD(1)	9	18.6	0.03
					ADE(2)	22	35.2	0.04
NE	66	−1075.7	−1200.0	2019.4	AD(2)	36	31.7	0.67

grados de libertad rechaza el modelo ARH(1) en favor del modelo AD(1). Los distintos contrastes realizados en la Tabla 4 sugieren que un modelo adecuado para estos datos sería el modelo AD(3) con $q = 34$ parámetros, muchos menos parámetros de los que tiene el modelo NE. Para la realización de estos contrastes, hemos comparado sucesivamente el modelo AD(3) con los modelos NE, AD(4) and AD(2), de tal forma que el modelo AD(3) fue el seleccionado. Mencionaremos que, en este caso, los modelos estructurados no proporcionaron un buen ajuste y no creemos necesario compararlos con los modelos no estructurados. Como ilustración, Zimmerman y Núñez-Antón (1997) escriben de forma explícita las ecuaciones recursivas para el ajuste de un modelo AD(2) con especificación autorregresiva.

En resumen, debemos mencionar que: (*i*) Los modelos de coeficientes aleatorios CAL y CAC no son adecuados para estos datos y no deben utilizarse; (*ii*) si los comparamos con los modelos AD o con otros modelos alternativos, ninguno de los modelos ADE da un ajuste adecuado para estos datos. Sin embargo, podríamos pensar que, dada la naturaleza de los modelos de coeficientes aleatorios, una comparación más justa entre estos modelos y los demás habría tenido que considerar un modelo de coeficientes aleatorios con media estructurada (por ejemplo, lineal o cuadrática en el tiempo). Por simple comparación, hemos ajustado los modelos CAL y CAC estructurando la media con una dependencia lineal en el tiempo y, dado que estos modelos tienen un peor

ajuste que los modelos con medias no estructuradas, no pueden competir con los mejores modelos ajustados para estos datos. Finalmente, indicaremos a nivel interpretativo que las estructuras antedependientes de primer orden son fácilmente motivables por el simple hecho de observar si las primeras subdiagonales o superdiagonales en la matriz de correlaciones presentan correlaciones muestrales que crecen o decrecen con el paso del tiempo. Hemos comprobado que esta intuición no es fácilmente extendible a los modelos de órdenes superiores. A nuestro entender, la única forma de «evaluar» si el modelo realmente ajusta los datos es obtener la matriz de varianzas y covarianzas para el modelo seleccionado y compararla con la empírica correspondiente. No hemos incluido esta comparación aquí, pero al igual que en los trabajos previos en este área, la hemos realizado y el ajuste, creemos, es bastante aceptable.

Tabla 6. Resultados del análisis de las distintas propuestas de modelos para la estructura de covarianzas intra-individuo para el *cattle data*-tratamiento B.

Estructura	q	AIC	BIC	$-2l_R$	MC	ν^*	χ^2	P
SC	2	−1188.9	−1192.7	2373.9				
AR(1)	2	−1104.8	−1108.6	2205.6				
AD(0)	11	−1345.9	−1366.6	2669.9				
WI	1	−1097.9	−1099.8	2193.9				
SCH	12	−1140.9	−1163.5	2257.8	SC	10	116.1	0.00
HF	12	−1184.6	−1207.1	2345.1	SC	10	28.8	0.00
CAL	4	−1128.7	−1136.2	2249.4	SC	2	124.5	0.00
CAC	7	−1099.8	−1113.0	2185.6	CAL	3	63.8	0.00
ARH(1)	12	−1054.7	−1077.2	2085.2	AR(1)	10	120.4	0.00
ADE(1)	4	−1074.0	−1081.5	2139.9	AR(1)	2	66.7	0.00
ADE(2)	8	−1077.3	−1092.3	2138.5	ADE(1)	4	1.4	0.84
ADE(3)	13	−1076.4	−1100.9	2126.9	ADE(2)	5	11.7	0.04
AD(1)	21	−1043.9	−1083.5	2045.8	ARH(1)	9	39.4	0.00
					ADE(1)	17	94.1	0.00
AD(2)	30	−1046.8	−1103.3	2033.6	AD(1)	9	12.3	0.20
					ADE(2)	22	105.0	0.00
AD(3)	38	−1045.0	−1116.5	2013.9	AD(2)	8	19.6	0.01
					AD(1)	17	31.9	0.02
					ADE(3)	25	113.0	0.00
NE	66	−1055.6	−1179.9	1979.2	AD(3)	28	34.7	0.18

Concluimos este análisis mencionando que los modelos antedependientes no estructurados son modelos adecuados para estos datos lo que, de alguna forma, ya fue indicado en Zimmerman y otros (1998). Si nos fijamos en el número de parámetros y el ajuste dado por los criterios AIC y BIC, éstos son alternativas válidas y relevantes para el

modelo NE u otros modelos. Indicaremos además que todos los modelos con muchos parámetros son, en este caso, más adecuados que lo modelos más simples, tales como los modelos SC, AD(0) o AR(1).

5.3. Análisis del Cattle Data

Contrastamos, en primer lugar, la hipótesis de igualdad de las dos matrices de covarianzas intra-individuos para los distintos grupos utilizando el contraste clásico de razón de verosimilitudes. La hipótesis de igualdad se rechaza de forma clara ($P = 0.02$). Consecuentemente, utilizamos estructuras paramétricas distintas para modelizar las matrices de covarianzas intra-individuo para cada uno de los dos grupos en estos datos. Los análisis posteriores que se realicen con estos datos (por ejemplo, análisis que permitan reducir la estructura de medias), deben utilizar, para cada grupo, la estructura de covarianzas intra-individuo que haya dado el «mejor» ajuste. Dado que éste no es el enfoque de este trabajo, proponemos este análisis para futuros trabajos y, por tanto, no lo incluimos en el presente.

Kenward (1987) utilizó un modelo AD para analizar estos datos pero a lo largo de todo su análisis utilizó una estructura de covarianzas común para los dos grupos, lo cual no es correcto. Es por eso que no haremos mayores comentarios a los resultados o modelos ajustados, aunque si mencionaremos que sus conclusiones indicaron que un modelo AD de orden dos es el adecuado para la estructura común de covarianzas en los dos grupos para estos datos. Es decir, usando el contraste de Gabriel (1962), los modelos de órdenes inferiores a dos no son adecuados, y el modelo de orden dos lo hace tan bien como los de órdenes superiores. Además, Kenward (1987) encontró que existía una diferencia significativa entre los distintos tratamientos.

Zimmerman y Núñez-Antón (1997) han ajustado modelos AD y ADE a estos datos. En este trabajo extenderemos el de Zimmerman y Núñez-Antón (1997) al centrarnos en ajustar estructuras de covarianzas alternativas, incluyendo las estructuras de coeficientes aleatorios y algunas otras alternativas que ya hemos mencionado en la Sección 4. El contraste de Gabriel (1962) sugiere que los órdenes adecuados de antedependencia para los modelos UAD que pueden usarse en los tratamientos A y B son dos y tres (al igual que modelos de órdenes superiores a éstos), respectivamente. Por esto, ajustamos modelos AD y ADE de órdenes menores o iguales a los sugeridos por este contraste. Los modelos ADE utilizan una especificación autorregresiva con varianzas específicas constantes (ver ecuación (5)). Es decir, en este caso tendremos que, por ejemplo, para el modelo ADE(1):

$$\begin{aligned} \phi_j &= \phi_1^{f(t_j;\lambda_1)-f(t_{j-1};\lambda_1)} \quad (j=2,\ldots,n) \\ \sigma_j^2 &= \sigma^2 \quad (j=2,\ldots,n) \end{aligned}$$

En este caso, tenemos que el modelo anterior tiene 4 parámetros, de tal forma que $\theta = (\phi_1, \lambda_1, \sigma^2, \sigma_1^2)$.

En las Tablas 5 y 6, tenemos la información de los modelos ajustados para el *cattle data*, grupos A y B, respectivamente. Para el cálculo de BIC, $N = 330$ y $p = 11$. En el caso del grupo A, si la selección del modelo se basa en el criterio AIC, el mejor modelo ajustado es el ADE(1), con $q = 4$ parámetros, aunque también los modelos AR(1) y ADE(2) . Si, por el contrario, usamos el criterio BIC, el mejor modelo ajustado es el WI, con $q = 1$ parámetro, aunque también los modelos ADE(1) y AR(1) tienen un buen ajuste. El contraste de CRVR entre estos modelos también sugiere el uso del modelo ADE(1). Sin embargo, este modelo no puede compararse con el modelo WI usando este contraste, ya que no son modelos anidados. Además, los contrastes escalonados que se realizan en la Tabla 5 sugieren que el mejor modelo para estos datos es el AD(2), con $q = 30$ parámetros. Un contraste entre el modelo ADE(1) y el modelo AD(2) rechaza el primero de ellos. Para la realización de estos contrastes, hemos comparado sucesivamente el modelo AD(2) con los modelos NE, AD(1) y ADE(2), de tal forma que el modelo AD(2) fue el seleccionado.

En el caso del grupo B, si la selección del modelo se basa en el criterio AIC y CRVR, los mejores modelos ajustados son los modelos AD, mientras que los modelos ARH(1), ADE(1) y AD(1) son los mejores si nos basamos en el criterio BIC. Los contrastes escalonados que se realizan en la Tabla 6 sugieren que el mejor modelo para estos datos es el AD(3), con $q = 38$ parámetros. Para la realización de estos contrastes, hemos comparado sucesivamente el modelo AD(3) con los modelos NE, AD(2) y ADE(3), de tal forma que el modelo AD(3) fue el seleccionado. Además, debemos mencionar que, en este caso, los modelos de coeficientes aleatorios CAL y CAC no son adecuados para estos datos y no deben utilizarse.

En resumen, los modelos antedependientes no estructurados son modelos adecuados para estos datos lo que, de alguna forma, ya fue indicado en Zimmerman y Núñez Antón (1997). Si nos fijamos en el número de parámetros y el ajuste dado por los criterios AIC y BIC, éstos son alternativas válidas y relevantes para el modelo NE u otros modelos. Indicaremos además que, en este caso, todos los modelos con muchos parámetros son más adecuados que lo modelos más simples.

Finalmente y como detalle adicional comentaremos que si utilizamos como modelos de covarianzas para cada uno de los grupos los modelos AD(2) y AD(3), respectivamente, y llevamos a cabo un contraste de diferencias significativas en las respuestas de los distintos grupos (tratamientos), concluiremos que existe una diferencia estadísticamente significativa ($P = 0.00$) en las respuestas de los distintos tratamientos.

6. CONCLUSIONES

Hemos estudiado los diferentes modelos que se pueden proponer para la estructura de covarianzas intra-individuos en el contexto de datos longitudinales. Uno de nuestras principales motivaciones era la de ilustrar la posible existencia de no estacionariedad en varianza y/o en correlación en esta matriz de covarianzas y los posibles modelos que permiten explicar estos comportamientos no estacionarios. Entre ellos, el más utilizado en la práctica es el modelo de coeficientes aleatorios, aunque no se le ha reconocido el mérito adecuado en la modelización de comportamientos de tipo no estacionario. Así, comparamos los ajustes de estos modelos con los de otros que permiten la modelización de estos comportamientos y con los modelos estacionarios más utilizados a través de dos estudios longitudinales que han sido frecuentemente citados en la literatura de datos longitudinales.

Hemos motivado el uso de todos los modelos en la Sección 4 a través de las características presentes en la matriz de covarianzas intra-individuos para cada uno de ellos, lo que nos lleva a aconsejar o desaconsejar su uso en algunos casos. Hemos hablado de modelos simples y sencillos, como los modelos SC, WI o AR(1), entre otros, además de hablar de modelos flexibles y en cierto sentido simples, como los modelos ADE, que pueden tener una utilidad importante en ciertas aplicaciones. También hemos mencionado los modelos más complicados o generales, como los modelos NE o AD, entre otros, que siempre representarán una posibilidad ya conocida de modelizar esta matriz en datos longitudinales. Esto es, si no se tiene una idea clara del modelo que se puede utilizar para modelizar el comportamiento de esta estructura en los datos, estos modelos son suficientemente generales y, por tanto, pueden utilizarse, aunque, obviamente, el precio a pagar es la alta dimensionalidad del vector de parámetros.

No hemos modelizado la estructura de medias de forma paramétrica explícita debido al interés especial que tenemos en este trabajo en la modelización de la estructura de covarianzas intra-individuos. A este respecto indicaremos que, en nuestra opinión, un análisis completo de este tipo de datos requiere plantear un modelo inicial bastante general para la estructura de medias, modelo que se puede basar en un análisis previo de las medias en los distintos tiempos. A continuación, y con esta estructura de medias, se deben proponer distintas estructuras de covarianza intra-individuos y seleccionar la que mejor ajuste los datos. Finalmente, con la estructura seleccionada se realizan contrastes de reducción en la estructura de medias, obteniendo de esta forma un modelo final para ambas estructuras. En este trabajo hemos utilizado una estructura bastante general para la media y nos hemos centrado en la modelización de la estructura de covarianzas intra-individuos de los datos.

Hemos ajustado distintos modelos a dos conjuntos de datos en los que los de coeficientes aleatorios demostraron no ser adecuados. Estas mismas conclusiones se han obtenido previamente para otros conjuntos de datos distintos (Núñez-Antón, 1993).

En los distintos modelos ajustados tenemos que mencionar su comparación no sólo en términos de ajuste, sino también en términos de flexibilidad y número de parámetros. El modelo NE es el más flexible y es el que más parámetros tiene, con $O(n^2)$ parámetros. Los modelos AD son un poco menos flexibles pero tienen menos parámetros (es decir, $O(n)$ parámetros). Los modelos ADE, ARIMA y de coeficientes aleatorios (CA), al igual que todos los modelos estacionarios, son muy estructurados (con $O(1)$ parámetros) y muy poco flexibles.

El espaciamiento irregular entre observaciones y la no rectangularidad en el diseño de medición no presentan problema alguno para los modelos ADE o CA, pero hay que tener cuidado con una o ambas condiciones cuando se ajustan los modelos NE, AD o ARIMA. Hay que mencionar que existe un claro problema para ajustar algunos de estos modelos, especialmente cuando se trata de observaciones irregularmente espaciadas. Cuando éste es el caso, PROC MIXED sólo se puede usar para los modelos en los que la estructura no dependa de los tiempos de observación (por ejemplo, CA, HF y NE). El resto de modelos deben ajustarse a través de programas que se deben escribir en algún lenguaje específico, en nuestro caso FORTRAN.

Los modelos antedependientes son pocos conocidos, muy útiles y claramente superiores a los modelos de coeficientes aleatorios en contextos no estacionarios para datos longitudinales. Su uso debe al menos empezar a ser considerado por los estadísticos como una alternativa real, cuando así lo sean, a los modelos CA.

AGRADECIMIENTOS

El trabajo de Vicente Núñez Antón ha sido financiado por los proyectos de investigación PB98-0149 de la Dirección General de Enseñanza Superior e Investigación Científica del Ministerio Español de Educación y Cultura, UPV 038.321-HA129/99 de la Universidad del País Vasco/Euskal Herriko Unibertsitatea y PI-1999-46 del Gobierno Vasco. El trabajo de Dale L. Zimmerman ha sido financiado parcialmente por el proyecto de investigación 9628612 de la National Science Foundation. Los autores agradecen los comentarios del editor y de un evaluador que han mejorado de forma importante la presentación de este trabajo.

REFERENCIAS

Bergstrom, A. R. (1985). «The estimation of parameters in nonstationary higher-order continuous-time dynamic models». *Econometric Theory*, 1, 369-385.

Byrne, P. J. & Arnold, S. F. (1983). «Inference about multivariate means for a nonstationary autoregressive model». *Journal of the American Statistical Association*, 78, 850-855.

Crowder, M. J. & Hand, D. J. (1990). *Analysis of Repeated Measures*. London: Chapman & Hall.

Cryer, J. D. (1986). *Time Series Analysis*. Boston: PWS-Kent.

Diggle, P. J. (1988). «An approach to the analysis of repeated measures». *Biometrics*, 44, 959-971.

Diggle, P. J. (1990). *Time Series: A Biostatistical Introduction*. Oxford: Oxford University Press.

Diggle, P. J., Liang, K. Y. & Zeger, S. L. (1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.

Everitt, B. S. (1994*a*). «Exploring multivariate data graphically: a brief review with examples». *Journal of Applied Statistics*, 21, 63-94.

— (1994*b*). *A Handbook of Statistical Analysis Using S-Plus*. London: Chapman & Hall.

Gabriel, K. R. (1962). «Ante-dependence analysis of an ordered set of variables». *Annals of Mathematical Statistics*, 33, 201-212.

Huynh, H. & Feldt, L. S. (1970). «Conditions under which mean square ratios in repeated measurements designs have exact $\mathcal{F}$-distributions». *Journal of the American Statistical Association*, 65, 1582-1589.

IMSL, Inc. (1991*a*). *Fortran Subroutines for Mathematical Applications. MATH/LIBRARY Version 2.0*. Houston, Texas: IMSL, Inc.

— (1991*b*). *Fortran Subroutines for Mathematical Applications. STAT/LIBRARY Version 2.0*. Houston, Texas: IMSL, Inc.

Jennrich, R. L. & Schluchter, M. D. (1986). «Unbalanced repeated-measures models with structured covariance matrices». *Biometrics*, 42, 805-820.

Johnson, K. L. (1989). «Higher-order antedependence models». Unpublished Ph. D. Thesis. Department of Statistics, Pennsylvania State University.

Jones, R. H. (1990). «Serial correlation or random subject effects?» *Communication in Statistics, Simulation and Computation*, 19, 1105-1123.

— (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach*. London: Chapman & Hall.

Jones, R. H. & Boadi-Boateng, F. (1991). «Unequally spaced longitudinal data with AR(1) serial correlation». *Biometrics*, 47, 161-175.

Kenward, M. C. (1987). «A method for comparing profiles of repeated measurements». *Applied Statistics*, 36, 296-308.

Laird, N. M. (1988). «Missing data in longitudinal studies». *Statistics in Medicine*, 7, 305-315.

Laird, N. M. & Ware, J. H. (1982). «Random effects models for longitudinal data». *Biometrics*, 38, 963-974.

Macchiavelli, R. E. & Arnold, S. F. (1994). «Variable order ante-dependence models». *Communications in Statistics, Theory and Methods*, 23, 2683-2699.

Muñoz, A., Carey, V., Schouten, J. P., Segal, M. & Rosner, B. (1992). «A parametric family of correlation structures for the analysis of longitudinal data». *Biometrics*, 48, 733-742.

Nelder, J. A. & Mead, R. (1965). «A simplex method for function minimization». *The Computer Journal*, 7, 308-313.

Núñez-Antón, V. (1993). «Analysis of longitudinal data with unequally spaced observations and time-dependent correlated errors». Unpublished Ph. D. Thesis. Department of Statistics and Actuarial Science, The University of Iowa.

— (1997). «Longitudinal data analysis: non-stationary error structures and antedependent models». *Applied Stochastic Models and Data Analysis*, 13, 279-287.

Núñez-Antón, V. & Woodworth, G. G. (1994). «Analysis of longitudinal data with unequally spaced observations and time-dependent correlated errors». *Biometrics*, 50, 445-456.

Núñez-Antón, V., El Barmi, H. & Zimmerman, D. L. (1995). «Una nota sobre matrices de covarianzas con inversas tridiagonales». *Estadística Española*, 37, 139, 201-215.

Rao, C. R. (1959). «Some problems involving linear hypotheses in multivariate analysis». *Biometrika*, 46, 49-58.

Reinsel, G. (1982). «Multivariate repeated-measurement or growth models with multivariate random-effects covariance structure». *Journal of the American Statistical Association*, 77, 190-195.

Rutter, C. M. & Elashoff, R. M. (1994). «Analysis of longitudinal data: random coefficient regression modelling». *Statistics in Medicine*, 13, 1211-1231.

SAS Institute Inc. (1996). *SAS/STAT Software: Changes and Enhancements through Release 6.11*. Cary, North Carolina: SAS Institute Inc.

Verbeke, G. & Molenberghs, G. (1997). *Linear Mixed Models in Practice. A SAS-Oriented Approach*. Lecture Notes in Statistics N°. 126. New York: Springer-Verlag.

Wolfinger, R. D. (1996). «Heterogeneous variance-covariance structures for repeated measures». *Journal of Agricultural, Biological, and Environmental Health*, 1(2), 205-230.

Zimmerman, D. L. & Núñez-Antón, V. (1997). «Structured antedependence models for longitudinal data». In *Modelling Longitudinal and Spatially Correlated Data. Methods, Applications, and Future Directions*. (T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russell-Cohen, W. G. Warren, and R. Wolfinger, Eds.) Lecture Notes in Statistics N°. 122, 63-76. New York: Springer-Verlag.

Zimmerman, D. L., Núñez-Antón, V. & El Barmi, H. (1998). «Computational aspects of likelihood-based estimation of first-order antedependence models». *Journal of Statistical Computation and Simulation*, 60, 67-84.

ENGLISH SUMMARY

MODELLING LONGITUDINAL DATA WITH NONSTATIONARY COVARIANCE STRUCTURES: RANDOM COEFFICIENTS MODELS VERSUS ALTERNATIVE MODELS

VICENTE NÚÑEZ-ANTÓN*
DALE L. ZIMMERMAN**

An important theme of longitudinal data analysis in the past two decades has been the development and use of explicit parametric models for the data's variance-covariance structure. However, nonstationary covariance structures had not been analyzed in detail for longitudinal data mainly because the existing applications did not require their use. There has been a large amount of recently proposed models but most of them are second-order stationary. A few, however, are flexible enough to accommodate nonstationarity, that is, nonconstant variances and/or correlations which are not only a function of the elapsed time between measurements. We study some of these proposed models and compare them to the random coefficients models, evaluating the relative strengths and limitations of each model, emphasizing when it is inappropriate or unlikely to be useful. We present two examples to illustrate the fitting and comparison of the models and to demonstrate that nonstationary longitudinal data can be modelled effectively and, in some cases, quite parsimoniously. In these examples the antedependence models generally prove to be superior and the random coefficients models prove to be inferior.

Keywords: Antedependence, Arima models, AIC, BIC, covariance structures, residual maximum likelihood, mixed models

AMS Classification (MSC 2000): 62J05, 62F10, 62P10

* Departamento de Econometría y Estadística (E.A. III). Universidad del País Vasco. Avenida Lehendakari Aguirre, 83. 48015 Bilbao. E-mail: vn@alcib.bs.ehu.es.

** Department of Statistics and Actuarial Science. The University of Iowa. Iowa City, Iowa 52242. Estados Unidos.

- Received April 2000.
- Accepted January 2001.

1. INTRODUCTION

An important theme of models in regression analysis and, specially, about their use to analyze longitudinal data, have been the parallel development and use of explicit parametric models for the data's variance-covariance structure and random coefficients models. Longitudinal data consists of measuring along time a given characteristic on each of the experimental units, normally allocated to different groups or treatments.

However, one of the main issues we want to mention has to do with the fact that random coefficients models have been usually regarded as models for regressions of response on time (and, possibly, on other covariates) that vary across subjects, instead of using the possibility of considering these models as a way to explain different phenomena in the within-subject structure of the data. In any case, random coefficients models can be used and, indeed, be very useful in longitudinal data analysis.

Compared to various analysis-of-variance methods, which ignore the covariance structure, and to the classical multivariate approach, which estimates the covariance matrix but imposes no structure on it (beyond that required for positive definiteness), parametric covariance modelling has several advantages. First, it generally results in more efficient estimation of parameters in the data's mean structure, which are usually of primary interest. Second, it yields more appropriate estimates of the standard errors of those estimated mean parameters. Third, in many cases it can deal effectively with missing data and with data for which the measurement times are not common across subjects. Finally, it can be employed even when the number of measurement times is large relative to the number of subjects.

Perhaps the most prevalent kind of covariance structure exhibited by longitudinal data is serial correlation, i.e. within-subject sample correlations that decrease as the elapsed time between measurements increases. The most popular parametric models for serial correlation are stationary autoregressive (AR) models and other parsimonious second-order stationary models. In these models, variances are constant over time and correlations between measurements equidistant in time are equal. When the sample variances and correlations comport with these assumptions, stationary models are generally very useful. When this is not so, however, the use of a stationary model is inadvisable. Instead, the researcher should consider a model flexible enough to accommodate nonstationarity.

If nonstationarity is manifested by nonconstant variances only, options for analysis include transforming the data to stabilize the variance or generalizing stationary models to allow for heterogeneous variances. Heterogeneous extensions of several stationary models are described and fit to data by Wolfinger (1996). These options may not be sufficient, however, when nonstationarity is also manifested by the correlations. There

are, nevertheless, several alternative models that are applicable to longitudinal data that exhibit nonstationarity in their correlations and variances. Perhaps the most obvious of these is the completely unstructured model of the classical multivariate approach. In many cases, however, the data's nonstationarity may possess an structure capable of being modelled with relatively few parameters, and to ignore this would forego the advantages of parametric modelling noted previously. One family of parametric models that can accommodate nonstationary correlations is a generalization of AR models known as antedependence models: general or unstructured and structured. Another, more well-known, nonstationary generalization of AR models are the autoregressive integrated moving average (ARIMA) models (e.g., see Diggle, 1990). One final possibility, we have already briefly mentioned, are the random coefficients models. These models are typically used for regressions on a continuous response variable that changes with time and that varies between individuals. Thus these models can actually explain certain types of nonstationarity.

Each of the models just described, besides the stationary models, has been proposed for use with nonstationary longitudinal data by at least one author, but such proposals have almost always occurred in isolation, apart from consideration of the other models. The general idea has been that, when there is a problem with the fitting of a given proposed model possibly due to the existing nonstationarity in the data, a random coefficients model is automatically adjusted instead. Consequently, the merits of each have never been systematically evaluated and virtually no guidelines exist as to their relative usefulness. In this article, we examine and compare these models. We consider their strengths and limitations, emphasizing when each is inappropriate or problematic and, specially, comparing these alternative models to the random coefficients model. We present two examples that illustrate the fitting and comparison of the models. The examples also demonstrate that nonstationary longitudinal data can be modelled effectively, and in some cases quite parsimoniously, with appropriate parametric models.

2. DATA SETS

Data from two longitudinal studies serve to motivate the consideration of nonstationary models. These data will also be used to illustrate the fitting and comparison of the different models for the within-subjects data structure. The *race data* consist of the «split» times for each of 80 competitors in each 10-km section of a 100-km race held in 1984 in the United Kingdom. Measurement times are evenly spaced and common to all subjects in the study. Thus, the data are rectangular. The objective of our analysis is to find a parsimonious model that adequately describes how competitor's performance on each 10-km section is related to the section number ($t = 1,2,\ldots,10$) and to the performance on previous sections.

The *cattle data* come from an experiment reported by Kenward (1987). Cattle receiving one of two intestinal parasite treatments, say A and B, were weighed 11 times over a 133-day period. Thirty animals received treatment A and thirty received treatment B. The first 10 measurements on each animal were made at two-week intervals and the final measurement was made one week after the tenth. No observations are missing. Although times are not equally spaced (due to the shorter interval before the last measurement), the measurement schedule is rectangular. We wish to study how cattle growth is affected by the treatments.

The sample variances and correlations corresponding to these longitudinal data sets show several interesting characteristics. First, the variances are not homogeneous, but instead tend to increase over time. Second, the correlations are all positive. Third, serial correlation appears to be present, as correlations within any given column tend to decrease towards zero (unless they are close to zero initially). Finally, correlations lagged the same number of observations apart are not constant. Rather, they tend to increase early in the study before levelling off, or in some cases (e.g. treatment B cattle data) decreasing slightly, later in the study.

A battery of power transformations was attempted for each data set with the aim of variance stabilization. These efforts met with only limited success, which is not surprising given that in several cases the variances do not appear to be smooth functions of the mean. Even in those cases where a transformation successfully stabilized the variance, the nonstationary behavior of the correlations persisted after transformation.

3. THE GENERAL MODEL

Suppose that repeated measurements of a continuous response variable are observed over time on each of m «subjects». Let $\mathbf{y}_i$ be the vector of n_i measurements on the ith subject and let $\mathbf{t}_i$ be the corresponding vector of measurement times. Suppose also that we observe a p-vector of covariates, $\mathbf{x}_{ij}$, associated with y_{ij}. Put $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_m')'$, $\mathbf{t} = (\mathbf{t}_1', \ldots, \mathbf{t}_m')'$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})'$, $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_m')'$, and $N = \sum_{i=1}^{m} n_i$. We refer to the set of measurement times in the study as the measurement schedule. We impose no restrictions on the measurement schedule; in general the measurement times may be unequally spaced within a subject and may differ across subjects. If measurement times are common across subjects, we call the measurement schedule *rectangular*. Thus, the measurement schedule of the race and the cattle data is rectangular. The extent of the measurement schedule's departure from rectangularity has important implications for modelling the covariance structure, as will be seen subsequently.

Several general modelling assumptions now provide a framework for parametric modelling of the covariance structure. These assumptions yield the model $\mathbf{y} \sim MVN(\mathbf{X}\beta, \Sigma)$, where β is a p-vector of fixed, unknown, and typically unrestricted parameters, and

the $N \times N$ unknown covariance matrix Σ is block diagonal, with non-zero blocks $\mathbf{V}_i = \text{var}(\mathbf{y}_i)$. In the case of a rectangular measurement schedule, the $\mathbf{V}_i$ are all equal. The parametric modelling approach now proceeds with the postulation of a parametric model for Σ: $\mathbf{y} \sim MVN(\mathbf{X}\beta, \Sigma(\mathbf{t}, \theta))$, where θ is a q-vector of unknown parameters, restricted to a parameter space Θ which is either the set of all θ-vectors for which Σ is positive definite or some subset of that set. Note that the elements of Σ are permitted to be functions of $\mathbf{t}$ but not of $\mathbf{X}$. Note further that Σ is positive definite if and only if $\mathbf{V}_i$ is positive definite for every i. Estimation of the parameters β and θ of this model is carried out by maximum likelihood (ML) or residual maximum likelihood (REML), which often must be implemented using numerical optimization routines. Further specifics on the fitting of the models and on comparing the fitted models are deferred to the examples. We now briefly describe the random coefficients and the antedependence models.

4. PARAMETRIC MODELS FOR THE WITHIN-SUBJECTS COVARIANCE STRUCTURE

Random Coefficients (RC) Models

A rather general random coefficients model (Laird and Ware, 1982), is $\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i \quad (i = 1, \ldots, m)$, where the $\mathbf{Z}_i$ are specified matrices, the $\mathbf{u}_i$ are vectors of random coefficients distributed independently as $MVN(\mathbf{0}, \mathbf{G}_i)$, the $\mathbf{G}_i$ are positive definite but otherwise unstructured matrices, and the $\mathbf{e}_i$ are distributed independently (of the $\mathbf{u}_i$ and of each other) as $MVN(\mathbf{0}, \sigma^2\mathbf{I}_{n_i})$. Typically the $\mathbf{G}_i$ are assumed to be equal; hence the covariance matrix of $\mathbf{y}_i$ is taken as $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \sigma^2\mathbf{I}_{n_i}$. Special cases include the linear random coefficients (RCL) and quadratic random coefficients (RCQ) models. In the linear case, $\mathbf{Z}_i = [\mathbf{1}_{n_i}, \mathbf{t}_i]$. In the quadratic case, $\mathbf{Z}_i = [\mathbf{1}_{n_i}, \mathbf{t}_i, (t_{i1}^2, t_{i2}^2, \ldots, t_{in_i}^2)']$.

Random coefficients models have often been considered as distinct from parametric covariance models, probably because the orogenesis of the covariance structure is typically a consideration of regressions that vary across subjects rather than a consideration of within-subject similarity. Nevertheless, they yield parametric covariance structures that generally have nonconstant variances and nonstationary correlations, a fact that does not appear to be widely appreciated.

Antedependence Models

The unstructured antedependence model of order s [UAD(s)] model is defined as: $y_1 = \mathbf{x}_1'\beta + \varepsilon_1$ and $y_j = \mathbf{x}_j'\beta + \sum_{k=1}^{s^*} \phi_{jk}(y_{j-k} - \mathbf{x}_{j-k}'\beta) + \varepsilon_j \; (j = 2, \ldots, n)$, where $s^* = \min(s, j-1)$, the ε_j's are independent normal random variables with zero means and possibly time-dependent variances $\sigma_j^2 > 0$, and the autoregressive coefficients $\{\phi_{jk}\}$ are completely unrestricted parameters. By the term «unstructured,» we mean that the parame-

ters ϕ_{jk} and σ_j^2 cannot be expressed as functions of a smaller number of parameters. The UAD(s) model generalizes the stationary AR(s) model by allowing the innovation variances and autoregressive coefficients to be time-varying. This greater generality makes UAD models useful for situations in which measurement times are unequally spaced or there is clear evidence of nonstationarity in the data's correlation structure. The cost of this extra flexibility is an increase in the number of parameters, which are $O(n)$ rather than $O(1)$. Furthermore, because the number of parameters increases with the number of distinct measurement times, rectangularity or approximate rectangularity is a practical necessity. The equation used to define an UAD(s) model is both a response equation specification and an autoregressive specification. The corresponding variance-correlation specification is interesting in its own right. In it, response variances and correlations between observations lagged s or less observations apart are arbitrary (subject to positive definiteness constraints), but correlations between observations lagged more than s observations apart are completely determined by those corresponding to lags s or less.

Although the UAD(s) model is more flexible than more specialized AD models, such as stationary AR models, its drawback is that sometimes it may have too many parameters to be useful. There is a way to reduce the number of parameters using the structured AD (SAD) models introduced by Zimmerman and Núñez-Antón (1997). In these models, the autoregressive coefficients or correlations (depending on whether an autoregressive or variance-correlation specification is used) of the UAD(s) model follow a Box-Cox power function of time, and the innovation variances or response variances (again depending on the specification) are polynomial or step functions of time. For example, the variance-correlation specification of this UAD(s) model is given by $v_{jj} = \sigma^2 g(t_j;\psi)$ $(j = 1,\ldots,n)$ and $\rho_{j,j-k} = \rho_k^{f(t_j;\lambda_k)-f(t_{j-k};\lambda_k)}$ $(j = k+1,\ldots,n;\ k = 1,\ldots,s)$, where g, to be most useful in practice, is a function of relatively few parameters (e.g. a low-order polynomial function), and

$$f(t;\lambda) = \begin{cases} (t^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log t & \text{if } \lambda = 0. \end{cases}$$

Exponentiation of the ρ_k's in by the Box-Cox form of f given above prescribes that the correlations on the kth-subdiagonal of the correlation matrix are monotone increasing if $\lambda_k < 1$, monotone decreasing if $\lambda_k > 1$, or constant if $\lambda_k = 1$ $(k = 1,\ldots,s)$. From another point of view, f effects a nonlinear deformation upon the time axis such that correlations between measurements equidistant in the deformed scale are constant.

5. EXAMPLES: FITTING OF THE MODELS

We fitted saturated models for the mean in each of the two data sets and the random coefficients models and several alternative models for the within-subject variance-

covariance structure. For the race data set, if we base our selection on AIC, the best fitting model was the UAD(3) with 34 parameters. However, if we use BIC as the selection criteria, the best fitting models were the ARH(1) and UAD(1). After carrying out several restricted likelihood ratio tests, the selected model was the UAD(3). For the cattle data set, using AIC, BIC and several restricted likelihood ratio tests led us to select as best fitting models for the groups A and B were, respectively, the UAD(2) and the UAD(3) models. For this data set, there was a significant group difference, as indicated by the corresponding test.

6. CONCLUSIONS

We have compared the RC models to several alternative models in the context of longitudinal data and when nonstationarity is present. First, there is the tradeoff between model flexibility and parsimony. The unstructured model with its $O(n^2)$ parameters is the most flexible and least parsimonious. The UAD model has $O(n)$ parameters and is the next most flexible. The SAD, ARIMA, and RC models all are highly structured, with $O(1)$ parameters, and thus are not as flexible as the others. Second, irregular spacing of measurements and non-rectangularity of the measurement schedule present no problems for the SAD and RC models, but one or both of these may require special care for the UN, UAD, and ARIMA models. A final comparison pertains to the existence of widely available software for fitting the models. The UN and RC models, and certain low-order UAD and ARIMA models, have the advantage on this score, for they can be fitted in PROC MIXED. Of all the parametric covariance structures which have been proposed for longitudinal data, stationary autoregressive and random coefficient models seem to receive the most attention; antedependence models, in contrast, get very little press. In our examples, however, stationary models generally did not fit as well as an antedependence model of some kind, and random coefficients models were not competitive at all. Thus, in these examples at least, it appears that some kind of antedependence model strikes the right balance between model flexibility and parsimony. Partly as a result of this, and partly due to their nice properties, we believe that antedependence models should be more routinely fit to longitudinal data exhibiting nonstationarity than they presently are.

QÜESTIIÓ, vol. 25, 2, p. 263-284, 2001

EL SESGO CONDICIONADO EN EL ANÁLISIS DE INFLUENCIA: UNA REVISIÓN

J. M. MUÑOZ-PICHARDO
J. L. MORENO-REBOLLO
T. GÓMEZ-GÓMEZ
A. ENGUIX-GONZÁLEZ
Universidad de Sevilla*

El sesgo condicionado se ha propuesto como diagnóstico de influencia en distintos modelos y técnicas estadísticas. Tratando de recoger una visión global de la utilidad del concepto, en este trabajo se hace una revisión general del mismo relacionándolo con la curva de sensibilidad y la curva de influencia muestral. Además, se señalan posibles líneas de trabajo que permitirán abordar el análisis de la influencia a través de este enfoque en una gran variedad de técnicas estadísticas.

Conditional bias in influence analysis: a review

Palabras clave: Análisis de influencia, modelos lineales, componentes principales, muestreo en poblaciones finitas, sesgo condicionado

Clasificación AMS (MSC 2000): 62J20, 62H25, 62D05

* Universidad de Sevilla. Facultad de Matemáticas. Departamento de Estadística e Investigación Operativa. Avda. Reina Mercedes s/n. 41012 Sevilla.

– Recibido en octubre de 2000.
– Aceptado en abril de 2001.

1. INTRODUCCIÓN

El objetivo de todo análisis estadístico es obtener conclusiones fiables a partir de los datos resultantes de una experimentación. Por tanto, la fiabilidad de las observaciones del proceso es de especial interés, ya que el análisis se realiza sobre codificaciones del fenómeno natural en estudio y las técnicas estadísticas que se apliquen pueden verse fuertemente afectadas por algunas de las observaciones realizadas. Este problema ha originado un gran número de métodos enfocados, bien al desarrollo de nuevas técnicas que no se vean influenciadas excesivamente por la modelización del fenómeno natural, bien al análisis de la calidad de los datos, o bien al estudio de aquellas observaciones que afectan considerablemente a los resultados del análisis. En este tercer enfoque se han propuesto un conjunto de métodos englobados en lo que genéricamente se conoce como el **Análisis de Influencia**.

La gran mayoría de las técnicas propuestas para el análisis de influencia están basadas en la aproximación genérica realizada por Cook y Weisberg (1982): para medir el efecto que sobre un aspecto de interés del análisis (estadístico, estimación,...) tiene una observación o un conjunto de ellas, se introducen pequeñas perturbaciones, que las afectan de alguna manera, y se cuantifica el cambio producido. Es decir, las técnicas surgen de conjugar adecuadamente *perturbación* del modelo y *comparación* de resultados. Cook (1987) trata de unificar el problema bajo la siguiente formulación general:

> Sea un conjunto de datos D, un modelo M postulado a priori, un resultado $R(D,M)$ seleccionado de una síntesis de los datos y el modelo, y sea $\mathbf{w}$ un vector de perturbaciones, perteneciente a un conjunto Ω de perturbaciones relevantes, siendo $M(\mathbf{w})$ el modelo perturbado, de forma que
>
> $$\exists \mathbf{w}_0 \in \Omega \ / \ M \approx M(\mathbf{w}_0) \, .$$
>
> Así, el Análisis de Influencia consiste en comparar los resultados $R(D,M(\mathbf{w}))$ y $R(D,M)$.

En consecuencia, son cuestiones claves la elección del esquema de perturbación y el método de comparación. De cada forma distinta de conjugar el binomio *perturbación-comparación* resultará una técnica para el análisis de influencia.

En cuanto al primer elemento del binomio, el más utilizado es el esquema de perturbación de la omisión de las observaciones a las que se le pretende evaluar su impacto. No obstante, la anterior formulación general del problema permite otros enfoques. Diversos autores, como Lawrance (1991) y Escobar y Meeker (1992), han propuesto clasificaciones de los distintos tipos de perturbaciones: *perturbaciones en los datos, perturbaciones en las hipótesis del modelo, perturbaciones en ponderaciones de casos.*

Este último esquema es utilizado en el denominado Análisis de Influencia Local (Cook (1986)).

Una vez seleccionado el esquema de perturbación, es necesario elegir el estadístico o resultado $R(D,M)$ del análisis y el método de comparación de resultados. La primera cuestión queda a criterio del investigador, en función del objetivo principal del análisis. Como ilustración de la gran variedad de métodos de comparación propuestos, pueden citarse los siguientes: *Curva de influencia muestral* (Cook y Weisberg (1982), Belsley *y otros* (1980)); *Razón de volúmenes de regiones de confianza*, cuando el resultado $R(D,M)$ seleccionado es una región de confianza (Andrews y Pregibon (1978)); *Razón entre determinantes*, para la estimación de matrices de varianzas y covarianzas (Barnett y Lewis (1994), Belsley *y otros* (1980)); *Distancias entre las distribuciones muestrales de los estimadores*, como la distancia de Rao (Muñoz-Pichardo y Fernández-Ponce (1997)); *Desplazamiento de verosimilitud*, como método de comparación respecto a los contornos de la log-verosimilitud del modelo postulado (Cook y Weisberg (1982), Cook *y otros* (1988)), o bien del modelo perturbado (Billor y Loynes (1993)). Brown y Lawrance (2000), a través de la verosimilitud, estudian este tópico en regresión múltiple sobre una amplia gama de esquemas de perturbación, extendiendo el estudio de la influencia a diferentes tests de hipótesis de interés en este modelo.

Muñoz-Pichardo *y otros* (1995) proponen el concepto de sesgo condicionado como enfoque genérico, válido en un gran número de las técnicas y modelos estadísticos. Tras la introducción anterior sobre el problema de la influencia, en este trabajo se pretende realizar una revisión de las distintas aplicaciones del sesgo condicionado (s.c.), así como su relación con otros conceptos estadísticos.

En la Sección 2, se recoge la relación con la curva de influencia muestral y la curva de sensibilidad. En la Sección 3, se recoge la aplicación del mismo a los modelos lineales, aportándose una estimación de la varianza del estimador del sesgo condicionado. A continuación, se aplica al análisis de componentes principales (Sección 4) y en el muestreo en poblaciones finitas (Sección 5). Finalmente, se concluye señalando otros campos en los que se puede abordar el análisis de influencia bajo la perspectiva del concepto de s.c..

2. EL CONCEPTO DE SESGO CONDICIONADO Y EL ANÁLISIS DE INFLUENCIA

Partiendo del Lema de Descomposición de Efron y Stein (1984), que expresa un estadístico T sobre una muestra aleatoria simple como una suma finita, cuyos términos son funciones de las esperanzas condicionadas de T dadas las observaciones muestrales, Muñoz-Pichardo *y otros* (1995) definen el concepto de *sesgo condicionado* y proponen su aplicación en el Análisis de Influencia. A continuación, se recoge el lema citado,

la definición del concepto de sesgo condicionado y la justificación de su aplicación al problema de la influencia.

Lema 1. (Efron y Stein). *Toda variable aleatoria $S(X_1, X_2, \ldots, X_n)$ función de n variables aleatorias independientes $X_1, X_2, \ldots, X_n$ puede expresarse como*

$$S(X_1 \ldots X_n) = E[S] + \sum_{i=1}^{n} A_i[X_i; S] + \sum\sum_{1 \le i < j \le n} B_{ij}[X_i, X_j; S] + \\ + \sum\sum\sum_{1 \le i < j < k \le n} C_{ijk}[X_i, X_j, X_k; S] + \cdots + H[X_1 \ldots X_n; S],$$

donde las $2^n - 1$ variables aleatorias del miembro derecho de la expresión tienen esperanza nula y están mutuamente incorreladas, siendo:

$$\begin{aligned} A_i[x_i; S] &= E[S \,|X_i = x_i] - E[S], \quad \text{«}i\text{-ésimo efecto medio»}, \\ B_{ij}[x_i, x_j; S] &= E[S \,|X_i = x_i,\, X_j = x_j] - E[S \,|X_i = x_i] - \\ &\quad - E[S \,|X_j = x_j] + E[S], \quad \text{«}(i,j)\text{-interacción de } 2^o \text{ orden»}, \end{aligned}$$

y así sucesivamente.

Dado un estadístico $T_n = T_n(Y_1 \ldots Y_n)$ definido sobre una muestra aleatoria, $Y_1 \ldots Y_n$, y su descomposición en términos del resultado anterior, se observa que $A_i[y_i; T_n]$ cuantifica la desviación que la realización muestral $Y_i = y_i$ provoca en el valor esperado de T_n. Por tanto, puede considerarse como una medida de la influencia que dicha realización muestral ejerce sobre T_n. Así, Muñoz-Pichardo *y otros* (1995) proponen la siguiente definición:

Definición 1. *Sea $Y_1 \ldots Y_n$ una m.a. de una v.a. Y, sea $T_n = T_n(Y_1 \ldots Y_n)$ un estadístico y sea $y_1 \ldots y_n$ una realización de la muestra. El sesgo condicionado (s.c.) de T_n dada la i-ésima observación se define como*

$$\mathcal{S}(y_i; T_n) = E[T_n \,|Y_i = y_i] - E[T_n].$$

Sobre esta definición se pueden realizar diversas consideraciones. En primer lugar, el s.c. depende de la distribución del estadístico T_n y del valor observado y_i. Por tanto, al contrario de la curva de influencia muestral, que se define a partir de una realización de toda la muestra, el s.c. mide la influencia del valor observado sobre el estadístico, en términos de la esperanza de su distribución muestral, y por tanto, es independiente de cualquier realización muestral concreta de los restantes elementos de la muestra. Por otra parte, no presupone ningún esquema de perturbación, salvo el condicionamiento

previo impuesto por el conocimiento de la observación bajo estudio. Además, es un parámetro cuya dimensión coincide con la dimensión del estadístico, al igual que ocurre con la curva de influencia muestral, por tanto, para cuantificarlo se ha de proceder a utilizar normas semejantes a las aplicadas sobre ésta. Finalmente, su dependencia de la distribución de T_n, puede provocar que el desconocimiento de algunos parámetros de la misma implique la necesidad de su estimación.

La definición 1 puede generalizarse como sigue:

Definición 2. *En las condiciones de la definición 1, el s.c. de T_n dado el conjunto de observaciones $\{y_{i_1} \ldots y_{i_m}\}$ se define como*

$$\mathcal{S}(y_{i_1} \ldots y_{i_m}; T_n) = E[T_n | Y_{i_1} = y_{i_1} \ldots Y_{i_m} = y_{i_m}] - E[T_n]. \tag{1}$$

En consecuencia, (1) puede considerarse como una medida de la influencia conjunta de las observaciones $\{y_{i_1} \ldots y_{i_m}\}$ sobre T_n.

De las dos definiciones anteriores se obtienen las siguientes expresiones:

$$B_{ij}[y_i, y_j; T_n] = \mathcal{S}(y_i, y_j; T_n) - \{\mathcal{S}(y_i; T_n) + \mathcal{S}(y_j; T_n)\}, \tag{2}$$

$$\begin{aligned} C_{ijk}[y_i, y_j, y_k; T_n] = {} & \mathcal{S}(y_i, y_j, y_k; T_n) - \\ & - \{\mathcal{S}(y_i, y_j; T_n) + \mathcal{S}(y_i, y_k; T_n) + \mathcal{S}(y_j, y_k; T_n)\} + \\ & + \{\mathcal{S}(y_i; T_n) + \mathcal{S}(y_j; T_n) + \mathcal{S}(y_k; T_n)\}. \end{aligned} \tag{3}$$

Por tanto, el efecto interacción de segundo orden debido a las observaciones (y_i, y_j), según (2), es la influencia conjunta de ambas observaciones sobre T_n menos la influencia individual de cada una de ellas. Análogo comentario se puede realizar observando la expresión (3).

Otras propiedades del s.c. son las relaciones con los conceptos de curva de sensibilidad y curva de influencia muestral. Hampel (1974), con el objetivo de estudiar la conducta infinitesimal de funcionales estadísticos, propone el concepto de función influencia sobre un funcional definido sobre el espacio de las distribuciones de probabilidad. A partir de dicha definición se han propuesto diversas versiones muestrales. Una de las de mayor interés es la *curva de sensibilidad,* propuesta por Tukey (1970): dada $Y_1 \ldots Y_{n-1}$ una muestra aleatoria de una v.a. Y, la curva de sensibilidad (CS) de T_n asociada a una realización muestral $y_1 \ldots y_{n-1}$ se define como:

$$CS_{n-1}(y; T_n) = n\{T_n(y_1, \ldots, y_{n-1}, y) - T_{n-1}(y_1, \ldots, y_{n-1})\}.$$

Cuando el objetivo es el estudio de la influencia de observaciones individuales sobre algún estadístico, se propone otra versión muestral, la *curva de influencia muestral*

(*CIM*): dada $Y_1 \dots Y_n$ una muestra aleatoria de una v.a. Y, y T_n un estadístico definido sobre la muestra, la curva de influencia muestral de un estadístico T_n asociada a la i-ésima observación de una realización muestral $y_1, \dots, y_n$, viene dada por

$$CIM_i(T_n) = -(n-1)\left\{T_{(i)} - T_n(y_1, \dots, y_n)\right\},$$

siendo $T_{(i)} = T_{n-1}(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$ el estadístico obtenido bajo la omisión de la i-ésima observación.

Generalizando la definición anterior, se define la curva de influencia muestral de T_n asociada a un conjunto de m observaciones B como sigue:

$$CIM_B(T_n) = -(n-m)\left\{T_{(B)} - T_n(y_1, \dots, y_n)\right\},$$

siendo $T_{(B)}$ el estadístico obtenido bajo la omisión de las observaciones contenidas en B.

La curva de sensibilidad puede considerarse como función aleatoria, asociada a la muestra $Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n$:

$$CS_{n-1}(y; T_n) = n\left\{T_n(Y_1 \dots Y_{i-1}, y, Y_{i+1} \dots Y_n) - T_{n-1}(Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n)\right\}.$$

Asimismo, la curva de influencia muestral puede considerarse como función aleatoria, asociada a la muestra $Y_1 \dots Y_n$:

$$CIM_i(T_n) = -(n-1)\left\{T_{n-1}(Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n) - T_n(Y_1 \dots Y_n)\right\}.$$

Bajo estas consideraciones, se pueden enunciar los siguientes resultados:

Teorema 1. *Sea $Y_1 \dots Y_n$ una muestra aleatoria, sea $T_n = T_n(Y_1 \dots Y_n)$ un estadístico y $T_{(i)} = T_{n-1}(Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n)$, y supóngase que $E[T_n] = E[T_{(i)}]$.*

1. *La curva de sensibilidad, asociada a la muestra $Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n$, verifica*

$$\frac{1}{n} E\left[CS_{n-1}(y_i; T_n)\right] = \mathcal{S}(y_i; T_n). \tag{4}$$

2. *La curva de influencia muestral, asociada a la muestra $Y_1 \dots Y_n$, verifica*

$$\frac{1}{n-1} E\left[CIM_i(T_n) \,|\, Y_i = y_i\right] = \mathcal{S}(y_i; T_n). \tag{5}$$

La igualdad (4), además de la relación entre los dos conceptos, viene a profundizar en el interés del s.c. como herramienta útil para el análisis de influencia, y la igualdad (5)

puede para fundamentar teóricamente el concepto de curva de influencia muestral como esquema de comparación válido para el análisis de influencia.

Por ello, Muñoz-Pichardo *y otros* (1995) proponen el siguiente estimador insesgado del s.c.

$$\widehat{\mathcal{S}}(y_i;T) = T_n(Y_1 \ldots Y_{i-1}, y_i, Y_{i+1} \ldots Y_n) - T_{n-1}(Y_1 \ldots Y_{i-1}\ Y_{i+1} \ldots Y_n) = T_n - T_{(i)},$$

y en general,

$$\widehat{\mathcal{S}}(y_{i_1} \ldots y_{i_m};T_n) = T_n - T_{(i_1,\ldots,i_m)},$$

que puede denominarse *sesgo condicionado estimado* (s.c.e.). Conviene hacer notar que en muchas ocasiones para el cálculo del s.c. y su estimador no es necesario conocer la distribución subyacente al modelo, algo que le ocurre a otras técnicas de diagnóstico de influencia como el desplazamiento de verosimilitud o la distancia de Rao.

Muñoz-Pichardo *y otros* (1998) aplican este concepto al Análisis de Influencia Local, en particular en el Modelo Lineal General. Un planteamiento genérico de dicha aplicación, siguiendo las pautas marcadas por Cook en su descripción del problema de la influencia, puede ser el siguiente:

> Sea un modelo M postulado a priori, sea $Y_1 \ldots Y_n$ una muestra aleatoria de la(s) variable(s) que intervienen en el mismo, sea un estadístico $T_n(Y_1 \ldots Y_n;M)$, $\mathbf{w}$ un vector perteneciente a un conjunto Ω de perturbaciones relevantes, siendo $M(\mathbf{w})$ el modelo perturbado, de forma que
>
> $$\exists \mathbf{w}_0 \in \Omega \ / \ M \approx M(\mathbf{w}_0) \ \text{ y } \ T_n(Y_1 \ldots Y_n;M) \approx T_n(Y_1 \ldots Y_n;M(\mathbf{w}_0)),$$
>
> y sea y_i una realización muestral de la i-ésima componente de la muestra aleatoria. El estudio de la función de $\mathbf{w} \in \Omega$:
>
> $$\mathcal{S}_w(y_i;T_n) = E_{M(\mathbf{w})}[T_n \ | Y_i = y_i] - E_{M(\mathbf{w})}[T_n],$$
>
> en un entorno de $\mathbf{w}_0$ permitirá realizar el análisis de influencia local que la realización muestral $Y_i = y_i$ ejerce sobre T_n, siendo $E_{M(\mathbf{w})}[-]$ la esperanza en el modelo perturbado $M(\mathbf{w})$.

Para este planteamiento general, son válidas las consideraciones realizadas anteriormente sobre el concepto de s.c.. En particular, en ocasiones se tendrá que buscar un estimador, pudiéndose considerar el estimador insesgado:

$$\widehat{\mathcal{S}}_w(y_i;T_n) = T_n(Y_1 \ldots Y_{i-1}, y_i, Y_{i+1} \ldots Y_n;M(\mathbf{w})) - T_{n-1}(Y_1 \ldots Y_{i-1}\ Y_{i+1} \ldots Y_n;M(\mathbf{w})).$$

Para ilustrar la utilidad de lo expresado hasta ahora, a continuación, se recoge una aplicación simple: el s.c. y diagnósticos de influencia sobre el estadístico media muestral.

2.1. Una aplicación ilustrativa: estadístico media muestral

Sea $Y_1 \ldots Y_n$ una muestra aleatoria de una v.a. Y, con $E[Y] = \mu$ y $Var[Y] = \sigma^2$, y consideremos el estimador BLUE de μ, el estadístico media muestral $T_n = \overline{Y}$. Puede obtenerse fácilmente el s.c. de dicho estadístico dada la i-ésima realización muestral y_i:

$$\mathcal{S}(y_i;\overline{Y}) = \frac{1}{n}(y_i - \mu).$$

Por tanto, como lógicamente se podía intuir, la influencia de una observación será mayor cuanto más diste de la media poblacional. Análogamente,

$$\mathcal{S}(y_{i_1} \ldots y_{i_m};\overline{Y}) = \frac{1}{n}\sum_{j=1}^{m}(y_{i_j} - \mu).$$

En ambas expresiones se observa su dependencia del parámetro μ, por lo que es necesario obtener las correspondientes estimaciones:

$$\widehat{\mathcal{S}}(y_i;\overline{Y}) = \frac{1}{n-1}(y_i - \overline{Y}) \tag{6}$$

y

$$\widehat{\mathcal{S}}(y_{i_1} \ldots y_{i_m};\overline{Y}) = \frac{1}{n-m}\sum_{j=1}^{m}(y_{i_j} - \overline{Y}), \tag{7}$$

respectivamente. Para evitar el problema del signo, a efectos prácticos, como medidas de influencia pueden considerarse los cuadrados de las expresiones (6) y (7):

$$S_i^2 = \left(\frac{y_i - \overline{Y}}{n-1}\right)^2 \quad , \quad S_{i_1 \ldots i_m}^2 = \left\{\frac{1}{n-m}\sum_{j=1}^{m}(y_{i_j} - \overline{Y})\right\}^2.$$

Con el objetivo de estudiar la influencia local sobre el BLUE de μ, se puede considerar el modelo perturbado determinado por:

$$\begin{aligned} E[Y_k] &= \mu, \quad \forall k \\ Var[Y_k] &= \sigma^2, \quad \forall k \neq i\,;\, Var[Y_i] = \sigma^2/w, \quad w > 0. \end{aligned} \tag{8}$$

El BLUE de μ en el modelo especificado en (8) viene dado por:

$$\overline{Y}_w = \frac{1}{n-1+w}\sum_{k \neq i} Y_k + \frac{w}{n-1+w} Y_i,$$

y por tanto,

$$\mathcal{S}_w^{(i)}(y_i;\overline{Y}_w) = \frac{w}{n-1+w}(y_i - \mu) = \frac{n\,w}{n-1+w}\mathcal{S}(y_i;\overline{Y}).$$

(El superíndice (i) se utiliza para indicar que sólo la i-ésima observación es perturbada).

Se puede observar que el s.c. en el modelo perturbado es proporcional al s.c. en el modelo postulado, siendo la razón de proporcionalidad:

$$\alpha(w) = \frac{n\,w}{n-1+w}\,.$$

En cuanto al estimador del s.c., se obtiene:

$$(9) \qquad \widehat{\mathcal{S}}_w^{(i)}\left(y_i;\overline{Y}_w\right) = \frac{1}{n-1}\frac{n\,w}{n-1+w}\left(y_i-\overline{Y}\right) = \frac{n\,w}{n-1+w}\widehat{\mathcal{S}}\left(y_i;\overline{Y}\right)\,.$$

Considerando nuevamente su cuadrado para evitar el signo, se puede utilizar con medida de influencia local la función de w:

$$S_i^2(w) = S_i^2\,\{\alpha(w)\}^2\,.$$

La siguiente igualdad redunda en la relevancia de la razón de proporcionalidad $\alpha(w)$:

$$(10) \qquad \widehat{\mathcal{S}}_w^{(i)}\left(y_j;\overline{Y}_w\right) = \alpha(w)\widehat{\mathcal{S}}\left(y_j;\overline{Y}\right) + (1-\alpha(w))\widehat{\mathcal{S}}\left(y_j;\overline{Y}_{(i)}\right), \quad j=1\ldots n.$$

Para $w \in (0,1)$, dicha razón verifica $0 < \alpha(w) < 1$, y en consecuencia, $\widehat{\mathcal{S}}_w^{(i)}\left(y_j;\overline{Y}_w\right)$ es combinación lineal convexa entre $\widehat{\mathcal{S}}\left(y_j;\overline{Y}\right)$ y $\widehat{\mathcal{S}}\left(y_j;\overline{Y}_{(i)}\right)$, donde los coeficientes no dependen de las observaciones i-ésima ni j-ésima, tan sólo del tamaño muestral y del factor de perturbación considerado.

Dado el papel relevante que juega esta razón de proporcionalidad, Muñoz-Pichardo *y otros* (1998) la denominan **potencial de influencia local de la i-observación**, interpretándola como la proporción de influencia de cualquier observación sobre el BLUE de μ en el modelo perturbado (8), explicada por la influencia que ejerce dicha observación sobre el BLUE de μ en el modelo postulado.

3. SESGO CONDICIONADO EN EL MODELO LINEAL GENERAL MULTIVARIANTE

En esta sección estudiamos el s.c. sobre el Modelo Lineal General Multivariante (MLGM), con objeto de tratar de forma unificada el análisis de influencia en los diversos modelos que se obtienen como casos particulares de él: Modelos de Regresión (múltiple y multivariante), Modelos de Análisis de la Varianza (univariante y multivariante), Modelos de Análisis de la Covarianza (univariante y multivariante), Análisis de Perfiles, Análisis de Medidas Repetidas, etc.

El modelo MLGM está definido por la igualdad matricial $\mathbf{Y} = \mathbf{XB}+\mathcal{E}$, donde $\mathbf{Y}$ es la matriz $(n \times q)$ de respuestas, $\mathbf{X}$ es una matriz $(n \times p)$ conocida, con rango $r(r \leq p \leq n)$, $\mathbf{B}$ es una matriz $(p \times q)$ de parámetros desconocidos y $\mathcal{E} = (\varepsilon_{ik})$ es la matriz de perturbaciones aleatorias que verifica:

$$\begin{aligned} E(\varepsilon_{ik}) &= 0, & 1 \leq i \leq n, 1 \leq k \leq q, \\ Cov(\varepsilon_{ik}, \varepsilon_{js}) &= \sigma_{ks}\, \delta_{ij}, & 1 \leq i, j \leq n, 1 \leq k, s \leq q, \end{aligned}$$

siendo δ_{ij} la delta de Kronecker.

Si $\mathbf{\Lambda B}$ es una función linealmente estimable (f.l.e.), siendo $\mathbf{\Lambda}$ una matriz de dimensiones $d \times q$, de rango d, y $\widehat{\mathbf{B}}$ cualquier estimador de mínimos cuadrados, Muñoz-Pichardo *y otros* (2000) obtienen el s.c. del BLUE, $\mathbf{\Lambda}\widehat{\mathbf{B}}$, asociado a la i-ésima observación de las respuestas $\mathbf{y}_i$,

$$\mathcal{S}(\mathbf{y}_i; \mathbf{\Lambda}\widehat{\mathbf{B}}) = \mathbf{\Lambda S}^{-}\mathbf{x}_i \left(\mathbf{y}_i' - \mathbf{x}_i'\mathbf{B}\right),$$

donde $\mathbf{x}_i'$ e $\mathbf{y}_i'$ son las i-ésimas filas de $\mathbf{X}$ e $\mathbf{Y}$, respectivamente, $\mathbf{S} = \mathbf{X}'\mathbf{X}$ y $\mathbf{S}^{-}$ es una inversa generalizada de $\mathbf{S}$. El s.c.e. puede expresarse en los siguientes términos:

$$\widehat{\mathcal{S}}(\mathbf{y}_i; \mathbf{\Lambda}\widehat{\mathbf{B}}) = \frac{1}{1 - v_{ii}} \mathbf{\Lambda S}^{-}\mathbf{x}_i \mathbf{e}_i', \tag{11}$$

donde $\mathbf{e}_i'$ es la i-ésima fila de la matriz de residuos $\mathbf{E} = \mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}} = (\mathbf{I} - \mathbf{V})\mathbf{Y}$, siendo v_{ii} el elemento i-ésimo diagonal de la matriz de predicción $\mathbf{V} = \mathbf{X}'\mathbf{S}^{-}\mathbf{X}$.

En el modelo univariante $(q = 1)$, $\underline{Y} = \mathbf{X}\beta + \varepsilon$, (MLG), donde $E[\varepsilon] = \mathbf{0}$, $Var[\varepsilon] = \sigma^2\mathbf{I}_n$, si $\mathbf{\lambda}'\beta$ es una f.l.e., el s.c.e. de $\mathbf{\lambda}'\widehat{\beta}$, dada la i-ésima observación, viene dado por (Muñoz-Pichardo *y otros* (1995))

$$\widehat{\mathcal{S}}(y_i; \mathbf{\lambda}'\widehat{\beta}) = \frac{1}{1 - v_{ii}} \mathbf{\lambda}'\mathbf{S}^{-}\mathbf{x}_i (y_i - \mathbf{x}_i'\widehat{\beta}).$$

Con objeto de analizar la precisión de dicho estimador, a continuación se obtiene su varianza y una estimación de la misma.

Teorema 2. *En el MLG, se verifica*

$$Var\left[\widehat{\mathcal{S}}(y_i; \mathbf{\lambda}'\widehat{\beta}) \,|\, Y_i = y_i\right] = \frac{v_{ii}}{1 - v_{ii}} (\mathbf{\lambda}'\mathbf{S}^{-}\mathbf{x}_i)^2 \sigma^2 .$$

La igualdad recogida en el teorema se obtiene de forma directa, dado que

$$\begin{aligned} Var\left[\widehat{\mathcal{S}}(y_i; \mathbf{\lambda}'\widehat{\beta}) \,|\, Y_i = y_i\right] &= \frac{(\mathbf{\lambda}'\mathbf{S}^{-}\mathbf{x}_i)^2}{(1 - v_{ii})^2} Var\left[\mathbf{x}_i'\widehat{\beta} \,|\, Y_i = y_i\right] \\ &= \frac{(\mathbf{\lambda}'\mathbf{S}^{-}\mathbf{x}_i)^2}{(1 - v_{ii})^2} Var\left[\mathbf{x}_i'\mathbf{S}^{-}\mathbf{X}_{(i)}'\underline{Y}_{(i)}\right]. \end{aligned}$$

En consecuencia, en el MLG un estimador de la varianza del s.c.e. viene dado por

$$\widehat{Var}\left[\widehat{\mathcal{S}}(y_i;\lambda'\widehat{\beta})\,|Y_i=y_i\right]=\frac{v_{ii}}{(1-v_{ii})^2}(\lambda'\mathbf{S}^{-}\mathbf{x}_i)^2\widehat{\sigma}^2_{(i)},$$

donde $\widehat{\sigma}^2_{(i)}$ es el estimador insesgado de σ^2 en el modelo bajo la omisión del i-ésimo caso.

3.1. Diagnósticos de influencia en el MLGM

En el modelo multivariante, la dimensión de la expresión (11) es $d \times q$. Por tanto, para cuantificar la influencia que la i-ésima observación ejerce sobre el BLUE de la f.l.e., $\mathbf{\Lambda}\widehat{\mathbf{B}}$, se ha de considerar una norma matricial.

Muñoz *y otros* (2000), como generalización de la norma propuesta por Cook y Weisberg (1982) y Belsley *y otros* (1980), proponen la siguiente: dada una matriz $\mathbf{A}$, de dimensiones $(d \times q)$,

$$\|\mathbf{A}\|_{(\mathbf{Q},\mathbf{C})}=\left[tr(\mathbf{A}'\mathbf{Q}\mathbf{A}\mathbf{C}^{-1})\right]^{1/2}, \tag{12}$$

donde $\mathbf{Q}$ y $\mathbf{C}$ son matrices simétricas d.p. de dimensiones $(q \times q)$ y $(d \times d)$, respectivamente.

Como casos particulares, se proponen las siguientes distancias como diagnósticos de influencia:

- D_i-distancia asociada a la i-ésima observación: $\mathbf{Q}=(\mathbf{\Lambda}\mathbf{S}^{-}\mathbf{\Lambda}')^{-1}$ y $\mathbf{C}=d\widehat{\mathbf{\Sigma}}$ (donde $\widehat{\mathbf{\Sigma}}=\frac{1}{n-r}\mathbf{E}'\mathbf{E}$ es un estimador insesgado de $\mathbf{\Sigma}=(\sigma_{ks})$),

  $$D_i(\mathbf{\Lambda}\widehat{\mathbf{B}})=\left\|\widehat{\mathcal{S}}(\mathbf{y}_i;\mathbf{\Lambda}\widehat{\mathbf{B}})\right\|^2_{((\mathbf{\Lambda}\mathbf{S}^{-}\mathbf{\Lambda}')^{-1},d\widehat{\mathbf{\Sigma}})}.$$

  Esta distancia debe considerarse como una generalización de la distancia de Cook (Cook y Weisberg (1982)), propuesta para el Modelo de Regresión Múltiple, posteriormente extendida al Modelo de Regresión Multivariante por Hossain y Naik (1989) y por Barret y Ling (1992).

- W_i-distancia asociada a la i-ésima observación: $\mathbf{Q}=(\mathbf{\Lambda}\mathbf{S}^{-}\mathbf{\Lambda}')^{-1}$ y $\mathbf{C}=\widehat{\mathbf{\Sigma}}_{(i)}$ (estimador insesgado de $\mathbf{\Sigma}$ en el modelo bajo la omisión de la i-ésima observación),

  $$W_i(\mathbf{\Lambda}\widehat{\mathbf{B}})=\left\|\widehat{\mathcal{S}}(\mathbf{y}_i;\mathbf{\Lambda}\widehat{\mathbf{B}})\right\|^2_{((\mathbf{\Lambda}\mathbf{S}^{-}\mathbf{\Lambda}')^{-1},\widehat{\mathbf{\Sigma}}_{(i)})}.$$

  Análogamente, esta distancia es una generalización de distancia de Welsch-Kuh (Belsley *y otros* (1980)), propuesta también en el modelo de regresión múltiple y posteriormente extendida al modelo de regresión multivariante por Hossain y Naik (1989).

- C_i-distancia asociada a la i-ésima observación: $\mathbf{Q} = (\mathbf{\Lambda S}^{-}\mathbf{\Lambda}')^{-1}$ y $\mathbf{C} = \frac{r}{n-r}\widehat{\mathbf{\Sigma}}_{(i)}$,

$$C_i(\mathbf{\Lambda}\widehat{\mathbf{B}}) = \left\| \widehat{\mathcal{S}}(\mathbf{y}_i;\mathbf{\Lambda}\widehat{\mathbf{B}}) \right\|^2_{((\mathbf{\Lambda S}^{-}\mathbf{\Lambda}')^{-1}, \frac{r}{n-r}\widehat{\mathbf{\Sigma}}_{(i)})}.$$

  Esta medida de influencia es una generalización de la distancia modificada de Cook, pro- puesta por Atkinson (1981) para el modelo de regresión múltiple.

La generalización de estas medidas para el estudio del análisis de influencia conjunta de una colección de observaciones es simple y directa.

3.2. Influencia Local en el MLGM

Sobre este modelo también se puede utilizar el s.c. para el análisis de influencia local. Muñoz-Pichardo *y otros* (1998) proponen medidas de influencia local en el modelo univariante ($q = 1$). Posteriormente Moreno-Rebollo *y otros* (2000) extienden dichos resultados al modelo multivariante.

Se considera el modelo perturbado, MLGM(i,w), bajo el esquema de ponderación de casos, en el que sólo el caso bajo estudio es ponderado con un peso $w > 0$, es decir,

$$Cov(\underline{\varepsilon}_j) = \begin{cases} \mathbf{\Sigma} & j = 1,\ldots,n;\ j \neq i \\ w\mathbf{\Sigma} & j = i, \end{cases}$$

donde $\underline{\varepsilon}'_j$ es la j-ésima fila de $\mathcal{E}$.

En MLGM(i,w), el s.c. del BLUE $\mathbf{\Lambda}\widehat{\mathbf{B}}_w$ de una f.l.e. viene dado por:

$$(13)\quad \mathcal{S}^{(i)}_w\left(\mathbf{y}_i;\mathbf{\Lambda}\widehat{\mathbf{B}}_w\right) = \frac{w}{1+(w-1)v_{ii}}\mathbf{\Lambda S}^{-}\mathbf{x}_i\left(\mathbf{y}_i - \mathbf{x}'_i\mathbf{B}\right) = \frac{w}{1+(w-1)v_{ii}}\mathcal{S}\left(\mathbf{y}_i;\mathbf{\Lambda}\widehat{\mathbf{B}}\right).$$

En (13) puede observarse la proporcionalidad entre el s.c. del BLUE de una f.l.e. en MLGM y en MLGM(i,w), de forma semejante a (9). En este caso, la razón de proporcionalidad viene dada por

$$\alpha^*_i(w) = \frac{w}{1+(w-1)v_{ii}},$$

que no depende de la f.l.e., sólo del elemento i-ésimo diagonal de la matriz de predicción.

Análogamente, el s.c.e. de $\mathbf{\Lambda}\widehat{\mathbf{B}}_w$ dada la i-ésima observacion y_i viene dado por:

$$\widehat{\mathcal{S}}^{(i)}_w\left(\mathbf{y}_i;\mathbf{\Lambda}\widehat{\mathbf{B}}_w\right) = \alpha^*_i(w)\,\widehat{\mathcal{S}}\left(\mathbf{y}_i;\mathbf{\Lambda}\widehat{\beta}\right).$$

Dado que estas funciones de w son vectoriales de dimensión d, Moreno-Rebollo *y otros* (2000) proponen las normas ya utilizadas en el modelo MLGM, y recogidas anteriormente, obteniéndose las siguientes igualdades:

$$\begin{aligned} D_i(w,\Lambda\widehat{\mathbf{B}}) &= [\alpha_i^*(w)]^2 D_i(\Lambda\widehat{\mathbf{B}}), \\ W_i(w,\Lambda\widehat{\mathbf{B}}) &= [\alpha_i^*(w)]^2 W_i(\Lambda\widehat{\mathbf{B}}), \\ C_i(w,\Lambda\widehat{\mathbf{B}}) &= [\alpha_i^*(w)]^2 C_i(\Lambda\widehat{\mathbf{B}}). \end{aligned}$$

En las tres expresiones anteriores se observa la proporcionalidad entre la medida de influencia local y la medida de influencia, con razón de proporcionalidad, función de w, idéntica para las tres, el cuadrado de la razón de proporcionalidad que relaciona el s.c. (y su estimación) en MLGM y MLGM(i,w). El análisis de estas funciones en un entorno de $w=1$ permitirá realizar el análisis de influencia local de la observación bajo estudio.

Finalmente, puede obtenerse una expresión semejante a (10),

$$\mathcal{S}_w^{(i)}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}_w\right) = \alpha_i^*(w)\mathcal{S}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}\right) + (1-\alpha_i^*(w))\mathcal{S}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}_{(i)}\right), \forall w>0 \text{ y } \ j=1,\ldots,n.$$

Para $w\in(0,1)$, $\mathcal{S}_w^{(i)}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}_w\right)$ es una combinación lineal convexa entre $\mathcal{S}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}\right)$ y $\mathcal{S}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}_{(i)}\right)$. Sus coeficientes no dependen de la j-ésima observacion ni de la f.l.e. $\Lambda\boldsymbol{\beta}$. Así, $\alpha_i^*(w)$ es la proporción de $\mathcal{S}_w^{(i)}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}_w\right)$ en la dirección de $\mathcal{S}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}\right)$ para cualquier j. Es decir, $\alpha_i^*(w)$ es la proporción de $\mathcal{S}_w^{(i)}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}_w\right)$ explicada por $\mathcal{S}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}\right)$. Análogamente, $[1-\alpha_i^*(w)]$ es la proporción en la dirección de $\mathcal{S}\left(\mathbf{y}_j;\Lambda\widehat{\mathbf{B}}_{(i)}\right)$. En consecuencia, podemos interpretar $\alpha_i^*(w)$ como la proporción de influencia de cualquier observación sobre el BLUE de cualquier f.l.e. en el modelo MLGM(w,i) explicada por la influencia que dicha observación ejerce sobre el BLUE de la f.l.e. en el modelo MLGM. De forma semejante puede interpretarse la razón $[1-\alpha_i^*(w)]$. Como la razón de proporcionalidad entre las distancias son los cuadrados de $\alpha_i^*(w)$ y $[1-\alpha_i^*(w)]$, Muñoz *y otros* (1998) definen $[\alpha_i^*(w)]^2$ como el Potencial de Influencia Local (*LIP*) de la i-ésima observación:

$$LIP_i(w) = \left(\frac{w}{1+(w-1)v_{ii}}\right)^2.$$

4. SESGO CONDICIONADO EN COMPONENTES PRINCIPALES

Dada la importancia de esta técnica estadística en distintas áreas de investigación, se han propuesto métodos para abordar el problema de la influencia en el Análisis de

Componentes Principales (ACP), basados en el esquema de perturbación de la omisión de observaciones y la curva de influencia muestral (Critchley (1985)).

También este problema puede abordarse a través del s.c.. Enguix-González *y otros* (2000) proponen medidas de influencia para las estimaciones de los autovalores y autovectores de la matriz de varianzas y covarianzas muestrales, resultados que se recogen a continuación.

Dado un vector aleatorio p-dimensional $X \sim N_p(\mu,\mathbf{\Sigma})$, y una muestra aleatoria $X_1 \ldots X_n$, sea

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j-\overline{X})(X_j-\overline{X})',$$

la matriz de varianzas y covarianzas muestrales, y denotemos por $\widehat{\lambda}_k$, $\widehat{\alpha}_k$ $(k=1\ldots p)$, a los autovalores y autovectores asociados, respectivamente.

Para una realización muestral $\mathbf{x}_1 \ldots \mathbf{x}_n$, se obtienen los siguientes resultados

$$\begin{aligned} S(\mathbf{x}_i,\widehat{\lambda}_k) &= \frac{1}{n-1} I(\mathbf{x}_i,\lambda_k) + O(n^{-2}), \\ S(\mathbf{x}_i,\widehat{\alpha}_k) &= \frac{1}{n-1} I(\mathbf{x}_i,\alpha_k) + O(n^{-2}), \end{aligned}$$

donde λ_k y α_k son los autovalores y autovectores asociados de la matriz $\mathbf{\Sigma}$, siendo $I(\mathbf{x},\lambda_k)$ e $I(\mathbf{x},\alpha_k)$ las funciones de influencia de dichos parámetros,

$$\begin{aligned} I(\mathbf{x},\lambda_k) &= \left\{\alpha_k'(\mathbf{x}-\mu)\right\}^2 - \lambda_k, \\ I(\mathbf{x},\alpha_k) &= -\alpha_k'(\mathbf{x}-\mu)\sum_{j\neq k}\frac{\alpha_j'(\mathbf{x}-\mu)}{\lambda_j-\lambda_k}\alpha_j. \end{aligned}$$

En base a (5), se propone los siguientes estimadores de $S(\mathbf{x}_i,\widehat{\lambda}_k)$ y $S(\mathbf{x}_i,\widehat{\alpha}_k)$,

$$\widehat{S}(\mathbf{x}_i,\widehat{\lambda}_k) = \widehat{\lambda}_k - \widehat{\lambda}_k^{(i)},$$

y

$$\widehat{S}(\mathbf{x}_i,\widehat{\alpha}_k) = \widehat{\alpha}_k - \widehat{\alpha}_k^{(i)},$$

donde $\widehat{\lambda}_k^{(i)}$ y $\widehat{\alpha}_k^{(i)}$ son los autovalores y autovectores asociados a $\widehat{\mathbf{\Sigma}}_{(i)}$. Ambas estimaciones pueden considerarse como estadísticos de diagnóstico de influencia en el ACP. Para evitar el problema del signo del s.c.e. en el caso de los autovalores, y el problema de la dimensión del s.c.e. en el caso de los autovectores, se pueden aplicar distancias, de forma análoga a lo realizado anteriormente en el MLGM. Así, se proponen:

- Para un autovalor, el cuadrado del s.c.e., normalizado con la estimación de $Var\left[\widehat{\lambda}_k\right]$,

$$(n-1)\frac{\left\{\widehat{S}(\mathbf{x}_i,\widehat{\lambda}_k)\right\}^2}{2\widehat{\lambda}_k^2}.$$

- Para un autovector, se pueden aplicar las normas anteriormente definidas de acuerdo a (12). En particular, dado que en este caso $d=1$, para $\mathbf{Q}=\mathbf{I}_p$ y $\mathbf{C}$ la unidad, se obtendría la norma euclídea del vector $\widehat{S}(\mathbf{x}_i,\widehat{\alpha}_k)$.

Este enfoque del estudio de la influencia permite generalizar los estadísticos de diagnóstico anteriores para un conjunto de autovalores, o un conjunto de autovectores, seleccionando adecuadamente las matrices que determinan la norma.

5. SESGO CONDICIONADO EN EL MUESTREO EN POBLACIONES FINITAS

Moreno-Rebollo *y otros* (1999) adaptan el concepto de s.c. como diagnóstico de influencia en el muestreo en poblaciones finitas. En particular, lo desarrollan para el estimador de Horvitz-Thompson del total poblacional.

Sea $U=\{u_1,\ldots,u_N\}$ una población finita y $\{\mathcal{M},p(.)\}$ un diseño muestral definido sobre U, donde $\mathcal{M}$ es el espacio muestral y $p(.)$ una distribución de probabilidad sobre $\mathcal{M}$. Sea Y una característica de la población, $Y=\{Y_1,\ldots,Y_N\}$, $\theta=\theta(Y)$ el parámetro de interés y $\widehat{\theta}=\widehat{\theta}(s)$ un estimador de θ basado en $s\in\mathcal{M}$. Se propone la siguiente definición.

Definición 3. *El sesgo condicionado sobre* $\widehat{\theta}$, *causado por la presencia de* u_i *en la muestra, se define por*

$$S(I_i=1;\widehat{\theta})=E[\widehat{\theta}\,|I_i=1]-E[\widehat{\theta}]\,,$$

donde $I_i(s)$ $(i=1\ldots N)$ *son las variables aleatorias*

$$I_i(s)=\begin{cases}1 & si\ u_i\in s\\ 0 & en\ otro\ caso.\end{cases}$$

Es decir, el s.c. mide la desviación en el valor esperado del estimador cuando el diseño muestral se perturba, restringiéndolo sobre las muestras que contienen a u_i.

En particular, si se considera el estimador de Horvitz-Thompson, $\widehat{T}_{HT} = \sum_s \frac{Y_i}{\pi_i}$, del total poblacional de la característica Y, $T(Y) = \sum_{i=1}^{N} Y_i$, se obtiene que

$$S(I_i = 1; \widehat{T}_{HT}) = Y_i \frac{1-\pi_i}{\pi_i} + \sum_{j \neq i} Y_j \frac{\Delta_{ij}}{\pi_i \pi_j} ,$$

donde $\pi_j = \Pr[I_j = 1]$, $j = 1, \ldots, N$, representan las probabilidades de inclusión de primer orden asociadas al diseño muestral $\{\mathcal{M}, p(.)\}$ y $\Delta_{ij} = Cov(I_i, I_j)$, $i, j = 1 \ldots N$.

Dado que el s.c. es un parámetro poblacional desconocido, se propone como estimador del mismo el estimador de Horvitz-Thompson sobre el diseño muestral restringido sobre las muestras que contienen a u_i:

$$\widehat{S}_{HT}(I_i = 1; \widehat{T}_{HT}) = Y_i \frac{1-\pi_i}{\pi_i} + \sum_{j \neq i} Y_j \frac{\Delta_{ij}}{\pi_{ij} \pi_j} I_i ,$$

donde $\pi_{ij} = \Pr[I_i = 1, I_j = 1]$, son las probabilidades de inclusión de segundo orden en el diseño muestral.

Un estimador insesgado de la varianza de $\widehat{S}_{HT}(I_i = 1; \widehat{T}_{HT})$ en el diseño restringido viene dado por:

$$\widehat{Var}\left[\widehat{S}_{HT}(I_i = 1; \widehat{T}_{HT})\right] = \sum_{j \neq i} \sum_{h \neq i} Y_j Y_h \frac{\Delta_{ij}\, \Delta_{ih}}{\pi_i \pi_j \pi_h} \left[\frac{\pi_i}{\pi_{ij}\, \pi_{ih}} - \frac{1}{\pi_{ijh}}\right] I_j I_h ,$$

donde $\pi_{ijh} = \Pr[I_i = 1, I_j = 1, I_h = 1]$, son las probabilidades de inclusión de tercer orden en $\{\mathcal{M}, p(.)\}$.

Es evidente que tanto el s.c., como el s.c.e. y el estimador de su varianza dependen del diseño muestral $\{\mathcal{M}, p(.)\}$, por lo que la influencia de cada unidad muestral seleccionada es función del comportamiento de dicha unidad respecto a la característica bajo estudio y del diseño muestral elegido por el investigador. La aplicación a cada tipo de diseño es directa, sin más que determinar los parámetros dependientes del mismo (probabilidades y covarianzas de las variables aleatorias indicadores), aunque en ocasiones pueda resultar compleja. No obstante, frente a la posible complejidad, el concepto de s.c. posee la característica de su generalidad. El planteamiento arriba recogido puede utilizarse en un amplio espectro del muestreo en poblaciones finitas, quedando abiertas un número considerable de líneas de investigación y desarrollo.

6. CONCLUSIÓN

Este trabajo recoge una revisión del concepto de s.c. y su aplicación al análisis de influencia en diversas técnicas estadísticas. Su definición genérica, su fácil interpretación y su cálculo no excesivamente complejo le permite tener un extenso campo de

aplicación y un amplio abanico de posibilidades en el análisis de influencia y análisis de influencia local de cualquier estadístico en cualquier modelo. Muestra de tales afirmaciones son los resultados anteriormente recogidos y los trabajos de Jiménez (1994) y Jiménez *y otros* (1995) en el área de las técnicas de remuestreo, en particular en el bootstrap, para detectar muestras bootstrap con un efecto considerable sobre los resultados de las estimaciones.

Las medidas de diagnóstico de influencia propuestas pueden generalizarse fácilmente al análisis de influencia conjunto de dos o más observaciones. Algunos resultados sobre tal aspecto están recogidos en los trabajos citados anteriormente.

Profundizando por la línea de los modelos lineales se puede abordar la influencia en los modelos lineales generalizados (univariantes y multivariantes), análisis de supervivencia, etc. En la línea del ACP, pueden obtenerse resultados de interés en Análisis Discriminante, Análisis Factorial, Análisis Canónico, etc. Finalmente, en el área del muestreo en poblaciones finitas, como se recoge en el apartado anterior, queda un amplio campo abierto para seguir desarrollando técnicas de diagnóstico de influencia, más aún cuando en este área la problemática de la influencia no ha sido, hasta ahora, tratada en profundidad.

REFERENCIAS

Andrews, D. F. y Pregibon, D. (1978). «Finding outliers that matter». *J. Royal Statistics Soc., Ser. B*, 40, 85-93.

Atkinson, A. C. (1981). «Two graphical displays for outliying and influential observations in regression». *Biometrika*, 68, 13-20.

Barnett, V. y Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley and Sons.

Barret, B. E. y Ling, R. F. (1992). «General classes of influence measures for multivariate regression». *J.A.S.A.*, (7), 184-191.

Belsley, D. A., Kuh, E. y Welsch, R. E. (1980). *Regresion diagnostics: Identifying influential data and sources of collinearity.* John Wiley and Sons.

Billor, N. y Loynes, R. M. (1993). «Local influence: a new approach». *Commun. Statist. Theory and Methods*, 22(6), 1595-1611.

Brown, G. C. y Lawrance, A. J. (2000). «Theory and illustration of regression influence diagnostics». *Commun. Statist. Theory and Methods*, 29 (9 & 10), 2079-2107.

Cook, R. D. y Weisberg, S. (1982). *Residuals and influence in regresion*, Chapman and Hall.

— (1986). «Assessment of Local Influence (with discussion)». *J. Royal Statistics Soc., Ser. B*, 48, 133-169.

— (1987). «Influence assessment». *Journal of Applied Statistics*, 14, 2, 117-132.

Cook, D., Peña, D. y Weisberg, S. (1988). «The likelihood displacement: a unifying principle for influence measures». *Commun. Statist.-Theory and Methods*, 17 (3), 623-640.

Critchley, F. (1985). «Influence in Principals Components Analysis». *Biometrika*, 43, 128-136.

Efron, B. y Stein, C. (1984). «The jackknife estimate of variance». *Ann. Statist.*, 9 (3), 586-596.

Enguix-González, A., Moreno-Rebollo, J. L., Jiménez-Gamero, M. D. y Muñoz-Pichardo, J. M. (2000). «Estudio de influencia en componentes principales a través del sesgo condicionado». *Actas XXV Congreso Nac. de Estadística e Investigación Operativa*, (Vigo, España).

Escobar, L. A. y Meeker, W. Q. (1992). «Assessing influence in regression analysis with censored data». *Biometrics*, 48, 507-528.

Hampel, F. R. (1974). «The influence curve and its role in robust estimation». *J. Am. Statist. Assoc.*, 69, 383-393.

Hossain, A. y Naik, D. N. (1989). «Detection of influential observations in multivariate regression». *Journal of Applied Statistics*, 16, 25-37.

Jiménez Gamero, M. D. (1994). *Análisis de las muestras generadas en el proceso de simulación bootstrap.* Tesis Doctoral. Universidad de Sevilla.

Jiménez Gamero, M. D., Muñoz Pichardo, J. M. y Muñoz Reyes, A. (1995). «Medida de influencia en las estimaciones bootstrap». *Actas del XXII Congreso Nac. de Estadística e I.O.*, (Sevilla).

Lawrance, A. J. (1991). «Local and deletion influence». En *Directions in robust statistics and diagnostics*, W. Stahel y S. Weisberg (eds.). Springer-Verlag.

Moreno-Rebollo, J. L., Enguix-González, A., Muñoz-Pichardo, J. M. y Alba, M. V. (2000). «Influencia Local en el Modelo Lineal General Multivariante». *Actas XXV Congreso Nac. de Estadística e Investigación Operativa*, (Vigo, España).

Moreno Rebollo, J. L., Muñoz Reyes, A. M. y Muñoz Pichardo, J. M. (1999). «Influence diagnostic in survey sampling: conditional bias». *Biometrika*, 86, (4), 923-928.

Muñoz Pichardo, J. M. y Fernández Ponce, J. M. (1997). «Distancias de Mahalanobis y Rao: Influencia en el Modelo Lineal General». *Actas IV International Meeting of Multidimensional Data analysis*, (Bilbao, Spain), 259-262.

Muñoz Pichardo, J. M., Muñoz García, J., Fernández Ponce, J. M. y Jiménez Gamero, M. D. (2000). «Influence analysis in multivariate linear general models». *Commun. Statist.-Theory and Methods*, 29 (aceptado para publicación).

Muñoz Pichardo, J. M., Muñoz García, J., Fernández Ponce, J. M. y López Blázquez, F. (1998). «Local Influence on the General Linear Model». *Sankhyã*, Ser. B, 60 (3).

Muñoz Pichardo, J. M., Muñoz García, J., Moreno Rebollo, J. y Pino Mejías, R. (1995). «A new approach to influence analysis in linear models». *Sankhyã*, Ser. A, 57 (3). 393-409.

Tukey, J. W. (1970). *Exploratory Data Analysis*, (1970/71: edición preliminar). Reading Mass. Addison-Wesley.

ENGLISH SUMMARY

CONDITIONAL BIAS IN INFLUENCE ANALYSIS: A REVIEW

J. M. MUÑOZ-PICHARDO
J. L. MORENO-REBOLLO
T. GÓMEZ-GÓMEZ
A. ENGUIX-GONZÁLEZ
Universidad de Sevilla*

Conditional bias has been proposed as an influence diagnostic on different models and statistical techniques. In this paper, we sum up these applications and we relate it to the sensitivity curve and the sample influential curve. Moreover, we point out some areas in which the Influence Analysis could be studied through this approach.

Keywords: Influence analysis, linear models, principal components, survey sampling, conditional bias

AMS Classification (MSC 2000): 62J20, 62H25, 62D05

* Universidad de Sevilla. Facultad de Matemáticas. Departamento de Estadística e Investigación Operativa. Avda. Reina Mercedes s/n. 41012 Sevilla.

- Received October 2000.
- Accepted April 2001.

From the Decomposition Lemma of Efron and Stein (1984), Muñoz - Pichardo *et al.* (1995) proposed the conditional bias (c.b.) as a general approach in Influence Analysis. In this paper, we gather together some applications of this concept in several models and statistical techniques.

Definition 1. *Let $Y_1 \ldots Y_n$ be a random sample of a random variable Y, let $T_n = T_n(Y_1 \ldots Y_n)$ be a statistic and let $y_1 \ldots y_n$ be a sample realization. The c.b. of T_n given y_i is defined as*

$$\mathcal{S}(y_i; T_n) = E[T_n | Y_i = y_i] - E[T_n].$$

From Definition 1, we note that the c.b. depends on the distribution of T_n and on the observed value y_i, and it assesses the influence of y_i on T_n in terms of its expected value. Therefore, it not depends on $y_2, \ldots, y_n$. The perturbation considered it is due to the knowledge of the observation under study, y_i. Moreover, if T_n is q-dimensional ($q > 1$), then a norm must be used in order to define an influence measure from $\mathcal{S}(y_i; T_n)$. Finally, we note that, in general, $\mathcal{S}(y_i; T_n)$ depends on unknown parameters, so it must be estimated.

In Section 2 we relate the c.b. to the expected value of the sensitivity curve (Tukey, 1970), and the sample influence curve. From these relations, Muñoz-Pichardo *et al.* (1995) proposed

$$\widehat{\mathcal{S}}(y_i; T) = T_n(Y_1 \ldots Y_{i-1}, y_i, Y_{i+1} \ldots Y_n) - T_{n-1}(Y_1 \ldots Y_{i-1}\, Y_{i+1} \ldots Y_n) = T_n - T_{(i)},$$

as estimator of $\mathcal{S}(y_i; T)$.

Following the guidelines laid down by Cook (1987), Muñoz-Pichardo *et al.* (1998) apply the c.b. in Local Influence Analysis through the study of the function

$$\mathcal{S}_w(y_i; T_n) = E_{M(\mathbf{w})}[T_n \,|Y_i = y_i] - E_{M(\mathbf{w})}[T_n],$$

being $M(\mathbf{w})$ a perturbation of the postulated model M.

As an illustration, in section 2.1 we obtain the c.b. and some influence measures on the sample mean. Also, we study the local influence on the sample mean of a random sample.

In section 3, we study the c.b. in the Multivariate General Linear Model (MGLM), in order to handle in an unified way the influence analysis on the models that are obtained as particular cases of it.

Given the MGLM, that is defined by the matrix identity $\mathbf{Y} = \mathbf{XB} + \mathcal{E}$, with the usual hypotheses, and a linear estimable function (l.e.) $\mathbf{AB}$, it is obtained that (Muñoz-Pichardo *et al.*, 2000)

$$\mathcal{S}(\mathbf{y}_i;\Lambda\widehat{\mathbf{B}}) = \Lambda\mathbf{S}^{-}\mathbf{x}_i\left(\mathbf{y}_i' - \mathbf{x}_i'\mathbf{B}\right), \quad \widehat{\mathcal{S}}(\mathbf{y}_i;\Lambda\widehat{\mathbf{B}}) = \frac{1}{1-v_{ii}}\Lambda\mathbf{S}^{-}\mathbf{x}_i\mathbf{e}_i',$$

being $\widehat{\mathbf{B}}$ a least squares estimator of $\mathbf{B}$, $\mathbf{x}_i'$ and $\mathbf{y}_i'$ the i-th rows of $\mathbf{X}$ and $\mathbf{Y}$, respectively, $\mathbf{S} = \mathbf{X}'\mathbf{X}$, $\mathbf{S}^{-}$ an inverse generalized of $\mathbf{S}$, $\mathbf{e}_i'$ the i-th row of the matrix of ordinary residuals, and v_{ii} the i-diagonal element of the hat matrix $\mathbf{V} = \mathbf{X}'\mathbf{S}^{-}\mathbf{X}$.

In order to assess the influence of y_i on the BLUE, $\Lambda\widehat{\mathbf{B}}$, we apply matrix norms, obtaining various influence diagnostics. The application of these diagnostics in the Multiple Regression Model, make possible to obtain, as particular cases, some of the influence measures defined in the literature. Moreover, we study the Local Influence Analysis in this model.

In section 4, we study the influence analysis in Principal Components, under hypothesis of normality. The c.b. of $\widehat{\lambda}_k$, $\widehat{\alpha}_k$, $k = 1\ldots p$, the eigenvalues and eigenvectors of the sample covariance matrix are given by (Enguix-González *et al.*, 2000),

$$S(\mathbf{x}_i,\widehat{\lambda}_k) = \frac{1}{n-1}I(\mathbf{x}_i,\lambda_k) + O(n^{-2}) \text{ and } S(\mathbf{x}_i,\widehat{\alpha}_k) = \frac{1}{n-1}I(\mathbf{x}_i,\alpha_k) + O(n^{-2}),$$

being $\mathbf{x}_1\ldots\mathbf{x}_n$ the sample realization and $I(\mathbf{x},\lambda_k)$, $I(\mathbf{x},\alpha_k)$ the influence functions of λ_k and α_k, the population eigenvalues and eigenvectors,

$$I(\mathbf{x},\lambda_k) = \left\{\alpha_k'(\mathbf{x}-\mu)\right\}^2 - \lambda_k \text{ and } I(\mathbf{x},\alpha_k) = -\alpha_k'(\mathbf{x}-\mu)\sum_{j\neq k}\frac{\alpha_j'(\mathbf{x}-\mu)}{\lambda_j-\lambda_k}\alpha_j.$$

From $\widehat{S}(\mathbf{x}_i,\widehat{\lambda}_k)$ and $\widehat{S}(\mathbf{x}_i,\widehat{\alpha}_k)$ several influence measure on $\widehat{\lambda}_k$ and $\widehat{\alpha}_k$ are proposed.

Finally, in section 5, we show that the c.b. can be applied in sampling from a finite population, when the inference is design-based. Moreno-Rebollo *et al. (1999)* adjust the definition of c.b. in order to obtain an influence measure in survey sampling. Let $U = \{u_1,\ldots,u_N\}$ be a finite population and $\{\mathcal{M},p(.)\}$ a sampling design defined on U, let π_i, $i = 1,\ldots,N$, be the first order inclusion probabilities. Let Y be a characteristic of the population, $Y = \{Y_1,\ldots,Y_N\}$, $\theta = \theta(Y)$ the parameter of interest and $\widehat{\theta} = \widehat{\theta}(s)$ an estimator of θ, for $s \in \mathcal{M}$.

Definition 3. *The conditional bias of $\widehat{\theta}$, caused by the presence of u_i in the sample s, is defined by $S(I_i = 1;\widehat{\theta}) = E[\widehat{\theta}\,|I_i = 1] - E[\widehat{\theta}]$, being $I_i(s) = 1$ if $u_i \in s$, $I_i(s) = 0$ otherwise.*

Particularly, if the Horvitz-Thompson (HT) estimator, $\widehat{T}_{HT} = \sum_s \frac{Y_i}{\pi_i}$, is considered, it is obtained that

$$S(I_i = 1;\widehat{T}_{HT}) = Y_i\frac{1-\pi_i}{\pi_i} + \sum_{j\neq i}Y_j\frac{\Delta_{ij}}{\pi_i\pi_j},$$

being $\Delta_{ij} = Cov(I_i, I_j)$, $i, j = 1 \ldots N$, on $\{\mathcal{M}, p(.)\}$.

Moreno-Rebollo *et al. (1999)* proposed to estimate $S(I_i = 1; \widehat{T}_{HT})$ by the HT-estimator based on the restricted sampling design, on the samples containing u_i,

$$\widehat{S}_{HT}(I_i = 1; \widehat{T}_{HT}) = Y_i \frac{1 - \pi_i}{\pi_i} + \sum_{j \neq i} Y_j \frac{\Delta_{ij}}{\pi_{ij}\pi_j} I_j.$$

We note that both the c.b. and its estimation depend on the sampling design, that is a distinctive feature of the influence measures in sample survey.

Investigació Operativa

QÜESTIIÓ, vol. 25, 2, p. 287-300, 2001

MODELIZACIÓN DE UN DSS PARA LA GESTIÓN DE PRODUCTOS PERECEDEROS

B. DÍAZ FERNÁNDEZ*
J. A. DEL BRÍO GONZÁLEZ
B. GONZÁLEZ TORRE

La gestión de inventarios de productos perecederos ha atraído desde hace tiempo la atención de los investigadores de Dirección de Operaciones. En este artículo se presenta la modelización e implementación de un sistema de apoyo a la toma de decisiones (DSS) para la gestión de productos perecederos, aplicado a la distribución interhospitalaria de hemoderivados. En estos casos se trata de satisfacer en lo posible las demandas, tratando de evitar a la vez la caducidad de los productos en manos de los clientes (los cuales teóricamente no sufren coste extra por sobrestock). Dada la naturaleza multicriterio del problema, para la modelización se ha usado goal programming, *con resultados que en casos concretos permiten reducciones considerables en la cantidad de productos inutilizables.*

DSS modelling for perishable inventory management

Palabras clave: Sistemas de apoyo a la toma de decisiones, programación por metas, gestión de inventarios

Clasificación AMS (MSC 2000): 90B05, 90B90, 90C29

* Autor para toda correspondencia: Belarmino Díaz Fernández. ETS Ingenieros Industriales. Universidad de Oviedo. Campus de Viesques. 33204 Gijón.

– Recibido en enero de 1998.
– Aceptado en mayo de 2001.

1. INTRODUCCIÓN

Desde hace muchos años, el aumento constante de competitividad en los mercados viene obligando a los responsables de Dirección de Operaciones a la búsqueda de soluciones para minimizar el coste global de colocación de los productos a disposición de los clientes. Inicialmente, los esfuerzos parecían orientarse hacia la reducción de los costes de producción (Sculty y Fawcett, 1993), campo en el que se lograron tan importantes avances que hoy en día parece difícil continuar obteniendo mejoras significativas. Es por eso que empiezan a buscarse nuevas vías para reducir el coste global de fabricación y comercialización, orientándose los esfuerzos hacia la minimización de los costes de transporte y distribución y diseñándose en consecuencia avanzados métodos para mejorar la gestión de stocks (Skjoett-Larsen, 2000).

Sin embargo, los modelos tradicionales de gestión de stocks han de ser modificados cuando se busca su aplicabilidad al caso de productos perecederos. Diversos autores, desde la perspectiva del marketing, están empezando a plantear la necesidad de modelos específicos (Henessy, 1999; Anónimo, 1998). Es así como surgen en el ámbito de la Dirección de Operaciones numerosos intentos de adaptación de los modelos clásicos, para contemplar la posibilidad de caducidad de los productos almacenados.

La política de pedidos elegida es clave dentro de la gestión de perecederos, intentando contar con suficiente cantidad disponible en stock (una vez evaluada la demanda), cumpliendo así el objetivo de que caduquen el menor número de unidades posible.

Toda la investigación analítica sobre este problema ha partido de la simplificación de que todas las unidades demandadas se usan. Debido a esto, los resultados que se obtienen no son directamente aplicables al inventario de un producto perecedero aunque sí pueden funcionar como interesantes aproximaciones. Uno de los trabajos más relevantes es el de Nahmias (1982). En este artículo el autor estudia dos alternativas: que la demanda sea determinista y que la demanda no sea determinista.

En el caso de demanda determinista, Nahmias considera el modelo EOQ para un producto con una vida útil de m períodos. El tamaño óptimo que minimiza el coste de mantenimiento y lanzamiento de un pedido viene dado por la fórmula de Wilson. Cuando la demanda no es determinista, los modelos son más complejos. Un caso especial dentro de los productos perecederos es cuando las unidades no pueden estar en stock más de un período. Si la vida útil del producto es exactamente un período, las decisiones de pedido en los períodos sucesivos son independientes y el problema se reduce a un caso del vendedor de periódicos (*newsboy problem*).

Williams (1999) relata cómo se modifican los periodos de pedido al considerar que los productos tienen una vida máxima de dos periodos. Paralelamente, Wee (1999) analizó el caso de productos con tiempo de vida fijo de cualquier duración, aplicando la teoría de sistemas de Markov. Estudios más avanzados, como el de Adachi *et al.* (1999),

plantean situaciones de demanda aleatoria y, más recientemente, Rahim *et al.* (2000) contemplan la posibilidad de que el producto se empiece a deteriorar en un momento aleatorio de su periodo de almacenaje.

El primer análisis de políticas óptimas para un producto perecedero con una vida útil fijada fue debido a Van Zyl (1964). Sin embargo, Nahmias y Pierskalla (1975), realizaron una aproximación mejorada, suponiendo que solamente hay costes de caducidad y de ruptura de stocks. Consideraron varias opciones para las políticas óptimas, como el caso multiperíodo, la posibilidad de que hubiera costes de pedido y de mantenimiento, y por último, también compararon la política de pedido óptima cuando el producto tiene una vida de dos períodos con la correspondiente política $y_\infty(x)$ de pedido óptima para el mismo problema en el caso de que el producto tuviera una vida infinita (es decir no fuera perecedero).

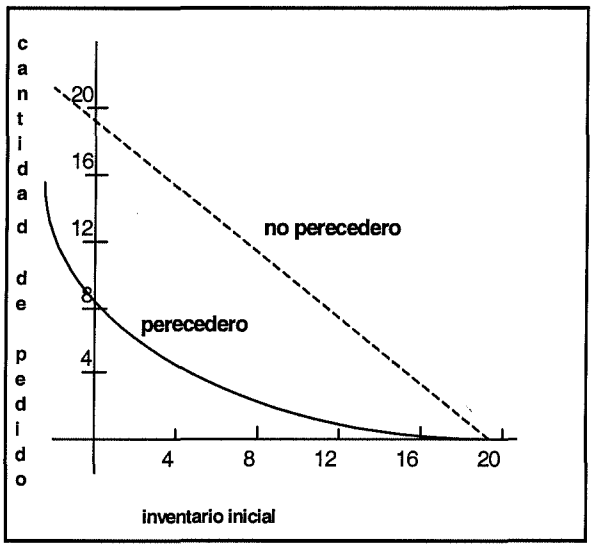


Figura 1. Cantidad de pedido óptimo para inventarios perecederos y no perecederos (Nahmias, 1982).

Con productos perecederos, la política óptima determinará pedir menor cantidad que cuando el producto no lo es, por el riesgo de la caducidad (figura 1). Además, la diferencia entre las dos políticas es mayor para valores pequeños del inventario inicial. La caducidad de los productos tiene el efecto de disuadir de los pedidos grandes, ya que si se solicitan grandes cantidades de estos productos, es muy probable que parte de ellos perezcan.

Pese a los intentos de adaptación, los modelos clásicos plantean grandes limitaciones de aplicabilidad por sustentarse sobre simplificaciones de problemas complejos de programación lineal, a través de la introducción de hipótesis altamente restrictivas. Es por ello que algunos autores optan por la búsqueda de soluciones a la gestión de inventario de productos perecederos utilizando directamente formulaciones de programación lineal (Klensorge y Schary, 1989). A pesar de todo, seguía existiendo el problema de que las restricciones eran demasiado exigentes. El cumplimiento estricto de las mismas mermaba en gran medida el beneficio obtenido. Para solventarlo, se empieza a considerar la posibilidad (como se hace en este trabajo) de modelización a través de criterios de programación por metas que, usando los mismos algoritmos de resolución que la programación lineal, sustituyen las restricciones por metas flexibles, cuyas desviaciones se pretende minimizar (Rifai, 1996).

En el presente trabajo se aplica un modelo basado en la metodología de programación por metas a la gestión de envíos de sangre desde un Centro de Transfusiones (que actúa como unidad central de análisis) a todos los hospitales regionales. La distribución de sangre ya había sido abordada con anterioridad a través de modelos tradicionales de gestión de inventarios, presuponiendo tanto periodos fijos de vida útil del producto como periodos de vida exponenciales. Aquí se pretende aprovechar la flexibilidad de la programación por metas para optimizar los envíos considerando la trascendencia de que la demanda de sangre pueda exceder al nivel diario de donaciones.

Para completar el trabajo, se precisó el desarrollo de una herramienta eficaz que facilitase al personal del Centro de Transfusiones la adopción de decisiones diarias acerca de las cantidades de sangre a suministrar. Por ello, se desarrolló un programa informático capaz de proporcionar automáticamente tal información, a partir de una serie de inputs o datos de entrada y un conjunto de datos históricos en permanente actualización. Por sus características, este programa puede ser entendido como un D.S.S.

2. UN MODELO DE PROGRAMACIÓN POR METAS PARA LA DISTRIBUCIÓN DE PRODUCTOS PERECEDEROS

El problema que en este artículo se aborda aparece en aquellas empresas cuya actividad es la fabricación y posterior venta de productos perecederos (farmacia, pastelería, carnes y pescados, libros de temporada, etc.), las cuales se apoyan en redes de distribuidores. La problemática de esta actividad radica en que, en muchas ocasiones, los distribuidores no compran el producto sino que funcionan como meros intermediarios recibiendo una comisión por unidad vendida, y por ello su interés es pedir la mayor cantidad posible de producto. Por su parte, la empresa productora debe tener en cuenta que aquellas unidades no vendidas por los suministradores corren el riesgo de caducar ya que se está trabajando con productos perecederos, y será ella la que deba asumir el coste correspondiente, al menos parcialmente. Por lo tanto a la empresa le interesará

enviar las unidades justas a los suministradores con el objetivo de que no tengan ni demasiadas unidades que puedan caducar, ni tengan roturas de stock.

Para modelar este problema se han identificado varios objetivos así como restricciones que se deben cumplir:

1) La empresa fabricante debe enviar las unidades adecuadas a los suministradores de modo que les caduque el menor número de unidades posible.

2) La empresa fabricante no podrá enviar más unidades de las que dispone. Ésta será una restricción que deberá satisfacer el modelo y no una meta ya que en los casos donde la demanda sea mayor que la disponibilidad se producirá una rotura de stock.

3) La empresa fabricante tiene que intentar satisfacer siempre las necesidades de los distribuidores. Aun sabiendo que normalmente las unidades que se solicitan son más que las que se venden, no debería enviarse nunca unidades por defecto debido a que en algún caso sí se pueden necesitar todas. Pese a esto, puede darse el caso de que la empresa no tenga suficientes unidades para cumplir las necesidades de los suministradores y eso no debe implicar que el problema no tenga solución sino que se buscará una desviación lo más pequeña posible. Por tanto, esta condición no se va a incluir como una restricción sino como una nueva meta. Por otra parte, si la empresa fabricante sí tiene suficientes unidades de producto para cumplir este segundo objetivo, deberá decidir con que edad envía esas unidades, ya que normalmente a los suministradores que les sobren más se les deberá enviar unidades más frescas para alargar la caducidad.

4) La empresa fabricante tiene que cubrir unas necesidades de producto fresco para todos los suministradores. Por la misma razón que en el apartado 3, esta condición deberá plantearse como un objetivo, de tal manera que si no se puede cumplir debido a que la empresa fabricante no tiene bastantes unidades frescas, se aproxime al máximo a las necesidades de los suministradores.

5) Por último, además de que los suministradores deban recibir una determinada cantidad de producto fresco, es ético que todos reciban producto con una media de edad aceptable. Todos los suministradores prefieren el producto más fresco, por ello no sería una política justa el que a aquellos que tengan más actividad se les envíen las unidades menos frescas, puesto que es una manera de penalizarlos. Por eso, aunque normalmente las unidades más frescas se enviarán a aquellos suministradores donde es más fácil que caduquen las unidades sobrantes (normalmente los de menores ventas), hay que mantener una cierta equidad en todos los envíos. Esta nueva condición se planteará como otra meta dentro de la política de inventarios de la empresa.

Como el problema tiene varios objetivos, un planteamiento de programación lineal clásico con una única función objetivo a maximizar o minimizar y una serie de restricciones, no es apropiado. En este caso se recurrirá a la *programación lineal por*

metas, donde los objetivos serán tratados como restricciones flexibles de forma que cada una de ellas va a tener una posible desviación por defecto o por exceso, tratando de conseguir que las desviaciones que no son interesantes sean lo menores posibles. Cada objetivo tiene por tanto su propia desviación que denominaremos Y, la cual puede ser positiva o negativa.

En relación a la definición de las variables, como anteriormente se expuso el primer objetivo será minimizar la cantidad de producto que le caduca a los distribuidores. Por tanto, si suponemos que hay n tipos de producto, y cada uno tiene una vida útil máxima de m días, definimos X_{ij}^l como la cantidad de producto de tipo $l = 1,2,\ldots,n$ con $i = 1,2,\ldots,m_l$ días de vida que se envía al distribuidor $j = 1,2,\ldots,k$.

Por tanto, el primer objetivo en forma de restricción consiste en que caduque el menor número de unidades posible;

$$(1) \qquad \sum_{i=1}^{m_l}\sum_{j=1}^{k} P_{ij}^l X_{ij}^l = Y_{1,l}^+ - Y_{1,l}^- \qquad \text{con } l = 1,\ldots,n$$

siendo P_{ij}^l la probabilidad de que una unidad de producto del tipo l con i días de vida caduque en el distribuidor j. Esta probabilidad la puede obtener la empresa a partir de sus propios datos históricos, pero si no se tuvieran suficientes datos, es posible realizar una aproximación a partir de la relación $P_{ij}^l = 1 - Q_{ij}^l$ siendo Q_{ij}^l la probabilidad de que una unidad de tipo l, con i días de vida, sea transferida por el distribuidor j. Estas Q_{ij}^l se pueden poner en relación con la probabilidad de que la unidad más fresca posible sea transferida (Kendall, Lee, 1980), mediante la relación $Q_{ij} = \gamma_j^{(m-i)} \cdot Q_{m,j}$, siendo $\gamma_j^{(m-i)}$ el parámetro de preferencia de edad para el suministrador j (grado en que se prefiere unidades más viejas sobre la más frescas). Si $\gamma_j^{(m-i)} \longrightarrow \infty$ estamos ante una situación FIFO pura.

El segundo objetivo consiste en que la empresa cubra siempre que pueda todas las demandas de los distribuidores. Si llamamos C_j^l la demanda del distribuidor j del producto tipo l (en el período de tiempo considerado), tendremos que:

$$(2) \qquad \sum_{i=1}^{m_l} X_{ij}^l - C_j^l = Y_{2,l,j}^+ - Y_{2,l,j}^- \qquad j = 1,2,\ldots,k; \quad l = 1,2,\ldots,n$$

El tercer objetivo trata de enviar un número mínimo de unidades frescas a los distribuidores. Si entendemos por unidades frescas las que tienen z_o días o menos, y llamamos F_j^l al nivel de unidades frescas de tipo l que requiere el suministrador j en el período de tiempo considerado, el objetivo en forma de restricción será:

$$(3) \qquad \sum_{i=1}^{z_o} X_{ij}^l - F_j^l = Y_{3,l,j}^+ - Y_{3,l,j}^- \qquad j = 1,2,\ldots,k; \quad l = 1,2,\ldots,n$$

Finalmente, el último objetivo busca que los distribuidores reciban una cantidad de producto con una media de edad aceptable. Para expresar esta restricción llamaremos, al igual que hicimos anteriormente, C_j^l a la demanda de producto tipo l del suministrador j (en el período de tiempo considerado), y A_j^l a la media de edad deseada por el suministrador j para el producto l. El objetivo en forma de restricción será:

$$(4) \qquad \sum_{i=1}^{m_l} i \cdot X_{ij}^l - A_j^l \cdot C_j^l = Y_{4,l,j}^+ - Y_{4,l,j}^- \qquad j = 1,\ldots,k; \quad l = 1,\ldots,n$$

Para completar el modelo, sólo queda añadir que la empresa productora no puede enviar más unidades a los distribuidores de las que efectivamente tiene. Por tanto, si llamamos K_i^l al número de unidades de producto tipo l con i días de vida que tiene la empresa en el período de tiempo considerado, tendremos que:

$$(5) \qquad \sum_{j=1}^{k} X_{ij}^l \leqslant K_i^l \qquad i = 1,\ldots,m_l, \quad \text{y} \quad l = 1,\ldots,n$$

Respecto a la función objetivo, como se expuso anteriormente, el objetivo general va a consistir en minimizar determinadas desviaciones Y_i. Por tanto, el modelo completo será:

$$(6) \qquad \min \sum_{l=1}^{n} Y_{1,l}^+ + \sum_{l=1}^{n}\sum_{j=1}^{k} Y_{2,l,j}^- + \sum_{l=1}^{n}\sum_{j=1}^{k} Y_{3,l,j}^- + \sum_{l=1}^{n}\sum_{j=1}^{k} Y_{4,l,j}^+$$

sujeto a las restricciones (1)-(5), y a la condición de no negatividad de todas las variables.

Para la búsqueda de la solución de este modelo se ha desarrollado un sistema informático de apoyo a la toma de decisiones, DSS, que se describe posteriormente.

3. IMPLEMENTACIÓN DEL MODELO EN LA GESTIÓN DE UN BANCO DE SANGRE

Se describe a continuación una aplicación del modelo presentado, al caso de la gestión de un banco de sangre para transfusiones, el cual presenta las características descritas en el epígrafe 2.

Se entiende por sistema de gestión de hemoderivados al proceso de control de la extracción de la sangre, el análisis y posterior suministro a los pacientes. En general, en España se utiliza para todo este proceso un sistema centralizado que se basa en la existencia de un único centro en donde se va a tratar la sangre recogida. Este centro es denominado en muchos casos *Centro Comunitario de Transfusiones* (figura 2).

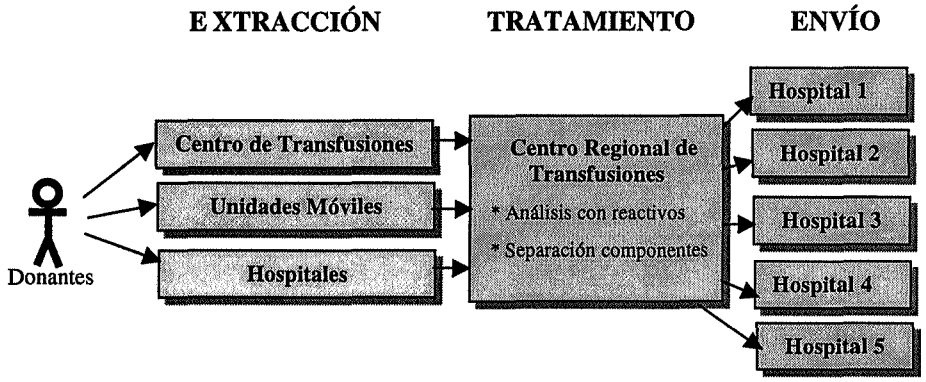


Figura 2. Flujo logístico de la sangre.

Como se observa, la sangre se obtiene de donantes voluntarios, ya sean personas inscritas a Hermandades de Donantes de Sangre o donantes no habituales, siendo el lugar de la extracción tanto el Centro Comunitario de Transfusiones como también unidades móviles u otros hospitales de la región.

Independientemente del lugar de la extracción, toda la sangre se envía al Centro Comunitario, en donde se analiza por medio de reactivos químicos. Con ello se pretende evitar enfermedades de transmisión sanguínea en las transfusiones (hepatitis, SIDA etc.). Posteriormente, a través de los procedimientos adecuados, casi todo el conjunto de la sangre entera se separa en componentes (sólo un 5% de la sangre recogida se deja en su mismo estado).

En la última fase, el Centro de Transfusiones funciona como un almacén que suministra la sangre necesaria para responder a las necesidades de los hospitales de la región, e incluso, en eventualidades, a los hospitales de otras Comunidades Autónomas.

Este mecanismo de gestión centralizado implica por tanto que puedan recogerse un número de bolsas de sangre en un hospital de la región, se envíen a analizar al Centro de Transfusiones y posteriormente esas bolsas vuelvan al mismo hospital. Aunque este proceso puede parecer lento y poco eficiente, es sin embargo el idóneo para Comunidades como el Principado de Asturias, donde las distancias entre el Hospital Central y el resto de hospitales no son grandes y por tanto los traslados no suponen ni un gran coste ni gran lapso de tiempo. En cambio, lo que tiene realmente un coste alto son los reactivos que se utilizan para analizar la sangre. Debido a que en cada análisis se comprueban unas treinta bolsas a la vez y que hay hospitales de la región que sólo consiguen tres o

cuatro bolsas en un día, el sistema centralizado permite obtener economías a escala en el análisis, ya que es prácticamente igual de costoso analizar esas cuatro que las treinta.

Por contra, el inconveniente que se puede plantear con este sistema es que cuando llega la sangre a los hospitales, ésta puede tener un avanzado estado de edad y, como normalmente el uso es menor que la demanda, mucha de la sobrante caduca. Para evitarlo, se ha adaptado el modelo lineal general descrito en el epígrafe 2 a este caso, de manera que se determine con qué edad máxima tendrá que enviarse la sangre desde el Centro de Transfusiones a los hospitales para minimizar la cantidad caducada. Hay que tener en cuenta que coexisten hospitales que tienen una actividad mucho mayor que otros y por tanto es más probable que caduquen menos unidades, aunque se les mande más de edad más avanzada.

En resumen, el objetivo es que el Centro realice una correcta política de envío en base a los datos que tiene de las caducidades históricas de los hospitales, de modo que el número de unidades de sangre que caduquen sean lo menores posibles. Se observa que el modelo descrito anteriormente se adapta al problema planteado de la gestión de sangre.

Siguiendo con la nomenclatura presentada en la sección 2, existen 8 tipos distintos de sangre ($l = 1,\ldots,8$), 9 hospitales donde distribuir ($j = 1,\ldots,9$), teniendo los días de vida del producto perecedero un rango de variación $i = 1,\ldots,35$ días. Pese a la simplicidad del modelo planteado, el tamaño en este caso concreto alcanza la cifra de 2968 variables y 504 restricciones (excluyendo las de no negatividad), por lo que se hace interesante tratar de automatizar su resolución a través de un programa informático para su evaluación diaria.

Este sistema de ayuda a la toma de decisiones, DSS, permite identificar cómo se debe distribuir la sangre desde la unidad central de análisis a los diversos hospitales, a partir de la introducción de unos datos sencillos (figura 3). La probabilidad de que caduquen los distintos tipos de sangre en función de su antigüedad y del hospital al que ésta es enviada puede ser calculada en base a los datos históricos disponibles en el Centro de Transfusiones, ya que éste obliga a los hospitales receptores a enviar un parte diario de situación. La cantidad de sangre de cada tipo disponible al comenzar la jornada en el Centro de Transfusiones y las demandas diarias de sangre, total y fresca, son los inputs que habrán de ser día a día introducidos al modelo. Los parámetros de edad media admisible para cada tipo de sangre y hospital, así como el número de días por debajo de los cuales la sangre se considera fresca, son también entradas del modelo, cuyo valor habrá de ser decidido por el responsable del servicio y periódicamente revisado en función de los datos reflejados por los partes diarios que envían los hospitales.

En la figura 4 se observa la pantalla inicial del programa en donde se introducen los datos de cantidad de sangre, clientes y demandas. Posteriormente el sistema resuelve el modelo ofreciendo la política de distribución recomendada (figura 5).

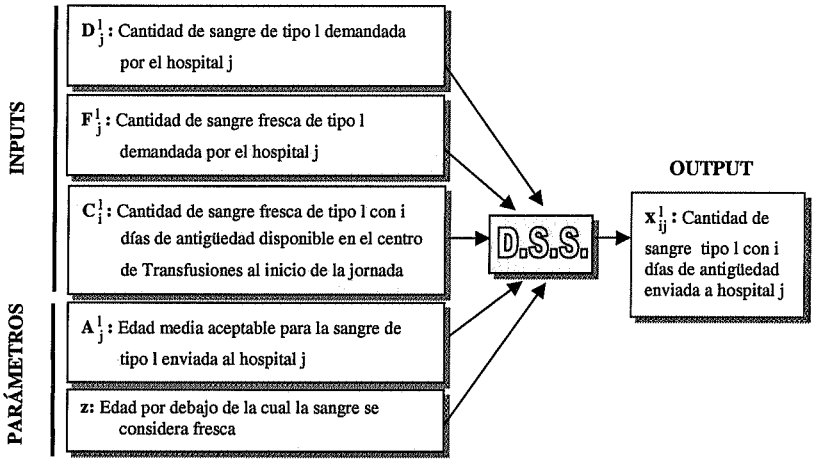


Figura 3. Esquema del D.S.S. desarrollado para el Centro Regional de Transfusiones.

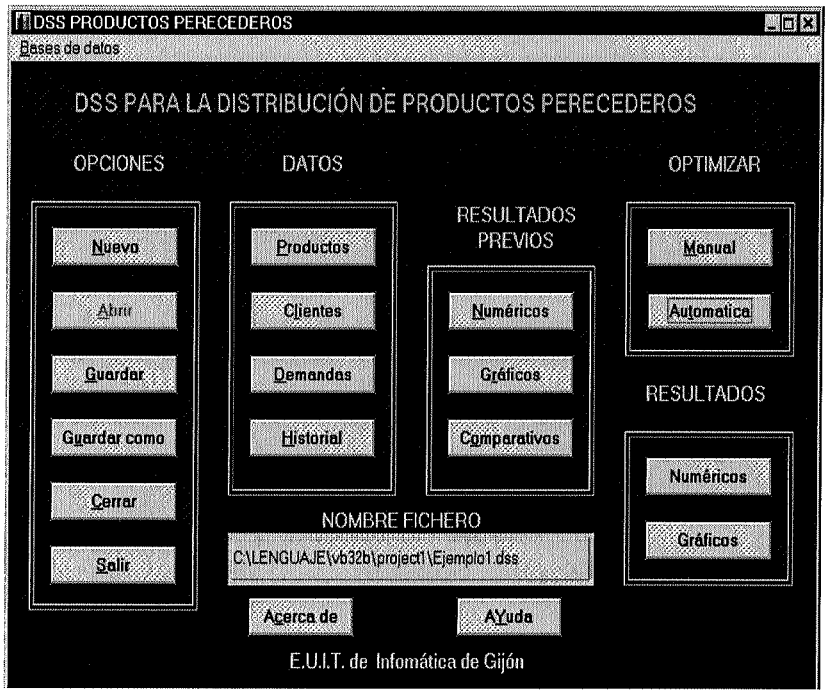


Figura 4. Pantalla inicial del DSS para la gestión de productos perecederos.

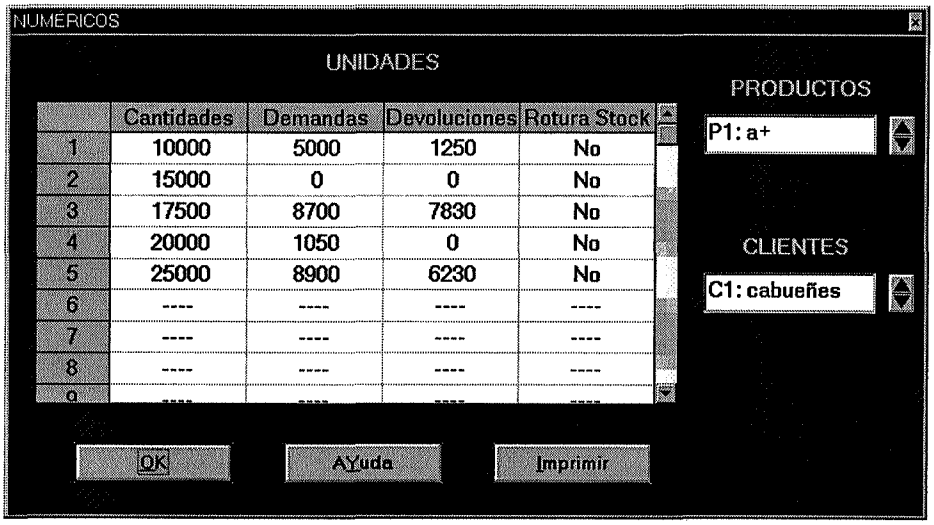


Figura 5. Salida con las recomendaciones de gestión del stock obtenidas del modelo. Cada línea representa un pedido.

4. CONCLUSIONES

El problema tradicional de suministro de productos desde un almacén central a los centros distribuidores se ve notablemente dificultado cuando se trata de gestionar artículos perecederos, ya que en este caso es preciso contemplar el riesgo de pérdida de unidades por caducidad.

Un caso particular de gestión de productos perecederos es el constituido por el envío de sangre a los hospitales regionales desde una unidad central, donde es recibida toda la sangre extraída a los donantes para su análisis y posterior utilización. Tal es el caso del Centro Regional de Transfusiones del Principado de Asturias, que ubicado en Oviedo abastece a los principales hospitales de la región. Para resolver el problema de distribución diaria de sangre, en este artículo se aplicó la metodología de programación por metas, desarrollándose un DSS para facilitar la evaluación diaria del problema por el personal habitual del centro.

Además, el propio programa permite acumular datos históricos, que sirven para actualizar la información del sistema y así revisar periódicamente el valor de los parámetros utilizados. A la vista de esos datos se observa que se ha reducido notablemente la cantidad de sangre caducada en los hospitales desde que se comenzó a aplicar el modelo (la figura 6 muestra una de las pantalla del sistema informático que permite comparar

los resultados), pasando, en el ejemplo, de una pérdida en promedio de 15.610 litros diarios de sangre de tipo A^+ antes a 5.960 litros desde la aplicación del mismo en uno de los hospitales abastecidos.

5. BIBLIOGRAFÍA

Adachi, Y., Nose, T. & Kuriyama, S. (1999). «Optimal inventory control policy subject to different selling prices of perishable commodities», *International Journal of Production Economics*, 60-61, 389-394.

Anónimo (1998). «Getting the fresh piece started», *Supermarket Business*, Nov. 1998, 13-16.

Hennessy, T. (1999). «Getting to the core», *Progressive Grocery*, 78, 5, 105-116.

Kendall, K. E. & Lee, S. M. (1980). «Formulating Blood Rotation Policies With Multiple Objetives», *Management Science*, 26, 1145-1157.

Klelnsorge, K. & Schary, B. (1989). «Evaluating Logistics Decisions», *International Journal of Physical Distributions & Logistics Management*, 19, 12.

Nahmias, S. (1982). «Perishable inventory theory: an overview», *Operations Research*, 30, 4, 680-708.

Nahmias, S. & Pierskalla, W. P. (1975). «Optimal Ordering Policies for Perishable Inventory I», *The Institute of Management Sciences*, 2, 485-493.

Rahim, M. A., Kabadi, S. N. & Barnerjee, P. K. (2000). «A single period perishable inventory model where deteriorating begins at a random point time», *International Journal of Systems Science*, 131-136.

Rifai, A. K. (1996). «A note on the structure of the goal-programming model: assessment and evaluation», *International Journal of Operation & Production Management*, 16, 1, 40-49.

Sculty, J. & Fawcett, S. E. (1993). «Comparative Logistics and Production Costs for Global Manufacturing Strategy», *International Journal of Operation & Production Management*, 13, 12.

Skjoett-Larsen, T. (2000). «European logistics beyond 2000», *International Journal of Physical Distributions & Logistics Management*, 30, 5, 377-387.

Van Zyl, G. J. J. (1964). *Inventory Control for Perishable Commodities*, Ph. D. Thesis, University of North Carolina, Chapel Hill, N.C.

Wee, H. (1999). «Deteriorating inventory model with quantity discount pricing and partial backordering», *International Journal of Production Economics*, 59, 511-518.

Williams, C. & Patuwo, B. E. (1999). «A perishable inventory model with positive order lead times», *European Journal of Operational Research*, 116, 2, 352-373.

ENGLISH SUMMARY

DSS MODELLING FOR PERISHABLE INVENTORY MANAGEMENT

B. DÍAZ FERNÁNDEZ*
J. A. DEL BRÍO GONZÁLEZ
B. GONZÁLEZ TORRE

Inventory management of perishable products has focused attention of researchers in Operation Management during the last decades. In this paper, a Decision Support System for perishable products distribution is developed and implemented, being applied to the blood delivery to different hospitals in a regional network from a central transfusion centre. In this case, the objective is not only satisfying the demands, but also trying to reduce the amount of blood that expires at the regional units (notice that clients do not suffer an additional cost by exceeding the expected level of inventory). For modelling this situation the approach selected was goal programming, due to the multicriteria nature of the problem. Important reduction of the quantity of wasted blood was detected in further application.

Keywords: Decision support systems, goal programming, gestión de inventarios

AMS Classification (MSC 2000): 90B05, 90B90, 90C29

* Autor para toda correspondencia: Belarmino Díaz Fernández. ETS Ingenieros Industriales. Universidad de Oviedo. Campus de Viesques. 33204 Gijón.

–Received January 1998.
–Accepted Mai 2001.

During the last few decades, increasing of competitiveness in the markets has forced researchers to look for ways of reducing the total cost of carrying a product to the market. At first, efforts were made on discovering patterns for reducing production costs, but nowadays, new research lines are focused on distribution and transport costs as well as in designing advanced methods for improving inventory management.

However, most of theses methods lack on applicability due to the introduction of an important number of hypothesis which are difficult to find in real environments. Besides, the complexity of the model gets increased in the case of perishable products, whom possibility of expiring must to be modelled as well as the need of finding a balance between the cost of higher inventory and the risk of loosing clients by no having available units. In these cases, general inventory models are not valid, so it becomes necessary to resort to tools such as linear programming, whose problem in this case is that constraints are very strictly established. For solving this problem, linear programming was substituted in our case by goal programming which.

Afterwards, the model was applied for implementing a decision support system for managing the supply of blood from a central transfusion unit to a regional net of hospitals.

Blood can be extracted in any of the hospitals, but it has to be carried to the transfusion centre, which operates as a central warehouse in which the analysis and treatment of the product takes place. After that, blood can be supplied to the hospitals according to the availability and their necessities. By centralising the analysis of the blood an important reduction in costs is obtained, thus the quantity of chemical products involved in the analysis does not depend on the quantity of blood processed. That means that if the process was not centralised, the necessities of chemicals (and consequently the costs) would be multiplied by the number of centres doing the analysis. The reduction in costs is much superior to the increase due to major needs of transport, at least if the central unit is strategically located and there are not long distances between the hospitals. That is the case of Asturian hospitals network, where the transfusion centre is quite well communicated and not very far from each particular unit.

In any case, decision of centralising the analysis, makes longer the time for getting the blood available for use. Therefore, it becomes much more important to find efficient models for distributing the blood between the transfusion centre and the hospitals, considering demands, levels of activity and expected period of life for an specific kind of product.

The inventory model for perishable product previously introduced was used for this particular case of distributing blood among a net of hospitals. A decision support system was built and in its further application it was found that an important reduction in expiring blood could be obtained applying the model. Therefore, goal programming seems to be quite a good methodology for creating efficient tools in the field of perishable inventory management.

Estadística Oficial

QÜESTIIÓ, vol. 25, 2, p. 303-334, 2001

TRATAMIENTO DE DATOS TERRITORIALIZADOS DE VIVIENDA EN EL INVENTARIO DE CAPITAL RESIDENCIAL

R. VERGÉS ESCUÍN
Universidad de Montreal*
Red Vergés, S.L.**

Después de plantear la filosofía de la contabilidad de capital fijo, el artículo examina la metodología necesaria para levantar el inventario permanente residencial de cada territorio. Al nivel de flujos, se ordenan y comparan las distintas fuentes de datos sobre edificación que existen en España y se describe el tratamiento requerido por cada una de ellas. Al nivel de stocks, se propone un procedimiento de ajuste y de desglose territorial del parque de viviendas por fecha de construcción. También se desarrolla la metodología para proyectar la supervivencia de los estratos de mismo período de construcción a partir de funciones de agotamiento. Se integra asimismo la variable de suelo urbanizable según el planeamiento vigente como indicador del desarrollo de nuevos estratos en las proyecciones del inventario. Por último se analizan algunos problemas acerca de la utilización de datos sobre clases de vivienda. El artículo concluye con la perspectiva de una valoración del capital fijo residencial al nivel territorial gracias a la emergencia de nueva información acerca de los mercados de vivienda.

Local housing data processing for the perpetual inventory of residential capital

Palabras clave: Capital, depreciación, inventario permanente, planeamiento, suelo, vivienda

Clasificación AMS (MSC 2000): 62P20

* Catedrático de economía inmobiliaria (profesor honorario). Faculté d'Aménagement. Coordinador de Estadística. Consejo Superior de Colegios de Arquitectos de España.
** Modelo RED 3. Correspondencia: Beatriz de Suabia, 152 B, 5° izda. 41005, Sevilla. Tel. (34) 954 921 829. redverges@arquired.es

- Recibido en febrero de 2001.
- Aceptado en mayo de 2001.

1. INTRODUCCIÓN

Pocas ciencias dependen tanto de la estadística como el saber económico, tal vez por ser más una ciencia de la vida social que una ciencia experimental. Y como que el objeto de la actividad económica es no sólo supervivencia sino también desarrollo de cara al cumplimiento del contrato social, es importante llevar medida estadística del nivel y distribución de la riqueza, tanto del capital productivo como del no productivo, es decir del capital residencial.

Se da el caso que la medida estadística de la riqueza residencial y de sus servicios derivados goza de gran tradición, lo cual se explica en gran parte porque su conocimiento ha sido desde siempre esencial en las labores de levantamiento, enumeración, recaudación y distribución de servicios públicos. Sin embargo, la dificultad de reunir durante años nueva información, ha frenado su desarrollo teórico, de forma que para poner en aplicación el arsenal estadístico actualmente disponible hay que revisitar los trabajos pioneros, lo cual explica la preponderancia de referencias de postguerra en el presente artículo.

La necesidad de reanudar los trabajos de medida de la producción y acumulación de riqueza residencial es cada vez más evidente, debido a la inesperada aceleración de su proceso de formación, el cual está desembocando por cierto en una inquietante sobreproducción edificatoria (Vergés, 2000). En esta línea, examinaremos en el presente artículo los métodos de medida y de previsión del nivel de riqueza que se corresponde con el capital fijo residencial, haciendo especial hincapié en la información disponible para llevar a cabo el análisis cuantitativo desde un enfoque espacio-temporal. Este enfoque es esencial por tres razones:

La primera razón es conceptual: los procesos acumulativos de capital a lo largo del tiempo se caracterizan fundamentalmente por su volumen físico y su distribución territorial antes que por su valor. La segunda razón es práctica: los sistemas de información geográfica han abierto posibilidades inéditas de representación territorial en tiempo real para este tipo de procesos, cuya figuración cartográfica era hasta hace poco aquí onerosa y prolongada. La tercera razón es teórica: el sector residencial y de servicios conexos tiende a substituir al sector agrícola como representante de la actividad de carácter extensivo, fijo y de baja tecnología en los modelos territoriales de *competitividad monopolística* de tipo Dixit-Stiglitz (1977), con los que se intenta desarrollar una nueva *economía espacial* vinculada con las actividades de exportación (Fujita, Krugman, Venables, 1999).

Veremos que los componentes espacio-temporales del capital residencial son de dos órdenes: el de las *viviendas* y el del *suelo* para construir de nuevas. Veremos también que en ambos existen dos vertientes de un mismo proceso: la formación por *flujos* y la acumulación en *stocks*. Veremos por fin, de qué métodos disponemos no sólo para medir sino también para prever la evolución de todos estos componentes.

2. FLUJOS Y STOCKS DE CAPITAL FIJO

En la tradición del *Income & Wealth*, la información relativa a flujos de riqueza incluye la formación y el consumo de capital, mientras que en lo relativo a su acumulación aparecen los stocks de capital o parques. La articulación entre unos y otros viene dada por una sencilla cadena de Markov popularizada por Goldsmith (1951) bajo el término de *Inventario Permanente*. En un momento t_1, el stock de capital es igual al stock anterior t_0, más la formación de nuevo capital, menos el consumo de capital existente ocurridos entre t_0 y t_1.

Bajo esta claridad conceptual se esconden numerosas dificultades. La más importante es sin duda la de medir el consumo de capital de forma aceptable. Esta dificultad viene compensada por la existencia de «hitos» (*benchmarks*), es decir de observaciones periódicas de stocks o parques. Como además suele llevarse contabilidad de la formación de capital, la vía generalmente utilizada para cerrar el inventario, consiste en despejar residualmente el consumo tratándolo como variable dependiente:

$$(1) \qquad CCF(t_1 - t_0) = SCF(t_0) - [SCF(t_1) - FBCF(t_1 - t_0)]$$

donde *CCF* es el consumo, *SCF* es el stock y *FBCF* es la formación bruta de capital fijo, todos estos conceptos medidos en los instantes o períodos indicados.

La instrumentación del stock de capital no sólo como medida de la riqueza sino también como ingrediente de las funciones de producción y productividad, ha constituido una aportación decisiva del neoclasicismo a la ciencia económica. Pero no ha sido tarea fácil, sobre todo a la hora de aplicar el concepto de inventario permanente a las cuentas macroeconómicas de capital, como lo atestigua la frecuencia y la intensidad de los debates (ver por ejemplo, la controversia entre Denison y Jorgenson-Griliches en 1969, reeditada por el SCB, 1972). Hay que decir que la mayor parte de críticas cruzadas se ha centrado una y otra vez sobre el problema de la *agregación*, o sea sobre el problema clave que es el de sumar valores de bienes heterogéneos y mutantes sometidos además a fluctuaciones de precios en los mercados (Usher, 1980).

Ahora bien, existen inventarios cuyos bienes poseen la propiedad de ser relativamente homogéneos, como los de viviendas familiares, automóviles de turismo, etc. los cuales parecen quedar bastante al margen de la crítica. La razón de ello es que los stocks suelen prestarse hasta cierto punto al procedimiento de la valoración unitaria, es decir, a la valoración de stocks por producto de cantidades por precios. Ello no significa que este que este tipo de valoración esté exento de problemas, ya que homogeneidad no significa identidad. Lo que queremos decir es que el problema de sumar distintas clases de manzanas es más fácil de resolver que el de sumar distintas clases de manzanas con distintas clases de peras, que es el problema que tiene que resolver, por ejemplo, la cesta del IPC. Por tanto también la valoración de stocks de bienes homogéneos deberá levantar las sempiternas dificultades de *vintage*, precios *hedónicos* o *números índices*, típicas del citado problema de la agregación (Hicks, 1973).

Lo que sí es una ventaja para los stocks de bienes homogéneos es de poder contarlos antes de buscar a saber cuanto valen. Por ejemplo, el censo de edificios puede decirnos cuántas viviendas hay en España y de qué clase, etc. y podemos sacar provecho de ello. En cambio, ningún censo puede decirnos cuántas plantas de producción hay en la industria alimentaria, por ejemplo, porque las hay de tantos tipos, tamaños y usos específicos que sería imposible reducirlas a un común denominador y que de lograrlo, tampoco nos sería de utilidad. Por todo ello, la estimación del capital fijo residencial puede y debe empezar por su inventario físico.

3. CAPITAL FIJO RESIDENCIAL

En la tradición del inventario de capital, lo residencial (en muchos casos llamado llanamente capital no productivo) ha ocupado un lugar preeminente en los grandes trabajos clásicos de Goldsmith (1956), Denison (ver *SCB*, 1969), Kendrick (1976), etc. Ello es debido a tres razones principales. Primera, la vivienda constituye sin duda el mayor sector de capital fijo de cualquier país. Segunda, la contabilidad física es relativamente sencilla, como acabamos de ver. Tercera, el inventario puede ser llevado en términos de unidades solas, en términos de superficies e incluso en términos de valor, es decir en términos de producto de cantidad de unidades por precios.

3.1. Inventario físico por unidades

El inventario físico por cantidad de viviendas es el más utilizado, aunque no esté exento de problemas. En efecto, no se dispone de datos acerca de las unidades reunidas ni divididas. Tampoco se dispone de los cambios de uso, que sean viviendas que dejan de serlo o que sean locales que pasan a ser viviendas. Por fin, no existen datos fiables acerca del flujo de derribos, ni de la desaparición contable de viviendas rehabilitadas. Éstas últimas suelen figurar conjuntamente con las nuevas unidades en el flujo de formación, pero no consta contrapartida alguna en forma de «acta de defunción» de las antiguas viviendas supuestamente llegadas al término de su vida útil. Así pues, para cerrar la ecuación (1) utilizada en el inventario físico, no queda más remedio que considerar al saldo neto de todas estas mutaciones como a un solo concepto que sería el equivalente cuantitativo del *CCF* o consumo de capital fijo.

3.2. Inventario físico por superficies

Aunque poco utilizado, el inventario físico por superficie podría ser más preciso, puesto que suma unidades métricas obviamente homogéneas, mientras que el inventario anterior sumaba unidades funcionalmente homogéneas pero de distintas tipologías. Pero las

series en superficie, que son relativamente fiables en los stocks gracias a la declaración censal de los ocupantes relativa a superficie habitable de la vivienda, no lo son tanto en las series de flujos ya que estas suelen incluir en la superficie construida total de los edificios a espacios comunes y a locales de uso comercial, garajes en sótano, etc. Por esto, el dividir superficie total por número de viviendas conduce a sesgar al alza la superficie media de la vivienda, amén de que una cosa es la superficie construida utilizada para calcular costes de construcción y otra cosa es la superficie habitable que figura en los catálogos de venta.

3.3. Inventario de capital fijo

Para confeccionar el inventario de capital propiamente dicho, se utilizan en principio dos métodos, el longitudinal y el transversal. El método longitudinal acumula simplemente la *FBCF* adoptando un hipotético *CCF* a partir de una estimación inicial situada generalmente en el origen temporal de las series de flujos. El resultado refleja un valor de substitución de las estructuras existentes, cuyo tendón de Aquiles es, como siempre, la estimación de una depreciación tanto física como económica.

De hecho, ésta ha sido estudiada sobretodo en los sectores industriales y de equipamientos. Aunque la depreciación fue formulada ya en 1925 por Hotelling y desarrollada operativamente en 1955 por Grant y Norton, y a pesar de los trabajos históricos concluidos por Winfrey en 1926, la escasez de información actualizada acerca de las duraciones de vida útil y de la pérdida de valor físico, productivo o económico (obsolescencia) del capital fijo, el inventario de capital ha acabado por transformar el inventario longitudinal en un procedimiento puramente académico. Así y todo, numerosos sistemas de contabilidad nacional suelen llevar alguna forma de inventario ante la necesidad de las administraciones por justificar partidas de inversión para consumo de capital fijo tan tangibles como substituciones, mejoras o rehabilitaciones. En cuanto al capital residencial, cabe mencionar los trabajos del autor para el gobierno francés, (Vergés, 1989) cuya metodología fue luego recomendada por el *Housing, Building and Plannig Committee* (ECE) de Naciones Unidas (Lujanen, 1985).

En cuanto al método transversal, consiste simplemente en encuestas puntuales (ver el clásico estudio de la Universidad de Deusto, 1968) o bien en valoraciones de series físicas derivadas de los hitos censales (Naredo, 2000), en ambos casos referenciadas a precios de mercado. El problema es que generalizar precios de mercado a un stock del cual sólo una pequeña parte sale al mercado a la vez, puede parecer ilógico, como señaló ya en su día Hicks (1946). En efecto, si todo un barrio se pusiera en venta al mismo tiempo, seguro que los precios bajarían. El principio de que los precios dependen de las cantidades, conocido como *ley de Gossen* (Kauder, 1965) es el mismo que se transgrede cuando se impone cualquier «*catastrazo*», lo cual no deja de provocar, acto seguido, el natural y consiguiente rechazo social. En todo caso, para llegar a estimaciones más

precisas del capital fijo residencial, sería necesario adentrarse en el arduo problema de los precios y de su agregación antes citado.

Por tanto, nos limitaremos en primero a abordar aquí lo que conocemos mejor, es decir las cantidades, concretamente las unidades de vivienda, tanto al nivel de flujos como al nivel de stocks. Veamos ahora de qué material se dispone en España para construir el inventario permanente del parque de viviendas, empezando por las series más populares que son los flujos de edificación.

4. FLUJOS DE EDIFICACIÓN

El conocimiento de los flujos de edificación residencial es esencial para estructurar la variable formación bruta de capital fijo (*FBCF*), particularmente en el caso la del inventario físico por unidades de vivienda. En términos contables, sólo las estructuras constituyen capital, puesto que ni el suelo ni el derecho de ocupación son productos. Esta especificidad no es muy útil a la hora de confeccionar el inventario de viviendas. En cambio debe ser tenida en cuenta a la hora de estimar la *FBCF* en la contabilidad nacional, puesto que el índice de precios implícitos a utilizar es el de las estructuras solamente, muy diferente del índice del precio de venta el cual incluye también al suelo (Sirmans & Redman, 1979).

Por otro lado, de lo que es obra de edificación residencial en España, sólo se conocen las cantidades agregadas gracias a los proyectos de arquitecto, como veremos enseguida. En cambio, los presupuestos de ejecución material que acompañan a dichos proyectos están muy sesgados a la baja y de momento no hay perspectiva de que rectifiquen (Vergés, 1992 y s.).

4.1. Edificación iniciada

En España, la principal fuente de información acerca del flujo de edificación proviene de los proyectos de arquitecto visados por sus propios Colegios Oficiales. Las series trimestrales se iniciaron en 1960 y, hasta 1990, se limitaron al número de viviendas por régimen de libres y de protección oficial, proyectadas en cada demarcación o provincia. Hasta 1992, las series no eran del todo fiables, entre otras cosas porque podían contener anteproyectos, etc. Sin embargo, constituyen la única fuente histórica sobre la edificación legal de los años de mayor crecimiento residencial de España. En todo caso, es a partir de estas series que la Dirección General para Vivienda y Arquitectura del MOPU (hoy Ministerio de Fomento) ha calculado desde siempre las viviendas iniciadas y terminadas aunque mediante una metodología de desfase que no ha sido nunca publicada.

A partir de 1992 las series de visados de arquitecto han sido mensualizadas y se ha ampliado el número de sus variables con la superficie de edificios tanto residenciales como no residenciales, esperando poderlas ofrecer a partir de 2002 a un nivel municipal. Además, se recogen solamente los proyectos de ejecución cuyos honorarios han sido ya abonados al arquitecto por el cliente, el cual ha debido adquirir previamente el solar (Vergés, 1992 y s.). Por estas razones, la probabilidad de que se abandone ulteriormente el proyecto es ínfima, aunque a veces se posponga el inicio de obra por razones económicas o logísticas.

Por su lado, Fomento recoge las lecturas que los aparejadores hacen de los proyectos de arquitecto cuando el cliente les contrata para dirigir la obra (Sánchez de Rivera, 1990 y s.). Naturalmente, la información que aparece en las respectivas estadísticas es la misma, salvo que su difusión es más tardía. En efecto, el aparejador suele ser contratado unos meses después del visado de arquitecto. Además, Fomento tiene que procesar todos los datos a partir de cero. Su cobertura es también algo inferior, puesto que en algún tipo de proyecto, sobre todo no residencial, el cliente acostumbra a prescindir de los servicios de aparejadores. En cambio, las cifras de estos últimos se pueden obtener municipalizadas, característica que los Colegios de Arquitectos ofrecerán tan solo a partir de 2002.

Fomento publica también una serie sobre licencias de obra concedidas por los ayuntamientos (Toro, 1990 y sig.). Esta serie recoge los cuestionarios que el «técnico competente» (en principio el arquitecto) rellena a la solicitud. Por tanto, su cronología debería situar la información dentro del plazo que separa el visado de arquitecto del contrato de dirección de obra del aparejador. Uno de los problemas que se presentan de cara a la comparabilidad cronológica es que los ayuntamientos pueden conceder licencia a un proyecto dicho «básico», que suele ser un simple estudio edificabilidad y que por supuesto no entra en la estadística de visado a pesar de llevar sello del Colegio Oficial de Arquitectos. Así, de tener lugar, este estudio es obviamente anterior al visado del proyecto de ejecución, que es el que recoge la estadística.

Además, existe toda una tipología de licencias de obra cuya selección y tramitación a fines estadísticos exige una dedicación y celeridad que no todos los ayuntamientos están en medida de asumir. Subsisten pues algunas dudas acerca de la cobertura y comparabilidad entre esta serie y las otras, ya sea de los Colegios de Arquitectos, ya sea de Fomento (aparejadores).

De todo ello se deduce que, en materia de flujos de unidades de vivienda y de superficie residencial o no residencial iniciada, lo más seguro es utilizar ya sea los visados de arquitecto cuando se busca antelación y cobertura, ya sea los visados de aparejador cuando se busca desglose municipal, por lo menos hasta que los Colegios de Arquitectos hayan completado el proceso antes mencionado de municipalización de su propia estadística.

4.2. Edificación terminada

Puesto que una estructura entra a formar parte del inventario sólo cuando se completa, hay interés en disponer de datos seguros acerca de las terminaciones de viviendas u de otros edificios. Al respecto, Fomento produce también series municipalizadas relativas a certificación de fin de obra de los aparejadores, aunque las publique solamente al nivel de provincias y de tamaño de municipios. Estas series parecen infravaloradas, ya que su acumulación desvía de forma creciente con respecto a la de visados de dirección de obra.

Una manera de superar las insuficiencias de estas series es de calcular una función de desfase que se pueda aplicar al flujo de visados, presumiblemente exhaustivos. Esta función ha sido establecida a partir del análisis de desfase (*lags*) sobre largo período efectuado mediante una explotación de la base de datos del Col·legi d'Arquitectes de Catalunya que también posee datos acerca de certificación de fin de obra. En esta función, el desfase se expresa en porcentaje de viviendas terminadas en cada trimestre con respecto al número total de visadas en un trimestre inicial «cero».

Los resultados han sido redistribuidos sobre 16 trimestres ya que a partir de este plazo aparece una «cola» poca significativa en volumen, pero larga en el tiempo debido a tardanzas técnicas o administrativas. Así, es frecuente certificar finales de obra de edificios ocupados desde hace meses. También es frecuente certificar finales de obra de edificios visados hace tiempo pero iniciados más recientemente. Finalmente, los porcentajes de finales de obra propuestos para cada trimestre a partir del visado de obra en un trimestre *cero* son los siguientes:

Trim.	*%*	*Trim.*	*%*
1	2,1	9	8,5
2	2,0	10	7,2
3	4,4	11	5,7
4	9,0	12	4,8
5	12,0	13	3,8
6	13,1	14	2,9
7	12,1	15	1,9
8	9,5	16	1,0

Esta función puede aplicarse a los visados de arquitecto. Por ejemplo, aplicando estos porcentajes a un flujo de 1000 viviendas visadas en el último trimestre del año 1996, se obtiene que 21 debieron terminarse en el primer trimestre de 1997, 20 en el 2°, 44 en el 3°, 90 en el 4°, etc. y así hasta el último trimestre del 2000 en el que se terminarían las 10 últimas viviendas. Luego, para obtener la previsión del número total de viviendas

terminadas *VT* en un trimestre y de un año determinado, deberán sumarse todas aquellas viviendas *vt* iniciadas en los 16 trimestres anteriores *i* que se terminan en el trimestre *y* considerado.

$$VT_y = \sum_{i=y-1}^{i=y-16} vt_y^i \qquad (2)$$

Por tanto, esta previsión es esencialmente retrospectiva, destinándose a paliar la cobertura insuficiente de la serie de certificación de fin de obra. Además sólo comienza a ser válida a partir del trimestre 16 después del trimestre de arranque de la serie de visados. Sin embargo, puede utilizarse para prever a corto plazo el nº de terminadas suponiendo que el nº de visadas de los próximos trimestres sea más o menos de mismo orden que el de los trimestres actuales. Conviene señalar también, que el resultado aparece siempre muy suavizado, contrastando con los frecuentes altibajos de las certificaciones documentadas. Ello es debido a que la configuración de la función efectiva de desfase puede variar de un trimestre a otro, mientras que en el cálculo se supone que es invariable.

4.3. Viviendas terminadas en las últimas décadas censales

Los censos de edificios sirven ante todo para actualizar el callejero de recorrido previo al censo de población y vivienda. Se aprovecha entonces dicho recorrido para recoger las características de los edificios, enumerando entre otras cosas sus viviendas e indagando la década de construcción. Si ésta es la última década, se indica también el año de construcción, dato que se puede enlazar con el de censos anteriores proporcionando así una serie anual continua de viviendas terminadas en cada municipio que podría retroceder hasta 1971 e incluso hasta 1961 en los grandes municipios.

Estos datos se publican en porcentajes estructurales para los municipios de más de 10.000 habitantes en los censos de edificios de 1980 y de 1990. Su transformación en cantidades absolutas presenta ciertas dificultades que deben solventarse recurriendo a los resultados cuantitativos del censo de viviendas, aun cuando éste se haya administrado desde 1981 con 4,5 meses de posterioridad con respecto al de edificios (del 15 de Octubre de 80 o 90 al 1 de Marzo de 81 o 91).

En Cataluña donde desde 1987 se recogen las mencionadas certificaciones de fin de obra, el número de viviendas terminadas según el censo de edificios aparece superior al cúmulo de certificaciones de fin de obra. Ello es debido a que el censo rastrea el territorio y que documenta a todas las viviendas existentes aunque provengan de alguna *mutación*, incluso a aquellas construidas sin el recurso ni del arquitecto, ni del aparejador, ni del ayuntamiento, es decir a las viviendas no declaradas o ilegales.

Puede incluso que algunas de dichas viviendas no aparezcan en el callejero censal a pesar de estar ya ocupadas y que aparezcan en el censo siguiente por haber sido legalizadas entretanto. Esto puede dar lugar a un fenómeno conocido como «inflación de la

penúltima cohorte». Su corrección es delicada puesto que debería implicar corrección del censo anterior, cosa a la que el estadístico se muestra naturalmente reacio. Consecuencia de ello es que el número de viviendas terminadas en los últimos años de la última década de cada censo puede estar sesgado a la baja.

Otro problema añadido es que la serie decenal de viviendas terminadas según el censo aparece a menudo sesgada por el efecto «cero» y «cinco». Ello es debido a que cuando el agente no logra dar con el año exacto de terminación de un edificio reciente, tiende a estimarlo él mismo redondeando el año, por supuesto. A pesar de ello, las series del censo de edificios son mucho más objetivas que las del censo de vivienda, donde se pregunta por el año de construcción esta vez al ocupante, el cual suele confundirlo con el de su propia entrada en la vivienda (la pregunta al ocupante desaparece en el censo del 2001).

El enlace entre las series 1971-80 del censo de edificios de 1980 y 1981-90 del mismo censo de 1990 está disponible y se prolongará hasta el 2000 cuando aparezcan los resultados del censo de edificios de 2001. Pero hasta 1990, estas series sólo cubren las viviendas terminadas en los municipios de más de 10.000 habitantes, además del conjunto formado por los municipios de menor población de cada provincia.

Una vez completadas las series de viviendas terminadas, conviene proceder a su corrección territorial en aquellos casos donde ha habido fusión o segregación de municipios. Pero en este último caso no se sabe en qué parte del municipio se terminaron las viviendas antes de la segregación. Por tanto, si las cifras son pequeñas, lo mejor es proceder por prorrateo referido al volumen del parque según los *Nomenclátors*, siendo el caso más corriente aquel en que se segregan entidades enteras de población. En cambio, si las cifras son importantes, lo mejor es recabar información circunstancial en el ayuntamiento.

En definitiva, la única serie aconsejable de viviendas terminadas entre 1971 y 1990 es la que deriva de los censos de edificios según el procedimiento descrito, incluso para los años 1987-90 en Cataluña. Para la década de los 90, hay que esperar a disponer de la serie del censo del 2001. Mientras tanto, para esta última década, deben utilizarse los flujos procedentes de visados de arquitectos o de aparejadores, trabajados en las condiciones anteriormente descritas.

5. PARQUE DE VIVIENDAS

Hasta aquí hemos examinado el material retrospectivo disponible para construir la variable *FBCF* en la ecuación (1), de forma a desarrollar el inventario permanente de vivienda en su componente físico del capital residencial. El siguiente paso consiste en construir la variable stock de capital fijo (*SCF*), y en derivar el consumo del mismo (*CCF*), no sólo retrospectivamente sino también en proyección.

5.1. Censos de edificios

5.1.1. Estratos por fecha de construcción

La pregunta acerca de la fecha de construcción de los edificios residenciales aparece en los censos de la mayoría de países industrializados. Sin embargo, en pocos lugares se ha utilizado el potencial de conocimiento ofrecido por el tratamiento estadístico de su enumeración, con vistas a confeccionar de forma objetiva un inventario permanente del parque de viviendas. Esto es debido ante todo a la dificultad de pasar del universo estocástico del censo al universo contable del inventario. Sin embargo, una buena depuración de las respuestas a la pregunta de la fecha de construcción permite llegar a resultados satisfactorios y de gran utilidad para la confección del inventario.

Para ello, no hay duda que el mayor problema por resolver reside en la diferencia fundamental entre la demografía humana y la llamada demografía residencial. En efecto, a partir del origen que es el nacimiento, un ser puede cambiar de territorio pero no puede desdoblarse ni cambiar de sexo (en todo caso no de forma habitual...). En cambio, desde su construcción, un edificio puede cambiar de uso, de tamaño o de contenido pero no de lugar. Así, un *estrato* formado por viviendas terminadas dentro de un período determinado de construcción, puede verse disminuida por la transformación de parte de ellas en otra cosa, o bien aumentada por división (de una vivienda pueden salir dos...). En definitiva, los individuos de una *cohorte* se mueven pero no mutan mientras que las viviendas de un *estrato* mutan pero no se mueven.

A causa de su inmovilidad, las series de estratos de viviendas de mismo período de construcción deben seguir un doble principio. Primero: un estrato no puede aumentar entre un censo y el censo siguiente porque, de hacerlo, contabilizaría viviendas terminadas entre tanto y por tanto asignables a un estrato más reciente. Segundo, el saldo neto de mutaciones incluyendo la desaparición o derribo debe ser negativo, salvo si se documenta lo contrario. En este caso, debe reducirse la distorsión analizando la casuística y adoptando una criterología apropiada: si hay adición de capital como sobreelevación de edificios o locales transformados en viviendas, lo nuevamente creado debe asignarse al estrato de creación, no al del edificio de origen. Si hay subdivisión, debe reducirse el número de unidades resultantes al de unidades equivalentes de origen, etc.

Esta casuística y su criterología han sido estudiadas en el caso de Francia (Vergés, 1989), gracias a la explotación de la encuesta del empleo. Dicha encuesta da lugar a un tratamiento periódico de máximo interés para enumerar cambios acaecidos entre las visitas repetidas a las viviendas de la muestra: desaparición, cambio de uso (vivienda principal en secundaria o vacante...), cambio de destino (en local...), etc. En España no consta haberse intentado algo semejante con la *EPA*, tal vez porque ello hubiera requerido ciertos retoques al cuestionario. Por tanto, cuando aparecen estratos crecientes, debe aplicárseles un método *hot-deck* de mínima corrección.

Este método se justifica además por el hecho de que sólo es necesario aplicarlo en estratos aislados, no habiéndose observado caso alguno en el que el conjunto del parque ya existente en el censo anterior fuera superior al mismo en el censo siguiente. Tampoco se observa crecimiento en los estratos de antes de 1940. Entonces, en el censo de edificios de 1990 se puede suponer que las aparentes variaciones positivas del saldo neto de mutaciones en los estratos que van de la postguerra hasta 1980, son debidas ya sea a imprecisiones en la respuesta a la pregunta acerca de la fecha de construcción del edificio, ya sea a problemas de depuración discutidos en su momento con el INE.

Para resolver este problema se ha confeccionado un programa *EDIF* que optimiza el ajuste de estratos en 1980 y 1990 en los municipios de más de 10.000 habitantes, que son los que se desglosan en los correspondientes censos. Concretamente, se definen unas evoluciones estructurales mínimas por estrato que sólo se aplican cuando hay crecimiento o decrecimiento insuficiente del mismo. Si todos decrecen, no se aplican las mínimas, incluso si los hay que estén por encima. Cuando se reduce un estrato a la mínima, el excedente se transfiere discrecionalmente al o a los estratos contiguos y de mayor decrecimento intercensal. En ciertos casos, el importe de la transferencia puede resultar más «económico» si se corrige el censo anterior. Cabe mencionar asimismo que como que los estratos cubren decenios enteros, conviene trasladar los hitos censales al 31 de diciembre procediendo por interpolación entre el 15 de octubre anterior y el 1 de marzo siguiente en el caso del parque total y entre las fechas de dos censos de edificios consecutivos en el caso de los estratos o fechas de construcción.

5.1.2. Municipios de menos de 10.000 habitantes

De los municipios de menos de 10.000 habitantes, los censos de edificios sólo ofrecen el período de construcción del conjunto. Este conjunto puede ser desglosado en 1990 mediante un analizador de correspondencia para cuadros rectangulares como *ACR* (ver anexo). Su objetivo es crear una *matriz de correspondencia* de la que sólo se conocen los totales horizontales en *ladillo*, los totales verticales en *cabecera*, así como la *estructura de base* que vincula a las informaciones de la matriz entre sí. Una vez resuelto el problema, los totales de la *estructura resultante* coinciden con *cabecera* y *ladillo* y satisface además la condición de mínimos cuadrados con respecto a la *estructura* de base propuesta. El algoritmo de *ACR* viene descrito en anexo.

En el caso concreto de 1990, la cabecera es la fecha de construcción del total de municipios menores de 10.000 antes mencionados y el *ladillo* es el total de viviendas de cada uno de estos municipios extraído del censo de viviendas y ajustado al total del censo de edificios. La *estructura de base* viene proporcionada por los datos informatizados del censo de viviendas de 1991 y difundidos por el INE bajo el acrónimo *SAETA*, produciendo así el desglose en forma de *matriz de correspondencia*, es decir el número ajustado de viviendas de cada municipio según de fecha de construcción. No se puede realizar el desglose equivalente de 1980 puesto que no se editó entonces algo parecido

a *SAETA*. Lo que sí se puede utilizar es el programa *EDIF* para estimar la matriz de correspondencia de 1980 a partir de la de 1990, lo que a efectos del desarrollo presentado en §5.2 es ampliamente suficiente.

Pero antes deben tenerse en cuenta a algunas incoherencias observadas en los estratos municipales de 1991 proporcionados por *SAETA*. Estas incoherencias son difíciles de corregir mediante *ACR*. Ellas son debidas a que la depuración de datos no consideró el hecho de que casi la totalidad de las numerosas viviendas censadas sin fecha de construcción eran secundarias o desocupadas, o sea que no había nadie en la vivienda para responder a la pregunta. Ahora bien, por lógica, puede deducirse que estas viviendas eran de la última década, puesto que sin ellas no se podría explicar el crecimiento observado en el parque entre 1980 y 1990, dada la insuficiencia de viviendas principales en el estrato 1981-90. Para resolver este problema, deben solicitarse las explotaciones autonómicas del censo de viviendas de 1991 que contengan el desglose cruzado por uso y por fecha de construcción y que especifiquen el campo «no consta», de manera a proceder a una reasignación por estratos antes de aplicar el método antedicho.

5.1.3. Actualización territorial

Como en el caso de las series de flujos, debe procederse también a la corrección territorial en aquellos casos donde ha habido fusión o segregación de municipios. En este último caso, si se han segregado entidades enteras de población, la consulta de los *Nomenclátors* permite resolver el problema, sino conviene recabar información en cada ayuntamiento.

En definitiva, el tratamiento por *hot-deck* del material censal permite un ajuste «suave» de la estadística del parque de viviendas por fecha de construcción desde el 1-1-1981 hasta el 1-1-1991 en cada municipio de más de 10.000 habitantes y en los grupos de municipios más pequeños. Esta estadística puede completarse con una serie no publicada del parque total de viviendas por municipio en el padrón de 1996, lo cual permite además ajustar provisionalmente la serie de viviendas acabadas según Fomento (aparejadores) desde 1991 hasta 1995.

5.2. Análisis longitudinal por fecha de construcción

Hemos visto que, como en todo fenómeno demográfico, un parque está constituido por estratos de viviendas de mismo período de construcción. Estos estratos están sometidos a procesos de agotamiento que pueden diferir dentro de un mismo parque (las viviendas antiguas pueden ser más duraderas que las actuales o lo contrario) o de un lugar a otro (las viviendas de un país nórdico afrontan condiciones climáticas más duras que las del sur), etc., etc.

Por consiguiente, el conocimiento de las leyes estadísticas que rigen estos procesos es de máxima importancia para prever la evolución del parque de vivienda existente. Veamos pues como formalizar las funciones recíprocas de agotamiento y subsistencia de dichos estratos, utilizando la misma notación que en la ecuación (1) y empezando por el *subestrato* formado por las viviendas terminadas en un solo año t_0.

5.2.1. Tasa relativa de agotamiento y tasa de subsistencia

Supongamos un capital *SCF* formado por viviendas terminadas en t_0 y de duración máxima de vida igual a E. La *tasa relativa de agotamiento* ν_t (*depletion*) es la proporción de capital subsistente al final de $t-1$ que se agota durante el período t. Esta tasa es nula en t_0 puesto que no existía parque en t_{-1} e igual a 1 en E, puesto que sea cual sea el valor de *SCF* en t_{E-1}, no quedará capital a partir de E.

$$SCF_{t-1} - SCF_t = SCF_{t-1}\,\nu_t, \qquad t = 0,1,\ldots,E \tag{3}$$
$$\nu_{E+1} = 1$$

Llamemos entonces *tasa absoluta de agotamiento* μ_t a la relación entre el agotamiento durante el período t y el capital de origen:

$$\mu_t = (SCF_{t-1} - SCF_t) \div SCF_0 = SCF_{t-1}\,\nu_t \div SCF_0, \qquad t = 0,1,\ldots,E+1 \tag{4}$$

Por otro lado, la *tasa de subsistencia* v_t es la proporción de capital subsistente al final del período $t-1$ con respecto al capital de origen SCF_0.

$$v_t = SCF_{t-1} \div SCF_0, \qquad t = 0,1,\ldots,E+1 \tag{5}$$

$$\mu_t = \nu_t\, v_t, \qquad t = 0,1,\ldots,E+1 \tag{6}$$

Si consideramos a t infinitamente pequeño, las tasas pueden ser representadas por funciones continuas de t, de intervalo $[0,E]$ y cumpliendo la condición (6) además de las siguientes:

$$\int_0^E \mu_t\, dt = 1 \tag{7}$$

$$v_t = 1 - \int_0^t \mu_x\, dx \tag{8}$$

Estas funciones continuas ν_t y v_t se desconocen. Sin embargo, si se dispone de hitos (*benchmarks*) pueden calcularse ciertos puntos de v_t mediante (5). Es necesario, además que, sea cual sea la función μ_t, ésta respete siempre la condición (6).

Para determinar estas funciones, debe recurrirse al procedimiento heurístico. Supongamos provisionalmente que ν_t pueda ser representada por una función monótona ν_t^*

aunque no necesariamente uniforme, ya que la diferencial $d\nu^*/dt$ será también una función monótona positiva creciente, estacionaria o decreciente. Por ejemplo, puede asignarse a la variable t con un exponente positivo constante:

$$d\nu^* \div dt = k^* t^{m-1} \tag{9}$$

donde k^* es una constante de ajuste. Considerando m constante, se obtiene por integración:

$$\nu_t^* = (k^* t^m \div m) + c\,(k^* - 1) \tag{10}$$

Como que $\nu_0^* = 0$ entonces $c = 0$. Además, como que $\nu_E^* = 1$ entonces $k^* \div m = E^{-m}$ y por fin:

$$\nu_t^* = t^m E^{-m}, \qquad t \in [0,E] \tag{11}$$

Supongamos, de forma igualmente provisional, que la tasa de subsistencia esté representada por una función monótona v_t^* tal que $v_o^* = 1$ y $v_E^* = 0$. Supongamos asimismo que la función diferencial dv^*/dt sea monótona decreciente, estacionaria o creciente según una potencia n positiva y constante de t. Según las mismas operaciones que en (9), (10) y (11) aplicadas a la variable independiente $E - t$, obtendremos:

$$v_t^* = (E-t)^n E^{-n}, \qquad t \in [0,E] \tag{12}$$

Finalmente, según (6), podemos escribir:

$$\mu_t^* = t(E-t)^n E^{-(m+n)}, \qquad t \in [0,E] \tag{13}$$

5.2.2. Tasa absoluta de agotamiento y funciones V

Se observa que en la ecuación (13) sólo es necesario determinar los parámetros m,n y E. El papel desempeñado por el conjunto de ν_t^* y v_t^* es heurístico, ya que permite engendrar una gama de funciones μ_t de las cuales sólo una cumple con la condición (8). Para demostrarlo, es suficiente observar que μ_t no cambia si se multiplica ν_t^* y se divide v_t^* por una misma función $g_t = \nu_t^*/\nu_t = v_t/v_t^*$ en las ecuaciones (11) y (12).

En realidad, tampoco es necesario conocer g_t puesto que se puede determinar μ_t iterativamente haciendo variar los parámetros m,n y E en la ecuación (13). Pero antes debe determinarse $\mu_t = f(\mu_t^*)$, de forma $\mu_t = k\mu_t^*$, donde k es la constante que permite cumplir la condición $SCF_E = 0$.

$$k = 1 \div \int_0^E \mu_x^*\, dx \tag{14}$$

$$\mu_t = t^m(E-t)^n \div \int_0^E x^m(E-x)^n\, dx \tag{15}$$

La *tasa absoluta de agotamiento* o *función V* puede representarse por una función *cuasi beta* (Johnson & Kotz, 1970). Es interesante notar que una función μ_t hipotética (no ajustada) con $m = n = 2$ sirvió durante años para estimar la subsistencia en la mayoría de inventarios de capital evocados al principio del artículo (Schiff, 1958) y sistematizados más tarde para actualizar las cuentas de capital residencial y no residencial publicadas periódicamente por el *SCB* (Musgrave, 1976).

5.2.3. Algoritmo de determinación de la función V por estrato

Antes de 1961, no se disponía de hitos censales desglosados por año de construcción, ni siquiera al nivel nacional. Por ejemplo, no se sabe cuántas viviendas se construyeron en 1915 y cuántas de ellas subsistían en 1940, 1970, 1980, 1990 o 2000. Por tanto, es difícil modelizar una función V para el subestrato de 1915 y proyectarla luego para saber lo que subsistirá de él en los años 2010, 2020, etc. En cambio, podemos saber, por ejemplo, cuál era en 1940 el volumen del estrato de viviendas formado entre 1900 y 1940 y cuántas de ellas subsistían en 1970, 1980, 1990 o 2000.

Incluso podemos saber cómo se distribuyó la formación de dicho estrato a lo largo de los cuatro decenios que separan 1900 de 1940, aunque al precio de una investigación censal bastante laboriosa (Vergés, 1990). Este tipo de investigación es posible debido a la alta cobertura y calidad de los censos españoles desde 1860, incluso por (grandes) municipios, con el condicionante de corregir las fusiones o divisiones de algunos de ellos, no sólo durante el período de formación sino también más tarde.

Supongamos pues que disponemos de la antedicha información acerca del volumen del estrato en el año de fin de su formación en 1940 y en los censos ulteriores de 1970, 1980, 1990 y 2000. Suponiendo que una misma función V se aplique a todos sus subestratos, debemos encontrar la combinación óptima de sus parámetros m, n y E cuya predicción agregada de 1940 coincida con el dato correspondiente y cuyas ulteriores predicciones también agregadas cumplan la condición de mínimos cuadrados con respecto a los valores observados en 1970, 1980, 1990 y 2000.

Esta optimización se obtiene mediante el algoritmo *Retropackage* (Vergés y Ordaz, 1994), el cual consiste en llenar selectivamente una matriz tridimensional definida por un amplio espectro de valores paramétricos de m, n y E, con los resultados correspondientes en término de mínimos cuadrados. Una vez seleccionada la casilla de la matriz con mejores resultados, se amplia su entorno con un nuevo espectro de mayor fineza y se repite la operación hasta que no se logre ya mejora significativa de los resultados. La combinación de parámetros retenida es la que se utiliza luego para proyectar los valores del estrato.

Obviamente, la disponibilidad de hitos recientes es determinante, por lo que este tipo de análisis recobrará toda su importancia cuando se disponga de los datos del censo de edi-

ficios de 2001. Mientras tanto, pueden utilizarse los resultados al nivel de Comunidades Autónomas (Vergés, 1990).

5.3. Funciones V y *backlogs*

No sería procedente concluir el presente apartado sin evocar el debate acerca de la posible existencia de *backlogs* en el agotamiento del stock de capital (Lioukas, 1982), debate que incide sin duda en el abordaje del consumo de capital residencial. Este debate tuvo lugar entre los neoclásicos, partidarios de considerar la depreciación y el agotamiento de capital como una función de la edad del bien y los «solovianos», partidarios de la idea que los bienes se vuelven obsoletos y tienden a desaparecer cuando otros bienes más productivos salen al mercado (Hall, 1968), o cuando hay *reswitching*, es decir desactivación o reactivación de bienes de producción porque las condiciones financieras incitan a descapitalizar o a recapitalizar para recolocar la inversión (Samuelson, 1966).

Según los partidarios de la función de edad, el consumo de capital sería *endémico* y razonablemente distribuido a lo largo de la duración de vida útil del bien, mientras que para los segundos, partidarios del *sudden exit*, el mismo consumo sería más bien *epidémico*, pudiendo presentar en cualquier momento alteraciones en la tendencia de su propio proceso (*backlogs*).

Al abordar el análisis del parque de viviendas, surgen serios interrogantes acerca de la continuidad en el proceso de la renovación urbana, es decir en el proceso de substitución de estructuras obsoletas por otras nuevas. Está claro que no se acostumbra a derribar a los edificios que están todavía en buen estado y que mantienen sus funciones, pero tampoco faltan edificios pasablemente degradados que continúan funcionando. Al contrario, han podido derribarse edificios y hasta barrios enteros por razones ajenas a su estado, a su función o a su funcionamiento.

En el caso de España, las funciones *V* observadas desde que se tienen datos al respecto, no parecen detectar presencia «estadística» de *backlogs* en el agotamiento del capital residencial, pero tampoco es imposible que el actual proceso de dispersión periurbana (*sprawl*) provoque fenómenos acelerados de abandono y, a la larga, de degradación en barrios centrales de nuestras áreas metropolitanas, lo cual podría alterar las previsiones realizadas desde un punto de vista *endémico*.

Esta reserva permite matizar la conveniencia de utilizar los resultados de las proyecciones de estratos según las funciones V como si de datos se tratara en la proyección general del inventario. Dicha utilización tiene la ventaja de despejar el problema de la *reposición* del parque y de centrarse en el problema del desarrollo de nuevos estratos vía el planeamiento vigente, como también veremos más adelante. La condición pa-

ra poder utilizar las proyecciones de estratos como datos de supervivencia del parque existente es de actualizar el cálculo cada vez que emerge nueva información censal.

6. PLANEAMIENTO DE SUELO

El desarrollo anterior permite avanzar hacia una proyección satisfactoria del parque de viviendas existente. Queda por prever el volumen que alcanzarán en el futuro los estratos formados por la nueva *FBCF*, es decir, por las viviendas que se terminarán en los años venideros, de manera a poder agregarla a dicha proyección.

Tradicionalmente, la previsión del flujo futuro hace hincapié en el análisis ciclotendencial. Sin embargo, los comportamientos actuales en materia de edificación obedecen más al instinto de *depredación territorial* del «sector», ya explícito en los planteamientos de Brady (1973), que a la lógica del *esfuerzo-recuperación* tanto de la oferta como de la demanda, subyacente en las teorías cíclicas de Keynes o de Kuznets, por ejemplo. Del punto de vista urbanístico, la ciudad no se desarrolla ya *partiendo de la producción de base* (destinada a la exportación) sino *avanzando hacia la saturación de los servicios residenciales*, vivienda incluida, por supuesto.

6.1. Suelo urbanizable

Hoy en día es mucho más realista prever la edificación de futuras viviendas partiendo del cómputo de suelo disponible al efecto. Una vez conocido este dato, sólo queda por asignarle un calendario, el cual dependerá, obviamente, de la coyuntura económica y financiera. Sin embargo, la información acerca del planeamiento vigente, tan esencial para la gestión territorial, es todavía escasa y deficiente. En efecto, la mayoría de ayuntamientos conocen generalmente la superficie de suelo calificado, así como su uso y densidad, pero no llevan registro acerca de su progresiva edificación a pesar de ser ellos mismos quienes otorgan licencia de obra.

Ante esta carencia, distintas Comunidades han procedido a la recogida de datos de planeamiento en las Comisiones de Urbanismo. La Generalitat de Catalunya ha desarrollado un método de puesta al día periódica, en el que se aprecia en una fecha de referencia el porcentaje ya edificado en cada sector de planeamiento (Mitjavila,1999). Este porcentaje se aplica luego al número de viviendas previsto en los diferentes sectores por el planeamiento municipal. A partir de dicha fecha, se van deduciendo las unidades de crecimiento que se estiman terminadas según las series municipales de flujos. Por unidades de crecimiento se consideran a las viviendas terminadas en el municipio menos la reposición, la cual absorbe en principio al equivalente de viviendas desaparecidas en los centros urbanos, aunque también este principio puede variar, como veremos más adelante.

Por su lado, el Gobierno de Euskadi ha publicado ya la 2ª edición de su banco de datos territoriales (*UDALPLAN'99*) y lo mismo ha ocurrido en la Generalitat Valenciana (1999) en ambos casos con resultados exhaustivos. Otras Comunidades, como Asturias, tienen previstas publicaciones al respecto.

6.2. Saturación de suelo urbano

Otra fuente de información acerca de suelo disponible es la estimación del suelo urbano no saturado, como pueden ser las parcelas todavía libres en urbanizaciones existentes. Esta estimación suele llevarse a cabo de forma circunstancial, como ha sido el caso para la Región 1 de Catalunya (Área Metropolitana de Barcelona) mediante un modelo estimativo basado en observaciones en ciertos municipios (Carreras, 1999). Nótese que en este modelo se observa la existencia virtual de una asíntota al proceso de saturación de este tipo de suelo, situada alrededor de un 75% de la superficie disponible según el planeamiento urbanístico.

6.3. Previsión de flujos

El ritmo de transformación de suelo urbanizable en suelo urbano y de suelo urbano en suelo edificado, sigue los aleas de la coyuntura económica, administrativa y sobre todo financiera. Por tanto, su previsión debe comprender varias alternativas, incluyendo la de formación de franja de baldío (*fringe*). Para ello, se sugiere utilizar el concepto de número de años necesario para saturar el suelo contenido en el conjunto vigente de sectores de planificación y utilizar dicho número como parámetro variable en las proyecciones.

6.4. Representación gráfica del inventario

El esquema siguiente expresa los resultados tanto del tratamiento de datos como de las proyecciones del parque existente. En ordenadas figura el número de viviendas de cada estrato de lo que podría ser un parque de viviendas de una ciudad de unos 10.000 habitantes. En abcisas figuran los hitos censales desde 1970 hasta 2020. Los estratos de antes de 1971 están acumulados y aparecen íntegramente heredados de períodos anteriores, pero no se pueden desglosar antes de 1980 por carecer aún del ajuste por *hot-deck* anterior a de dicha fecha. Las observaciones por estratos empiezan en 1980 y pronto se dispondrá de la de 2000. Los estratos de 1971-80 y 1981-90 se corresponden con la acumulación de las series decenales del flujo de terminación de viviendas de ambos períodos según los censos de edificios (§4.3).

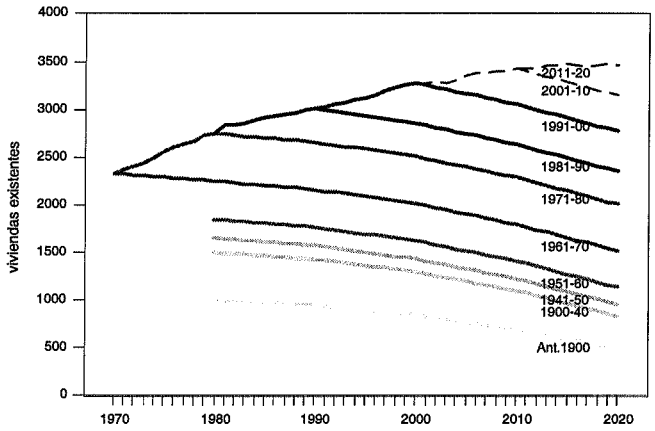


Figura 1. Inventario permanente de vivienda. Estratos según fecha de construcción. Representación 1971-2000. Proyección 2001-2020.

En el mismo esquema aparecen los resultados de la proyección del parque de antes de 1991 por estratos hasta 2020 según §5.2.3, a la cual se ha añadido la acumulación de la serie 1991-2000 de terminación de viviendas según los aparejadores (§4.2), en la hipótesis de que dicho estrato no sufra agotamiento ninguno antes del año horizonte. Por supuesto, la proyección resultante del nivel y distribución por estratos del parque al final del 2000 deberá ser substituida en su momento por la observación y su correspondiente tratamiento según §5.1. Por fin, se expresan también en el esquema los estratos 2001-2010 y 2011-2020 (en líneas de puntos).

El volumen que adquirirán los estratos 2001-2010 y 2011-2020 dependerá de dos factores: la reposición necesaria del parque existente y el plazo dentro del cual se cuenta edificar sobre el suelo actualmente disponible. Si este plazo es muy corto, una década por ejemplo, el estrato 2001-2010 comprenderá no sólo a las viviendas necesarias para compensar el agotamiento y mantener el parque a su nivel actual, sino también a la totalidad de las viviendas edificables sobre nuevo suelo. Obviamente, el estrato puede crecer más allá de lo previsto si se califica nuevo suelo.

Si al contrario el plazo es muy largo, entonces cada uno de los nuevos estratos se dotará con sólo una pequeña fracción de las viviendas edificables sobre el suelo disponible, más la correspondiente cuota de reposición. Las opciones al respecto son innumerables, dado además que el ritmo de reposición puede ser también variable. En efecto, no todo lo que se derriba o abandona por vetustez tiene porque ser reedificado de inmediato. Recíprocamente, lo que se reedifica puede ser más de lo que se repone, si la edificabilidad urbanística reciente lo permite. Nótese asimismo que al disponer del dato del parque total según el padrón de 1996, se puede proceder al ajuste provisional del estrato 1991-2000, por lo menos hasta el 31-XII-95.

7. USOS DEL PARQUE

Aunque el uso de la vivienda interfiera escasamente en la confección del inventario del parque, conviene mencionar las fuentes sobre las distintas clases de vivienda, así como algunos de los problemas que presentan.

7.1. Vivienda principal

El concepto de vivienda principal está vinculado con los conceptos de hogar o de familia desde los censos del siglo *XIX* e incluso antes (Vergés, 1990). Actualmente, la cohabitación de familias en una vivienda es cuantitativamente insignificante. Además, salvo en periferias de alguna capital, existe muy poca vivienda precaria. Por tanto, cuando los censos enuncian el concepto de vivienda principal, se refieren prácticamente a la vivienda ocupada por una familia a título de residencia principal en un edificio, independientemente de la consistencia, calidad o intensidad de ocupación del local según variables socio-demográficas del propio censo o de otras fuentes.

Sin embargo, el censo de vivienda principal no equivale al censo de familias porque, 1°, no incluye a personas o familias que viven en vivienda colectiva estando capacitadas para vivir en vivienda familiar, dado que si éste fuera el caso, estarían censadas como población que vive en familia, 2°, tampoco incluye a personas o familias empadronadas o no, que se ven privadas de vivienda o de alojamiento (*homeless*) o que ocupan ilegalmente viviendas desocupadas (*okupas*) y 3°, no incluye a personas o familias empadronadas que viven en viviendas que por error o por razones inciertas han sido censadas como desocupadas (Vergés, 1998*b*).

Estos problemas deben ser tenidos en cuenta a la hora de efectuar estimaciones de una variable a partir de otra. Ahora bien, en la corrección de diferencias entre viviendas principales y familias en pequeños y medianos municipios del censo de vivienda de 1981, la norma es el acercamiento de la segunda a la primera variable.

Otro fenómeno a tener en cuenta es la ocupación de viviendas como de uso principal por personas empadronadas en otro lugar, por ejemplo, por estudiantes no emancipados de sus respectivas familias pero que residen en una vivienda alquilada o adquirida por sus padres, etc. cerca del centro de estudios. Así, en el censo francés aparece la clase *segunda vivienda principal*, pero todavía no en el censo español.

7.2. Vivienda no principal

7.2.1. Uso secundario y desocupación

El concepto de vivienda secundaria parece bastante claro en su definición, pero presenta ciertas dificultades contables. Las principal fuente de datos al respecto es también el

censo de vivienda, el cual ha venido administrándose en una fecha (el 1° de marzo) con escasa probabilidad de presencia del hogar en su vivienda secundaria. Lo que puede ocurrir entonces es que el agente censal la enumere como desocupada y viceversa.

Dados los numerosos casos de duda, el censo de 1991 ha introducido una nueva clase que recoge a las viviendas de las que *no consta* el uso. Pero esta innovación parece haber complicado algo más las cosas. Veamos por ejemplo, como han sido enumeradas las viviendas de tres conocidas entidades turísticas entre 1981 y 1991:

Tabla 1. Cambios en la clase de viviendas de municipios turísticos entre 1981 y 1991.

entidad	censo	viviendas familiares	principales	secundarias	desocupadas	no consta
LLançà	1981	4.505	889	3.559	59	–
	1991	5.847	1.213	4.116	518	0
Mojácar	1981	1.327	456	490	381	–
	1991	4.864	1.340	93	3.215	0
L'Escala	1981	8.008	1.165	6.520	323	–
	1991	11.106	1.661	78	3	9.364

En el caso de LLançà, no parece existir problema alguno: 20% de viviendas eran principales y 80% secundarias, tanto en 1981 como en 1991, con un número razonable de desocupadas en ambos censos. En Mojácar, el desarrollo turístico tuvo lugar entre ambos censos. Con sólo 27% de viviendas principales en 1991, el 72% restante se clasificó como vivienda desocupada cuando en realidad era secundaria. Lo que no se sabe es cuántas de ellas estaban realmente desocupadas en el momento censal, es decir por vender o por alquilar y cuántas permanecían realmente ocupadas, aunque su ocupante propietario o arrendatario hubiera estado ausente en el momento del censo. En el ejemplo de L'Escala, el problema es el mismo que en Mojácar, salvo que al parecer, los agentes censales recibieron la instrucción de utilizar en este caso la casilla «no consta» de preferencia a la casilla «desocupada».

7.2.2. Conversión de secundarias en principales

Este importante fenómeno es cada vez más frecuente, especialmente en zonas periféricas de áreas metropolitanas y también en zonas turísticas que se van convirtiendo en aglomeraciones residenciales. Al nivel de stocks, el fenómeno se puede medir comparando las clases de vivienda de un censo a otro. Al nivel de flujos, se dispone de la estadística de *variación residencial* elaborada por cada Comunidad a partir del archivo

central del cambio de residencia padronal recogido por el INE. Pero para extraer información explicativa, hay que solicitar estudios a medida con especificación de la edad y sexo de los migrantes a los institutos autonómicos de estadística.

Ahora bien, puede ser suficiente consultar información más inmediatamente accesible como puede ser la del flujo de edificación de viviendas (§4). En efecto, el aumento de población en un municipio con urbanizaciones de segunda residencia puede explicarse por la conversión de estas últimas en primera residencia, cuando dicho aumento no encuentra contrapartida suficiente en la nueva construcción. Estudios de esta índole se realizan frecuentemente por los servicios de urbanismo municipales o autonómicos, aunque por lo general se quedan sin publicar.

7.2.3. Uso y fecha de construcción

Aunque el cruce de uso y de fecha de construcción permita levantar ciertas indeterminaciones al nivel estadístico, como hemos visto en el §5.1.2, no conviene utilizarlo diacrónicamente, es decir de un censo a otro, debido a los cambios de uso acontecidos en cada estrato. Sin embargo, sí puede ser utilizado con las debidas precauciones en el estudio del *filtraje*, ya que la edad de la vivienda suele ser un indicador del estado sino de la calidad del edificio (deterioro, ascensor, etc.) y se convierte en factor explicativo de la rotación por compraventa de las viviendas del parque[1].

8. CONCLUSIÓN

El artículo ha desarrollado una metodología para el tratamiento analítico y proyectivo de los tres componentes del parque de viviendas: 1°, parque por estratos según fechas

[1]El mecanismo *filtering-up* descrito en 1949 por Ratcliff es el siguiente: cuando un hogar de una clase determinada necesita cambiar de vivienda, busca una mejor en función de su mayor nivel de renta, liberando a mismo tiempo la antigua suya que es ocupada y mejorada por otro hogar de la clase inmediatamente inferior pero que también ha visto aumentar su renta. Éste último hogar libera a su vez su antigua vivienda que es ocupada por alguien de la clase siguiente, etc. Se forma así una cadena cuyo eslabón superior es una vivienda nueva de mayor nivel, puesto que por definición la clase más elevada no encuentra en el parque unidad a su medida. Al extremo inferior, una vivienda es abandonada o derribada, puesto que también por definición, el que la ocupaba puede ahora acceder a la de la clase inmediatamente superior en cuanto ésta se libera.

El buen funcionamiento de este mecanismo está supeditado a tres condiciones: 1° que la renta se eleve sin demasiada desigualdad, 2° que el mercado local sea estable, 3° que el parque se renueve al mismo ritmo que la elevación de la renta. Si la renta baja en lugar de subir o si sube para unos pero baja para otros, el proceso puede invertirse, frenando la renovación del parque y empeorando las condiciones de ocupación (*fitering-down*). Si se perturba el mercado con una oferta desleal por parte de un suelo alejado de menor repercusión debido a que éste no ha internalizado todavía sus costes urbanos, la población responde a dicha oferta y el barrio se vacía (*sprawl*). Por fin, si se renueva el barrio más allá de lo que pueden costear sus ocupantes, estos tienen que desplazarse y suelen ser substituidos por una clase socioeconómica más elevada: es el retorno de la «*gentry*» (*gentrification*).

de construcción, por lo menos desde 1980, 2°, proyección de cada estrato hasta el año horizonte y 3°, proyección de futuros estratos a partir del año 2001. Esta metodología es aplicable con un desglose territorial que puede llegar al nivel municipal. Por consiguiente, disponemos de los ingredientes necesarios para confeccionar el inventario físico por unidades de viviendas enunciado en el §3.1. Dicho inventario es de máximo interés para la ordenación y gestión territorial del parque de viviendas y para el justo desarrollo del planeamiento de suelo.

El inventario físico es también un ingrediente necesario (aunque no suficiente) para calcular el capital fijo residencial propiamente dicho enunciado en el §3.3. Ahora bien, para alcanzar tal objetivo es necesario disponer de cantidades y precios no sólo de vivienda nueva sino también de vivienda existente. En efecto, éste es el punto de partida del largo proceso que conduce a la *agregación* evocada en el §2. De hecho, se dispone de bastante buena información acerca de la vivienda nueva, pero no acerca de la vivienda existente puesto que los precios disponibles derivan de tasaciones pasablemente sesgadas al alza y no de la observación del propio mercado. Por consiguiente, las estimaciones de repercusión de suelo sufren del mismo sesgo. En cuanto al número de transacciones, elemento imprescindible para confeccionar el deflactor de precios hedónicos, tampoco está disponible (Vergés, 1998*a*). En resumen, desconocemos casi todo del mercado de compraventa lo cual, dicho sea de paso, no es una situación idónea para intentar regularlo.

Para superar esta dificultad, va a ser necesario disponer de más información para construir series apropiadas que cubran por lo menos la década de los 90. Para ello contamos, entre otros, con el material aportado por las respuestas a la nueva pregunta censal sobre la fecha de entrada en la vivienda ocupada, pregunta que el autor sugirió incluir en el censo de 2001. De la respuesta a ésta y a otras nuevas preguntas (ubicación de la vivienda secundaria, etc.) se derivará información acerca de permanencia, compraventa, secundaridad, etc. toda ella necesaria para construir las variables cuantitativas del *SCF* residencial. En materia de precios, contamos también con la explotación de la *EPF 2001*, explotación que ya se hizo en 1990-91 pero que no se pudo desagregar geográficamente porque la muestra, especialmente pensada para la *cesta del IPC*, no es geográficamente representativa de variables censales como las proporcionará precisamente y por vez primera el censo de 2001.

Con todos estos trabajos a la vista, es de esperar que las Comunidades puedan disponer en el futuro de mejores instrumentos de medida y de previsión con el fin de desempeñar sus funciones de ordenación territorial y de planificación residencial, orientando en definitiva a los agentes hacia un desarrollo territorial que internalice los costes de compensación en aras de un desarrollo más acorde con el contrato social (Stiglitz, 1977).

REFERENCIAS

Brady, E. (1973). «An Econometric Analysis of the U.S. Residential Housing Market», in R.B. Ricks (Ed.). *National Housing Models: Applications of Econometric Techniques to Problems of Housing Research*. Mass.: Lexington, 1-47.

Carreras i Quillis, J.M. (1999). «Model de saturació de sòl urbà. Ámbit Metropolità de Barcelona». *Informe. Direcció General d'Urbanisme*. D.P.T.O.P. Generalitat de Catalunya.

Dixit, A.K. & Stiglitz, J.E. (1977). «Monopolistic competition and optimum product diversity». *e American Economic Review*, 67, 3, 297-308.

Fujita, M., Krugman, P. & Venables, A.J. (1999). *The Spatial Economy. Cities, Regions and International Trade*. Cambridge (Mass.): MIT Press.

Generalitat Valenciana (1999). *El Planeamiento Urbanístico de la Comunidad Valenciana*. Conselleria d'Obres Públiques, Urbanisme i Transports.

Goldsmith, R.W. (1951). «The Perpetual Inventory of National Wealth, in NBER». *Studies in Income and Wealth*, 14, 5-61.

Grant, E.L. & Norton, P.T. (1955). *Depreciation*. N.Y.: Ronald.

Hall, R.E. (1968). «Technical Change and Capital from the Point of View of the Dual». *Review of Economics and Statistics*, 35, 35-46.

Hicks, J.R. (1946). *Value and Capital*. Oxford: Clarendon Press.

— (1973). *Capital and Time*. Oxford: Oxford University Press.

Hotelling, H. (1925). «A General Mathematical Theory of Depreciation». *American Statistical Association Journal*, 20, 340-353.

INE (1983). *Censo de Edificios de 1980. IV. Resultados por provincias*. Madrid: INE Artes Gráficas.

— (1993). *Censo de Edificios de 1990. IV. Resultados por provincias*. Madrid: INE.

Johnson, N.L. & Kotz, S. (1970). *Continuous Univariate Distributions, II*, Bos.: Houghton Mifflin, 37-56.

Kauder, E. (1965). *A History of Marginal Utility Theory*. Princeton: Princeton University Press.

Kendrick, J.W. (1976). *The Formation and Stocks of Total Capital*. N.B.E.R. General Series, 100. N.Y.: Columbia University Press.

Lioukas, S.K. (1982). «The Cyclical Behavior of Capital Retirment: Some New Evidence». *Applied Economics*, 14, 73-79.

Lujanen, M. (1985). *Forecasting and Programming of Housing*. N.Y.: United Nations, ECE/HBP/51, 10-12.

Mitjavila i Garcia, M. (1999). «Planejament urbanístic i usos del sòl. Província de Barcelona». *Direcció General d'Urbanisme*. D.P.T.O.P. Generalitat de Catalunya.

Musgrave, J.C. (1976). «Fixed Nonresidential Business and Residential Capital in the U.S., 1926-1975». *Survey of Current Business*, 56, april, 46-52.

Naredo, J.M. (2000). *Composición y valor del patrimonio inmobiliario en España. 1990-1997*. D.G.P.E.P. Madrid: Ministerio de Fomento.

Ratcliff, R.U. (1949). *Urban Land Economics*, McGraw-Hill.

Samuelson, P.A. (1966). «A Summing Up». *The Quaterly Journal of Economics*, 80, 568-583.

Sánchez de Rivera, R. (1990 y s.) *Obras en Edificación*. D.G.P.E.P. Madrid: Ministerio de Fomento. Anual.

Schiff, E. (1958). «Gross Stock Estimated from Past Installations». *The Review of Economics and Statistics*, 40, 174-177.

Stiglitz, J.E. (1977). «The Theory of Local Public Goods», in M.S. Feldstein & R.P. Imnan (Eds.). *The Economics of Public Services*. London: MacMillan.

Survey of Current Business, 52, (mai 1972):
Jorgenson, D.W. & Griliches, Z. (1969). «The Explanation of Productivity Change». *The Review of Economic Studies*, 34, 249-283.
Denison, E.F. (1969). «Some Major Issues in Productivity Analysis: An Examination Estimates by Jorgenson and Griliches». *SCB*, 49, mai, II, 1-27.
D.W.J. & Z.G. (1972). «Issues of Growth Accounting: A Reply to Edward F. Denison». *SCB*, 52, II, 65-94.
E.F.D. (1972). «Final Comment». *SCB*, 52, II, 95-110.
D.W.J. & Z.G. (1972). «Final Reply». *SCB*, 52, II, 111.

Toro Valverde, G. (1990 y s.) *Edificación y Vivienda.* D.G.P.E.P. Madrid: Ministerio de Fomento. Anual.

UDALPLAN'99. *Banco de Datos Territoriales. Suelo Residencial y de Actividades Económicas de la C.A.P.V. (2000)*. Gobierno Vasco. Departamento de Ordenación del Territorio, Vivienda y Medio Ambiente.

Universidad Comercial de Deusto (1968). *La Riqueza Nacional de España*. Bilbao.

Usher, D. (1980). «The Measurement of Capital, Studies in Income and Wealth», *NBER*, 45. Chicago: The University of Chicago Press.

Vergés Escuín, R. (2000). «Vivienda: el conocimiento de la demanda o la espera de Godot». *Análisis Local*, 29, 5-12.

— (1998*a*). «El precio de la vivienda urbana», in R. Vergés (Ed.). *El precio de la vivienda y la formación del hogar*. Col·lecció Urbanitats 6. Centre de Cultura Contemporània de Barcelona, 117-144.

— (1998*b*). «L'enquesta sobre l'habitatge desocupat en els districtes de Barcelona». *Informe. Gabinet d'estudis urbanistics*. Ajuntament de Barcelona, 29 pp.

— (1992 y s.) «Informes Trimestrales de Coyuntura». *Consejo Superior de Colegios de Arquitectos de España*. Informes publicados a partir de 1995 en Directivos Construcción, mensual.

— (1990). «Inventario permanente de vivienda y modelo de previsión de demanda. Vol IV». *Informe. Banco Hipotecario de España*. Dirección General para Vivienda y Arquitectura (MOPU), 400 pp.

— (1989). «Le capital-logements en France». *Rapport. I. Concepts et mesure*. Direction de la Construction. M.E.L. Gouvernement Français, 153 pp.

Vergés Escuín, R. & Ordaz Sanz, J.A. (1994). «Retropackage. Algoritmo GLS para funciones de agotamiento y subsistencia de stocks». *Estudios de economía aplicada. II. VIII Reunión anual de ASEPELT-España*. Universitat de les Illes Balears. Palma, 2-3 junio, 71-78.

Winfrey, R. (1935). «Statistical Analysis of Industrial Property Retirement». *Bull.*, 125, Iowa Engineering Experimental Station.

ANEXO. ALGORITMO DE CUADROS RECTANGULARES ACR

Dada una matriz de referencia $A_{m\cdot n}$, determinar una matriz virtual $B_{m\cdot n}$ cuyas diferencias $b_{ij} - a_{ij}$ sean mínimas.

1. Matriz de referencia

$$A = (a_{aij}), \qquad \begin{matrix} i = 1,\ldots,m \\ j = 1,\ldots,n \end{matrix}$$

donde:

$$a_{i+} = \sum_{j=1}^{n} a_{ij}$$

$$a_{+j} = \sum_{i=1}^{m} a_{ij}$$

$$A = \sum_{i=1}^{m} a_{i+} = \sum_{j=1}^{n} a_{+j} = \sum_{i=1}^{m}\sum_{i=1}^{m} a_{ij}$$

2. Matriz virtual

$$B = (b_{ij}), \qquad \begin{matrix} i = 1,\ldots,m \\ j = 1,\ldots,n \end{matrix}$$

donde:

$$b_{i+} = \sum_{j=1}^{n} b_{ij}$$

$$b_{+j} = \sum_{i=1}^{m} b_{ij}$$

$$B = \sum_{i=1}^{m} b_{i+} = \sum_{j=1}^{n} b_{+j} = \sum_{i=1}^{m}\sum_{i=1}^{m} b_{ij}$$

y en $t = 0$ (momento inicial) conocemos:

b_{i+}

b_{+j}

$$B = \sum_{i=1}^{m} b_{i+} = \sum_{j=1}^{n} b_{+j}$$

Determinar entonces b_{ij}

3. Algoritmo

La determinación de la matriz virtual se consigue mediante un algoritmo iterativo de cuadros rectangulares, comportando tantos pasos $t = 1, \ldots, h$ como sea necesario hasta que b_{ij} en h no sea significativamente distinto de b_{ij} en $h-1$. Se demuestra que al término de este procedimiento, se consigue la condición de mínimos cuadrados de la diferencia entre A y B.

$$\begin{aligned}
t &= 0 \qquad b_{ij}^{0} = a_{ij} \\
t &= 1' \qquad b_{ij}^{1'} = b_{ij}^{0} \left(b_{i+} \div b_{i+}^{0}\right) \left(b_{+j} \div b_{+j}^{0}\right) \\
t &= 1 \qquad b_{ij}^{1} = b_{ij}^{1'} \left(B \div B^{1'}\right) \\
t &= 2' \qquad b_{ij}^{2'} = b_{ij}^{1} \left(b_{i+} \div b_{i+}^{1}\right) \left(b_{+j} \div b_{+j}^{1}\right) \\
t &= 2 \qquad b_{ij}^{2} = b_{ij}^{2'} \left(B \div B^{2'}\right) \\
&\vdots \qquad\qquad\qquad \vdots \\
t &= h' \qquad b_{ij}^{h'} = b_{ij}^{h-1} \left(b_{i+} \div b_{i+}^{h-1}\right) \left(b_{+j} \div b_{+j}^{h-1}\right) \\
t &= h \qquad b_{ij}^{h} = b_{ij}^{h'} \left(B \div B^{h'}\right)
\end{aligned}$$

Se considera que $t = h$ cuando la diferencia $B_t - B_{t'}$ es suficientemente próxima de cero, lo cual indica que b_{ij} en h no es significativamente distinto de b_{ij} en $h-1$. En el límite, se cumple la condición de mínimos cuadrados entre a_{ij} y b_{ij}.

ENGLISH SUMMARY

LOCAL HOUSING DATA PROCESSING FOR THE PERPETUAL INVENTORY OF RESIDENTIAL CAPITAL

R. VERGÉS ESCUÍN
University of Montreal*
Red Vergés, S.L.**

Once the philosophy of the Fixed Capital Accounts is established, the article examines the methodology necessary to design the Local Perpetual Inventory of Housing. With respect to flows, the different data sources on building which exist in Spain have been classified and compared, and the data processing for each of these is described. In terms of stocks, a fitting procedure is proposed and a local disaggregation process of housing stock by construction date. The methodology to forecast the length of life of the housing cohorts by depletion functions is also developed. Likewise, building land variables according to the zoning as emerging new stratus in inventory forecasts is introduced. Lastly, some problems regarding the use of housing classification data are analyzed. The article concludes with a vision towards an appraisal of local fixed housing capital using the emerging new information on housing markets.

Keywords: Capital, depreciation, perpetual inventary, planning, land, housing

AMS Classification (MSC 2000): 62P20

* Professor of Building Economics. Faculty of Planning.
Coordinator of Statistics. Spanish Council of Arquitect's Colleges.
** Model RED 3. Mail adress: Beatriz de Suabia, 152 B, 5° i, 41005, Sevilla.
E-mail: redverges@arquired.es

- Received February 2001.
- Accepted May 2001.

1. INTRODUCTION

The statistical measures of residential wealth delight in great tradition. Measuring methods and forecast of fixed housing capital are examined in this article by studying the available information for quantitative analysis with an emphasis on local and time series.

The study of fixed capital (*FC*) is composed of two phases: the issue of flows and their accumulation on stocks. Both join in a calculation using an inventory technique in which the unknown is usually the capital consumption:

$$FCC(t_1 - t_0) = FCS(t_0) - [FCS(t_1) - GFCF(t_1 - t_0)]$$

where *FCC* is the *FC* consumption, *FCS* is the *FC* stock and *GFCF* is the gross *FC* formation, all of which have been measured at the instances and periods indicated.

This equation can also be applied to physical inventory when the goods are homogeneous as in the case of family housing. The stock of capital is obtained later as a product of the components of the physical inventory for its respective prices. The article studies the usage of information related to flows and stocks of physical inventory, in this case the housing flows and stocks.

2. DATA ON FLOWS

In Spain, the main source of information on housing development flows since 1960 have been the projects which have been registered by the professional associations of archictects. The Ministery of Development (Fomento) also issues the series gathered by the architect's technicians, in particular on the completed projects. Building permits are also provided by the same Department. This diversity of time series constitutes a problem of chronological interpretation that could be solved by means of a lag functions.

3. HOUSING STOCK

Spain, like the majority of countries with a census tradition, disposes of statistics on the number of houses and their variables, including construction dates. This variant is being issued every ten years since 1950 and allows one to know the evolution of the housing cohorts according to construction date. This evolution can be analyzed by means of beta functions whose parametres can be adjusted to predict the length of life of the houses. This way, there is an instrument available to measure and predict the life span of the housing stock and the number of houses needed to replace them.

To determine the parametres of the functions, a computation with RETROPACK method is used by which iteration by steps, comes to determine a least squares fitting with respect to census observations. This article also examines suitability for the housing stocks, in particular the differences between second and unoccupied homes as well as the transformation from second to main residences.

4. PROJECTING THE LAND

The reservation of land for development in each geographic unit is the main decisive for the construction of new homes. The calculation is very recent and some Spain regions have come up with updated information. This allows one to forecast the development of new stratus depending on the hypothetical variables of construction cycles and saturation limits.

5. CONCLUSION

The article concludes with an overall view on the development perspectives of residential capital due to better knowledge of housing market which will be provided by the census of the year 2001.

Biometria

QÜESTIIÓ, vol. 25, 2, p. 337-363, 2001

ANÁLISIS DE DURACIÓN MEDIANTE UN MODELO LINEAL GENERALIZADO SEMIPARAMÉTRICO

JESUS ORBE*

Aitkin y Clayton (1980) proponen el análisis de modelos de duración mediante modelos lineales generalizados. En este trabajo extendemos esta metodología permitiendo que el efecto de alguna de las variables explicativas pueda no ser especificado. Así, el modelo propuesto es un modelo lineal generalizado semiparamétrico, con una componente paramétrica donde se especifica la forma funcional concreta del efecto de las variables explicativas sobre la duración, y una componente no paramétrica donde recogemos el efecto de una variable explicativa sin asumir forma funcional alguna. Desarrollaremos el proceso de estimación así como un procedimiento bootstrap para realizar inferencia. Como aplicación, analizaremos con la metodología propuesta el tiempo de supervivencia para una muestra de pacientes diagnosticados de SIDA.

Lifetime data analysis using a semiparametric generalized linear model

Palabras clave: Modelos de duración, censura, bootstrap, modelos lineales generalizados, estimación semiparamétrica

Clasificación AMS (MSC 2000): 62J12, 62N05, 62G09

* Departamento de Econometría y Estadística (E.A. III). Universidad del País Vasco/Euskal Herriko Unibertsitatea. Avenida Lehendakari Aguirre, 83. 48015 Bilbao. E-mail: jo@alcib.bs.ehu.es.

- Recibido en junio de 2000.
- Aceptado en febrero de 2001.

1. INTRODUCCIÓN

Proponemos una extensión al trabajo presentado por Aitkin y Clayton (1980), quienes presentan la posibilidad de estimar una serie de modelos de duración paramétricos utilizando un modelo lineal generalizado para la variable indicador de censura, cuya función de verosimilitud es proporcional a la correspondiente del modelo de duración original. Tomando como base esta idea, extendemos la metodología de estos autores a un conjunto de situaciones más general. Esta extensión permite modelizar aquellas situaciones en las que la forma funcional del efecto de alguna de las variables explicativas sobre la variable de interés es desconocida o simplemente la especificación de una determinada forma funcional nos parece restrictiva. Esta flexibilización se puede realizar de un modo bastante natural basándonos en un modelo lineal generalizado. Así, vamos a extender el trabajo de Aitkin y Clayton a un contexto semiparamétrico. Como ilustración aplicamos la metodología para analizar el efecto de ciertas variables sobre el tiempo de supervivencia, desde el momento del diagnóstico, en una muestra de enfermos diagnosticados de SIDA.

2. CONEXIÓN ENTRE MODELOS DE DURACIÓN Y MODELOS LINEALES GENERALIZADOS

Sea $T_1,\ldots,T_n$ una muestra aleatoria simple para la variable duración, la cual no es observada en su totalidad debido a la existencia de censura[1] y en su lugar observamos

$$Y_i = \min(T_i, C_i), \qquad \delta_i = \begin{cases} 1; & \text{si } T_i \leq C_i \\ 0; & \text{si } T_i > C_i \end{cases},$$

donde $C_1,\ldots,C_n$ son los valores que toma la variable censura C, la cual suponemos independiente de la variable duración T. Además, δ_i es la variable indicador de censura, tomando valor 0 si la observación correspondiente está censurada o valor 1 si no lo está[2].

Supongamos que esa variable duración puede ser explicada con un modelo que pertenece a la clase de modelos de duración con función de riesgo proporcional propuesto por Cox (1972), en el cual se especifica la siguiente modelización para la función de riesgo:

$$\lambda(t;x) = \lambda_0(t)\exp(x^T\beta) = \lambda_0(t)\exp\eta,$$

[1] Situación habitual cuando se analiza esta clase de datos. Si consideramos el caso más habitual, censura por la derecha, tenemos una observación censurada cuando su tiempo de fallo no ha sido observado al finalizar el estudio (para un detallado análisis de los distintos tipos de censura, vease, por ejemplo, Lawless, 1982).

[2] Con esta especificación estamos suponiendo el caso de censura aleatoria, el habitualmente utilizado. Además, añadiremos que este tipo de censura engloba a otros tipos de censura más restrictivos.

donde $\eta = x^T\beta$ es el predictor lineal y $\lambda_0(t)$ la función de riesgo básica. La estimación de los parámetros del modelo puede realizarse maximizando la función de verosimilitud

$$L = \prod_{i=1}^{n} f(y_i, x_i)^{\delta_i} S(y_i, x_i)^{1-\delta_i}. \tag{1}$$

Utilizando las relaciones existentes entre las funciones de supervivencia, riesgo, riesgo acumulado y de densidad, obtenemos las siguientes expresiones:

$$S(t,x) = \exp(-\Lambda_0(t)e^{\eta})$$

$$f(t,x) = \lambda_0(t)\exp(\eta - \Lambda_0(t)e^{\eta}),$$

donde $\Lambda_0(t) = \int_0^t \lambda_0(t)dt$ es la función de riesgo acumulado básica.

Sustituyendo las expresiones anteriores en (1), tomando logaritmos y reordenando términos obtenemos,

$$\ln L = \sum_{i=1}^{n} \delta_i [\ln \Lambda_0(y_i) + \eta_i] - \Lambda_0(y_i)e^{\eta_i} + \delta_i \ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right). \tag{2}$$

Tomando $\mu_i = \Lambda_0(y_i)e^{\eta_i}$ tenemos que,

$$\ln L = \underbrace{\sum_{i=1}^{n} (\delta_i \ln \mu_i - \mu_i)}_{(a)} + \underbrace{\sum_{i=1}^{n} \delta_i \ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right)}_{(b)}. \tag{3}$$

Se puede verificar que el sumando (a) de la expresión anterior es proporcional al logaritmo de la función de verosimilitud correspondiente a una muestra de n variables aleatorias independientes δ_i con distribución de Poisson[3] de media μ_i. Por otra parte, el término (b) no depende de los parámetros β, sólo depende de la función de riesgo básica, la cual puede depender de parámetros de la distribución.

De esta forma, siguiendo el trabajo de Aitkin y Clayton (1980), y dada la función de riesgo acumulado básica $\Lambda_0(t)$, podemos estimar los coeficientes β del modelo tratando a la variable indicadora de censura δ_i como una variable aleatoria con distribución de Poisson de media $\mu_i = \Lambda_0(y_i)e^{\eta_i}$. Es decir, podemos construir un modelo log-lineal de Poisson «auxiliar» al modelo de duración, tal que

$$\ln(\mu_i) = \ln \Lambda_0(y_i) + x_i^T\beta,$$

con una función de verosimilitud proporcional al término (a) de la expresión (3).

[3] $\sum_{i=1}^{n}(\delta_i \ln \mu_i - \mu_i) - \sum_{i=1}^{n} \ln \delta_i!$

El modelo log-lineal de Poisson, es un caso particular de los Modelos lineales generalizados (MLG). Estos modelos son introducidos por primera vez en el trabajo de Nelder y Wedderburn (1972) y pueden considerarse como una generalización del modelo lineal clásico. En este caso concreto tenemos una componente aleatoria con una distribución de Poisson, y una función de enlace logarítmica que nos relaciona a la media con el predictor lineal. Además en este modelo tenemos un término adicional, $\ln \Lambda_0(y_i)$, denominado «offset». Como MLG puede ser maximizado utilizando el procedimiento de estimación habitual en éstos, maximizando la función de verosimilitud o, equivalentemente, aplicando mínimos cuadrados ponderados iterativos (IRLS) (McCullagh y Nelder 1983).

En Aitkin y Clayton (1980) se describe en detalle el proceso de estimación mediante MLG para los modelos de duración con distribución exponencial, Weibull y valor extremo. Por otra parte, Whitehead (1980) propone un procedimiento de estimación análogo a los anteriores pero para aquellos casos en que desconocemos la forma funcional de la función de riesgo básica; es decir, propone la estimación mediante MLG para modelos de función de riesgo proporcional de Cox (1972).

El resto del trabajo se organiza de la siguiente forma. En la Sección 3 mostramos una aplicación utilizando la metodología tradicional en análisis de duración. Posteriormente, y motivándolo en la aplicación anterior, en la Sección 4, presentamos una extensión al trabajo de Aitkin y Clayton (1980) desarrollando el proceso de estimación de este nuevo modelo. En la Sección 5 proponemos un nuevo procedimiento bootstrap para realizar inferencia en el modelo propuesto. Finalmente, la Sección 6 presenta los resultados y conclusiones más importantes.

3. APLICACIÓN

3.1. Datos

Para ilustrar la metodología descrita hemos aplicado ésta para analizar el tiempo de supervivencia desde el momento del diagnóstico, en una muestra compuesta por 461 enfermos diagnosticados de SIDA desde 1984 hasta el comienzo de 1991, residentes en las comunidades autónomas del País Vasco y Navarra. Utilizando la fecha de diagnóstico y la fecha de fallecimiento, o en el caso de las observaciones censuradas, la fecha final del seguimiento (diciembre de 1992), obtenemos la variable de interés, la duración o tiempo de supervivencia desde el momento del diagnóstico medida en número de trimestres. A diferencia de la mayoría de los trabajos realizados en este área, los cuales se han interesado en estudiar la duración del periodo de incubación, nosotros nos hemos centrado en el estudio de la duración de la última etapa. En el desarrollo del virus VIH tenemos tres etapas. La primera de ellas, la conocida como fase «pre-anticuerpos», es la

más corta con una duración de varios meses (aproximadamente el 50% de los enfermos genera anticuerpos antes de los dos meses después de la infección). Esta etapa va desde el momento en que se produce la infección hasta el desarrollo de los anticuerpos o punto de seroconversión, y es el periodo de tiempo donde al enfermo se clasifica como seronegativo. La segunda etapa, etapa de incubación, es la más larga de las tres (aproximadamente la mitad de los infectados desarrollaban la enfermedad antes de los 10 años). Este periodo parte desde el momento de la seroconversión hasta el diagnóstico de SIDA. Durante esta etapa el individuo es clasificado como seropositivo. Y por último, la tercera etapa, que recoge el tiempo de supervivencia desde el diagnóstico del SIDA. El comienzo de esta etapa tiene lugar en el momento en que el individuo desarrolla alguna enfermedad clasificada dentro de las enfermedades relacionadas con el SIDA.

Para ayudar a describir esta variable disponemos de una serie de variables que nos recogen ciertas características de los enfermos. Así, la variable **Edad** recoge la edad del enfermo en el momento del diagnóstico. **Sexo** es una variable ficticia, que toma valor 1 si el enfermo es varón y valor 0 si es mujer. Tenemos información sobre la enfermedad con la cual se le diagnostica el SIDA. Así, la variable **Enfer1** toma valor 1 si la enfermedad de diagnóstico es una infección oportunista, **Enfer2**, si es un linfoma o un sarcoma de Kaposi y **Enfer3** si es debido a una encefalopatía VIH o al síndrome de «agotamiento» VIH. Además, tenemos información sobre la vía de transmisión de la enfermedad: la variable indicador **Sexual** toma valor 1 si la vía de transmisión es sexual, **Drogas** toma valor 1 si la infección se produce por consumo de drogas, **Sanguínea** toma valor 1 cuando el enfermo es infectado por transmisión sanguínea, **Madre-hijo** toma valor 1 si la transmisión se produce de la madre al hijo, y **Otras** cuando se desconoce la vía de transmisión. Por último, la variable **Periodo** es una variable indicador que toma valor 1 cuando la fecha del diagnóstico es posterior a 1987. El motivo de introducir esta variable ficticia es estudiar el posible efecto de la introducción, a mediados de 1987, del fármaco Zidovudine (también conocido como AZT) sobre la supervivencia del enfermo.

3.2. Análisis de duración tradicional

A continuación, analizamos el efecto que tiene cada una de las variables descritas en la sección anterior sobre el tiempo de supervivencia desde el diagnóstico de la enfermedad. Para ello, comenzamos suponiendo que la variable duración sigue una distribución de Weibull, una de las distribuciones más importantes y más utilizadas en la práctica. La distribución de Weibull es lo suficientemente flexible como para englobar distintos tipos de funciones de riesgo (crecientes, decrecientes o constantes en función del valor que tome el parámetro de forma p). Por tanto, comenzamos ajustando un modelo de regresión Weibull.

Tabla 1. Estimación del modelo de regresión Weibull

Variable	Coeficiente	desv. típica	T-ratio	P-valor
Constante	1.6864	0.4265	3.954	0.00007
Sexo	0.0585	0.1285	0.455	0.64913
Periodo	0.2024	0.1029	1.967	0.04915
Enfer1	0.1881	0.2455	0.766	0.44364
Enfer2	−0.0247	0.3057	−0.081	0.93558
Sexual	−0.1829	0.2372	−0.771	0.44070
Drogas	−0.0392	0.2031	−0.193	0.84711
Sanguínea	−0.0114	0.2735	−0.042	0.96671
Madre−hijo	0.4005	0.4429	0.904	0.36593
Edad	−0.0172	0.0065	−2.621	0.00876
σ	0.9899	0.0366	26.99	0.00000

El efecto de las variables X sobre la función de supervivencia y riesgo puede incluirse a través del parámetro de escala λ. Para ello, utilizamos la forma funcional habitual

$$\lambda = e^{-x^T\beta}, \tag{4}$$

donde x^T es el vector (1×10) de valores que toman los regresores, incluyendo la constante, para cada individuo, y β el vector (10×1) de coeficientes asociados a cada regresor. Por tanto, la función de supervivencia quedará especificada como,

$$S(t,x) = \exp[-(e^{-x^T\beta}t)^p], \qquad p > 0, \quad t > 0,$$

y la función de riesgo como

$$\lambda(t,x) = e^{-x^T\beta}p(e^{-x^T\beta}t)^{p-1}, \qquad p > 0, \quad t > 0.$$

Para la especificación (4), este modelo puede reescribirse en términos log-lineales; es decir,

$$\ln(T) = X\beta + \sigma\varepsilon, \qquad \text{donde} \qquad \sigma = p^{-1},$$

y donde ε tiene una distribución valor extremo estándar. La estimación del modelo se realiza maximizando la función de verosimilitud (1). Los resultados de la estimación se muestran en la Tabla 1.

Analizando los resultados de la Tabla 1 podemos apreciar una estimación del parámetro σ igual a 0.9899, prácticamente 1, y además, significativo. Esto nos está indicando que la distribución de la duración puede ser una exponencial. Para contrastar esta hipótesis

podemos construir el estadístico correspondiente al contraste de la razón de verosimilitudes. Es decir; realizamos el siguiente contraste dentro de la clase de modelos de regresión Weibull: $H_o: \sigma = 1$ (distribución exponencial) frente a $H_a: \sigma \neq 1$ (distribución no exponencial).

Ajustamos los modelos bajo la hipótesis nula y bajo la hipótesis alternativa y calculamos el máximo del logaritmo de la función de verosimilitud para cada caso. Obtenemos unos valores de -713.29, en el modelo exponencial, y -713.25 en el modelo no exponencial.

Si construimos el estadístico tenemos que,

$$\Lambda = 2\{\ln[L(\hat{\beta},\hat{\sigma})] - \ln[L(\tilde{\beta},\sigma = 1)]\} \xrightarrow{d} \chi^2_1, \tag{5}$$

donde $(\tilde{\beta},\sigma = 1)$ son las estimaciones del modelo restringido, en nuestro caso el modelo exponencial, y $(\hat{\beta},\hat{\sigma})$ son las del modelo general, es decir, del modelo Weibull. Por tanto no encontramos evidencia estadística contraria a la especificación de un modelo de regresión exponencial.

Como consecuencia del contraste, parece razonable ajustar un modelo de distribución exponencial a nuestros datos. Estimamos de nuevo por máxima verosimilitud y obtenemos los resultados recogidos en la Tabla 2.

Si comparamos los resultados de las Tablas 1 y 2, vemos que apenas varían, lo que refuerza la idea de la distribución exponencial.

Podríamos pensar en llevar a cabo un contraste de significación conjunto de todas las variables del modelo, excepto la constante, para estudiar la contribución conjunta de todas las variables sobre el ajuste del modelo. En el caso de que rechazáramos la no significatividad conjunta del modelo, ésta no sería una condición suficiente para considerar al modelo especificado como válido, habríamos de contrastarla con algún contraste de diagnóstico basado en los residuos, lo que realizaremos posteriormente.

Pasamos a realizar el contraste utilizando el estadístico formado por la razón de verosimilitudes. Ajustamos el modelo restrictivo en el que sólo tenemos como variable regresora la constante y obtenemos un valor máximo del logaritmo de la función de verosimilitud de -724.87. Para el modelo menos restrictivo, donde incluimos todas las variables regresoras obtenemos un valor de -713.29. Si calculamos el valor del estadístico para este contraste, que en este caso se distribuye como una χ^2 con 9 grados de libertad, obtenemos un valor de 23.16, superior incluso al cuantil que deja una probabilidad del 1% a su derecha (21.7). Por tanto, podemos concluir que, aún con un nivel de significación del 1%, rechazamos que las variables regresoras en conjunto no contribuyen a la explicación del modelo.

Tabla 2. Estimación del modelo de regresión exponencial

Variable	Coeficiente	desv. típica	T-ratio	P-valor
Constante	1.6844	0.4306	3.912	0.00009
Sexo	0.0577	0.1298	0.445	0.65700
Periodo	0.2049	0.1035	1.980	0.04770
Enfer1	0.1873	0.2479	0.755	0.45000
Enfer2	−0.0257	0.3087	−0.083	0.93400
Sexual	−0.1835	0.2396	−0.766	0.44400
Drogas	−0.0397	0.2052	−0.193	0.84700
Sanguínea	−0.0109	0.2763	−0.040	0.96800
Madre−hijo	0.3982	0.4472	0.890	0.37300
Edad	−0.0172	0.0066	−2.607	0.00913
σ	1	-	-	-

En cuanto al efecto de cada variable, tenemos que el tiempo de supervivencia del individuo se verá afectado por el periodo en el que se le diagnosticó el SIDA, si el diagnóstico del individuo es posterior a 1987, influirá positivamente en su duración. Por tanto, parece que el uso del fármaco zidovudine, más conocido como AZT, alarga el tiempo de supervivencia y reduce el riesgo, obteniendo unos tiempos de supervivencia, a partir del diagnóstico, superiores.

En cuanto a la variable edad, también parece ser relevante para explicar el tiempo de supervivencia del individuo. A mayor edad, menor será el tiempo de supervivencia y, por tanto, mayor el riesgo.

Una vez tenido en cuenta el efecto de estas variables sobre el tiempo de supervivencia, parece que variables como sexo, el tipo de enfermedad con la que se le diagnostica el SIDA o la categoría de transmisión a la que pertenece no influyen significativamente sobre el tiempo que sobrevive el enfermo.

Además, para los modelos de regresión exponencial y para la especificación $\lambda(x,\beta) = \exp(x^T\beta)$, es posible interpretar los coeficientes en términos de la duración media. La media de una variable aleatoria con distribución exponencial y función de densidad $f(t) = \lambda e^{-\lambda t}$, es $1/\lambda$. En nuestro modelo habíamos especificado $\lambda(x,\beta) = e^{-x^T\beta}$, entonces $1/\lambda = e^{x^T\beta}$. Si tomamos logaritmos tenemos que ln(duración media)$= x^T\beta$, de donde

$$\frac{\partial \ln(\text{duración media})}{\partial x} = \beta.$$

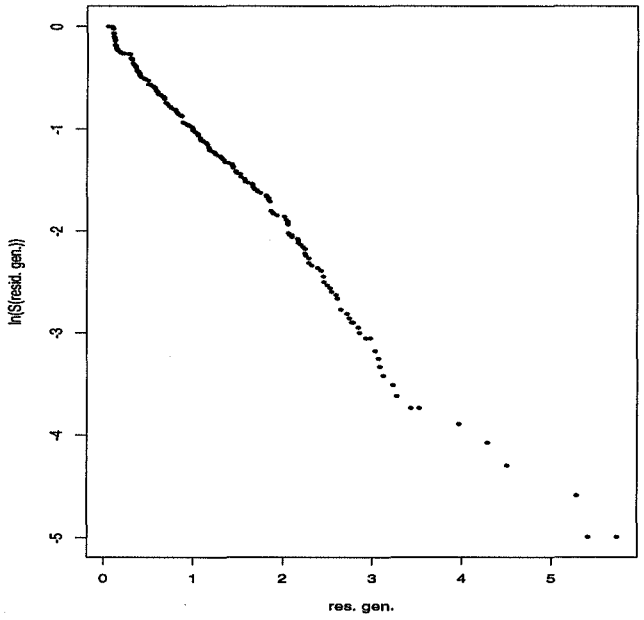


Figura 1. Contraste de diagnóstico

Por tanto, β recogerá la variación porcentual de la duración media ante variaciones de la variable regresora x. En el caso de nuestro estudio, esta interpretación sólo tiene sentido para la variable edad (el resto son variables ficticias). Como $\beta_{10} = -0.0172$, este valor nos indica que un aumento de un año en la edad del individuo en el momento de diagnóstico provocará un descenso del 1.7% en el tiempo de supervivencia medio.

Para que los contrastes y conclusiones sobre las estimaciones tengan validez, tenemos que asegurarnos de que el modelo de regresión exponencial se ajusta a nuestros datos realizando un contraste de diagnóstico. Un contraste de diagnóstico sencillo y frecuentemente utilizado es el basado en los residuos generalizados $(\hat{\Lambda}(y_i, x_i))$ o las estimaciones de la función de riesgo integrado[4].

Comenzaremos con un contraste gráfico, donde representamos el logaritmo de la función de supervivencia estimada para los residuos generalizados, mediante el método de

[4]Los residuos generalizados $\hat{\Lambda}(y_i, x_i)$ tienen una distribución exponencial estandar bajo la hipótesis nula de especificación correcta del modelo.

Kaplan-Meier, frente a los residuos generalizados. Bajo la hipótesis nula de especificación correcta, debemos obtener una línea recta que pase por el origen, y de pendiente menos uno. De la Figura 1 podemos suponer que la especificación es aproximadamente la correcta.

Para corroborar el contraste gráfico, y basandonos en la misma idea, realizamos un contraste más formal utilizando el estadístico propuesto por Kiefer (1988)

$$\frac{[\sum_{i=1}^{n} \hat{\Lambda}(y_i, x_i)^2] - 2n}{\sqrt{20n}},$$

con distribución asintótica $N(0,1)$, bajo la hipótesis de correcta especificación del modelo. Antes de computar el estadístico tenemos que ajustar las observaciones censuradas, sumándoles el valor medio[5], en este caso un 1. El valor del estadístico es de 0.13. Por tanto, no encontramos evidencia contraria a la especificación de un modelo de regresión exponencial.

Para finalizar esta sección señalar que, como es lógico, se obtienen las mismas estimaciones utilizando un MLG auxiliar siguiendo la propuesta de Aitkin y Clayton (1980).

4. ANÁLISIS DE DURACIÓN MEDIANTE UN MLG SEMIPARAMÉTRICO

Si analizamos el modelo ajustado en la sección previa nos encontramos un modelo donde el efecto de las variables explicativas es introducido de una forma paramétrica. Es decir, estamos imponiendo una determinada relación (lineal) entre éstas y el logaritmo de la función de riesgo, en el caso de expresar el modelo como caso particular de los modelos de riesgo proporcional, o con el logaritmo de la duración, en el caso de especificar un modelo log-lineal. En algunas situaciones podríamos considerar que la relación paramétrica especificada para alguna de las variables explicativas es muy restrictiva, pudiendo resultar más adecuado introducir el efecto de esta variable de una forma no paramétrica. Así, tendríamos una especificación semiparamétrica, con una parte en la que se recogen variables relacionadas de una forma lineal con la variable a explicar, y otra parte, en la que no se especifique una particular dependencia paramétrica sobre la variable a analizar. Es decir, permitiríamos que los datos reflejaran esta relación mediante una curva de suavizado no paramétrica. Con esta generalización o extensión ampliamos de forma considerable el campo de aplicación de la metodología anterior. Esta extensión nos permite modelizar situaciones en las que no conocemos la forma funcional del efecto de una variable explicativa sobre la variable a explicar, o situaciones en las que suponer una dependencia lineal, u otra cualquiera, entre alguna

[5]Para más detalles sobre este contraste consultar Kiefer (1988).

de las variables explicativas y la variable a analizar sea un supuesto bastante fuerte, o incluso carezca de sentido.

Por tanto, como se puede apreciar la propuesta que vamos a presentar a continuación podría aplicarse en un importante número de situaciones. Un ejemplo ilustrativo del tipo de situaciones que podrían estimarse bajo esta propuesta se recoge en la Sección 3.

En el modelo de la Sección 3 la variable explicativa periodo intenta recoger el efecto de la introducción, a mediados del año 1987, del fármaco zidovudine. Esta variable está construida como una variable ficticia que toma dos valores: valor 1 indicándonos que el diagnóstico del individuo se ha producido con posterioridad a 1987, y valor 0 en caso contrario. Resulta bastante restrictivo dividir el efecto periodo de diagnóstico en dos grupos (antes y después de 1987). Además, parece más lógico o adecuado suponer que el efecto no va ser tan brusco como queda especificado por esa variable ficticia. Por tanto, en esta sección introducimos una componente adicional en el modelo compuesta por una función que depende del periodo de diagnóstico y no especificamos su forma funcional. Así podemos recoger el efecto que tratábamos de recoger, ahora de una forma gradual, además de la evolución completa del efecto que tiene el periodo en que se le diagnostica la enfermedad sobre la supervivencia del individuo.

Por tanto, la extensión que estamos proponiendo al trabajo de Aitkin y Clayton (1980) consiste en considerar modelos de duración con la siguiente función de riesgo

$$\lambda(t;x) = \lambda_0(t)\exp(x^T\beta + h(r)), \tag{6}$$

donde $h(r)$ es una función sin especificar con la cual se recoge el efecto de la variable explicativa R. Así, hemos pasado de un predictor lineal paramétrico $\eta = x^T\beta$ a uno semiparamétrico $\eta = x^T\beta + h(r)$.

Para la estimación de este modelo, al igual que para su equivalente paramétrico, proponemos maximizar la función de verosimilitud, aunque, en este caso consideramos una función de verosimilitud penalizada. Es decir, introducimos un término adicional que penaliza la no suavidad de la función h, y cuyo objetivo es hacer identificable la estimación de β. Por lo tanto, en este proceso de estimación queremos un buen ajuste y una función h lo más suave posible. Consideramos como funciones candidatas a todas las funciones pertenecientes al espacio de Sobolev de orden m ($W_2^m[a,b]$); es decir, todas aquellas funciones cuya derivada m-ésima al cuadrado es integrable en el intervalo $[a,b]$.

La bondad de ajuste dependerá del criterio de optimización elegido. Para el caso de la estimación por máxima verosimilitud, éste estará recogido en la función de verosimilitud.

La medida de suavidad para funciones $h \in (W_2^m[a,b])$ puede estar recogida por $\int_a^b \left[h^{(m)}(r)\right]^2 dr$. En la práctica lo habitual es considerar el caso $m = 2$.

Así, la estimación del modelo puede realizarse mediante la función de verosimilitud penalizada (Good y Gaskins, 1971), la cual considera o tiene en cuenta estas dos características. Para el modelo que estamos considerando, modelo (6), el logaritmo de la función de verosimilitud penalizada tiene la siguiente expresión,

$$\Pi = \sum_{i=1}^{n} \left\{ \delta_i \left[\ln \Lambda_0(y_i) + x_i^T \beta + h(r_i) \right] - \Lambda_0(y_i) e^{x_i^T \beta + h(r_i)} + \delta_i \ln \left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)} \right) \right\} -$$

$$(7) \qquad - \frac{1}{2} \alpha \int_a^b [h''(r)]^2 dr.$$

El parámetro de suavizado α, refleja la importancia que damos a la suavidad de la función y a la bondad del ajuste del modelo. Para un valor de α grande estamos dando más importancia a la suavidad, penalizando fuertemente las funciones estimadoras con segunda derivada elevada. Para un valor pequeño estamos dando mayor importancia al buen ajuste del modelo.

De forma análoga al tipo de modelo considerado en la Sección 2, la estimación puede realizarse construyendo un MLG auxiliar, en este caso semiparamétrico, con un logaritmo de la función de verosimilitud proporcional a (7).

El MLG semiparamétrico, auxiliar a este modelo de duración concreto, es un modelo log-lineal de Poisson para la variable indicador de censura δ, con una función de enlace logarítmica y el siguiente predictor lineal:

$$\ln \mu_i = \ln \Lambda_0(y_i) + x_i^T \beta + h(r_i)$$

El logaritmo de la función de verosimilitud penalizada de este modelo auxiliar viene dado por:

$$(8) \qquad \Pi = \sum_{i=1}^{n} \delta_i \ln \mu_i - \mu_i - \sum_{i=1}^{n} \ln \delta_i! - \frac{1}{2} \alpha \int_a^b \left[h''(r) \right]^2 dr.$$

Si se sustituye μ_i por su valor $e^{\ln \Lambda_0(y_i) + x_i^T \beta + h(r_i)}$, se puede comprobar que esta expresión es proporcional a (7). Por tanto, las estimaciones en una y otra expresión son las mismas.

Antes de pasar a maximizar la expresión (7), señalaremos que se puede demostrar que la solución, para la función h, al problema de maximizar (7) es una función «spline» cúbica natural. De esta forma y utilizando las propiedades de este tipo de funciones podemos reexpresar (7) (y de forma equivalente (8)) como

$$\Pi = \sum_{i=1}^{n} \left\{ \delta_i \left[\ln \Lambda_0(y_i) + x_i^T \beta + (Nh)_i \right] - \Lambda_0(y_i) e^{x_i^T \beta + (Nh)_i} + \delta_i \ln \left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)} \right) \right\} -$$

$$(9) \qquad - \frac{1}{2} \alpha h^T K h,$$

donde ahora h es el vector de valores $h_j = h(r_j)$, para $j = 1,\ldots,d$ donde d indica el número de valores distintos que toma la variable R, la matriz N se conoce como la matriz incidencia y su función consiste en asignar a cada elemento el valor que le corresponde de la variable que hemos introducido de forma no paramétrica y K es una matriz que se construye utilizando ciertas propiedades de las funciones spline cúbicas naturales[6].

Para maximizar respecto a los coeficientes β y a h podemos utilizar el algoritmo de Fisher scoring. La aplicación de este algoritmo, como se puede demostrar (la demostración en detalle puede encontrarse en Orbe, 2000, pag. 144-146), es equivalente a la resolución del siguiente sistema de ecuaciones simultáneas

$$
\begin{aligned}
X^T W X\beta &= X^T W(Z^* - Nh) \quad (a) \\
(N^T W N + \alpha K)h &= N^T W(Z^* - X\beta) \quad (b)
\end{aligned}
\tag{10}
$$

donde el elemento i-ésimo del vector Z^* es:

$$
z_i^* = \ln \Lambda_0(y_i) + x_i^T\beta + (Nh)_i + (\delta_i - \mu_i)\frac{1}{\mu_i} \tag{11}
$$

donde $\mu_i = \Lambda_0(y_i)e^{x_i^T\beta+(Nh)_i}$, W es una matriz de ponderaciones donde los elementos de la diagonal principal son de la forma[7]

$$
w_{ii} = \Lambda_0(y_i)e^{x_i^T\beta+(Nh)_i}, \tag{12}
$$

Para obtener las estimaciones de β y h podemos realizar un procedimiento de «backfitting» (Buja, Hastie y Tibshirani, 1989) entre las ecuaciones (10a) y (10b), hasta alcanzar la convergencia. Así, por una parte, si en la ecuación (10a) conocemos h, el vector de coeficientes β se obtiene regresando por mínimos cuadrados ponderados las diferencias $(Z^* - Nh)$ sobre la matriz de variables regresoras X (las variables de la componente paramétrica). Las ponderaciones serán las anteriormente indicadas. Por otra parte, si conocemos β en (10b) podemos obtener h mediante un suavizador spline cúbico natural aplicado a las diferencias $(Z^* - X\beta)$.

Resumiendo, el procedimiento de estimación completo comienza con la construcción de la matriz incidencia N. Posteriormente, iniciamos el proceso iterativo tomando $\hat{\beta} = 0$, calculamos las estimaciones iniciales del vector h regresando por mínimos cuadrados ordinarios el logaritmo de la variable indicadora de censura $\ln\delta$ sobre la matriz de incidencia. Es decir, $\hat{h} = (N^T N)^{-1}N^T(\ln(\delta))$. Con estas dos estimaciones iniciales, construimos la estimación inicial de $\hat{\mu}_i = \Lambda_0(y_i)e^{x_i^T\hat{\beta}+(N\hat{h})_i}$ y, aplicando la función de enlace

[6]Para más detalles véase, por ejemplo, Green y Silverman (1994), Cap. 2.

[7]Como es lógico μ_i y w_{ii} coinciden, puesto que, por una parte, tenemos una función enlace logarítmica y, por otra parte, en una distribución de Poisson la media y varianza coinciden.

logarítmica, obtenemos el valor inicial del predictor lineal $\hat{\eta}_i = \ln \Lambda_0(y_i) + x_i^T \hat{\beta} + (N\hat{h})_i$. Utilizando las estimaciones de μ y η, obtenemos el vector Z^* siguiendo la expresión (11) y la matriz de ponderaciones W siguiendo la expresión (12). Una vez obtenidos estos valores iniciales, comenzamos con el procedimiento de backfitting, sustituyendo de forma alternativa las estimaciones de (10a) y (10b) hasta que se produzca la convergencia.

En nuestro caso hemos visto que es adecuado proponer una distribución exponencial para la variable T. Por lo tanto, en la expresión (9) sustituimos las expresiones de las funciones de riesgo y riesgo acumulado por las correspondientes de una variable con distribución exponencial. Así, $\Lambda_0(t) = t$ y $\lambda_0(t)/\Lambda_0(t) = 1/t$. Y para maximizar el sistema (10) previamente debemos de realizar la misma sustitución en las expresiones (11) y (12).

En cuanto a la elección del parámetro de suavizado, existen dos aproximaciones al problema diferentes. Por una parte, una aproximación subjetiva que comtempla la posibilidad de que este parámetro sea libremente escogido por el investigador. Esta aproximación es la más utilizada en la práctica. Por otra parte, tenemos una alternativa automática donde el parámetro de suavizado es elegido por los datos. Entre estos criterios automáticos quizá el método más conocido sea el de validación cruzada. Una posibilidad interesante (realizada en este trabajo) sería la de combinar ambas aproximaciones. Como punto de partida utilizaremos un criterio automático como el de validación cruzada generalizada y posteriormente estimaremos el modelo con otros valores.

5. ANÁLISIS DE LAS ESTIMACIONES

Una vez estimados los parámetros del modelo y la función no paramétrica, se nos presenta el problema del análisis de la significatividad o, en general, de realizar inferencia. Dada la componente no paramétrica, podríamos pensar en utilizar contrastes asintóticos (Hastie y Tibshirani, 1990). En lugar de utilizar este tipo de contrastes hemos optado por realizar el estudio de las estimaciones obtenidas mediante técnicas bootstrap. Una de las ventajas que presenta el bootstrap es la posibilidad de analizar las propiedades y realizar inferencia incluso con tamaños de muestra reducidos. Sin embargo, no existe un método bootstrap específico adaptable al modelo propuesto, por lo que procedemos a la elaboración de uno.

Aplicaremos un bootstrap en regresión, ya que disponemos de un conjunto de observaciones no homogéneas, donde la heterogeneidad la tratamos de recoger a través de una serie de variables explicativas utilizando un modelo de regresión. Además, al suponer una distribución concreta para la variable de interés, desarrollaremos un bootstrap paramétrico.

La idea del bootstrap en regresión es la misma que la del bootstrap para modelos homogéneos. Dado que el modelo que estamos considerando parece el adecuado para nuestros datos, realizamos un bootstrap en regresión basado en el modelo. Este procedimiento consiste en obtener la remuestra bootstrap para la perturbación del modelo y, siguiendo la especificación del modelo, construir la remuestra bootstrap para la variable respuesta (para más detalles sobre las técnicas bootstrap, véase, por ejemplo, Efron y Tibshirani, 1993, y Davison y Hinkley, 1997).

Por otra parte, dado que en la muestra tenemos observaciones censuradas y esto tiene que reflejarse en las remuestras bootstrap, tenemos que aplicar un bootstrap adecuado para datos censurados. Tenemos dos posibilidades de remuestreo en el caso de muestras con censura. Efron (1981) propone estimar las funciones de distribución Kaplan-Meier (Kaplan y Meier, 1958) para la variable de interés $\hat{F}_n$ y lo mismo para la variable censura $\hat{G}_n$, posteriormente generar con ambas funciones de distribución sendas muestras para la variable de interés $t_1^*, \ldots, t_n^*$ y para la variable censura $c_1^*, \ldots, c_n^*$, y considerar la siguiente remuestra bootstrap,

$$y_i^* = \min(t_i^*, c_i^*), \qquad \delta_i^* = \begin{cases} 1; & \text{si } t_i^* \leq c_i^* \\ 0; & \text{si } t_i^* > c_i^* \end{cases} .$$

La otra posibilidad, presentada por Reid (1981), consiste en tomar una muestra de observaciones independientes e idénticamente distribuidas con la función de distribución, estimada mediante el estimador Kaplan-Meier, de la variable de interés y considerar la correspondiente función de distribución empírica.

Akritas (1986) demuestra que el plan de remuestreo de Efron es mejor que el de Reid. Además, para el caso de censura aleatoria, Efron demuestra que realizar lo anterior es equivalente a remuestrear con reemplazamiento sobre los pares de variable observada e indicador de censura $(y_1, \delta_1), \ldots, (y_n, \delta_n)$.

Hay que señalar que estos dos procedimientos de generación de muestras bootstrap, para muestras con observaciones censuradas, están pensados para el caso de muestras homogéneas; es decir, para situaciones en las que no tenemos variables explicativas que influyen sobre la variable a explicar, y para el caso en que desconocemos las funciones de distribución de la variable de interés y de la variable censura. Sin embargo éste no es nuestro caso. En nuestro problema, estamos suponiendo una distribución para la variable de interés T (distribución exponencial) y no estamos suponiendo distribución alguna para la variable censura C y, además, tenemos variables explicativas en el modelo. Por lo tanto, para solucionar este problema, tenemos que proponer un nuevo procedimiento generador de muestras bootstrap, adecuado a los supuestos del modelo. Hay que señalar que la propuesta de Efron (para el caso de no suponer la distribución de la variable duración y la variable censura), aún seguiría siendo valida para el caso heterogéneo siempre y cuando supongamos que la variable censura siga el mismo modelo de regresión propuesto para la variable duración.

El modelo concreto que estamos considerando es un modelo de regresión exponencial semiparamétrico, es decir, tenemos la siguiente función de densidad para la variable de interés T:

$$f(t;x) = \lambda e^{-\lambda t}; \quad \text{donde} \quad \lambda = e^{-(x^T\beta + h(r))}.$$

Para la variable censura C no estamos suponiendo distribución alguna.

Para este tipo de modelos proponemos el siguiente procedimiento para generar las remuestras bootstrap:

Paso 1: Ajustar el modelo (6) para el caso de una distribución exponencial.

Paso 2: Generar las perturbaciones bootstrap $\varepsilon_1^*, \ldots, \varepsilon_n^*$, con una distribución valor extremo mínimo.

Paso 3: Obtener la muestra bootstrap para la variable de interés basándonos en el modelo

$$\ln T_i^* = x_i^T \hat{\beta} + \hat{h}(r_i) + \varepsilon_i^*; \quad \text{para} \quad i = 1, \ldots, n.$$

Paso 4: Obtener la muestra bootstrap para variable censura generando una muestra de n observaciones a partir de la función de distribución G de la variable censura.

Paso 5: Comparando las remuestras bootstrap para la variable de interés (paso 3) y la variable censura (paso 4), obtenemos la variable observada bootstrap Y^*, y la correspondiente variable indicador bootstrap δ^*,

$$y_i^* = \min\{t_i^*, c_i^*\}, \qquad \text{y} \qquad \delta_i^* = \begin{cases} 1; & \text{si } t_i^* \leq c_i^* \\ 0; & \text{si } t_i^* > c_i^* \end{cases}.$$

Paso 6: Estimar el modelo (6) (para el caso exponencial) utilizando la información disponible en la remuestra bootstrap.

Paso 7: Volver al paso 2 y repetir el proceso M veces.

Para obtener las estimaciones del modelo (6), en el paso 1, desarrollamos el proceso de estimación propuesto en la sección anterior para el caso particular de una distribución exponencial para la variable T. En el paso 2 estamos considerando el modelo lineal para el logaritmo de la duración[8]. Por lo tanto, en este paso, obtenemos la remuestra

[8] Como suponemos una distribución exponencial de parámetro $\lambda = e^{-(x^T\beta + h(r))}$ para la variable duración, al tomar la transformación logarítmica podemos reescribir el modelo en términos log-lineales como $\ln T = X\beta + h(r) + \varepsilon$ donde, entonces, ε tiene una distribución valor extremo mínimo.

bootstrap de las perturbaciones realizando un bootstrap paramétrico, donde consideramos una distribución valor extremo para las perturbaciones. En el paso 3, como acabamos de comentar, utilizamos la expresión log-lineal de nuestro modelo (6) (ver pie de página 8) para obtener la remuestra bootstrap de la variable de interés. En el paso 4, generamos la variable censura sin considerar ningun supuesto adicional al modelo, que contemple una relación determinada entre las variables explicativas y ésta. La función de distribución G de la variable censura es desconocida y la estimamos utilizando el estimador de Kaplan-Meier, $\hat{G}_n$, adecuado para esta variable. En el paso 6, y como en el paso 1, utilizamos el procedimiento de estimación descrito en la Sección 4. Por último, indicaremos que el número de remuestras bootstrap a considerar depende del objetivo del estudio, si únicamente deseamos calcular las desviaciones típicas de las estimaciones obtenidas, un valor de $M = 200$ puede ser suficiente para obtener unos valores fiables. En cambio, si nuestro objetivo es más ambicioso, y deseamos construir intervalos de confianza, tenemos que considerar un número sensiblemente superior (al menos $M = 1000$), para tener una buena estimación de los percentiles en las colas de la distribución.

6. RESULTADOS Y CONCLUSIONES

Como ilustración de las dos secciones anteriores aplicamos la metodología descrita al conjunto de datos presentados en la Sección 3. Así, la motivación de la extensión del modelo paramétrico, ajustado en la Sección 3, a uno semiparamétrico, tiene su fundamento en el supuesto más razonable de un efecto real, de la introducción del fármaco AZT, más suave o más gradual que el especificado en la Sección 3, utilizando una variable ficticia. Por lo tanto, ahora consideramos un modelo semiparamétrico, concretamente ajustamos el modelo (6) para el caso particular de una variable T con distribución exponencial. El efecto de las variables explicativas quedará dividido en dos términos. El paramétrico, donde recogemos el efecto de todas las variables explicativas excepto la variable periodo de diagnóstico[9] cuyo efecto sera recogido de una forma no paramétrica a través de una función h. El parámetro de suavizado toma un valor igual a 50.

Los resultados de esta estimación y del posterior análisis, de las estimaciones obtenidas mediante técnicas bootstrap, son presentados en las Tablas 3, 4 y Figura 2. Indicaremos que el número de remuestras bootstrap considerado para este análisis es de $M = 1999$.

[9]Ahora, a diferencia de la Sección 3, la variable periodo no va a ser una variable ficticia. En su lugar vamos a crear una nueva variable periodo de diagnóstico que toma valor 1 para los individuos diagnosticados de SIDA en el primer trimestre de la muestra, valor 2, para los diagnosticados en el segundo trimestre, y así sucesivamente, hasta el último trimestre de diagnóstico existente en la muestra.

Tabla 3. Estimaciones de los coeficientes β de la componente paramétrica

Variable	Coeficiente	Desv. típica
Constante	1.37063	0.40330
Sexo	0.02259	0.13103
Enfer1	0.19599	0.25522
Enfer2	0.13172	0.31948
Sexual	−0.23432	0.24722
Drogas	−0.10928	0.20709
Sanguínea	−0.00008	0.28563
Madre-hijo	0.17904	0.47801
Edad	−0.01870	0.00643

Tabla 4. Intervalos de confianza al 95% para las estimaciones de los coeficientes β

	Intervalos bootstrap		Intervalos asintóticos	
Variable	**Lim. Inf.**	**Lim. Sup.**	**Lim. Inf.**	**Lim. Sup.**
Constante	0.5391	2.1134	0.5171	2.2241
Sexo	−0.2176	0.2659	−0.2342	0.2794
Enfer1	−0.2796	0.7088	−0.2893	0.6813
Enfer2	−0.4584	0.7571	−0.4907	0.7542
Sexual	−0.6794	0.2768	−0.7090	0.2404
Drogas	−0.4636	0.3413	−0.5143	0.2957
Sanguínea	−0.5258	0.5947	−0.5451	0.5449
Madre−hijo	−0.6230	1.1253	−0.7045	1.0626
Edad	−0.0308	−0.0056	−0.0317	−0.0056

La Tabla 3 muestra la estimación de los coeficientes β para aquellas variables introducidas en la componente paramétrica del modelo junto a la estimación bootstrap de sus desviaciones típicas. La Tabla 4, además de presentar los intervalos de confianza bootstrap percentil BC (al 95%) para estos coeficientes β, también incluye los intervalos de confianza asintóticos, que habitualmente se calculan en el contexto de los modelos aditivos generalizados[10]. Los resultados son similares, aunque en líneas generales se puede observar que la amplitud de los intervalos bootstrap es menor. Además, estos

[10]Para más detalles veáse Hastie y Tibshirani (1990).

son válidos incluso para muestras de tamaño reducido. La Figura 2 nos presenta la estimación de la componente no paramétrica, la función $h(r)$, así como, las bandas de confianza bootstrap percentil al 95% (para una detallada descripción sobre intervalos de confianza bootstrap ver, por ejemplo, Efron, 1987 y Efron y Tibshirani, 1986).

Antes de pasar a interpretar los resultados obtenidos tenemos que señalar que las estimaciones presentadas en las tablas y figura mencionadas anteriormente indican el efecto de esas variables sobre el logaritmo de la duración. El efecto sobre la función de riesgo va a ser el mismo pero de signo contrario al presentado en las tablas y figura. Una vez aclarada esta cuestión pasamos a reseñar los resultados más relevantes.

En cuanto a las variables introducidas en la componente paramétrica del modelo, indicaremos que únicamente la variable edad resulta significativa para explicar el tiempo de supervivencia del enfermo. A mayor edad en el momento del diagnóstico tenemos un tiempo de supervivencia menor para el enfermo. El resto de las variables de la componente paramétrica resultan no significativas para explicar la supervivencia. Estos mismos resultados se han obtenido en otros trabajos como se recoge en la síntesis de resultados, obtenidos por diferentes autores aplicando diferentes metodologías, presentada en Brookmeyer y Gail (1993).

En cuanto a la componente no paramétrica, propuesta para flexibilizar la más restrictiva aproximación realizada en la Sección 3 (donde se dividía el periodo de estudio en dos partes mediante una variable ficticia), podemos apreciar además del efecto de la introducción del fármaco AZT, la evolución del efecto del periodo de diagnóstico sobre el tiempo de supervivencia. Así, podemos observar una tendencia ligeramente creciente, mayores tiempos de supervivencia, a medida que nos desplazamos de los primeros periodos de diagnóstico. Esta suave tendencia creciente puede venir provocada por el cada vez mayor conocimiento de la enfermedad con el paso del tiempo, lo cual puede originar diagnósticos cada vez más precoces, aumentando así el tiempo de supervivencia desde el momento del diagnóstico. Posteriormente, observamos una fuerte aceleración, en este efecto positivo, sobre la supervivencia, para finalmente mantenerse en niveles máximos. Aquí habría que recordar que la introducción del AZT se produce a mediados de 1987 (alrededor del trimestre 13).

Por lo tanto, la Figura 2 parece mostrarnos un efecto beneficioso de la introducción del fármaco, provocando una importante mejora en el tiempo de supervivencia del enfermo. Como se puede apreciar en la figura, la aceleración de este efecto positivo se produce varios trimestres antes de la introducción del fármaco, lo cual resulta bastante lógico, ya que individuos diagnósticados de SIDA, antes de la introducción del fármaco, también van a recibir el fármaco (aunque no desde un principio) y por tanto, también se benefician de los resultados positivos de éste. Señalaremos que este efecto positivo del AZT también se obtiene en el modelo de la Sección 3 y en otros trabajos como se señala en Brookmeyer y Gail (1993). Entre ellos podemos citar, por ejemplo, Lemp y otros (1990) y Moore y otros (1991). Sin embargo, tenemos que añadir que con la especi-

ficación semiparamétrica que proponemos en la Sección 4 somos capaces de capturar el efecto del AZT de una forma gradual y más flexible, cosa que no podemos hacer bajo una especificación con variable ficticia, puesto que esta especificación está considerando un efecto repentino o brusco. Además, la especificación semiparamétrica nos permite, aparte del efecto de la introducción del AZT, analizar la evolución total del efecto periodo de diagnóstico sobre la supervivencia.

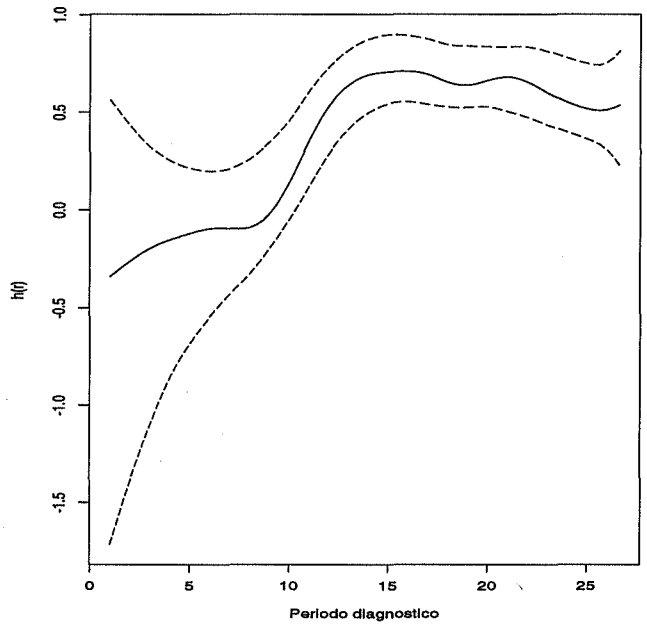


Figura 2. Estimación e intervalo de confianza bootstrap (95%) para la componente no paramétrica

Antes de finalizar, indicaremos que la principal motivación del trabajo presentado ha sido la propuesta de extensión del trabajo de Aitkin y Clayton (1980). Por tanto, el análisis empírico llevado a cabo, prentende, básicamente, ilustrar la metodología propuesta. Aún así, se han extraido una serie de resultados interesantes que, además, nos pueden ayudar a entender mejor la relevancia de la propuesta que estamos realizando. La extensión propuesta amplía el campo de aplicación de la metodología de esos autores, permitiendo considerar aquellas situaciones donde la forma funcional del efecto de alguna de las variables explicativas sobre la variable de interés es desconocida o situaciones en las que la especificación de una determinada forma funcional resulta un supuesto bastante restrictivo o carece de sentido. Para finalizar, señalaremos que

la inferencia del modelo se ha realizado mediante técnicas bootstrap, para lo cual hemos propuesto un procedimiento de generación de remuestras bootstrap adecuado a las características del modelo.

AGRADECIMIENTOS

Este trabajo ha sido financiado por los proyectos de investigación UPV 038.321-HA129/99 de la Universidad del País Vasco/Euskal Herriko Unibertsitatea, PB98-0149 de la Dirección General de Enseñanza Superior e Investigación Científica del Ministerio Español de Educación y Cultura y PI-1999-70 del Gobierno Vasco/Eusko Jaurlaritza. El autor agradece tanto los comentarios de la editora como de los evaluadores que han servido para mejorar de forma importante el trabajo realizado.

REFERENCIAS

Aitkin, M. & Clayton, D. (1980). «The Fitting of Exponential, Weibull and Extreme Value Distributions to Complex Censored Survival Data using GLIM». *Applied Statistic*, 29, 156-163.

Akritas, M. G. (1986). «Bootstrapping the Kaplan-Meier Estimator». *Journal of the American Statistical Association*, 81, 1032-1038.

Brookmeyer, R. & Gail, M. H. (1993). *AIDS Epidemiology a Quantitative Approach.* Oxford University Press: Oxford.

Buja, A., Hastie, T. J. & Tibshirani, R. J. (1989). «Linear Smoothers and Additive Models (with Discussion)». *Annals of Statistic*, 17, 453-555.

Cox, D. R. (1972). «Regression Models and Life-Tables». *Journal of the Royal Statistical Society-Series B*, 34, 187-220.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press: Cambridge.

Efron, B. (1981). «Censored Data and Bootstrap». *Journal of the American Statistical Association*, 76, 312-319.

— (1987). «Better Bootstrap Confidence Intervals». *Journal of the American Statistical Association*, 82, 171-200.

Efron, B. & Tibshirani, R. (1986). «Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy». *Statistical Science*, 1, 54-77.

— (1993). *An Introduction to the Bootstrap.* Chapman and Hall: New York.

Good, I. J. & Gaskins, R. A. (1971). «Non-parametric Roughness Penalties for Probability Densities». *Biometrika*, 58, 255-277.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall: London.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall: London.

Kaplan, E. L. & Meier, P. (1958). «Nonparametric Estimation from Incomplete Observations». *Journal of the American Statistical Association*, 53, 457-481.

Kiefer, N. M. (1988). «Economic Duration Data and Hazard Functions». *Journal of Economic Literature*, 26, 646-679.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons: New York.

Lemp, G. P., Payne, S. F. & Neal, D. (1990). «Survival Trends for Patients with AIDS». *Journal of the American Medical Association*, 263, 402-406.

McCullagh, P. & Nelder, J. A. (1983). *Generalized Linear Models*. Chapman and Hall: London.

Moore, R. D., Hidalgo, J., Sugland, B. W. & Chaisson, R. E. (1991). «Zidovudine and the Natural History af the Acquired Inmunodeficiency Syndrome». *New England Journal of Medicine*, 263, 1412-1416.

Nelder, J. A. & Wedderburn, R. W. M. (1972). «Generalized Linear Models». *Journal of the Royal Statistical Society-Series A*, 135, 370-384.

Orbe, J. (2000). *Un Modelo de Regresión Parcial Censurado para Análisis de Supervivencia*. Tesis Doctoral, Universidad del País Vasco, Bilbao.

Reid, N. (1981). «Estimating the Median Survival Time». *Biometrika*, 68, 601-608.

Whitehead, J. (1980). «Fitting Cox's Regression Model to Survival Data using GLIM». *Applied Statistics*, 29, 268-275.

ENGLISH SUMMARY

LIFETIME DATA ANALYSIS USING A SEMIPARAMETRIC GENERALIZED LINEAR MODEL

JESUS ORBE*

Aitkin and Clayton (1980) propose to analyze duration models using generalized linear models. In this work, we extend that methodology by allowing the introduction of the effect of some covariable in a nonparametric way. Thus, the proposed model is a semiparametric generalized linear model, with a parametric component where we specify the functional form of the effect of the covariables on the duration variable, and a nonparametric component where we capture the effect of some covariable without assuming any functional form. We develop the estimation process and we use a bootstrap procedure to infer on the estimates for the parameters. As an application, we study the survival time for a sample of AIDS diagnosed patients.

Keywords: Duration models, censorship, bootstrap, generalized linear models, semiparametric estimation .

AMS Classification (MSC 2000): 62J12, 62N05, 62G09

* Departamento de Econometría y Estadística (E.A. III). Universidad del País Vasco/Euskal Herriko Unibertsitatea. Avenida Lehendakari Aguirre, 83. 48015 Bilbao. E-mail: jo@alcib.bs.ehu.es.

–Received June 2000.
–Accepted February 2001.

1. INTRODUCTION

In this paper we propose an extension of the work presented by Aitkin and Clayton (1980). These authors put forward the possibility of estimating some duration models using generalized linear models. We use this idea and extend their methodology to a semiparametric case. By using this extension, we can consider situations where we do not know the functional form of the effect of some covariate on the duration or situations where considering some specific parametric functional form for this effect may be very restrictive.

We first describe the link between duration models and generalized linear models. Thus, we consider the class of proportional hazard models

$$\lambda(t;x) = \lambda_0(t)\exp(x^T\beta),$$

where $\eta = x^T\beta$ is the linear predictor. For a sample of n observations, we can obtain this log-likelihood for the model

$$\ln L = \sum_{i=1}^{n} \delta_i [\ln \Lambda_0(y_i) + \eta_i] - \Lambda_0(y_i)e^{\eta_i} + \delta_i \ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right)$$

by taking $\mu_i = \Lambda_0(y_i)e^{\eta_i}$, we can rewrite this equation as

$$\ln L = \underbrace{\sum_{i=1}^{n}(\delta_i \ln \mu_i - \mu_i)}_{(a)} + \underbrace{\sum_{i=1}^{n} \delta_i \ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right)}_{(b)}.$$

It can be verified that the (a) is proportional to the logarithm of the likelihood function of a sample of δ_i random variables with Poisson distribution with mean value μ_i. Therefore, the parameters of the original duration model can be estimated using a generalized linear model (GLM), the log-linear Poisson model,

$$\ln \mu_i = \ln \Lambda_0(y_i) + x_i^T\beta$$

We present a dataset of AIDS diagnosed patients and analyze the effect of some covariables on the survival time from the diagnosis moment using the traditional methodology in survival analysis. Using this application we motivate the extension of the methodology proposed by Aitkin and Clayton (1980).

2. DURATION ANALYSIS USING A SEMIPARAMETRIC GLM

We propose to extend the work of Aitkin and Clayton (1980) by introducing a nonparametric term in the model. Thus, we consider duration models with hazard function

$$\lambda(t;x) = \lambda_0(t)\exp(x^T\beta + h(r)),$$

where $h(r)$ is a smooth function that is used to capture the effect of the covariable R. Therefore, we have passed from a parametric linear predictor $\eta = x^T\beta$ to a semiparametric one $\eta = x^T\beta + h(r)$.

In order to estimate the parameters of the model and the function $h(r)$, we can use the penalized log-likelihood function

$$\Pi = \sum_{i=1}^{n} \left\{ \delta_i \left[\ln \Lambda_0(y_i) + x_i^T\beta + h(r_i) \right] - \Lambda_0(y_i) e^{x_i^T\beta + h(r_i)} + \delta_i \ln \left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)} \right) \right\} -$$

$$- \frac{1}{2}\alpha \int_a^b [h''(r)]^2 dr.$$

Here, as in the parametric case, we can use a generalized linear model to obtain the estimators but, in this case, a semiparametric one. Thus, we can use a log-linear Poisson model for the censorship δ indicator variable, with a logarithmic link function and the following semiparametric linear predictor

$$\ln \mu_i = \ln \Lambda_0(y_i) + x_i^T\beta + h(r_i)$$

The penalized log-likelihood function for this «auxiliar» semiparametric generalized linear model is

$$\Pi = \sum_{i=1}^{n} \left[\delta_i \ln \mu_i - \mu_i - \sum_{i=1}^{n} \ln \delta_i! \right] - \frac{1}{2}\alpha \int_a^b \left[h''(r) \right]^2 dr.$$

If we substitute μ_i by $e^{\ln \Lambda_0(y_i) + x_i^T\beta + h(r_i)}$, we can see that this expression is proportional to the previous one.

It can be demonstrated that the solution to maximize the penalized log-likelihood function for $h(r)$ function is a natural cubic spline. Therefore, using some properties of these functions, the penalized log-likelihood can be rewritten as

$$\Pi = \sum_{i=1}^{n} \left\{ \delta_i \left[\ln \Lambda_0(y_i) + x_i^T\beta + (Nh)_i \right] - \Lambda_0(y_i) e^{x_i^T\beta + (Nh)_i} + \delta_i \ln \left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)} \right) \right\} -$$

$$- \frac{1}{2}\alpha h^T K h$$

The same steps can be carried out in the penalized log-likelihood of the «auxiliar» semiparametric generalized linear model and, then, using the Fisher scoring algorithm, we can obtain the equations system

$$X^T W X \beta = X^T W (Z^* - Nh) \qquad (a)$$

$$(N^T W N + \alpha K) h = N^T W (Z^* - X\beta) \qquad (b)$$

where the i-th element of vector Z^* is

$$z_i^* = \ln \Lambda_0(y_i) + x_i^T \beta + (Nh)_i + (\delta_i - \mu_i)\frac{1}{\mu_i}$$

The elements of the main diagonal in W are

$$w_{ii} = \Lambda_0(y_i) e^{x_i^T \beta + (Nh)_i}$$

In order to obtain the estimators of the model, we can apply a backfitting algorithm between (a) and (b) until convergence is achieved.

3. INFERENCE AND MAIN RESULTS

Once the estimation procedure is finished, we are interested in doing inference. This analysis is done by using bootstrap resampling techniques. In order to do this, we propose a new procedure to obtain the bootstrap resamples that are adequate to the characteristics of our model. Considering our case, a semiparametric exponential regression model, this procedure consists on the following steps:

Step 1: Fit the original model using the proposed methodology.

Step 2: Generate the bootstrap sample for the error variable, $\varepsilon_1^*, \ldots, \varepsilon_n^*$, using a minimum extreme value probability distribution.

Step 3: Obtain the bootstrap sample for the duration variable doing a model-based bootstrap. That is,

$$\ln T_i^* = x_i^T \hat{\beta} + \hat{h}(r_i) + \varepsilon_i^*; \quad \text{for} \quad i = 1, \ldots, n.$$

Step 4: Obtain the bootstrap sample for the censoring variable through the estimation of the distribution function of censoring variable, G.

Step 5: Compare the bootstrap samples of the duration and censoring variables and, thus, obtain the bootstrap sample of the observed variable Y^*, and the corresponding bootstrap indicator variable δ^*,

$$y_i^* = \min\{t_i^*, c_i^*\}, \qquad \text{and} \qquad \delta_i^* = \begin{cases} 1; & \text{if } t_i^* \leq c_i^* \\ 0; & \text{if } t_i^* > c_i^* \end{cases}.$$

Step 6: Estimate the model for the bootstrap sample.

Step 7: Go back to Step 2 and repeat the procedure M times.

After we estimate the model and, using the bootstrap techniques, calculate the standard deviations and confidence intervals, we can summarize the most relevant results obtained from the empirical analysis. With regard to the covariates introduced in the parametric component, the age of the patient has a negative significant effect on the survival time. As for the estimation of the nonparametric component, we can observe that the introduction of AZT treatment has a positive effect on the survival, increasing the survival time of patients. Finally, I would like to add that, with the extension proposed here, it is possible to capture the gradual effect on survival of this medicine, which is not possible by using a dummy variable specification.

Secció Docent i Problemes

SECCIÓ DOCENT I PROBLEMES

La «Secció docent i problemes» té l'objectiu de publicar articles de caire docent, difícilment publicables en revistes de recerca. A cada número de *Qüestiió* s'inclouen d'un a tres problemes i les solucions es donen en el número següent.

Els lectors poden proposar problemes amb les solucions pertinents i enviar-los a *Qüestiió,* que farà una selecció i en publicarà els més adequats, fent la corresponent referència a l'autor.

També seran ben rebudes solucions alternatives a les propostes fetes per l'autor dels problemes. L'editorial es reservarà, però, el dret a publicar-les.

SOLUCIÓ AL PROBLEMA PROPOSAT AL VOLUM 25 N. 1

PROBLEMA N. 88

Admitiendo que el número X de artículos evaluados por cada experto sigue la distribución de Poisson

$$f_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad k = 0,1,2,\ldots$$

como sólo tenemos datos de los expertos que evaluaron algún artículo, debemos considerar la distribución de X condicionada a $X > 0$, es decir

$$f_\lambda(k/X>0) = \frac{e^{-\lambda}\cdot\lambda^k}{(1-e^{-\lambda})k!} \quad k = 1,2,\ldots$$

El valor medio de esta variable es

$$\mu = \sum_{k=1}^{\infty}\frac{k\cdot e^{-\lambda}\cdot\lambda^k}{(1-e^{-\lambda})k!} = \frac{\lambda\cdot e^{-\lambda}}{(1-e^{-\lambda})}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}$$

$$= \frac{\lambda\cdot e^{-\lambda}}{(1-e^{-\lambda})}e^{\lambda} = \frac{\lambda}{1-e^{-\lambda}}.$$

La probabilidad de $X = 1$ condicionada a $X > 0$ es

$$p(\lambda) = \frac{\lambda\cdot e^{-\lambda}}{1-e^{-\lambda}}$$

La probabilidad de $X = 1$ condicionada a $X > 0$ es $1 - p(\lambda)$. Hemos observado las frecuencias siguientes:

$a = 80$ expertos que evaluaron un artículo,
$b = 66$ expertos que evaluaron más de un artículo.

A fin de estimar λ, la función y la ecuación de verosimilitud son:

$$L = p(\lambda)^a(1-p(\lambda))^b$$

$$\ln L = a\ln p(\lambda) + b\ln(1-p(\lambda))$$

$$\frac{\partial \ln L}{\partial\lambda} = \frac{a\,p'(\lambda)}{p(\lambda)} - \frac{b\,p'(\lambda)}{(1-p(\lambda))} = 0$$

Como $p'(\lambda) > 0$ para todo $\lambda > 0$, la ecuación se reduce a

$$\frac{a}{p(\lambda)} - \frac{b}{(1-p(\lambda))} = 0 \quad \Rightarrow \quad p(\lambda) = \frac{a}{a+b}.$$

La solución de la ecuación

$$\frac{\lambda \cdot e^{-\lambda}}{1-e^{-\lambda}} = \frac{80}{80+66}$$

es $\hat{\lambda} = 1.10$. La estimación máximo verosímil de μ es

$$\hat{\mu} = \frac{\hat{\lambda}}{1-e^{-\hat{\lambda}}} = 1.65$$

Respuesta: los expertos (referees) consultados por *Qüestiió* (1987-1997) evaluaron, por término medio, 1.65 artículos cada uno.

C.M. Cuadras
Universitat de Barcelona

PROBLEMA PROPOSAT

PROBLEMA N. 89

Consider n independent $(p \times 1)$ vectors x_i $(i = 1, \ldots, n)$,$n \geqslant p$, each with the distribution $\mathcal{N}_p(\mu, V)$. Define then $S := (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$, where $\bar{x} := n^{-1} \sum_{i=1}^{n} x_i$. Find $E(S^{-1})$.

Heinz Neudecker

Comentaris de llibres

MODELOS DE ECUACIONES ESTRUCTURALES

Joan Manuel Batista Foguet y Germà Coenders Gallart

Editorial LA MURALLA, Madrid, 2000
Colección: CUADERNOS DE ESTADÍSTICA, 174 pp + tablas + figuras

Este libro es el sexto de la colección Cuadernos de Estadística, colección dedicada al análisis multivariante, regresión y análisis de la varianza. «Modelos de Ecuaciones Estructurales» es una interesante monografía sobre el tema del análisis factorial confirmatorio y sus modernas extensiones. En ciencias experimentales el científico trata de explicar a partir de lo que puede observar y medir, pero en las ciencias esencialmente no experimentales, como la psicología, la sociología y la economía, hay ciertas «variables latentes», que a menudo forman parte del lenguaje común (la inteligencia general, la satisfacción, la subversión política, las derechas y las izquierdas, etc.), que no son directamente medibles pero que existen al menos por conveniencia humana y como abstracción de conceptos muy amplios. Admitiendo la existencia de las variables latentes, se hace necesario medirlas indirectamente a través de variables observables bajo un determinado modelo.

El libro empieza con una breve exposición de los dos enfoques del análisis factorial, exploratorio y confirmatorio, introducidos por Ch. Spearman y K. Joreskog, respectivamente, aunque también deberíamos mencionar los nombre de L. L. Thurstone y D. N. Lawley. Otro notable precedente es el «path analysis» de S. Wright, que permitía estudiar las relaciones entre variables, con aplicación a la genética cuantitativa. Los modelos de ecuaciones estructurales empiezan en 1970, cuando una conferencia entre psicómetras, sociómetras y económetras permitió poner de relieve la parte común de diversas metodologías y elaborar una nueva que hacía posible modelar, entender e interpretar las relaciones lineales entre variables observables y variables latentes, a partir de una matriz de covarianzas y de un modelo de relación. Un aspecto fundamental es la posibilidad de aceptar o rechazar el modelo mediante un test. Aparecieron entonces diversos programas para tratar tales modelos, siendo Lisrel y EQS los más conocidos.

La monografía, en su capítulo 2, motiva al lector con un ejemplo de investigación de mercados que permite introducir los conceptos principales, con la descripción de las variables de medida y propuesta de las dimensiones latentes.

El capítulo 3 está dedicado a explicar las reglas y símbolos del «diagrama de caminos», para relacionar las variables observables con las variables latentes, incluyendo los términos de error. La ecuación fundamental

$$\text{observación} = \text{modelo} + \text{error}$$

es explicada y representada en el contexto de las ecuaciones estructurales, con diagramas simples e intuitivos, que deben permitir al investigador expresar fácilmente la estructura de sus variables a partir del conocimiento que tiene de la realidad. Los autores ilustran con varios ejemplos este primer paso de la modelización.

En los capítulos 4 y 5 se generalizan y profundizan los conceptos anteriores, permitiendo abordar modelos más complejos. Las diferentes etapas (especificación, identificación, estimación, diagnóstico y modificación) se exponen ordenadamente, pasando de lo particular a lo general. Los autores no se limitan a los casos estándar, basados en normalidad multivariante y estimación máximo verosímil, sino que discuten métodos y criterios alternativos, como la utilización de variables ordinales, métodos asintóticamente a libre distribución y de remuestreo.

El capítulo 6 se dedica a describir la utilización del programa Lisrel, de Joreskog y Sorbom, lo que se lleva a cabo con numerosas figuras, que representan cuadros y ventanas que permiten al usuario la implementación práctica, a partir de unos datos, de los pasos explicados en los capítulo anteriores. El capítulo 7 expone los resultados de la utilización de Lisrel y es por lo tanto la continuación del capítulo 6. Ambos capítulos se ilustran adecuada y exhaustivamente con el ejemplo previamente introducido en el capítulo 2.

El capítulo 8 está dedicado a presentar, elaborar y comentar dos nuevos ejemplos (estabilidad de actitudes y evaluación de la calidad de medida en modelos multirasgo-multimétodo), que los propios autores han estudiado, utilizando como medida de ajuste el estadístico ji-cuadrado de Satorra-Bentler.

Finalmente, el capítulo 9, con el título «Resumen y Conclusiones», contiene un rosario de opiniones, consejos y observaciones de tipo científico, dirigidas al usuario de los modelos de ecuaciones estructurales, advirtiendo del uso equivocado de los mismos y de que la falta de cumplimiento de las condiciones del modelo puede inducir a falsas relaciones causales. En mi opinión, el pincipal escollo de esta metodología es «acertar» con el modelo, pues a menudo la variabilidad explicada por el modelo es demasiado baja. Las recomendaciones contenidas en este capítulo, en el que se pone de manifiesto la formación en el campo de la psicología del primer autor, son dignas de tener en cuenta.

«Modelos de Ecuaciones Estructurales», escrito por dos reputados especialistas, no debe considerarse un manual para usuarios, sino más bien una monografía que introduce al lector en el tema de una forma erudita y actualizada (pues contiene numerosas referencias bibliográficas), y que es muy recomendable para todos aquellos estudiantes, profesores, profesionales e investigadores en ciencias sociales, que en sus estudios necesiten modelar las relaciones entre diversas variables.

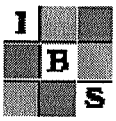
C.M. Cuadras
Universitat de Barcelona

Ressenyes d'activitats institucionals

Sociedad Española de Biometría

Región Española
de la Sociedad Internacional de Biometría

Spanish Region
of the International Biometric Society



http://www.iata.csic.es/ibsresp

La **Sociedad Española de Biometría/Región Española de la Sociedad Internacional de Biometría** (abreviadamente **SEB** o ***REsp***) tiene como objetivos promover, impulsar y difundir el desarrollo y la aplicación de los métodos matemáticos y estadísticos a la biología, medicina, psicología, farmacología, agricultura y otras ciencias afines (ciencias relacionadas con los seres vivos). Cualquier profesional o alumno de estas disciplinas puede ser miembro de la SEB.

Consejo Directivo

Presidenta:	María Jesús Bayarri García (Medicina)
Vicepresidenta:	Guadalupe Gómez Melis (Biología)
Secretario y Tesorero:	Fernando López Santoveña (Agronomía)
Vocal en calidad de Miembro del Consejo de la IBS:	Emilio A. Carbonell Guevara (Agronomía)
Vocales:	Juan Luis Chorro Gascó (Psicología)
	Juan Ferrándiz Ferragud (Biología)
	Purificación Galindo Villardón (Medicina)
	Eduardo García Cueto (Psicología)
	José Luis González Andújar (Agronomía)
	Alex Sánchez Plá (Biología)
Corresponsal de la REsp en el «Biometric Bulletin» de la IBS:	María Luz Calle Rosingana

VIII CONFERENCIA ESPAÑOLA DE BIOMETRÍA

Pamplona, 28, 29 y 30 de marzo de 2001

Información: www.unavarra.es/directo/congresos/apoyo/biometria.htm

The TES Institute
Training of European Statisticians

GENERAL INTRODUCTION

The TES Institute is a non-profit making association of 17 European National Statistical Institutes and the University Centre of Luxembourg. It offers a post-graduate vocational training programme for statisticians and other groups working in the statistical environment. The TES Institute today boasts a solid foundation of statistical training experience, built up since the beginning of the nineties.

Our mission is to provide world-wide cutting edge lifelong international training for professional statisticians. The training emphasises the inter country know-how transfer and focuses on the dissemination of best practices in terms of providing statistics compliant with both European and International Statistical System standards.

The training programmes of the TES Institute are designed for both public and private sector statisticians in the broadest sense of the word i.e. university graduates of any discipline involved in statistical work within Statistical Institutes, European Institutions, Government Agencies, National or private Banks, Enterprises, ...

Training methods are problem-oriented and based on a twofold track approach combining both teaching of theoretical concepts and practical sessions using real life cases. Moreover, the «learning-by-doing» process provides an opportunity to respond in an effective way to recent scientific and technological developments as well as new requirements from the labour market.

The training programmes offered by the TES Institute provide both theoretical and practical background but the courses have all a very strong applied character.

These programmes also offer participants the opportunity to meet colleagues form all over Europe and other countries since the TES Institute has extended its activities to the Central European, Mediterranean Basin and TACIS countries.

The above characteristics represent the basic conditions to acquire sharper competence in their work environment and highlight the European dimension of their activity.

After ten years of existence, the programme became entire part of the statistical world. For the time being, around 500 participants coming from more than 30 countries are trained every academic year. Such an interest is mainly due to the large number of courses on offer. Indeed, the TES portfolio comprises more than 80 courses of short duration all at post-graduate level.

The TES Institute offers a coherent set of interrelated courses in the following domains:

- Official Statistics: General Issues (OSG)
- Official Economic Statistics (ECO)
- Official Social Statistics (SOC)
- Data Collection and Survey Methodology (DAT)
- Applied Statistical Analysis (ASA)
- Publication and Dissemination of Statistics (PDS)
- Management in a Statistical Institute (MSI)
- Statistical Information Systems (SIS)

To reach the widest audience possible, the TES Institute currently offers courses in English, French, German, Arabic and Russian.

After a few years of co-operation with the Central European countries, the TES Institute has recently extended the co-operation to the MEDSTAT and TACIS region. Such an internationalisation is the direct result of the growing importance of training as a part of the current technological and intellectual development. Therefore, as far as statistics and economics are concerned, it is of the utmost importance to extend the best national practices to an international level.

It is obvious that the TES programmes should be considered as a complement and not a substitute to the training provided at national level.

In brief, one may say that by offering training opportunities which are complementary to the ones provided at national level, the TES Institute is offering a new approach of the subsidiarity concept.

TRAINING

The Core Programme 2001 started in January will end in November 2001. The overview of the remaining courses can be found at the end of the present paper.

The TES Institute is currently preparing the Core Programme 2002 that will start in January 2002 and offer 28 courses. The final programme will be available by September 2001.

Beside the execution of the subsidised annual vocational training programme –which is only one of our many activities– the TES Institute also offers Special Courses that may be repetition of courses in the Core Programme or tailor-made and organised at the request of a country or a group of countries. In this context the TES Institute is active through the PHARE programme for Central European Countries, through the MEDSTAT programme in the countries of the Mediterranean Basin and through the TACIS programme in a number of CIS countries.

CONSULTING

In addition to the vocational training activities, we will continue to develop and extend our consulting activities in the above mentioned regions and elsewhere in order to maintain and solidify our position in the market of professional training and staff development for statisticians. These consulting activities consist in either *competence building* by the training of «in-country» trainers with multiplicative effects of the training or *institutional building* by the development of training centres, improvement of vocational training system and curricula. For the time being, the TES Institute has been involved in consulting activities with the Russian Federation, Ukraine and Kazakhstan.

RESEARCH

In various Research Institutes, Universities and National Statistical Institutes of the Member States of the European Union and other countries, research is taking place covering statistical methodology and statistical techniques such as data collection methods, data transmission, storage and warehousing, statistical data analysis, etc. TES is confident that, being closely involved in dissemination of the results of activities in the field of research and technological development, will not only help the Commission in its future policy on R&D but will also contribute to a better communication between the various players in statistical research. As a result, the TES Institute has decided to play an active role in European research projects in various fields of statistics.

Within this framework, the TES Institute is participating in a consortium, in the 5th Framework Programme of the Commission, focusing on the creation and the development of on-line training courses (VL-CATS Project). On the basis of possibilities offered by existing tools, the VL-CATS project will provide access over the Internet, to teaching and educational material with a special focus on Official Statistics. Through this project, experts will be given the opportunity to share and disseminate their statistical knowledge on a wider scale. Along with EU and EFTA countries, statisticians from accessing countries and other Central European countries will also be given the opportunity to produce statistics compliant with the European Statistical System standards.

The main objectives of VL-CATS Project are:

- To create a virtual library of all reference material available in Official Statistics;
- To define and implement standards for the development of distance courses in the various fields of Official Statistics;
- To develop a controlled environment that will support the virtual library and give selective access to training courses;
- To develop a service for update and quality assurance of VL-CATS.

PUBLICATIONS

The TES Institute is regularly publishing articles on its current activities in periodic Newsletter of several Statistical Institutes and some statistical journals as Qüestiió (Quaderns d'Estadística i Investigació Operativa), edited by the Institut d'Estadística de Catalunya.

The TES Institute has started the production of TES Manuals on subjects covered by the vocational training programme:

- The first manual available at the TES institute presents *The role of statistics in a democracy.*
- The second manual to be published soon will cover the *Index numbers for spatial and temporary comparisons.*
- Further manuals will cover topics of *sampling techniques, seasonal adjustment methods.*

TES NEWSLETTER

Beginning of February, the TES Institute has disseminated the first issue of its Newsletter named «Facts & Visions». This newsletter should be considered as an important

information tool between the TES Institute and its partners from all over Europe focusing on matters related to training and staff development in the statistical world.

We hope that our partners will use «Facts & Visions» to report on their own activities. So, may we encourage you to use your keyboards for European information purposes.

Copies of «Facts & Visions» are available at the TES Institute as well as guidelines for the submission of articles.

GENERAL INFORMATION

For further details on any of the above mentioned topics, please contact directly Ms. Valérie Vandewalle (co-ordinator for *Evaluation and Information*):

Phone: (352) 29.85.85.34
Fax: (352) 29.85.29
E-mail: vvandewalle@tes-institute.lu

Code	Course Title	Days	Course Leader	Location	From	To
DAT-202/2001	Use of Auxiliary Information in Sampling Surveys	4	Olivier Sautory	Paris	20-feb-01	23-feb-01
ASA-201/2001	Seasonal Adjustment Methods	5	Agustin Maravall	Luxembourg	26-feb-01	2-mar-01
PDS-001/2001	Basic Principles of Publication and Dissemination of Statistical Products	5	Ed Swires Hennessy	London	5-mar-01	9-mar-01
OSG-001/2001	The European Statistical System	3	Eurostat Experts	Luxembourg	19-mar-01	21-mar-01
SIS-203/2001	Geographical Information Systems	4	Marco Painho	Lisbon	26-mar-01	29-mar-01
MSI-101/2001	Adding Value through Strategic Management	5	John Scrivener	London	2-abr-01	6-abr-01
ECO-150/2001	Business Cycle Statistics	n 3	Klaus Reeh	Luxembourg	17-sep-01	19-sep-01
ECO-101/2001	Nomenclatures, Classifications and their Harmonisation	4	Niels Langkjaer	Vienne	23-abr-01	26-abr-01
SIS-201/2001	Statistical Disclosure Control	4	Peter Paul de Wolf	Voorburg	24-sep-01	27-sep-01
ECO-204/2001	The European System of Accounts (ESA95) - Goods and Services	3	Paul Konijn	Luxembourg	2-may-01	4-may-01
ECO-151/2001	Measuring the Economic Activity on the Internet	n 3	Pierre Dybman	Luxembourg	7-may-01	9-may-01
PDS-101/2001	Towards User-friendly Statistical Reporting	n 3	John Wright	London	9-may-01	11-may-01
ECO-203/2001	The European System of Accounts (ESA95) - Financial Accounts	3	Christine Coin	Luxembourg	21-may-01	23-may-01
PDS-105/2001	Marketing and Sales of Statistical Products and Services	3	Klaus Haagensen	Copenhagen	28-may-01	30-may-01
DAT-002/2001	Sampling Techniques and Practice	10	Prof. T.M.F. Smith	Southampton	18-jun-01	29-jun-01
ECO-001/2001	National Account Statistics in Practice (French)	10	François Lequiller	Paris	18-jun-01	29-jun-01
MSI-150/2001	Quality Management in Statistics	n 3	Lilli Japec Werner Grünewald	To be announced	25-jun-01	27-jun-01
DAT-105/2001	The Use of Administrative Sources for Statistical Purposes	n 3	Steve Vale	Helsinki	3-sep-01	5-sep-01
SOC-102/2001	Labour Cost and Labour Price Statistics	3	Steve Clarke	Luxembourg	12-sep-01	14-sep-01
DAT-207/2001	Advanced Sampling Techniques	n 3	Yves Tillé	Neuchatel	24-sep-01	27-sep-01
SIS-201/2001	Statistical Disclosure Control	4	Peter-Paul de Wolf	Voorburg	24-sep-01	27-sep-01
ECO-105/2001	Theory and Application of Enterprise Panel Surveys	5	Pierre Lavallée	Madrid	1-oct-01	5-oct-01
ECO-201/2001	Environmental Expenditure Statistics and Accounts (SERIEE)	n 4	Anton Steurer	Luxembourg	1-oct-01	4-oct-01
ASA-101/2001	Introduction to Applied Time Series Analysis	5	Agustin Maravall	Brussels	8-oct-01	12-oct-01
OSG-003/2001	Confidentiality and Protection of Privacy	3	Photis Nanopoulos	Luxembourg	15-oct-01	17-oct-01
ECO-103/2001	Enterprise Statistics	5	Johan Lock	Voorburg	5-nov-01	9-nov-01
SOC-001/2001	Systems of Social Statistics	5	Pieter Everaers	Heerlen	12-nov-01	16-nov-01

Institut Européen
d'Administration Publique –Centre Européen des Régions (IEAP-CER)
European Institute
of Public Administration –European Centre for the Regions (EIPA-ECR)

Generalitat de Catalunya
Institut d'Estadística de Catalunya

European Workshop

on

"Regional Data and Statistics in Europe: A Necessary Support in Facing this New Century"

Barcelona (E), 8-9 October 2001

Jointly organized by

the ***European Centre for the Regions (EIPA-ECR)*** and
the ***Institut d'Estadística de Catalunya (Idescat)*** **in Barcelona (E)**

With the support of

Eurostat - European Commission

INTRODUCTION

Information is the most important resource. To be meaningful it has to be organised. Some of the main actors in the organisation of data and information for territorial entities and governmental bodies in Europe are the institutes/departments of statistics, at supra-national, national as well as regional level. Especially, the regional centres/units of statistics in the process of the production, provision as well as distribution of information, due to the on-going changes at national and European Union level, has significantly been increased in recent years.

The *European Centre for the Regions (EIPA-ECR)*, the European Institute of Public Administration's Antenna in Barcelona (E), together with the *Institut d'Estadística de Catalunya (Idescat)* has therefore designed an activity in which representatives of these institutions as well as public officials, academics and policy-makers of subnational authorities will have the opportunity to examine the development, the new challenges and necessity of regional data in order to support both the statistical European system and, in general, the European economy in the 21st century.

Presentations, discussions as well as 'best practises' on these issues will give participants the chance to exchange views and experiences with experts from other European regions and organisations. The two-day workshop, which will be conducted in English and French (with simultanous interpretation) will be highly interactive and involve full participation.

ORGANISATIONAL ASPECTS:

Design and Organisation:	Idescat and EIPA-ECR Barcelona
Coordinating Responsibles:	
Mr Alexander Heichlinger Lecturer & Project Leader EIPA-ECR Barcelona Calle Girona, 20 E-08010 Barcelona E-mail: a.heichlinger@eipa-ecr.com Tel: +34-93-567 2404	Mr Enric Ripoll Subdirector of Technical Statistical Assistance Idescat Via Laietana, 58 E-08003 Barcelona E-mail: eripoll@idescat.es Tel: +34-93-412 0924
Registration:	Mrs Miriam Escolà Programme Organisation EIPA-ECR Barcelona E-mail: m.escola@eipa-ecr.com Tel: +34 93 567 2400 Fax: +34 93 567 2399
Languages:	English & French (with interpretation)
Location:	Auditorium, Escola d'Administració Publica de Catalunya (EAPC)

PROGRAMME

Monday, 8 October 2001

09.00 **Arrival and Registration of Participants**

09.30 **Inauguration of Workshop and Welcome of Participants**
Eduard Sánchez Monjo, Director, EIPA-ECR, Barcelona (E);
Coordinator of Regional Cooperation, Maastricht (NL)
Jordi Oliveres, Director of Idescat, Barcelona (E)
Francesc Homs, Minister of Economy and Finance;
President of Idescat, Generalitat de Catalunya, Barcelona*

I. Structural, Institutional and Organisational Aspects of Regional Institutes of Statistics in Europe

10.00 **A) Harmonization versus Decentralisation versus Cooperation:**
Indicators such as inflation rate, GDP, governments public finance, the unemployment quote etc. are essential data nessecary to understand regional economic challenges and devise new policies to face them. The significance of a direct interlocution and a well established coordination between Eurostat and their national as well as regional counterparts will be examined in this context.
Berthold Feldmann, Head of Section 'Regional Accounts and Indicators', Eurostat, Luxembourg (L)
Lourdes Llorens, Director of EUSTAT-Statistical Office of the Basque Country, Vitoria (E)

11.30 Coffee break

12.00 **B) The Legal Conceptual Framework:**
Different legal and institutional remits provide the institutes different starting position in the planification of statistical acticivities for the respective regional public administration. A comparative overview of the 'official' status of the institutes as well as their portfolio of competences and functions will be analysed in depth.
Jordi Oliveres, Director of Idescat, Barcelona (E)
Hans Loreth, Vice-President of the State Office for Statistics and Data Processing of Baden-Württemberg, Stuttgart (D)

13.30 Joint Lunch

* to be confirmed

15.30 **C) Availability and Distribution of Data from Europe's Regions and Municipalities:**
How can both the public and private sector best benefit from the information (e.g. demographic, infrastructural etc.) generated by the regional and local statistical institutes? How can this information be accessed and distributed? And what instruments might be best used to successfully promote and make known the statistical information on Europe's regions and municipalities?
Pierre Joly, Regional Director, INSEE, Languedoc-Roussillon, Montpelier (F)
Jean Houard, Director of Research, Studies and Statistics Service, Ministry of the Wallonia Region, Jambes (B)

II. The European Statistical System

17.00 **Dissemination of European Regional Data: the Data Shop Network**
– Organisation and functioning
– Databases and diffusion of statistics
– Publication series and electronic sites
Anna M. Martínez, Responsible of Support to Data Shop Network-Unit 'Information and Dissemination', Eurostat, Luxembourg

18.30 End of first day

20.30 Joint Dinner at a Restaurant in Barcelona

Tuesday, 9 October 2001

III. Panel Discussions on 'Best Practices'

Chaired by: Enric Ripoll, Subdirector of Technical Statistical Assistance, Idescat

09.30 **A) Demographic & Social Statistics of Regions: Migration Statistics**
Experiences and challenges in the application of instruments and methodologies are discussed on issues about the statistics on migration and other related statistical sources at regional level (i.e. mobility, population census, etc.), as well as their impacts on the regional and local public authorities.
Joaquín Arango, Fundación Ortega y Gasset, Madrid (E)
Philippe Wanner, Swiss Forum for Migration Studies, Neuchatel (CH)

11.00 Coffee break

Chaired by: Alexander Heichlinger, Lecturer & Project Leader, EIPA-ECR

11.30	**B) Economic Statistics of Regions: Short Term Economic Data** The development and mapping of economic cycles and changes will be examined from the regional statistics of short term economic data point of view which seems to be more and more necessary to support both the regions and the European Union area. *Jordi Galter*, Department of Economic Statistics Production, Idescat, Barcelona (E) *Arend Steenken*, Director, State Office for Statistics and Data Processing of Brandenburg, Potsdam (D)
13.00	**Debate**
13.30	**Closure of Workshop** *Josep Maria Guinart*, Director, School of Public Administration of Catalonia (EAPC), Barcelona *Eduard Sánchez Monjo*, Director, EIPA-ECR, Barcelona; Coordinator of Regional Cooperation, Maastricht *Jordi Oliveres*, Director of Idescat, Barcelona
14.00	Joint Lunch
After lunch (optional)	**Visiting Tour of the City of Barcelona**

*****

Informació per als autors i lectors

NORMES PER A LA PRESENTACIÓ D'ARTICLES A *QÜESTIIÓ*

La revista accepta, per a la seva publicació, articles originals no sotmesos en cap altra revista dins els àmbits de l'Estadística, la Investigació Operativa, l'Estadística Oficial i la Biometria. Els articles poden ser teòrics o aplicats, incloent aspectes computacionals i/o de caire docent, i poden presentar-se en anglès, francès, català o qualsevol altre llengua oficial a l'Estat espanyol.

Tots els originals destinats a les esmentades seccions temàtiques de *Qüestiió*, així com els de la «Secció docent i problemes», seran sotmesos sistemàticament a una avaluació prèvia a càrrec d'especialistes independents i/o membres del Consell Editor, llevat d'aquells que siguin invitats per la revista i les reimpressions d'articles. El resultat de l'avaluació serà comunicat a l'autor principal als efectes d'eventuals correccions formals o dels seus continguts.

Per a totes les trameses d'originals, la revista emetrà un acusament de recepció la data del qual figurarà com a «data de rebuda» en la publicació de l'article. Per la seva banda, la «data d'acceptació» de l'article serà la data de recepció de la versió definitiva.

Per a la presentació d'articles l'autor trametrà a l'adreça postal de la Secretaria de *Qüestiió* (Institut d'Estadística de Catalunya) dues còpies impreses del treball complet en DIN A4, a una sola cara i a doble espai i, paral·lelament, per correu electrònic (questiio@idescat.es) la primera pàgina de l'article en format PDF o PS que contindrà el títol, el nom de l'autor o autors, l'afiliació i l'adreça completa, així com un resum de 75-100 paraules seguit de les principals paraules clau (en l'idioma original) i la seva adscripció a la classificació MSC2000 de la American Mathematical Society. Abans de sotmetre els articles a la revista, s'aconsella els autors que revisin la correcció lingüística dels textos d'acord amb l'idioma original i les eventuals traduccions a l'anglès.

Les referències bibliogràfiques es faran indicant el cognom de l'autor seguit de l'any de la publicació entre parèntesi [i.e.: Mahalanobis (1936), Rao (1982b)] i seran llistades alfabèticament al final de l'article; les referències múltiples d'un mateix autor s'ordenaran cronològicament. Les notes explicatives es numeraran correlativament i han d'aparèixer al peu de la pàgina corresponent. Les taules i figures també es numeraran correlativament en el text i seran reproduïdes directament dels originals tramesos en cas que no sigui possible la seva autoedició.

Una vegada avaluat satisfactòriament l'article caldrà que l'autor el trameti en versió impresa i suport electrònic, segons les instruccions que li seran indicades pel responsable de l'avaluació del seu treball. Es recomana que aquesta versió final es trameti preferiblement en el processador de textos LaTeX 2ε [subsidiàriament, es poden trametre els textos i les taules en Word Perfect —versió 6.0A o anterior— o ASCII]; en el cas de figures, diagrames o gràfics es recomanen els formats adients per als programes-editors PS, EPS o PCX. Els autors han de garantir la correspondència exacta entre la versió impresa i la còpia electrònica. D'altra banda, si l'article no està escrit en llengua anglesa s'haurà d'adjuntar la traducció del títol original, de l'abstract i de les paraules clau, així com un ampli resum en anglès (amb una extensió d'entre 2 i 5 pàgines i amb la mateixa estructura de l'article original).

La Secretaria de *Qüestiió* posa a disposició dels autors que ho sol·licitin plantilles en format LaTeX 2ε per a la seva edició i les referències adequades de la classificació de l'AMS.

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE SUBMISSION OF ARTICLES FOR *QÜESTIIÓ*

The journal welcomes submission of articles and contributions that are not being considered for publication in any other journal in the fields of Statistics, Operational Research, Official Statistics or Biometrics. Articles may be theoretical or applied, including teaching aspects and applications, and will be accepted in English, French, Catalan or any of the other official languages in Spain.

All originals assigned to the thematic sections of Qüestiió, including articles for the «Teaching section and problems» will be systematically reviewed by independent referees and/or members of the Editorial Board, who will send a report to the main author of the article in order to correct, if necessary, any formal or content aspects. The articles invited by the journal and articles reprinted will be excluded from this evaluation process.

For all submissions, the journal will issue a receipt corresponding to the submission date, which will appear as «date received» in the final publication of the article. The «acceptance date» of the article, which will appear in its final publication, will be the date of sending the final version to the journal.

For the presentation of original articles, the author should send, to the postal address of the Secretary of Qüestiió (Institut d'Estadística de Catalunya), two copies of the complete article, typed on A4 sheets, one side of the paper only, double spaced and with wide margins. At the same time, the author should send by electronic mail (questiio@idescat.es) the first page of the article on PDF or PS format including the title, the name of the author or authors, their affiliation, full address as well as an abstract of the paper (75-100 words) followed by the keywords (in the original language) and its assignation to the AMS classification. Before submitting the papers to the magazine, authors are advised to revise their linguistic correctness, according to their original language and their possible translation into English.

Bibliographical references should state the author's name followed by the year of publication in brackets [e.g.: Mahalanobis (1936), Rao (1982b)] and they should be listed at the end of the article in alphabetical order; multiple references to the same author should be given in chronological order. Footnotes should be numbered in the article and appear at the foot of the corresponding page. Figures and tables are to be numbered in consecutive order in the text using Arabic numerals and will be directly reproduced from the originals submitted if it is not impossible to print them electronically.

Once the evaluation of the article had been successful, the author is required to send it on a diskette (a 3.5-inch disk on MS-DOS format) together with its paper copy; it should be a brand new diskette and contain very clearly written the names of the authors and the title of the article. This final version should be processed by LaTeX 2_ε, preferably, or, failing that, by Word Perfect (6.0A or earlier) or ASCII for text and tables; for figures, diagrams or graphs, the appropriate formats of PS, EPS or PCX software tools are strongly recommended. Authors must ensure that the electronic copy version and the one on paper are exactly the same. Furthermore, if the article is not written in English, the translation of its original title, short abstract and keywords should be enclosed, as well as a full summary of the article in English (that is, 2-5 pages with the same structure as the original one).

The Secretary of Qüestiió can send, by request of the authors, the LaTeX 2_ε for manuscript preparation and the appropriate AMS classification references.

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS INSTITUCIONALS A *QÜESTIIÓ*

Qüestiió convida les entitats patrocinadores, les institucions col·laboradores, els organismes públics i privats, i tota la comunitat científica vinculada a l'estadística o la investigació operativa, a la publicació d'anuncis institucionals sobre cursos, seminaris, congressos i activitats similars que, preferentment, tinguin lloc en el nostre país. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les entitats interessades, de manera que Qüestiió no fa una cerca sistemàtica d'esdeveniments d'aquesta naturalesa, ni té cap ànim d'exhaustivitat en les ressenyes d'activitats finalment publicades.

Una vegada aprovada la inclusió dels anuncis sol·licitats es procedirà a la seva publicació, i es reproduirà directament dels originals tramesos amb les mides adequades i la màxima qualitat tipogràfica possible; en aquest cas, Qüestiió no procedeix a cap mena de procés d'autoedició de la versió impresa que l'anunciant hagi tramès. Si els originals es trameten en els mateixos termes electrònics exigits per als articles (vegeu «Normes per a la presentació d'articles a Qüestiió»), la revista procedirà a la seva autoedició. Si es desitja una qualitat superior a la reproducció simple o l'autoedició, o bé la seva publicació en color, els sol·licitants hauran de posar-se en contacte amb la Secretaria de Qüestiió per tal de trametre els fotolits dels textos originals corresponents.

La disposició dels textos i les figures adjuntes dels anuncis han de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de Qüestiió es reserva la decisió final pel que fa a la seva publicació.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la inserció en els números/volums de Qüestiió que prèviament s'hagi establert. La revista no es fa responsable dels retards, per part de l'anunciant, que impedeixin la publicació de l'anunci en els termes previstos.

Mònica M. Jaime
Secretaria de Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR INSTITUTIONAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió invites all sponsor entities, collaborating institutions, other public and private bodies and the entire scientific community related to Statistics or Operations Research to submit institutional advertisements on courses, seminars, congress and similar activities that will be held, preferably in our country. These will be accepted in English, French, Catalan or any of the other official languages in Spain. The initiative should always come from the entities interested in advertising them so that Qüestiió's aim is not to do a systematic search of these events and therefore does not publish a comprehensive list of such activities.

Once their insertion is approved the advertisements will be reproduced from the most accurate photocopy of the originals sent by the advertiser to Qüestiió in paper copy, with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editing process to the printed version that the advertiser has sent. If the original advertisements are sent in the same electronic format requested by the articles (please see «Guidelines for the submission of articles for Qüestiió») the journal will print it directly from the file. If a better quality than the simple reproduction or automatic printing or a colour version of the adverts is desired, the authors should contact the Secretary of Qüestiió in order to negotiate this.

The typesetting of texts and figures in the advertisement should have maximum intelligibility and clearness, neither compressing the information too much nor using formats or letter fonts that are too small. Furthermore, the information has to be reliable, without errors and respectful of the people and institutions. The management of Qüestiió has the right to a final decision concerning the insertion of the advertisement.

Advertisers commit themselves to give the text/materials on request in order to insert them in the issues of Qüestiió that have been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published on the agreed terms.

Mònica M. Jaime
Secretaria de Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS PRIVATS O AMB FINALITAT COMERCIAL A *QÜESTIIÓ*

Qüestiió accepta la publicació d'anuncis privats o amb finalitat comercial sobre productes, serveis o altres eines promocionals a l'entorn de l'estadística o la investigació operativa. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les organitzacions que hi estiguin interessades, de manera que Qüestiió no fa una cerca sistemàtica de novetats o productes d'aquesta naturalesa ni té cap ànim d'exhaustivitat en els anuncis finalment publicats.

Els anuncis en **blanc i negre** s'elaboren a partir de la fotocòpia més acurada possible dels originals que trameti l'anunciant en versió impresa, amb les mides adequades i la màxima qualitat tipogràfica. Per tant, en aquest cas la revista no efectua cap procés d'edició ulterior respecte de la versió impresa que l'anunciant hagi tramès. Alternativament, si els anuncis originals es trameten en els mateixos termes formals exigits per als articles (vegeu «Normes per a la presentació d'articles a Qüestiió»), la revista procedirà a la seva autoedició. Igualment, si es desitja una qualitat superior a la reproducció simple, els sol·licitants hauran de trametre els fotolits dels originals corresponents o encarregar-los a Qüestiió, que els facturarà separadament.

Els anuncis **en color** requereixen els fotolits dels textos originals, que poden ser subministrats directament per l'anunciant o bé encarregats per la revista a compte de l'anunciant; en el segon cas, l'anunciant ha de trametre a la revista els originals impresos en color amb la màxima qualitat, per tal de filmar-los amb les millors garanties i condicions. El cost dels fotolits realitzats per Qüestiió serà sempre a càrrec de l'anunciant, a qui se li repercutirà l'import i l'IVA d'aquests, juntament amb les tarifes que corresponen a la modalitat d'anunci per la qual hagi optat.

La disposició dels textos i figures adjuntes dels anuncis ha de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de Qüestiió es reserva la decisió final de la seva inclusió.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la seva inserció en el(s) número(s)/volum(s) de Qüestiió que prèviament s'hagi establert. La revista no es fa responsable dels retards per part de l'anunciant que impedeixin la publicació de l'anunci en els termes previstos.

Imports:

1 pàgina en color (un número aïllat):	125.000 PTA + IVA
1 pàgina en color (tres números consecutius):	200.000 PTA + IVA
1 pàgina en blanc i negre (un número aïllat):	30.000 PTA + IVA
1 pàgina en blanc i negre (tres números consecutius):	50.000 PTA + IVA
1/2 pàgina en blanc i negre (un número aïllat):	20.000 PTA + IVA
1/2 pàgina en blanc i negre (tres números consecutius):	35.000 PTA + IVA

Mides opcionals dels anuncis:

1 pàgina sencera (espai intern):	19.0 cm. × 12.3 cm.
1 pàgina sencera (espai extern):	23.8 cm. × 17.0 cm.
1/2 pàgina (espai intern):	9.5 cm. × 12.3 cm.
1/2 pàgina (espai extern):	11.9 cm. × 17.0 cm.

Forma de Pagament:

- Transferència bancària al compte: 2013-0100-53-0200698577
- Xec bancari nominatiu a l'Institut d'Estadística de Catalunya
- Pagament amb targeta de crèdit

El pagament serà per l'import total de la factura corresponent, on hi figurarà el cost dels fotolits en el cas que l'edició de l'anunci hagi estat a càrrec de l'Institut. En el cas que l'anunciant necessiti una factura proforma, només cal que ho faci saber amb l'antelació suficient.

Correspondència:

Mònica M. Jaime
Secretaria de Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE PRIVATE OR COMMERCIAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió accepts for their publication both private and commercial advertisements on products, services or other promotional tools related to statistics or operational research and will be accepted in English, French, Catalan or any of the official languages in Spain. The initiatives should always come from entities interested in advertising them so that Qüestiió's aim is not to do a systematic search of news and therefore does not publish a comprehensive list of such private or profit activities.

The **black and white** advertisements are made out from the most accurate photocopy of the originals sent by the advertiser to Qüestiió in paper copy with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editorial process to the printed version that the advertiser has sent. Alternatively, if the original advertisements are sent in the same formal terms required by the articles (please see «Guidelines for the submission of articles for Qüestiió»), the journal will proceed to its autoedition. In the same way, if a better quality than the simple reproduction is wanted, the authors should send the photolits of the corresponding original texts or, on the other hand, order to Qüestiió their fulfilment, which will be invoiced separately from the rates charged as advertisements.

The advertisements **in colour** need the photolits of the original texts, which can be provided directly by the advertiser or requested by Qüestiió to the advertiser charge; in the second case, the advertiser must sent to the journal the originals printed in colour with the best possible quality, so that they can be filmed at the best conditions and guarantees. The cost of the photolits made by Qüestiió will always be charged to the advertiser together with the VAT derived from it, plus the prices corresponding to the type of the advertisement that has been chosen.

The set up of texts and figures of the advertisement should provide the maximum intelligibility and clearness, neither squeezing together the information nor using set ups or letter types that are too small. On the other hand the publicity has to be reliable, without fraud and respectful to the persons and institutions. The direction of Qüestiió has the right of the last decision concerning the insertion of the advertisement.

The advertiser committs himself to give the texts/materials on request, in order to insert them in the issue(s) of Qüestiió that had been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published in the agreed terms.

Rates:

1 colour page (only one issue):	125.000 PTA + VAT
1 colour page (three consecutive issues):	200.000 PTA + VAT
1 black and white page (only one issue):	30.000 PTA + VAT
1 black and white page (three consecutive issues):	50.000 PTA + VAT
1/2 black and white page (only one issue):	20.000 PTA + VAT
1/2 black and white page (three consecutive issues):	35.000 PTA + VAT

Advertisement sizes (optional):

1 full page (internal space):	19.0 cm. × 12.3 cm.
1 full page (external space):	23.8 cm. × 17.0 cm.
1/2 page (internal space):	9.5 cm. × 12.3 cm.
1/2 page (external space):	11.9 cm. × 17.0 cm.

Payment:

- A bank transfer to account number: 2013-0100-53-0200698577
- A bank cheque to Institut d'Estadística de Catalunya
- Charge on a credit card

The payment should be for the amount shown at the invoice, where it will be shouwn the total cost of the photolits, in case that Qüestiió would be in charge of the filmation of the advertisement. If advertiser need a pro-forma invoice, he should let us know some time in advance so that Qüestiió could send it to the proper address.

Mail address:

Mònica M. Jaime
Secretaria de Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

Butlleta de subscripció a la revista *Qüestiió*

Nom i cognoms _____

_____

Empresa/Institució _____

_____

Adreça _____

_____

Codi postal _____ Ciutat _____

Tel. _____ Fax _____ NIF _____

Data _____

Signatura

Desitjo subscriure'm a ***Qüestiió*** per a l'any 2001

Preu de subscripció vigent:

– Estat espanyol: 3.600 Pta/any (21,64 €) (IVA inclòs)
– Estranger: 4.000 Pta/any (24,04 €) (IVA inclòs)

Forma de pagament

☐ Transferència al compte 2013-0100-53-0200698577

☐ Domiciliació bancària al compte número

☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

☐ Xec nominatiu a l'Institut d'Estadística de Catalunya

☐ Gir postal

☐ En efectiu

Retorneu aquesta butlleta (o una fotocòpia) a:

Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Preu de números solts (actuals i endarrerits):

– Estat espanyol: 1.500 Pta/exemplar (9,02 €) (IVA inclòs)
– Estranger: 1.700 Pta/exemplar (10,22 €) (IVA inclòs)

Exemplar per a l'entitat bancària

Autorització de domiciliació bancària per al pagament de les subscripcions anuals de la revista ***Qüestiió***

El sotasignat _____

autoritza el Banc/Caixa _____

Adreça _____

Codi postal _____ Ciutat _____

a abonar les subscripcions a la revista ***Qüestiió*** amb càrrec al seu compte

número ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Data _____

Signatura

Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Novetats editorials en matèria estadística de la Generalitat de Catalunya
gener-agost 2001

- **Anuari Estadístic de Catalunya, 1992-2000 CD-ROM**
  gener 2001, 3.000 PTA (18,03 €)
  ISBN 84-393-5332-4

- **Xifres de Catalunya 2001**
  (quadríptic divulgatiu en català, castellà, francés, anglès i alemany)

- **Estadística de població 1996 Vol. 6 Lloc de naixement i nacionalitat de la població. Dades comarcals i municipals**
  octubre 2000, 1.500 PTA (9,02 €) 274 pp.,
  ISBN 84-393-5264-6

- **Estadística de població 1996. Vol. 11 Professions de la població. Dades comarcals i municipals**
  febrer 2001, 1.250 PTA (7,51 €) 173 pp.,
  ISBN 84-393-5360-X

- **Estadística de població 1996. Vol. 14 Estructures familiars de la població. Dades comarcals i municipals**
  febrer 2001, 1.500 PTA (9,02 €) 247 pp.,
  ISBN 84-393-5361-8

- **Moviments migratoris 1998 Dades comarcals i municipals**
  gener 2001, 1.600 PTA (9,62 €) 184 pp.,
  ISBN 84-393-5341-3

- **Planificació i coordinació de l'estadística catalana**
  desembre 2000, 1.900 PTA (11,42 €) 350 pp.,
  ISBN 84-393-5201-8

- **Estadística de biblioteques 1998 Característiques bàsiques**
  febrer 2001, 1.250 PTA (7,51 €) 131 pp.,
  ISBN 84-393-5359-6

- **Localització de l'activitat econòmica 1998 Empreses, professionals, establiments i superfícies**
  febrer 2001, 1.700 PTA (10,22 €) 365 pp.,
  ISBN 84-393-5364-2

- **Anuari Estadístic del Departament de Política Territorial i Obres Públiques 1998 (CD-ROM)**
  Departament de Política Territorial i Obres Públiques
  Secretaria General
  2000, 2.000 PTA (9,02 €) 246 pp.,
  ISBN 84-393-5332-4

- **Mercat de treball 2000 Ampliació de resultats anuals de l'enquesta de població activa**
  juny 2001, 1.700 PTA (10,22 €)
  ISBN 84-393-5448-7

- **Moviments migratoris 1999 Dades comarcals i municipals**
  juny 2001, 1.600 PTA (9,62 €)
  ISBN 84-393-5447-9

- **Estadística de població 1996 Vol. 12 Fluxos de la mobilitat obligada per treball i estudi. Dades comarcals i municipals**
  maig 2001, 2.250 PTA (13,52 €)
  ISBN 84-393-5423-1

- **Estadística de població 1996 Vol. 13 Localització de l'ocupació laboral. Dades comarcals i municipals**
  maig 2001, 1.250 PTA (7,51 €)
  ISBN 84-393-5424-X

- **Comptes de les administracions públiques de Catalunya 1997**
  maig 2001, 1.300 PTA (7,81 €)
  ISBN 84-393-5414-2

LLIBRERIES DE LA GENERALITAT

Barcelona
Rambla dels Estudis, 118 (tel. 93 302 64 62)
llibrbcn@correu.cattel.com

Girona
Gran Via de Jaume I, 38 (tel. 972 22 72 67)
llibrgi@ibernet.com

Lleida
Rambla d'Aragó, 43 (tel. 973 28 19 30)
llibrlle@ibernet.com

Madrid
Blanquerna. Llibreria catalana.
Serrano, 1 (tel. 91 431 00 22)
blanquerna@nauta.es

PUNT DE VENDA
Puigcerdà
Plaça del Rec, 5 (tel. 972 88 05 14)

VENDA PER CORREU
Apartat 2800, 08080 Barcelona
eadop@correu.gencat.es
Plaça del Rec, 5 (tel. 972 88 05 14)

Publicacions de la Generalitat. Apartat de correus 2800, 08080 Barcelona

Nom i cognoms _____

Empresa / Institució _____

Professió _____ E-mail _____

Adreça _____

Població _____ CP _____

NIF / DNI _____ Telèfon _____

Desitjo rebre els volums

_____

_____

○ Carregueu l'import a la meva targeta de crèdit Signatura

☐ American Express ☐ 6000

☐ Master Charge ☐ Visa

○ Contra reemborsament

Núm. de targeta |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|

Data de caducitat |_|_|_|_|

DOGC