Caja 324

N° 25

MINISTERIO DE ECONOMIA
SECRETARIA DE COORDINACION Y PLANIFICACION ECONOMICA
INSTITUTO NACIONAL DE PLANIFICACION ECONOMICA

BIBLIOTECA
Hipólito Yrigoyen 250, piso 8º, of. 801/6.
Buenos Aires (Argentina)

SERIE DE INVESTIGACIONES N° 12

Dres. FERNANDO FERRERO Y CARLOS E. SANCHEZ

MODELO LINEAL PARA LA IDENTIFICACION DE INTERACCIONES

Presentado: Reunión de Centros de Investigaciones Económicas, La Plata, 1969

UNIVERSIDAD NACIONAL DE CORDOBA
FACULTAD DE CIENCIAS ECONÓMICAS
INSTITUTO DE ECONOMIA Y FINANZAS
1971

MINISTERIO DE ECONOMIA
SECRETARIA DE COORDINACION Y PLANIFICACION ECONOMICA
INSTITUTO NACIONAL DE PLANIFICACION ECONOMICA
BIBLIOTECA
Hipólito Yrigoyen 250, piso 8o, of. 801/C.
Buenos Aires (Argentina)

SERIE DE INVESTIGACIONES Nº 12

Base 06/98

Dres. FERNANDO FERRERO Y CARLOS E. SANCHEZ

MODELO LINEAL PARA LA IDENTIFICACION DE INTERACCIONES

UNIVERSIDAD NACIONAL DE CORDOBA
FACULTAD DE CIENCIAS ECONÓMICAS
INSTITUTO DE ECONOMIA Y FINANZAS
1971

MODELO LINEAL PARA LA IDENTIFICACION DE INTERACCIONES *

INTRODUCCION

El presente trabajo tiene por objeto desarrollar un método para identificar y estimar interacciones en un modelo lineal donde todas las variables son de naturaleza dicotómica. Lejos de ser ésta una simple cuestión de transformación de variables, en el sentido de que variables originariamente cuantitativas se transforman en cualitativas, el problema central que aquí se plantea es lograr un conjunto de variables en el que cada una de ellas controla todos los niveles posibles de las restantes, facilitando de esta manera la estimación del efecto puro o aislado que a cada una de ellas le corresponderá. Tales efectos, por el método de construcción adoptado, no dependerán en modo alguno de los valores o niveles que asuman las restantes variables.

Las variables que en este modelo se utilizan no son otra cosa que las que resultan de la conjunción de los diversos niveles de todas las variables y en la medida que la estimación puntual y la docimasia de hipótesis revelen la presencia de coeficientes significativos, tales combinaciones estarán denunciando un comportamiento diferencial por parte de las observaciones o individuos que las componen. Cuando una categoría cualquiera tiene precisamente estas características se la denomina una interacción.

* Serie de Investigaciones del Instituto de Economía y Finanzas Nº 12.

Conviene aclarar que el propósito del trabajo es esencialmente metodológico y aun cuando se ha tratado de probarlo empíricamente para investigar la posesión de un bien durable, en este caso automóviles, no se pretende con ello elaborar un marco teórico o conceptual de tal fenómeno.

ESTRUCTURA DEL MODELO

En el modelo lineal que analizaremos, que no es de regresión ni de varianza ni de covarianza, todas las variables intervinientes tanto exógenas como endógenas son de naturaleza dicotómica. Una determinada variable asume el valor uno o cero según que el individuo u observación pertenezca a una cierta categoría, o posea o no un determinado atributo. En este caso particular relacionamos la posesión de un cierto bien durable —automóviles— con un conjunto de variables, tales como ingreso familiar, edad del jefe, posición ocupacional, educación, etc., con lo que el modelo por ser lineal adoptará la forma

$$Y = XB + u, \quad \text{(sin formular por el momento supuesto alguno acerca de la variable aleatoria u)}$$

y en términos de matrices y vectores, la siguiente composición:

Posee automóvil	Ingresos			Edad		Ocupación del jefe			
	a_0-a_1	a_2-a_3	a_4-a_5	b_1-b_2	b_3-b_4	Asalariado	Intermedio	Jefe alto nivel	
0	1	0	0	1	0	0	1	0	. . .
1	0	1	0	0	1	0	0	1	. . .
.	.	.	.	.	.	.	.	.	. . .
.	.	.	.	.	.	.	.	.	. . .

De este modo el individuo ubicado en la primera fila se caracteriza por pertenecer al primer tramo de ingresos, al primer inter-

valo de edad y al segundo de ocupación y no posee automóvil. El segundo sí posee automóvil, pertenece al segundo tramo de ingresos, al segundo de edad y al tercero de ocupación. Se tendrán de esta forma tantas filas cuantas observaciones existan y un número de columnas coincidente con el número de variables que se analiza.

En forma abreviada,

$$Y_i = \begin{cases} 1 & \text{si el i-ésimo individuo posee automóvil} \\ 0 & \text{en caso contrario} \end{cases}$$

$$X_{ij} = \begin{cases} 1 & \text{si el i-ésimo individuo pertenece a la j-ésima categoría} \\ 0 & \text{en caso contrario} \end{cases}$$

Dado un conjunto de observaciones del tipo señalado, ha de existir ciertamente alguna relación entre las variables independientes y la variable que convencionalmente designamos como dependiente. Es necesario encontrar, en otras palabras, algo que nos explique porqué una mayor proporción de los individuos que pertenecen a X_j poseen automóvil, en tanto que los que pertenecen a $X_{j'}$ muestran una frecuencia menor en orden a la posesión de automóviles.

Puesto que las variables que en realidad se utilizarán son distintas a las indicadas en la página anterior y para evitar la confusión entre variables y sub-variables, se usará la notación: X^{k,m_k}_{ij} para representar la siguiente categoría: el primer supra índice indicará una de las variables originales; ingreso, educación, edad, etc.; el segundo supra índice indicará el número de niveles en que tal variable ha sido dividida o también el número de variables dicotómicas que da origen; el primer sub-índice individualizará al i-ésimo individuo; y el segundo, el nivel específico al que el individuo pertenece. De esta forma, en el ejemplo anterior la primera columna se representará por $X^{1,3}_{.1}$; la segunda, $X^{1,3}_{.2}$; la cuarta, $X^{2,2}_{.1}$; la séptima, $X^{3,3}_{.2}$, etc. A su vez, un elemento cualquiera de la matriz.

$$X_{ij}^{k,m_k} = \begin{cases} 1 & \text{si el i-ésimo individuo pertenece al j-ésimo nivel de la k-ésima variable (ésta ha sido dividida en } m_k \text{ niveles)} \\ 0 & \text{en caso contrario} \end{cases}$$

De hecho que si $X_{ij}^{k,m_{k=1}}$ ello implicará que $X_{ij'}^{k,m_{k=}}O$ para todo $j \neq j'$. Esta particular característica hace que la matriz sea de rango incompleto.

Es necesario destacar que si bien las variables que aquí aparecen son de naturaleza dicotómica —lo que restringiría el uso del método sólo a variables no cuantitativas—, éstas en su forma original pueden ser de cualquier tipo. Si el nivel de medición original es típicamente el de una variable cuantitativa, como en el caso de Ingresos del ejemplo anterior, siempre será posible dividirla en intervalos y asignar a cada uno de ellos una variable dicotómica específica. De aquí que en lo sucesivo, al referirnos a variables, estaremos más bien implicando categorías o niveles de una cierta variable.

ANALISIS DE RELACIONES

A primera vista podría aparecer acertado plantear un modelo de regresión del tipo $Y = XB + u$ y estimar el vector B con las correspondientes dócimas de hipótesis asociadas a B o a un subconjunto de sus elementos. Los coeficientes así estimados darían una medida del efecto aislado de cada variable cuando las restantes permanecen constantes. Tal procedimiento, empero, no es aconsejable ya que, 1) la matriz X es de rango inferior a Σ_{mk} (número de variables incluidas), lo cual obligará a prescindir de algunas variables o estimar combinaciones lineales de parámetros; 2) en la propia construcción del modelo de regresión lineal se parte del supuesto de que las variables sólo tienen efectos aditivos y de allí que el coeficiente de regresión estimado sea una medida apropiada del efecto aislado de cada variable. Sin embargo, el método de regresión múltiple es primordialmente eficaz en orden a la predicción de la variable dependiente, pero si lo que se busca es la deter-

minación de los parámetros de una relación estructural, la regresión múltiple no es en tal circunstancia el método más apropiado [1]; 3) cuando se combinan un grupo de variables económicas y demográficas el supuesto de aditividad se hace más insostenible; no es fácil aceptar que el ingreso de una persona pueda ser disociado de su posición ocupacional o de su edad y más aún admitir que cuando estas variables actúan aisladamente sus efectos son semejantes a cuando lo hacen conjuntamente. Cuando se maneja un conjunto de variables que presentan marcadas interacciones, carece de sentido investigar el efecto de cada una de ellas consideradas aisladamente [2].

Todas estas circunstancias aconsejan utilizar un modelo donde las variables intervinientes resultan de una conjunción de los diversos niveles de todas las variables. De esta forma las nuevas variables así construidas representarán el efecto conjunto y simultáneo de cada nivel de las restantes variables y podrán en consecuencia indicar un efecto puro o aislado que en modo alguno dependerá de si las otras variables permanecen o no constantes.

A las variables primitivas, que por lo expuesto conllevan los efectos confundidos de las restantes, las denominaremos "genéricas" o "complejas"; a las nuevas variables, que representan efectos exclusivos y atribuibles a ellas mismas, las denominaremos "simples" o "puras".

Para pasar de las complejas a las puras definiremos la operación de producto lógico de categorías que se simbolizará por (x), de forma que las nuevas variables serán

$$Z_{.1} = X^{1.m_1}_{.1} \ (x) \ X^{2.m_2}_{.1} \ldots\ (x) \ X^{K.m_K}_{.1} \qquad \text{(siendo K el número de variables complejas)}$$

$$Z_{.2} = X^{1.m_1}_{.1} \ (x) \ X^{2.m_2}_{.1} \ldots\ (x) \ X^{K.m_K}_{.2}$$

$$\ldots \quad \ldots \quad \ldots$$

1. TITNER, G.: *Econometrics*, New York, John Wiley and Sons; 1952.
2. MORGAN, J.H. and SONQUIST, J.A.: "Some results from a nonsymmetrical branching process that looks for interaccion effects". *Proceedings of the Social Statistics Section of the American Statistics Association*, 1963, 40-45.

donde $Z_{.1}$ representará la combinación de los primeros niveles de todas las variables complejas; $Z_{.2}$ la combinación de los primeros niveles de K-1 y el segundo de la K-ésima, etc., en general,

$$Z_{.j} = X^{1,m_1}_{.j_1} \ (x) \ X^{2,m_2}_{.j_2} \ ..... \ (x) \ X^{K,m_K}_{.j_K}$$

siendo $j_1 = 1, 2, \ldots\ldots\ldots m_1$

$j_2 = 1, 2, \ldots\ldots\ldots m_2$

$\ldots\ldots\ldots\ldots\ldots$

$j_K = 1, 2, \ldots\ldots\ldots m_K$

Con estas nuevas variables la representación del modelo cambia en el sentido de que ahora un individuo sólo puede pertenecer a una única variable $Z_{.j}$; la intersección entre $Z_{.j}$ y $Z_{.j'}$, siendo $j \neq j'$, es nula ya que una de ellas incluye algún nivel que la otra excluye. Obviamente a una cierta $Z_{.j}$ podrán pertenecer más de un individuo.

La matriz del modelo tendrá en consecuencia la siguiente composición:

Y	$Z_{.1}$	$Z_{.2}$	$Z_{.3}$	....	$Z_{.j}$ ...		$Z_{.m}$
1	0	1	0	...	0	.....	0
0	1	0	0	...	0	.....	0
0	1	0	0	...	0	.....	0
.	..	..	..	...	..	.....	..
1	0	0	1	...	0	.....	0
r	n_1	n_2	n_3	...	n_j	.....	n_m

en la cual r es el número de individuos que en la muestra poseen automóvil.

n_j es el número de individuos que pertenecen a $Z_{.j}$

y $\quad n = \sum^{m} n_j \qquad$ y $\quad m = \pi^{K} m_i$

ESTIMACION DE PARAMETROS

Para la variable j, el producto vectorial $Y' Z_j$ es el número de individuos que del total n_j de la categoría poseen automóvil, sea $p_j = Y' Z_j / n_j$ la proporción que poseen automóvil en la categoría j. A su vez la matriz Z tiene la característica de que $Z' Z = I . n_.$ (I es una matriz unidad de mxm y $n_.' = (n_1 , n_2 \ldots\ldots\ldots n_m)$ y por consiguiente su inversa será

$$(Z' Z)^{-1} = I . n_.^{-1},$$

El producto $Z' Y = n_. p_.$, donde $(n_. p_.)' = (n_1 p_1 , n_2 p_2 \ldots n_m p_m)$.

Los estimadores mínimo cuadráticos del modelo serán por tanto

$$B = (Z' Z)^{-1} Z' Y = I . n_.^{-1} n_. p_. = p_. \qquad (p_.' = p_1 , p_2 \ldots\ldots p_m).$$

De forma tal que por la propia estructura del modelo que estamos considerando los estimadores mínimo cuadráticos son directamente iguales a las frecuencias marginales que indican la proporción de automóviles que se poseen en cada categoría. Tal circunstancia facilita enormemente el cálculo y permite realizarlos cualquiera sea el tamaño de la matriz.

Al igual que se definió la operación de intersección, se puede definir la operación de unión o suma lógica de dos categorías, implicando con ello una nueva categoría, menos pura que sus componentes, y a la cual pertenecerán los individuos que correspondían a una u otra categoría. Así por ejemplo $Z_{j+j'} = Z_{.j} (+) Z_{.j'}$ constará de $n_j + n_{j'}$ individuos y de ellos $Y' (Z_{.j} + Z_{.j'})$ poseerán automóviles. Se puede comprobar que por la naturaleza de la matriz Z, el coeficiente estimado de esta nueva categoría es

$$p_{j+j'} = (n_j p_j + n_{j'} p_{j'}) / n_j + n_{j'}$$

que se puede extender a cualquier número de categorías. Además en base a esta operación se podrá más adelante estimar la contribución a la variación explicada de un subconjunto cualquiera de variables.

ANALISIS DE VARIANZA

Los coeficientes estimados oscilan entre cero y uno; si todos los individuos que pertenecen a una cierta categoría poseen automóvil, el valor de p será uno; si ninguno posee automóvil, el valor de p será 0. En base a las relaciones definidas, se puede comprobar que la variación total, $y'y = r - r^2/n$, la suma residual de cuadrados, es

$$e'e = Y'Y - B'Z'Y = r - p'.(n_j p_j) = r - \sum_{j=1}^{m} n_j p^2_j ,$$

la variación explicada,

$$= \sum_{j=1} n_j p^2_j - r^2/n$$

el coeficiente de correlación múltiple,

$$R^2 = (\Sigma n_j p^2_j - r^2/n) / (r - r^2/n)$$

la varianza de los coeficientes estimados,

$$\text{var}(B_j) = (r - \sum_{j=1} n_j p^2_j) / n_j (n-m)$$

$$\text{cov}(B_j B_{j'}) = 0 \qquad \text{para todo } j \neq j'$$

Por su parte, la variación explicada puede descomponerse del siguiente modo:

$$\Sigma n_j p^2_j - r^2/n = n_i p^2_i - (n_i p_i)^2/n + \sum_{i \neq j} n_j p^2_j - (\sum_{i \neq j} n_j p_j)^2/n -$$

$$- 2 (n_i p_i \sum_{i \neq j} n_j p_j) / n$$

$$= \sum_{i \neq j} n_j p^2_j - r^2_i / (n - n_i) + n n_i (p_i - r/n)^2/n - n_i$$

en la que el primer sumando representa la variación explicada cuando no se incluye en el modelo la categoría Z_i (siendo $r_i = \sum_{i \neq j} n_j p_j$), y el segundo la adición debida a la incorporación de tal categoría.

Formalmente la tabla de análisis de varianza adoptará la siguiente disposición:

Fuente de variación	Suma de cuadrados	Grados de lib.
Debido a Z_i	$nn_i\,(p_i - r/n)^2 / (n - n_i)$	1
Debido a Z_1, Z_2, ... Z_m, excepto Z_i	$\sum_{j \neq i} n_j p^2_j - r^2_i / (n - n_i)$	m—2
Variación explicada	$\sum_{j=1} n_j p^2_j - r^2/n$	m—1
Variación residual	$r - \sum_{j=1} n_j p^2_j$	n—m
Variación total	$r - r^2/n$	n—1

Vale la pena destacar que en el problema usual de la regresión lineal un coeficiente que no difiere significativamente de cero conduce por lo general a la afirmación de que tal variable no explica o no contribuye a la variación explicada por la regresión. Dejando de lado los reparos que tal generalización merece, la regla comúnmente aceptada es de que tal variable puede ser excluida del modelo sin que por ello éste pierda su valor predictivo. En el caso presente, sin embargo, un coeficiente que se aproxima a cero o a uno (supuesto que r dista de ambos extremos) es en cualquier caso significativo, bien sea porque está revelando una asociación positiva con la variable dependiente o bien porque está indicando una correlación negativa.

El valor de r representa en esencia un promedio y una categoría cualquiera es significativa en la medida que su coeficiente estimado se aparta del valor promedio; circunstancia ésta que evidenciará un comportamiento diferencial por parte de los miembros que pertenecen a esa categoría y existirá en consecuencia una interacción en la combinación de niveles que tal categoría supone. La detección de interacciones es precisamente el objetivo del método que desarrollamos.

Siguiendo un razonamiento similar, si lo que interesa es la variación atribuible a dos categorías, que representan desde luego un nivel de interacción más general, tal contribución se podrá estimar a través del valor de

$$(n_{i} + n_{j}) n ((r/n) - (n_{i}p_{i} + n_{j}p_{j}) / (n_{i} + n_{j})) / (n - n_{i} - n_{j})$$

DOCIMASIA DE HIPOTESIS

Para no sujetar la validez de las conclusiones a los supuestos que generalmente acompañan los modelos lineales —supuestos cuya inclusión es necesaria a los efectos de docimar hipótesis referentes a los parámetros— se ha optado por un test no paramétrico que se basará en los coeficientes estimados. Bajo los supuestos de la hipótesis nula que afirmarán la no significatividad de las categorías incluidas, el test o estrategia radicará en la distribución del escalar $Y' Z_j$, número de personas que del total n_j de la categoría Z_j poseen automóvil. Cualquiera sea la forma o los parámetros de la distribución de donde provienen las observaciones, el estadístico utilizado tendrá una distribución hipergeométrica de parámetro r/n.

El vector Y es una muestra de tamaño n en la cual cada individuo ha sido clasificado dicotómicamente conforme a la posesión de automóvil; el vector Z_j es una submuestra de la anterior, de tamaño n_j, y en la que cada individuo ha sido a su vez reclasificado según su pertenencia a dicha categoría. En cualquier caso $n_j \neq n$.

Podemos considerar a Z_j como una submuestra de Y o bien condicionar Z_j por lo observado en Y, en este último caso Z_j equivaldría a una muestra de la población que Y representa. Ciertamente lo ideal sería conocer R/N (siendo $E(r/n) = R/N$) que no es otra cosa que el parámetro de la población, pero la cuestión esencialmente no varía si se enfoca el problema combinatorio correctamente.

Consideremos primeramente un caso abreviado en el que intervienen 20 observaciones y tres variables y que presentan la siguiente disposición:

Obs. nº	Y	Z_1	Z_2	Z_3
1	1	0	1	0
2	0	0	0	1
3	0	0	0	1
4	1	0	1	0
5	0	0	0	1
6	0	1	0	0
7	1	1	0	0
8	1	0	1	0
9	0	0	0	1
10	1	0	1	0
11	0	0	0	1
12	1	0	1	0
13	0	1	0	0
14	1	0	1	0
15	1	1	0	0
16	0	0	0	1
17	0	0	0	1
18	0	0	0	1
19	0	0	0	1
20	0	0	0	1

$$r = 8 \quad Z'_1 Y = 2 \quad Z'_2 Y = 6 \quad Z'_3 Y = 0$$
$$n = 20 \quad n_1 = 4 \quad n_2 = 6 \quad n_3 = 10$$
$$p_1 = 0.5 \quad p_2 = 1 \quad p_3 = 0$$

Analizando los resultados de la variable que representa Z_3 cuyo coeficiente es $p_3 = 0$, podría parecer a primera vista que a pesar de ser tan bajo, tal resultado es compatible con una muestra de tamaño 10 proveniente de una población en la que 40% de los individuos tienen automóvil, o equivalentemente de una población de 20 individuos, 8 de los cuales poseen automóvil. Un resultado tan bajo como el observado podría ser consecuencia de fluctuaciones muestras, por lo que la hipótesis nula establecerá que tal muestra proviene en efecto de una población compuesta por 20 individuos y en la que un 40% de ellos poseen automóvil. A los fines de docimar tal hipó-

tesis utilizaremos el estadístico Z'_3Y cuya distribución bajo la hipótesis nula responde a una hipergeométrica de parámetro 0,40; la distribución de Z'_3Y será por consiguiente:

$$P(Z'_3Y=0/H_0) = Pr\ (p_3=0) = \binom{8}{0}\binom{12}{10}\Big/\binom{20}{10} = 0.00036$$

$$P(Z'Y=1/H_0) = P\ (p_3=0.1) = \binom{8}{1}\binom{12}{9}\Big/\binom{20}{10} = 0.00953$$

$$P(Z'Y=2/H_0) = P\ (p_3=0.2) = \binom{8}{2}\binom{12}{8}\Big/\binom{20}{10} = 0.07502$$

....

Si los individuos hubieran sido efectivamente extraídos de una población en la que 8 de cada 20 poseen automóvil, la evidencia muestral nos colocaría frente a la siguiente alternativa, a) si aceptamos la hipótesis nula, deberemos por fuerza admitir que un evento de probabilidad tan baja como 0.00036 ha podido ocurrir, b) o bien rechazamos la hipótesis nula fundándonos en que el valor observado queda comprendido dentro de la zona de rechazo. A un nivel de significación del 5% la misma decisión cabrá adoptar si en vez de observar el valor $p_3=0$ hubiéramos observado $p_3=0{,}1$.

Por lo que respecta a la variable Z_1 si se calculan las probabilidades para cada uno de los valores que puede asumir el estadístico Z'_1Y se encontrará que el valor observado $p_1=0.50$ no motivará el rechazo de H_0. Para la variable Z_2, un valor tan alto como el observado sólo tiene una probabilidad de 0,0072 de provenir de una población con parámetro 0,40. De un modo general, el test consistirá en establecer la siguiente hipótesis:

$$H_0 \ : \ p_j = r/n$$

$$H_1 \ : \ p_j \neq r/n$$

basándose en la distribución de n_jp_j que para muestras pequeñas tiene una distribución exacta,

$$n_j p_j \sim \binom{r}{n_i p_i} \binom{n-r}{n_j - n_i p_i} / \binom{n}{n_j}$$

y para grandes muestras su aproximación normal,

$$n_j p_j \sim (\sqrt{2\pi h})^{-1} \exp\left(-\tfrac{1}{2}\,(n_j p_j - n_j r/n)^2 / h^2\right)$$

siendo $h = (n_j(r/n)\ (1-r/n))^{1/2}$

APLICACION A UN CASO CONCRETO

En base a las encuestas que sobre empleo y desempleo realiza periódicamente el Instituto de Economía y Finanzas se han obtenido datos sobre posesión de bienes durables. A los fines del presente trabajo se adoptó la correspondiente a abril de 1967 y allí se extrajo información sobre 1364 unidades familiares. Las variables que se utilizaron fueron:

A) Ingresos del grupo familiar

- A.1 menos de $ 30.000 mens.
- A.2 30.000 - 70.000.—
- A.3 más de 70.000.—

B) Edad del Jefe

- B.1 30-39 años
- B.2 40-49 años
- B.3 50-59 años
- B.4 60 y más

C) Categoría ocupacional del Jefe

- C.1 No calificados (tareas manuales y no manuales)
- C.2 Calificados (tareas manuales y no manuales)
- C.3 Técnicos, profesionales, jefes y directivos

E) Educación del Jefe

- E.1 Sin ningún tipo de enseñanza
- E.2 Enseñanza primaria (completa o incompleta)

E.3 Ciclos medios (completos e incompletos) y universitario incompleto.

E.4 Universitario completo

Y) Variable dependiente

0 No posee automóvil

1 Posee automóvil

La tabla I presenta los resultados que se obtuvieron por aplicación del método descripto. Para una explicación mejor de la variable dependiente hubiera sido necesario incorporar otras variables, de forma tal que los coeficientes estimados de cada celda se aproximaran a cero o uno. Considérese, por ejemplo, la variable $A_3 B_3 E_4 C_3$ en la que aparecen 11 individuos, de los cuales 8 poseen automóvil y 3 no poseen. Es fácil conjeturar que las ocho unidades familiares que poseen automóvil tienen que ser diferentes a las restantes y que tales diferencias pueden revelarse a través de alguna variable no incluida en el modelo. O tal vez las mismas variables utilizadas podrían indicar tales diferencias si es que no se hubiera recurrido a un nivel de agregación (o de formación de niveles) tan alto como el que aquí se usó. El criterio que sirvió de base para la formación de niveles fue un tanto arbitrario; alternativamente y desde luego con mejores resultados podría utilizarse un método de construcción de niveles que maximice la homogeneidad interna de cada celda (o de cada variable), entendida ésta en el sentido que cada celda contenga el mayor número posible de individuos con automóvil y por consiguiente el menor número posible de individuos sin tal atributo. Como caso límite las celdas deberían formarse de modo que algunas incluyan únicamente unidades que poseen automóviles y las restantes celdas con todos los individuos que no poseen automóvil.

Las ventajas de este procedimiento pueden apreciarse en las fórmulas ya desarrolladas. La suma residual de cuadrados

$$e' e = r - \sum_{j=1} n_j p^2_j$$

tiene la propiedad de que si todos los coeficientes son iguales a cero o uno,

$$\sum n_j p'^2_j = \sum n_j p_j = r$$

y en consecuencia

$$e' e = 0.$$

El análisis de varianza de la tabla I dio los siguientes resultados:

Fuente de variación	Suma de cuadrados	G.L.
Variación explicada	73.070	144
Variación residual	169.185	1219
Variación total	242.255	1363

Dentro del esquema desarrollado se puede estimar la contribución a la variación explicada proveniente de las variables que en la tabla I aparecen con coeficientes significativamente altos. Para ello, sólo basta unir todas las categorías excepto aquellas cuya contribución se quiere determinar, obteniendo una nueva variable cuyo coeficiente será

$$p'_j = (\sum n_j p_j) / \sum n_j$$

(donde el subíndice j se extiende a todas las categorías excepto las significativas)

$$= 169/965 = 0{,}1751$$

Con este nuevo modelo, la tabla de análisis de varianza se convierte

Fuente de variación	Suma de cuadrados	G. L.
Variación explicada	69.950	44
Variación residual	172.305	1319
Variación total	242.255	1363

Para estimar la contribución de p'_j

$$n.n'_j (p'_j - r/n)^2 / n - n'_j = 1364.\ 965\ (0.2309 - 0.1751)^2 / 399$$
$$= 10{,}263$$

Luego la contribución de las restantes variables es de aproximadamente 85% de la variación explicada.

Para docimar individualmente cada coeficiente se construyó una tabla (Tabla II del Apéndice) con los valores críticos del 90 y 95% para muestras de tamaño inferior o igual a 30. Cuando los valores de n_j superaban tal límite, se utilizó la aproximación normal ya descripta.

INTERPRETACION DE LOS RESULTADOS

Del análisis de los coeficientes estimados (p_j) y de sus niveles de significación se desprende que, en primer lugar, todas las variables presentan una correlación positiva entre posesión de automóvil y su nivel, o sea que a medida que se asciende en el nivel de ésta, mayor es el valor del coeficiente. En el caso de edad, sin embargo, esto es observable sólo en dos tramos, menos de 40 años en el que la posesión evidencia un comportamiento significativamente bajo y, superada esa edad crítica, un tramo en el que los coeficientes son aproximadamente iguales para cualquier nivel de edad.

Con respecto a la contribución de cada categoría a la explicación total del modelo (30%), se observa que un 85% de esa variación explicada está dada por los extremos de las variables, es decir, alta educación, alto nivel de ingresos, alta índole ocupacional, etc. y también por ciertos niveles intermedios de estas variables que se interaccionan de un modo particular.

La no posesión de automóviles, por su parte, es explicada en su totalidad por el nivel más bajo de ingreso y los dos primeros niveles de educación, cualquiera sea la índole ocupacional y edad.

Finalmente, existe una zona en la que los coeficientes estimados no difieren significativamente del promedio de todas las categorías (0,2309). Esta zona de comportamiento no diferencial está primordialmente encerrada entre el tercer nivel de educación con ingresos bajos y el segundo nivel de educación con ingresos intermedios.

Las variables incluidas permiten formular algunas consideraciones sobre el tema, siguiendo los lineamientos teóricos dados en un

trabajo anterior (3) en el que se analiza la conducta del consumidor con respecto a la posesión de bienes de consumo durables.

Mediante el análisis de regresión se vincula allí el stock de bienes durables poseído por el consumidor con las variables ingreso corriente y edad. No se utiliza, sin embargo, una regresión múltiple que sólo permitiría medir el efecto de cada variable sobre el stock de durables poseído, aislando la influencia de la otra pero sin controlar el nivel al cual se la mantiene constante. Por el contrario, se intenta controlar a cada variable para niveles dados de la otra en base a regresiones simples en las que, para cada uno de los distintos niveles definidos, tanto para ingreso como para edad, se relaciona el stock con la variable "libre".

Partiendo del supuesto de que el transcurso del tiempo es un requisito necesario para la acumulación de riqueza, se deduce que las variaciones de edad, conjugadas con distintos niveles de ingreso corriente, dan lugar a otros tantos niveles de riqueza.

Esta variable, a su vez, es definida como el patrimonio neto del consumidor en el que se incluye la capacitación del individuo o riqueza humana. Este último aspecto sin embargo no es controlado dentro del análisis, ya que ni la variable ingreso, ni la variable edad, son "depuradas" del concepto capacitación.

Al incluir ahora las variables educación e índole ocupacional, la medición de la riqueza puede por lo tanto ser refinada un poco más a efecto de comprobar la hipótesis de que la riqueza acumulada, o capitalización del consumidor, actúa como factor determinante da la propiedad del bien durable automóvil.

En base al cuadro posterior pueden identificarse cuatro zonas diferentes. La primera, tramo 1, corresponde a una situación en la que, ingresos y educación bajos, impiden una acumulación de riqueza tal que permita tener acceso al bien automóvil.

La segunda, tramos 2 y 3, incluye a aquellas unidades de consumo que presentan una conducta diferencial con respecto al promedio. A diferencia del caso anterior, un mayor ingreso o un superior

3. SÁNCHEZ, C.E.: "Posesión de Bienes de Consumo Durables", *Revista de Economía y Estadística*, N° 1-2, primero y segundo trimestres de 1968; 37-59.

nivel de educación, conjugados con los distintos niveles de edad y de índole ocupacional, definen niveles de riqueza en los que las decisiones sobre posesión resultan positivas.

En la tercera zona, tramos 4 y 5, los niveles superiores o de educación hacen que las variaciones de edad o de ocupación produzcan una conducta diferencial, es decir, la mayor riqueza que esto implica lleva a frecuencias significativas y altas.

ANALISIS DE LOS RESULTADOS

Tramo	Variables Controladas	Nivel	Comportamiento de las Variables Edad y Ocupación
1	Ingreso	bajo	para cualquier nivel de edad u ocupación
	Educación	bajo	los coeficientes son significativos y bajos.
2	Ingreso	bajo	para cualquier nivel de edad u ocupación
	Educación	medio y alto	los coeficientes no son significativos (igual al promedio).
3	Ingreso	medio	ídem anterior
	Educación	bajo	
4	Ingreso	alto	coeficientes no significativos excepto
	Educación	bajo	para edad y ocupación altas que presentan coeficientes significativos y altos.
5	Ingreso	medio	los coeficientes de edad y ocupación al-
	Educación	medio y alto	tas pasan a ser significativos y altos
6	Ingreso	alto	para cualquier nivel de edad u ocupación
	Educación	alto	los coeficientes son significativos y altos.

Por último, en el tramo 6, con ingreso y con educación altos, las variables edad y ocupación vuelven a perder poder explicativo, esta

vez en forma opuesta al tramo 1. Es decir el consumidor ha podido ajustar su conducta más rápidamente que en los tramos 4 y 5 y en consecuencia, tanto las variaciones de edad como las de la índole ocupacional no producen un comportamiento diferencial.

Estas conclusiones corroboran la hipótesis formulada y sugieren que el método propuesto constituye un eficaz instrumento de análisis. Como ya se señalara anteriormente, la ampliación del marco conceptual permitirá hacer explícita la interacción de variables aquí no consideradas, aumentando el poder explicativo del modelo y brindando en consecuencia una mejor base empírica para la formulación de una teoría sistemática.

TABLA I

		E_1				E_2			
		C_1	C_2	C_3	total	C_1	C_2	C_3	total
A_1	B_1	0 0.000 (-) 5	0 0.000 (-) 2	0 0	0 0.000 (-) 7	1 0.025 (-) 39	1 0.029 (-) 33	0 0	2 0.027 (-) 72
	B_2	0 0.000 (-) 7	0 0.000 (-) 2	0 0	0 0.000 (-) 9	5 0.087 (-) 52	11 0.136 (-) 70	1 0.330 (-) 2	17 0.121 (-) 124
	B_3	0 0.000 (-) 9	1 0.250 (-) 3	0 0	1 0.083 (o) 12	7 0.097 (-) 65	10 0.133 (-) 65	1 0.250 (o) 3	18 0.119 (-) 123
	B_4	1 0.033 (-) 33	0 0.000 (-) 10	0 0	1 0.023 (-) 43	21 0.111 (-) 168	7 0.094 (-) 67	1 1.000 (o) 0	29 0.110 (-) 235
	total	1 0.018 (-) 54	1 0.059 (-) 17	0 0	2 0.028 (-) 71	34 0.095 (-) 324	29 0.110 (-) 235	3 0.375 (o) 5	66 0.105 (-) 564

Cada celda constituye una variable. La cantidad en el margen superior izquierdo es el número de personas que poseen automóvil; la del margen inferior izquierdo, el número de individuos que no poseen automóvil. A la derecha de ambas, el valor del coeficiente estimado. Los símbolos (—), (o), (+), indican que el coeficiente es significativamente bajo, no significativo y significativamente alto, respectivamente, al nivel del 95%.

TABLA I (Continuación)

		E_1				E_2			
		C_1	C_2	C_3	total	C_1	C_2	C_3	total
A_2	B_1	0 0	0 0	0 0	0 0	0 0.000 (o) 2	1 0.142 (o) 6	0 0.000 (o) 1	1 0.100 (o) 9
	B_2	0 0	0 0	0 0	0	5 0.277 (o) 13	9 0.346 (o) 17	3 0.600 (o) 2	17 0.347 (o) 32
	B_3	0 0	1 1.000 0	0 0	1 1.000 0	5 0.300 (o) 10	11 0.379 (o) 18	2 0.667 (o) 1	18 0.382 (o) 29
	B_4	0 0	0 0.000 (-) 2	0 0	0 0.000 (-) 2	1 0.083 (o) 11	10 0.455 (+) 12	4 0.400 (o) 6	15 0.341 (o) 29
	total	0 0	1 0.333 (o) 2	0 0	1 0.333 (o) 2	11 0.235 (o) 36	31 0.369 (+) 53	9 0.474 (+) 10	51 0.340 (+) 99

TABLA I (Continuación)

		E_1				E_2			
		C_1	C_2	C_3	total	C_1	C_2	C_3	total
A_3	B_1	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
	B_2	0 0	0 0	0 0	0 0	0 0.000 (o) 1	0 0.000 (o) 1	0 0 (o)	0 0.000 (o) 2
	B_3	0 0	0 0	0 0	0 0	0 0.000 1	1 0.500 (o) 1	0 0.000 1	1 0.250 (o) 3
	B_4	0 0	0 0	0 0	0 0	0 0	2 0.667 (o) 1	3 1.000 (+) 0	5 0.833 (+) 1
	total	0 0	0 0	0 0	0 0	0 0.000 (-) 2	3 0.500 (o) 3	3 0.750 (+) 1	6 0.500 (+) 6
	total	1 0.018 (-) 54	2 0.105 (-) 19	0 0	3 0.039 (-) 73	45 0.110 (-) 362	63 0.178 (-) 291	15 0.483 (+) 16	123 0.184 (-) 669

TABLA I (Continuación)

		E_3				E_4				
		C_1	C_2	C_3	total	C_1	C_2	C_3	total	total
A_1	B_1	4 0.167 (o) 20	0 0.000 (-) 31	0 0.000 (o) 4	4 0.068 (-) 55	0 0	0 0	0 0	0 0	6 0.043 (-) 134
	B_2	4 0.154 (o) 22	8 0.250 (o) 24	4 0.800 (+) 1	16 0.253 (o) 47	0 0	0 0	1 0.500 (o) 1	1 0.500 (o) 1	34 0.158 (o) 181
	B_3	1 0.067 (o) 14	4 0.210 (o) 15	2 0.500 (o) 2	7 0.184 (o) 31	0 0	0 0	0 0	0 0	26 0.128 (-) 176
	B_4	7 0.159 (o) 37	5 0.227 (o) 17	1 0.333 (o) 2	13 0.188 (o) 56	1 0.250 (o) 3	0 0	0 0.000 (o) 3	1 0.142 (o) 6	44 0.114 (-) 340
	total	16 0.147 (-) 93	17 0.163 (o) 87	7 0.438 (o) 9	40 0.175 (o) 189	1 0.250 (o) 3	0 0	1 0.200 (o) 4	2 0.226 (o) 7	110 0.116 (-) 831

TABLA I (Continuación)

		E_3				E_4				
		C_1	C_2	C_3	total	C_1	C_2	C_3	total	total
A_2	B_1	2 0.500 (o) 2	1 0.067 (o) 14	0 0	3 0.158 (o) 16	0 0	0 0	2 0.500 (o) 2	2 0.500 (o) 2	6 0.182 (o) 27
	B_2	5 0.454 (o) 6	12 0.343 (+) 23	7 0.778 (+) 2	24 0.436 (+) 31	0 0	1 1.000 (o) 0	6 0.667 (+) 3	7 0.700 (+) 3	48 0.421 (+) 66
	B_3	4 0.571 (o) 3	11 0.367 (+) 19	6 0.750 (+) 2	21 0.467 (+) 24	0 0	1 1.000 (o) 0	4 0.571 (+) 3	5 0.625 (+) 3	45 0.560 (+) 56
	B_4	2 0.250 (o) 6	12 0.521 (+) 11	5 0.833 (+) 1	19 0.513 (+) 18	1 1.000 (o) 0	0 0.000 (o) 1	7 0.636 (+) 4	8 0.615 (+) 5	42 0.438 (+) 54
	total	13 0.433 (+) 17	36 0.350 (+) 67	18 0.783 (+) 5	67 0.429 (+) 89	1 1.000 (o) 0	2 0.667 (o) 1	19 0.613 (+) 12	22 0.628 (+) 13	141 0.410 (+) 203

TABLA I (Continuación)

		E_3				E_4				
		C_1	C_2	C_3	total	C_1	C_2	C_3	total	total
A_3	B_1	1 0.500 (o) 1	2 1.000 (+) 0	0 0	3 0.750 (+) 1	0 0	0 0	0 0	0 0	3 0.750 (+) 1
	B_2	0 0	3 1.000 (+) 0	3 0.750 (+) 1	6 0.857 (+) 1	0 0	1 1.000 0	11 1.000 (+) 0	12 12.000 (+) 0	18 0.857 (+) 1
	B_3	2 1.000 (+) 0	5 0.714 (+) 2	5 1.000 (+) 0	12 0.857 (+) 2	0 0	1 1.000 0	8 0.727 (+) 3	9 0.750 (+) 3	22 0.733 (+) 8
	B_4	2 0.667 (o) 1	2 0.667 (o) 1	4 1.000 (+) 0	10 0.833 (+) 2	0 0	1 1.000 0	7 1.000 (+) 0	8 1.000 (+) 0	23 0.885 (+) 3
	total	5 0.714 (+) 2	12 0.800 (+) 3	12 0.923 (+) 1	29 0.837 (+) 6	0 0	3 1.000 (+) 0	26 0.897 (+) 3	29 0.906 (+) 3	64 0.810 (+) 15
	total	34 0.233 (o) 112	65 0.293 (+) 157	37 0.711 (+) 15	138 0.327 (+) 284	2 0.400 (o) 3	5 0.833 (+) 1	46 0.708 (+) 19	53 0.697 (+) 23	315 0.2309 1049

TABLA II

VALORES CRITICOS DEL TEST HIPERGEOMETRICO

	95%		90%	
n_j	valor crítico inferior	valor crítico superior	valor crítico inferior	valor crítico superior
2				2
3		3		3
4		3		3
5		4		3
6		4		4
7		5		4
8		5		4
9		5	0	5
10		6	0	5
11		6	0	5
12	0	6	0	6
13	0	7	0	6
14	0	7	0	6
15	0	7	1	7
16	0	8	1	7
17	0	8	1	7
18	0	8	1	7
19	1	8	1	8
20	1	9	1	8
21	1	9	2	8
22	1	9	2	9
23	1	10	2	9
24	1	10	2	9
25	2	10	2	9
26	2	10	2	10
27	2	10	3	11
28	2	11	3	11
29	2	12	3	11
30	2	12	3	11