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

Serie documentos de trabajo

EL USO DE MODELOS LOG-LINEALES PARA EL ANÁLISIS DEL CONSUMO RESIDENCIAL DE ENERGÍA

Enrique de Alba
ITAM

Yolanda Mendoza
Dirección General de Política Energética
SEMIP

DOCUMENTO DE TRABAJO

Núm. VI - 1984

EL USO DE MODELOS LOG-LINEALES PARA EL ANALISIS DEL CONSUMO RESIDENCIAL DE ENERGIA

Enrique de Alba

y

Yolanda Mendoza

Este trabajo está dirigido a quienes se interesen en la aplicación de métodos estadísticos en Economía. Se presenta primero una reseña breve de lo que es el modelo log-lineal, su definición, algunas características, métodos de estimación, selección de modelos, medidas de asociación y análisis de los resultados. Al final se aplica la metodología a un cuadro de resultados de la Encuesta de Ingreso y Gasto de los Hogares, 1977 de la Secretaría de Programación y Presupuesto, para ilustrar el uso del modelo. Los resultados sirven para verificar algunas hipótesis respecto al comportamiento de las familias en cuanto al gasto en energéticos.

ANALYSING RESIDENTIAL ENERGY CONSUMPTION
AN APPLICATION OF LOG-LINEAR MODELS

Enrique de Alba

and

Yolanda Mendoza

This article is intended for those who are interested in applying statistical methods in Economics. A brief presentation is made of the log-linear model, its definition, some properties, estimation methods, model selection, measures of association between variables and interpretation of the results. In the last part an application is made to a table of results from the 1977 Household Income and Expenditure Survey, carried out by Secretaría de Programación y Presupuesto, to illustrate the use of this kind of model. The results obtained verify some hipotheses on household behavior regarding expenditures on energy.

EL USO DE MODELOS LOG-LINEALES PARA EL ANALISIS DEL CONSUMO RESIDENCIAL DE ENERGIA

ENRIQUE DE ALBA

Instituto Tecnológico Autónomo de México

Y

YOLANDA MENDOZA

Dirección General de Política Energética
SEMIP

1984

INTRODUCCION

Los métodos econométricos que se utilizan con mayor frecuencia, o por lo menos los que ocupan un mayor espacio en los textos sobre el tema, se basan en el supuesto de variables dependientes contínuas; así tenemos el caso del modelo lineal y series de tiempo. En ocasiones pueden presentarse datos correspondientes a variables cualitativas, como la distinción entre zonas urbanas ó zonas rurales; también puede suceder que a pesar de que las variables sean continuas los resultados se presenten agrupados, como por ejemplo estratos de ingreso familiar. En estas situaciones los métodos anteriores no son adecuados y habrá que recurrir a métodos para el análisis de datos categóricos.

En esta nota se pretende hacer una exposición sencilla de los conceptos que se utilizan para el análisis estadístico de datos categóricos, concretamente mediante el uso del modelo conocido como log-lineal.

Se ha dicho que las variables se dividen en categorías. Las categorías se presentan cuando clasificamos a la población, digamos en Q grupos. Se dice una clasificación es exhaustiva cuando pueden clasificarse a todos los miembros de la población en alguna categoría; dichas categorías son mutuamente exclusivas cuando están definidas de tal manera que se pueda acomodar a un indi---

viduo correctamente en una y sólo una de ellas. Cuando se hace una clasificación se busca que tenga estas dos propiedades, para que no haya confusión al acomodar a un individuo de población y caiga siempre en algún grupo.

Al hablar de datos cualitativos es necesario un arreglo que refleje la estructura de los datos. Este arreglo debe definir las categorías de cada variable y las interrelaciones con las demás categorías de las otras variables en el modelo. Un arreglo rectangular es la estructura más adecuada para representar a 2 variables: los renglones del arreglo se hacen corresponder con los niveles de la la. variable mientras que las columnas se asocian con los niveles de la 2a. variable. Si se desea presentar más de 2 variables pueden utilizarse tantos arreglos rectangulares como sea necesario. Por ejemplo, supóngase que se tienen como variables el ingreso (alto, medio, bajo) y el sexo (masculino, femenino); se presenta la información en una tabla, los renglones correspondientes a cada nivel de ingreso y las columnas al sexo:

		SEXO	
		Masculino	Femenino
Ingreso	Alto		
	Medio		
	Bajo		

Si se considera la inclusión de una tercera variable por ejemplo el tipo de población en el que viven los individuos (rural, urbana) la información se resumiría en 2 arreglos idénticos al anterior, uno para la población rural y otro para la población urbana.

Población Rural

Ingreso	Sexo: Masculino	Sexo: Femenino
Alto		
Medio		
Bajo		

Población Urbana

Ingreso	Sexo: Masculino	Sexo: Femenino
Alto		
Medio		
Bajo		

TABLA 2

Al observar el arreglo vemos que la posición en las celdas nos refleja las características de los individuos que caen en ellas, a este arreglo se le conoce como tabla de contingencia, en datos cualitativos multivariados cada individuo se describe por una serie de atributos; al tomar una muestra todos los individuos con la misma descripción son enumerados y esta cuenta entra en la celda correspondiente de la tabla de contingencia, por ejemplo, todos los individuos en la muestra que tengan ingreso alto y sean hombres entran en la celda superior izquierda de la tabla 1.

Una vez que se tiene la tabla de contingencia el interés se enfoca

en analizar las relaciones que existen entre las variables. El modelo log-lineal proporciona un esquema para el análisis de dichas tablas.

En el análisis de tablas de contingencia se presentan dos situaciones:

i) Una variable se vé como de respuesta y las demás son explicativas, esto es lo que se conoce como tabla asimétrica.

ii) No se hace distinción entre variables dependientes e independientes, esto es lo que se llama tabla simétrica.

El objetivo del Análisis es plantear los dos tipos de situaciones, primero no distinguiendo las variables, o sea considerándolas conjuntamente y posteriormente planteando un modelo en el que se planteen relaciones de causalidad.

Además, para el análisis de las relaciones entre 2 variables se utiliza una medida de asociación, que se presenta para cada tabla bivariada y nos dá una idea de como están asociadas las variables de dicha tabla, la magnitud de dicha asociación, así como si es o no significativa.

I . Modelo bivariado

I.A) El caso 2 x 2

Se ha planteado la importancia del análisis de datos categóricos y la idea de la exposición es plantear un modelo con un número cualesquiera de variables, sin embargo, antes de generalizar el modelo, se expondrá la tabla de contingencia más simple, la tabla 2 x 2, basada en 4 celdas, con dos categorías para cada variable, pues esto facilitará la comprensión de un modelo con un mayor número de variables

Considérese la variable A_1 que puede clasificarse en las categorías i_1 ($i_1 = 1, 2$) y la variable A_2 que se clasifica en las categorías i_2 ($i_2 = 1, 2$). Supongamos que los dos renglones de la tabla corresponden a las categorías de la variable 1 (A_1) y que las columnas se asocian a la variable 2 (A_2). Se utilizan subíndices para referirse a la posición en la tabla, el 1o. le corresponde a la variable 1 y el 2o. a la variable 2.

El modelo está basado en probabilidades (p) pero para referirse a frecuencias se utiliza una transformación simple, como se verá más adelante. Considérese la tabla :

A_1 \ A_2	1	2
1	P_{11}	P_{12}
2	P_{21}	P_{22}

En donde $p_{i_1 i_2}$ es la probabilidad que tiene un individuo de estar en la categoría i_1 de la variable 1 y en la categoría i_2 de la variable 2.

El modelo log-lineal para esta tabla se escribe :

$$\log p_{i_1 i_2} = \mu + \alpha_1(i_1) + \alpha_2(i_2) + \beta_{12}(i_1, i_2)$$

$$i_1 = 1, 2$$

$$i_2 = 1, 2 \quad ,$$

a μ, $\alpha_1(i_1)$, $\alpha_2(i_2)$ y $\beta_{12}(i_1, i_2)$ se les conoce cómo parámetros o efectos del modelo. μ es la media general de los logarít - mos de las probabilidades, es decir :

$$\mu = \frac{1}{4}(\log p_{11} + \log p_{12} + \log p_{21} + \log p_{22})$$

$\mu + \alpha_1(i_1)$ es la media de los logarítmos de las probabili-

dades en el nivel i_1 de la variable 1 y se obtiene de :

$$\mu + \alpha_1(i_1) = \frac{1}{2} (\log p_{i_1 1} + \log p_{i_1 2})$$

$$i_1 = 1, 2$$

$\mu + \alpha_2(i_2)$ es la media de los logarítmos de las probabilidades en el nivel i_2 de la variable 2.

$$\mu + \alpha_2(i_2) = \frac{1}{2} [\log p_{1 i_2} + \log p_{2 i_2}]$$

$$i_2 = 1, 2$$

Nótese que $\beta_{12}(i_1, i_2)$ puede encontrarse mediante simples operaciones algebraicas una vez que se conocen los valores de μ, $\alpha_1(i_1)$ y $\alpha_2(i_2)$, pues

$$\beta_{12}(i_1, i_2) = \log p_{i_1 i_2} - \alpha_1(i_1) - \alpha_2(i_2)$$

$$= \log p_{i_1 i_2} - \mu - \frac{1}{2} \sum_{i_1} \log p_{i_1 i_2}$$

$$+ \mu - \frac{1}{2} \sum_{i_2} \log p_{i_1 i_2} + \mu$$

sustituyendo μ por su valor como media general, se obtiene :

$$\beta_{12}(i_1, i_2) = \log p_{i_1 i_2} - \frac{1}{2} \sum_{i_1} \log p_{i_1 i_2} - \frac{1}{2} \sum_{i_2} \log p_{i_1 i_2} + \frac{1}{4} \sum_{i_1 i_2} \log p_{i_1 i_2}$$

A continuación se presentan las restricciones a que estan sujetos estos parámetros.

I.A.1) Restricciones ANOVA

Como $\alpha_1(i_1)$ y $\alpha_2(i_2)$ representan desviaciones con respecto a la media general, al sumar sobre todos los valores de un subíndice de éstos efectos la suma es cero, lo cual se expresa :

$$\alpha_1(1) + \alpha_1(2) = \alpha_2(1) + \alpha_2(2) = 0$$

Si se define $\ell_{i_1 i_2} = \log p_{i_1 i_2}$ se escriben los parámetros como :

$$\mu = \frac{1}{4} \sum_{i_1 i_2} \ell_{i_1 i_2} = \frac{1}{4} \ell_{\cdot\cdot}$$

$$\alpha_1(i_1) = \frac{1}{2} \ell_{i_1 \cdot} - \frac{1}{4} \ell_{\cdot\cdot}$$

$$\alpha_2(i_2) = \frac{1}{2} \ell_{\cdot i_2} - \frac{1}{4} \ell_{\cdot\cdot}$$

en donde el punto indica suma sobre los subíndices donde aparece Esta notación permite que se muestre que se cumplen las restricciones, véase que al sumar sobre i_1 en α_1 se obtiene:

$$\alpha_1(\cdot) = \sum_{i_1=1}^{2} \left(\frac{\ell_{i_1\cdot}}{2} - \frac{\ell_{\cdot\cdot}}{4} \right) = 0$$

de la misma manera se cumple con que: $\alpha_2(\cdot) = 0$.

El término β_{12} representa una desviación con respecto a $\mu + \alpha_1(i_1) + \alpha_2(i_2)$ o sea:

$$\beta_{12}(i_1, i_2) = \ell_{i_1 i_2} - \frac{\ell_{i_1\cdot}}{2} - \frac{\ell_{\cdot i_2}}{2} + \frac{\ell_{\cdot\cdot}}{4} ,$$

así que,

$$\beta_{12}(i_1, \cdot) = \beta_{12}(\cdot, i_2) = 0 \quad ; \quad i_1 = 1, 2 \qquad i_2 = 1, 2$$

o dicho de otra manera

$$\beta_{12}(1,1) = -\beta_{12}(1,2) = -\beta_{12}(2,1) = \beta_{12}(2,2)$$

I.A.2) Modelos con frecuencias celdales

Los modelos pueden describir, como se dijo, frecuencias celdales esperadas en lugar de probabilidades; la diferencia en la formulación estriba en una constante. Si se considera una tabla con conteos esperados m, de tal manera que el tamaño de muestra es :

$$n = \sum_{i_1 i_2} m_{i_1 i_2}$$

donde

$$m_{i_1 i_2} = n\, p_{i_1 i_2} \quad ;$$

tomando logarítmo,

$$\log m_{i_1 i_2} = \log n + \ell_{i_1 i_2}$$

y entonces el modelo log-lineal considerando frecuencias esperadas celdales es ($\log m_{i_1 i_2} = \ell_{m_{i_1 i_2}}$):

$$\ell_{m_{i_1 i_2}} = \mu' + \alpha_1(i_1) + \alpha_2(i_2) + \beta_{12}(i_1 , i_2)$$

con

$$\mu' = \log n + \mu$$

Se utilizara μ en cualquiera de los casos, ya sea que se trate de frecuencias esperadas o probabilidades. En la siguiente sección se plantea la construcción del modelo a partir de la condición de independencia.

I.A.3) Formulación del modelo a partir de la condición de independencia

La tabla de contingencia para las variables 1 y 2, puede extenderse a incluir los totales marginales, los cuales se expresan :

$$\sum_{i_2} p_{i_1 i_2} = p_{i_1 \cdot}$$

$$\sum_{i_1} p_{i_1 i_2} = p_{\cdot i_2} \ ,$$

que son los totales marginales renglón y columna respectivamente, dan la probabilidad que tiene un individuo de caer en la categoría i_1 de la variable 1, y la probabilidad que tiene de caer en la categoría i_2 de la variable 2, respectivamente.

El arreglo completo es :

A_1 \ A_2	1	2	
1	$P_{1\,1}$	$P_{1\,2}$	$P_{1.}$
2	$P_{2\,1}$	$P_{2\,2}$	$P_{2.}$
	$P_{.1}$	$P_{.2}$	1

Para llegar a la formulación del modelo a la manera log-lineal se supone que A_1 y A_2 son independientes. La condición de independencia se expresa:

$$p_{i_1 i_2} = p(A_1 = i_1) * p(A_2 = i_2) = p_{i_1 \bullet}\, p_{\bullet i_2}$$

Utilizando conteos esperados,

$$m_{i_1 i_2} = \frac{m_{i_1 \bullet}\; m_{\bullet i_2}}{n}$$

Del modelo log-lineal en donde se ha omitido el parámetro $\beta_{12}(i_1, i_2)$ se sabe que;

$$m_{i_1} = e^{\mu + \alpha_2(i_2)} \sum_{i_1} e^{\alpha_1(i_1)}$$

$$m_{i_2} = e^{\mu + \alpha_1(i_1)} \sum_{i_2} e^{\alpha_2(i_2)}$$

y

$$n = e^{\mu} \sum_{i_1} e^{\alpha_1(i_1)} \sum_{i_2} e^{\alpha_2(i_2)}$$

de donde:

$$e^{\mu + \alpha_1(i_1) + \alpha_2(i_2)} = m_{i_1 i_2}$$

Por lo tanto, el modelo log-lineal que supone independencia es consistente con la formulación que se hace en tablas de contingencia de esta condición.

I.B) Extensión a tablas I_1 x I_2

Se puede extender la tabla para tener I_1 categorías para la variable 1 e I_2 categorías para la variable 2. Los parámetros en este contexto se calculan como sigue, la medida general es:

$$\mu = \frac{1}{I_1 I_2} \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \ell_{m_{i_1 i_2}}$$

los efectos principales se escriben

$$\alpha_1(i_1) = \frac{1}{I_2} \sum_{i_2=1}^{I_2} \ell_{m i_1 i_2} - \mu$$

$$\alpha_2(i_2) = \frac{1}{I_1} \sum_{i_1=1}^{I_1} \ell_{m i_1 i_2} - \mu$$

y el efecto bivariado se expresa:

$$\beta_{12}(i_1,i_2) = \ell_{m i_1 i_2} - \frac{1}{I_2} \sum_{i_2=1}^{I_2} \ell_{m i_1 i_2} - \frac{1}{I_1} \sum_{i_1=1}^{I_1} \ell_{m i_1 i_2} + \mu$$

I.C) Interpretación de parámetros

Los parámetros del modelo log-lineal son análogos a los efectos en Análisis de Varianza (ANOVA) cuando una variable se descompone en términos aditivos que representan una media general, efectos principales y efectos de orden mayor (bivariados, trivariados, etc.)

El término μ es un efecto normalizante que permite que:

$$\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} p_{i_1 i_2} = 1$$

o bien:

$$\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} m_{i_1 i_2} = n$$

El término $\alpha_1(i_1)$ representa el efecto principal de la log-frecuencia de estar en el nivel i_1 de la variable 1.

Similarmente $\alpha_2(i_2)$ representa el efecto principal de la log-frecuencia de estar en la categoría i_2 de la variable 2.

El término $\beta_{12}(i_1, i_2)$ representa el efecto conjunto de la log-frecuencia de estar en la categoría i_1 de la variable 1 y en la categoría i_2 de la variable 2.

La generalización a tablas de contingencia con más de dos variables sigue un desarrollo análogo al presentado, con la utilización de algebra matricial para facilitar la exposición.

II. Modelo Log-Lineal q-variado.

La formulación del modelo log-lineal se ha presentado como una descomposición de las log-frecuencias de una tabla de contingencia en varios componentes aditivos. Siguiendo esta idea se presenta a continuación la generalización de este modelo a una tabla multidimensional que incluya q variables.

Supóngase que se tienen $A_1, A_2, \ldots, A_q$ variables categóricas, una tabla de contingencia consiste de $x_{i_1 i_2 \ldots i_q}$ frecuencias donde

$$i_j = 1, 2, \ldots, I_j \quad , \text{ con } \quad j = 1, 2, \ldots, q \ ,$$

donde I_j es el total de categorías para la variable j. Se puede simplificar a un sólo subíndice i, que se forma de la siguiente manera:

$$i = (i_1 - 1) I_2 I_3 \ldots I_q + (i_2 - 1) I_3 I_4 \ldots I_q + \ldots$$

$$+ (i_{q-1} - 1) I_q + i_q \ .$$

El uso de este subíndice permite asignar un sólo número a cada una de las celdas, sin embargo, no nos indica las características de los individuos que caen en ellas, es un índice lexicográfi-

co, es decir, sigue un cierto órden (2). Se puede escribir entonces:

$$x_i = x_{i_1 i_2 \cdots i_q}, \qquad i = 1, 2, \ldots, Q$$

tal que:

$$Q = \prod_{k=1}^{q} I_k$$

es el número total de celdas que tiene la tabla.

Para facilitar la explicación de algunos conceptos que se plantearán más adelante, es conveniente poder referirse a un conjunto cualquiera de subíndices:

$$\eta = \{ i_1, i_2, \ldots, i_q \}$$

en donde los subíndices toma los valores

$$i_1 = 1, 2, \ldots, I_1$$

$$i_2 = 1, 2, \ldots, I_2$$

$$\vdots \quad \vdots \quad \vdots$$

$$i_q = 1, 2, \ldots, I_q$$

$$\sum_{\eta} p_{\eta} = 1$$

Para poder definir a los modelos log-lineales en un contexto general, se requiere su expresión en algebra matricial, por lo cual se ordena los log p_{η} ' s en un vector:

$$\underline{\ell} = \underline{\log\ p} = \begin{pmatrix} \ell_{11\cdots 1} \\ \ell_{11\cdots 2} \\ \\ \\ \\ \\ \ell_{i_1 i_2 \cdots i_q} \end{pmatrix}$$

En donde las ℓ 's se particionan a la manera ANOVA como:

$$\begin{aligned} \ell_{\eta} = \log p_{\eta} = \mu &+ \alpha_1(i_1) + \ldots + \alpha_q(i_q) \\ &+ \beta_{12}(i_1, i_2) + \ldots \\ &+ \beta_{q-1\,q}(i_{q-1}, i_q) + \\ &\ldots \\ &+ \omega_{12\ldots q}(i_1, i_2, \ldots, i_q) \end{aligned}$$

o bien:

$$\log \; p_{\eta} \;=\; \mu \;+\; z_{\eta}$$

donde μ denota un efecto global; $\alpha_1(i_1)$ es el efecto debido a A_1 en el nivel i_1, se le llama efecto principal de A_1; $\alpha_q(i_q)$ es el efecto debido a A_q en el nivel i_q; $\beta_{12}(i_1,i_2)$ denota la interacción de 2º orden o efecto bivariado entre A_1 y A_2 a los niveles i_1, i_2 respectivamente y así sucesivamente hasta que $\omega_{12\ldots q}(i_1,i_2,\ldots,i_q)$ es el efecto de interacción de orden q o efecto q - variado entre las q variables $A_1, A_2, \ldots, A_q$ a los niveles $i_1, \ldots, i_q$.

Como $p_{\eta} < 1$, $\log P_{\eta}$ es negativo, pero μ no es fijo y debido a esto los efectos no están restringidos en signo.

Las restricciones ANOVA exigen que si se suman los efectos sobre todos los valores de un subíndice, la suma debe ser cero, por ejemplo:

$$\alpha_1(\,\cdot\,) \;=\; \sum_{i_1=1}^{I_1} \alpha_1(i_1) \;=\; 0$$

El punto denota suma sobre el subíndice donde aparece.

Se utiliza η para denotar un conjunto que incluya a las q variables, y η_r un subconjunto que incluya a ciertas variables, donde **r** es un número . Se puede indexar una celda con η y hablar de la frecuencia observada en la celda como x_η . Por ejemplo supóngase que el conjunto de variables $\sigma = \{A_1, A_2, \ldots, A_q\}$ se particiona en 2 subconjuntos σ_1 , σ_2 de tal manera que $\sigma_1 = \{A_1, A_2, \ldots, A_s\}$, $\sigma_2 = \{A_{s+1}, A_{s+2}, \ldots, A_q\}$ Entonces puede utilizarse η_1 y η_2 para denotar subconjuntos de índices que se refieran a las categorías de las variables incluídas en σ_1 y σ_2 respectivamente.

Dadas las convenciones anteriores, la probabilidad de caer en una celda cualquiera se expresa:

$$p_\eta = p_{i_1 i_2 \ldots i_q} = Pr\left(A_1 = i_1, A_2 = i_2, \ldots, A_q = i_q\right)$$

y representa la probabilidad de caer en la celda η . La formulación del modelo permite que se cumpla con que la suma de las probabilidades sea 1, o sea:

Se requiere que se cumplan el siguiente conjunto de restricciones:

$$\alpha_1(\cdot) = \quad . \quad . \quad . \quad = \alpha_q(\cdot) = 0$$

$$\beta_{12}(\cdot, i_2) = \quad . \quad . \quad . \quad = \beta_{q-1\,q}(i_{q-1}, \cdot) = 0$$

$$\omega_{12\ldots q}(\cdot, \ldots, i_q) = \ldots = \omega_{12\ldots q}(i_1, \ldots, \cdot) = 0$$

Un ejemplo de la expresión de la matríz de diseño en el caso saturado 2 x 2 para las variables A_1 y A_2 es el siguiente:

$$\begin{Bmatrix} \ell_{m\,11} \\ \ell_{m\,12} \\ \ell_{m\,21} \\ \ell_{m\,22} \end{Bmatrix} = \begin{Bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{Bmatrix} \begin{Bmatrix} \mu \\ \alpha_1(1) \\ \alpha_1(2) \\ \alpha_2(1) \\ \alpha_2(2) \\ \beta_{12}(1,1) \\ \beta_{12}(1,2) \\ \beta_{12}(2,1) \\ \beta_{12}(2,2) \end{Bmatrix}$$

En esta expresión no se han incluído las restricciones tipo ANOVA por lo que se observa que la matriz de diseño, la que contiene ceros y unos, es singular. Al incorporar las restricciones el modelo queda de la siguiente forma:

$$\begin{pmatrix} \ell_{m\,11} \\ \ell_{m\,12} \\ \ell_{m\,21} \\ \ell_{m\,22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1(1) \\ \alpha_2(1) \\ \beta_{12}(1,1) \end{pmatrix}$$

III. Estimación

Introduccion.

El estimar implica obtener valo-- res ajustados para cada una de las celdas en la tabla de contingencia, asi como los valores de los parámetros del modelo log-lineal. Para calcular estos valores se supone cierto comportamiento teórico caracterizado por el valor de la probabilidad de caer en las celdas; este valor depende de los parámetros incluídos dentro del modelo.

La tabla completa consta de celdas elementales y en ella, puede sumarse sobre uno o varios subíndices, estas sumas se agrupan en tablas de celdas no elementales que tienen menos dimensiones que el arreglo completo. Llamamos a estas tablas de sumas configuraciones y se denotan mediante C . Por ejemplo, supóngase que se tienen tres variables, el arreglo bidimensional obtenido de sumar sobre la 3a. variable, el conjunto $\{x_{i_1 i_2 \cdot}\}$ es la configuración C_{12} . Cada una de las $i_1 \times i_2$ celdas no elementales está formada por la suma de i_3 celdas elementales, dando como resultado que la configuración C_{12} involucra unicamen-

te a las variables 1 y 2.

Para ilustrar lo anterior, supóngase que se tienen datos acerca del envenenamiento después de un día de campo. Las variables son:

1) Envenenamiento { Presente / No presente

2) Ensalada de papas consumida { si / no

3) Jaiba consumida { si / no

La muestra consiste de 304 personas que asistieron al día de campo y los resultados del cuestionario que llenaron se presentan en la siguiente tabla: (6)

	Comida consumida			
	Jaiba si		Jaiba no	
	Ensalada de papas		Ensalada de papas	
Envenenamiento	si	no	si	no
si	120	4	20	0
no	80	31	24	23

Esta tabla consta de celdas elementales; p a r a interpretarla puede observarse que según esta tabla 4 personas de la muestra comieron jaiba, no comieron ensalada y presentaron envenenamiento. La configuración C_{12} representa la suma sobre la 3a. variable, jaiba, y es:

	Ensalada de papas	
Envenenamiento	si	no
si	142	4
no	104	54

Nótese que ha desaparecido la variable jaiba, cada una de las celdas representa la suma sobre los que comieron y no comieron jaiba. Por ejemplo, el valor 142 son los individuos que comieron ensalada y presentaron envenenemiento o sea 120 + 22.

Otro concepto importante dentro del análisis de tablas de contingencia es el de tablas incompletas,que se definen en términos del tipo de celdas vacias (con valor cero) que tengan. Los ceros de una tabla pueden clasificarse en:

i) Ceros muestrales o aleatorios: Surgen cuando la muestra no es lo suficientemente grande para obtener respuesta en una combinación particular de caracterfsticas.

ii) Ceros fijos o estructurales: Cuando una combinación especffica de atributos no ocurre en una población, estas celdas deben permane-

cer vacias, y esto a menudo se sabe a priori.

Las tablas con ceros estructurales se llaman incompletas y su análisis es complicado ya que debe imponerse la restricción que ciertas probabilidades sean cero.

Se limitará el estudio a las tablas completas que pueden o no presentar ceros muestrales. La presencia de ceros muestrales puede provocar que un modelo planteado no pueda estimarse; existen reglas para detectar cuando un modelo no es estimable (1).

Por otro lado, al escoger un modelo debe tenerse en cuenta que hay 2 límites del rango de modelos a ajustar.

Estos límites son:

i) El modelo saturado: en donde las log-frecuencias ajustadas encajan exactamente en las log-frecuencias observadas. Se tienen cero grados de libertad y la estadística que mide la bondad del ajuste toma el valor de 0.

ii) El modelo mínimo: contiene el conjunto mínimo de parámetros permitido. Para ciertos esquemas muestrales (Poisson y multinomial) solo incluye a μ, mien---

---

(1) Estas reglas están basadas en el examen de algunas configuraciones que surgen a partir de la tabla completa. Véase Kawasaki (1979) Capítulo IV.

tras que para otros (multinomial condicional y otros) incluye solamente a los parámetros asociados con las marginales fijas por el modelo.

El objetivo del análisis es encontrar un buen modelo situado entre éstos extremos que nos describa las relaciones que existen entre las variables del modelo. Es deseable que este modelo posea parámetros estadísticamente significativos, que presente un buen ajuste y que sea simple, ésto último se refiere a incluir el menor número de parámetros posible.

III.A Método de Estimación de Máxima Verosimilitud.

En esta sección se plantea la estimación del modelo log-lineal mediante el Método de Máxima Verosimilitud. Este implica que debe haber algún supuesto acerca de la distribución que genera las observaciones (frecuencias) en las celdas, es decir se supone una forma específica para la función de densidad f (x_i) donde x_i es la frecuencia en la i - ésima celda y f es una función de densidad -- - (f.d.p.).

Los esquemas muestrales que más frecuentemente se utilizan son:

a) Muestreo Poisson independiente

b) Muestreo Multinomial simple

c) Muestreo Multinomial condicional

El primero, Poisson, surge cuando para un período fijo se observa un conjunto de procesos Poisson y no está fijo el tamaño de la muestra. El multinomial se presenta cuando se fija el tamaño de muestra. El último surge cuando se imponen condiciones a los totales marginales.

Una vez planteado el modelo, se deberá maximizar la función Verosimilitud lo cual se logra mediante algún método numérico. A continuación se presenta el de ajuste proporcional iterativo.

III.B) Método de Ajuste Proporcional Iterativo.

Para aplicar este método es necesario establecer las estadísticas suficientes que corresponden al modelo log-lineal que va a ajustarse, se limita el ajuste a modelos log-lineales jerárquicos, es decir, modelos en los que la presencia de un término de orden mayor implica que los términos asociados a este de menor orden esten presentes, o bién, si un término de menor orden no se incluye, entonces todos los correspondientes términos de mayor orden relacionados con él están ausentes.

Las estadísticas suficientes son las configuraciones asociadas a cada parámetro del modelo log-lineal. Por ejemplo en el caso de 2 variables la configuración asociada al parámetro β_{12} es C_{12}.

Como se trata de modelos log-lineales jerárquicos es posible establecer un conjunto mínimo de estadísticas suficientes en donde se incluyen todas aquellas configuraciones que correspondan a los términos de orden mayor (s) en el modelo y examinar si a partir de este conjunto es posible obtener las demás estadísticas suficientes asociadas a los parámetros.

Si no, se examinan los términos de orden s - 1 que no estén

asociados a los de orden s, y se incluyen en el conjunto las configuraciones asociadas a estos.

Se continúa así hasta tener en el conjunto mínimo de estadísticas suficientes las configuraciones necesarias, para que a partir de éstas se obtengan todas las estadísticas suficientes asociadas con los parámetros del modelo log-lineal.

Por ejemplo, a partir de la configuración C_{12} es posible obtener la configuración $C_{1.}$ y $C_{.2}$.

El procedimiento que se describe a continuación parte de que, una vez definido el conjunto de estadísticas mínimas suficientes, las frecuencias esperadas del modelo se calculan igualando los totales marginales esperados con los valores correspondientes de las estadísticas suficientes.

Aunque en ocasiones puede ocurrir que las estimaciones en las celdas se escriben como una función directa de las estadísticas suficientes, lo que se denomina estimación directa; si se trabaja con un conjunto razonablemente grande de datos, no se requiere determinar si existe o no estimación directa, ya que éste método da las estimaciones automáticamente. El procedimiento iterativo que se expone ajusta las configuraciones suficientes ciclo a ciclo y posee las siguientes propiedades:

i) Siempre converge al conjunto único de estimaciones de M-V.

ii) Permite especificar el grado de precisión de las estimaciones en las celdas elementales, además de que éstos valores dependen de las estadísticas suficientes.

iii) Se puede escoger cualquier conjunto de estimaciones iniciales y si hay estimaciones directas el procedimiento converge en un ciclo.

En general, mediante éste método las celdas internas de la tabla son proporcionalmente ajustadas a un conjunto de marginales. La convergencia queda asegurada ya que la verosimilitud es una función monótona decreciente por lo que siempre pueden obtenerse estimaciones con cierta precisión fija de antemano.

La presencia de celdas vacías causa que todas aquellas celdas que formen una configuración mínima con alguna celda vacía, tendrán como estimación cero. Supongáse por ejemplo un modelo con tres variables en donde se va a ajustar la configuración mínima C_{12} y $x_{11\bullet} = 0$, entonces los valores ajustados para cada una de las celdas elementales que sumadas son $x_{11\bullet}$ serán cero, es decir $m_{11i_3} = 0$ para $i_3 = 1, 2, \ldots, I_3$.

1) Descripción del método

Supóngase que se han escogido como estadísticas mínimas

suficientes las configuraciones: C_{η_t} , donde $t = 1, \ldots, s$, con conteos celdales marginales x_{η_t} respectivamente. Se escoge un conjunto de estimaciones iniciales $\hat{m}_{\eta}^{(0)}$ y se ajusta cada - configuración en turno, utilizando en cada paso del ciclo la estimación obtenida en el paso anterior. Como hay que ajustar cada una de las configuraciones en cada ciclo, en el primer ciclo se ajustan s configuraciones, en el segundo ciclo s configuraciones más, dando un total de 2 s configuraciones ajustadas. Por lo tanto las configuraciones que se han ajustado al término de r ciclos son rs. Es por esto que al iniciar un nuevo ciclo las relaciones son:

$$\hat{m}_{\eta}^{(rs+1)} = \hat{m}_{\eta}^{(rs)} \frac{x_{\eta_1}}{\hat{m}_{\eta_1}^{(rs)}}$$

Siendo C_{η_1} la primera configuración a ajustar, o sea, el procedimiento vuelve a ajustar las s configuraciones consecutivamente. Al comienzo de este ciclo se tiene la estimación recien obtenida que es la que se utiliza para obtener una nueva estimación. La estimación correspondiente será la $rs + 1$ y se vuelven a ajustar todas las configuraciones, empezando con la η_1 . En general el K - ésimo es:

$$\hat{m}^{(K)} = \hat{m}^{(K-1)} \frac{x_{\eta_t}}{\hat{m}^{(K-1)}}$$

se sugiere utilizar como estimaciones iniciales $\hat{m}^{(0)} = 1$

2) Convergencia del método

Para estipular en qúe momento parar, es necesario considerar el cociente de verosimilitud después de ajustar k configuraciones:

$$D_{\eta}^{(K)} = \Sigma\, x_{\eta} \log x_{\eta} - \Sigma\, x_{\eta} \log \hat{m}_{\eta}^{(K)}$$

lo que interesa es maximizar la verosimilitud y la prueba de convergencia depende del cambio que ocurra en ésta. Las estimaciones maximizan el kernel de la log-verosimilitud o sea $\Sigma\, x_{\eta} \log \hat{m}_{\eta}$

sujeto a restricciones, éste valor aumenta a medida que aumenta k, y la frontera superior es $\Sigma\, x_{\eta} \log x_{\eta}$; por lo tanto, un buen criterio de convergencia es $D_{\eta}^{(K)}$. Debe determinarse que tán pequeña debe ser la diferencia entre ciclos para ser considerada insignificante. La regla para parar es estipular que los cambios ciclo a ciclo no excedan un valor pequeño δ .

Para el caso considerado de s pasos, es decir, s configuraciones mínimas, el cambio en la función de verosimilitud entre ciclos es:

$$\Delta D = D_{\eta}^{(K-S)} - D_{\eta}^{(K)}$$

IV. Diversas Pruebas

Cuando se ajusta un modelo normalmente se quiere tener alguna idea de la bondad de ajuste del mismo, la confiabilidad de los parámetros estimados, grado de asociación entre variables, etc. En el caso de los modelos log-lineales también son de interés algunos indicadores relativos a la confiabilidad de las estimaciones celdales.

A continuación se presenta algunos criterios para llevar a cabo lo anterior en el caso de los modelos log-lineales.

IV.A) Estadísticas de bondad de ajuste.

Para conocer la significancia estadística de un modelo, es decir, qué tanto se aproximan las estimaciones a los datos observados, se requiere obtener ciertas estadísticas. Estas estadísticas nos describen que tan bueno es un modelo, o sea su bondad de ajuste, comparando las frecuencias observadas contra las esperadas bajo un esquema particular.

Las estadísticas X^2 y G^2 que se describen a continuación tienen una distribución χ^2 asintótica cuando el modelo es correcto. Los grados de libertad de estas estadísticas pueden obtenerse de 2 maneras:

i) Contando el número total de parámetros independientes que se hacen cero.

ii) Contando el número total de parámetros estimados y restando este número del total de celdas estimadas.

I.A) Ji - Cuadrada de Pearson.

Se define como:

$$X^2 = \sum \frac{(x_i - \hat{m}_i)^2}{\hat{m}_i^2}$$

donde:

x_i es la observación en la celda i - ésima

$\hat{m}_i$ es la estimación para la celda i - ésima.

Esta estadística es una medida del ajuste global del modelo y es muy utilizada en el estudio de tablas de contingencia.

IV.B) Cociente de Verosimilitud.

Se expresa como:

$$G^2 = -2 \sum x_i \log \left(\frac{\hat{m}_i}{x_i}\right)$$

que es menos dos veces el cociente de verosimilitud. El valor de G^2 se interpreta como la probabilidad de que las diferencias entre las frecuencias observadas y ajustadas hayan sido aleatorias dado que el modelo sea correcto.

La estadística G^2 tiene una propiedad interesante que es la de poder descomponerse condicionalmente. Esto es, supóngase que se tiene un modelo (2) anidado dentro de otro modelo (1), es decir, el modelo 2 contiene una parte de los parámetros del modelo 1. Entonces la G^2 del modelo 2, denotada $G^2(2)$ que mide el ajuste de las estimaciones bajo el modelo 2,

puede descomponerse en:

i) Una medida del ajuste del modelo 1,

ii) Una medida condicional del modelo 2 dado el modelo 1. Dicho de otra manera, una medida de las distancias de las estimaciones $\{\hat{m}_i^{(2)}\}$ a las estimaciones $\{\hat{m}_i^{(1)}\}$ que se denota $G^2(2/1)$.

Para ilustrar lo anterior, se expone a continuación la descomposición de $G^2(2)$:

$$G^2(2) = -2 \sum x_i \left(\frac{\hat{m}_i^{(2)}}{x_i}\right)$$

$$= -2 \sum x_i \log \hat{m}_i^{(2)} + 2 \sum x_i \log \hat{m}_i^{(1)}$$

$$-2 \sum x_i \log \hat{m}_i^{(1)} + 2 \sum x_i \log x_i$$

$$= -2 \sum x_i \log \left(\frac{\hat{m}_i^{(2)}}{\hat{m}_i^{(1)}}\right) - 2 \sum x_i \log \left(\frac{\hat{m}_i^{(1)}}{x_i}\right)$$

$$= G^2(2/1) + G^2(1)$$

en esta expresión se observan las partes ii) e i) mencionadas que forman a $G^2(2)$.

De este resultado puede obtenerse:

$$G^2(2/1) = G^2(2) - G^2(1)$$

que puede utilizarse para verificar si los parámetros del modelo 2 le añaden significancia al modelo 1. Mediante esta fórmula pueden compararse los 2 modelos y lo que se prueba es si la adición de parámetros resulta en una reducción significativa de la $G^2(2)$. Para este efecto se calcula $G^2(2) - G^2(1)$ que se distribuye como una χ^2 con $\nu_2 - \nu_1$ grados de libertad, si se supone que los grados de libertad respectivos de $G^2(2)$ y $G^2(1)$ son ν_2 y ν_1.

IV.C) Transformación de Freeman - Tukey.

Cuando la muestra es pequeña no puede decirse cómo se distribuyen las estadísticas anteriores, ya que X^2 y G^2 se distribuyen como una χ^2 asintóticamente, es decir se aproximan a una χ^2 cuando el tamaño de la muestra es lo suficientemente grande. Una transformación sugerida para remediar este problema es la de Freeman - Tukey que trata de hacer las desviaciones estandarizadas en cada celda cercanas a desviaciones estandar con media cero y varianza unitaria. Como es sabido

si se suman estas desviaciones al cuadrado se obtendrá una distribución χ^2 . La transformación de Freeman - Tukey se define:

$$z_{i\cdot} = \sqrt{x_i} + \sqrt{x_i + 1} - \sqrt{4\hat{m}_i + 1}$$

que se utiliza para probar si el modelo es bueno así como para detectar observaciones aberrantes.. Cada una de las z_i tiene - aproximadamente una distribución Normal estándar y como tal se puede utilizar para ver qué celdas son aberrantes.

IV.D) Pruebas acerca de los parámetros.

Las estimaciones de M-V se obtienen al maximizar el kernel de la log-verosimilitud, éstas estimaciones se encuentran mediante el método iterativo que se ha presentado. La estimación $\hat{\underline{\theta}}$ está distribuída asintóticamente Normal y la correspondiente matriz de varianza-covarianza es la negativa de la inversa de la matriz Hessiana; esto es debido a que se trata de una estimación de M-V.

Dividiendo cada estimación por su correspondiente desviación estándar estimada se obtienen las estadísticas t correspondientes a cada parámetro.

Las hipótesis acerca de los parámetros en conjunto se establecen mediante pruebas de cociente de verosimilitud para muestras grandes. Este cociente de verosimilitud se define como:

$$\lambda = -2 \log \frac{L(\hat{\theta}_0)}{L(\hat{\theta}_1)}$$

en donde $L(\hat{\theta}_0)$ y $L(\hat{\theta}_1)$ son las funciones de verosimilitud bajo los parámetros $\hat{\underline{\theta}}_0$ y $\hat{\underline{\theta}}_1$ respectivamente. En este caso $\hat{\underline{\theta}}_0$ involucra parámetros extra a los impues-

tos bajo $\hat{\underline{\theta}}_1$. λ se distribuye como una χ^2 bajo la hipótesis nula que $\underline{\theta}$ es igual a $\hat{\underline{\theta}}_0$. El valor de λ es el valor de la prueba $G^2(2 \mid 1)$ con el modelo 1 ($\underline{\theta}_1$) anidado en el modelo 2 ($\underline{\theta}_0$), esto es para comparar dos modelos pero si se quiere probar la bondad de ajuste, se puede suponer que $\underline{\theta}_0$ es el vector que contiene a todos los parámetros posibles, definiendo así un modelo saturado, y entonces $\hat{m}_i = x_i \quad \forall \quad i$.

Para comparar dos modelos, otra estadística que puede utilizarse es:

$$R^2 = \frac{G^2(1) - G^2(2)}{G^2(1)}$$

que es análoga a la R^2 utilizada en regresión y mide el éxito del modelo 2 al predecir la probabilidad.

IV.E) Una Medida de Asociación.

Goodman y Kruskal (1954) propusieron una medida de asociación entre dos variables categóricas ordenadas (por ejemplo, variables que presentan las categorías alto, normal, bajo, o bien crece, constante, decrece; etc.). Esta medida nos da, una idea de la magnitud de la asociación y de su dirección. La dirección guarda relación con la concentración de individuos en las celdas, por ejemplo, considérense las tablas siguientes para el caso 3 x 3:

x_{11}	0	0
0	x_{22}	0
0	0	x_{33}

y

0	0	x_{13}
0	x_{22}	0
x_{31}	0	0

la primera tabla muestra asociación positiva perfecta mientras que la segunda muestra asociación negativa perfecta. La medida propuesta toma por definición los valores de + 1 y - 1 en cada caso respectivamente.

La base del coeficiente gamma de Goodman - Kruskal es un modelo probabilístico de comportamiento que se define a continuación.

Supóngase que dos individuos se seleccionan independientemente y en forma aleatoria de una población (con reemplazo, lo cual es innecesario en una población lo suficientemente grande). Cada uno de ellos cae en una celda de la tabla. Sea (i_1 , i_2) la celda en donde cae el 1º individuo y (i_1' , i_2') la del segundo.

Si hay independencia se espera que el orden de las celdas no tenga conexión; si hay asociación positiva se espera que los individuos se concentren en la diagonal indicada, esto es, si $i_1 > i_1'$ entonces $i_2 > i_2'$, o si $i_1 < i_1'$ entonces $i_2 < i_2'$, a esto se le llama órdenes iguales. En caso de asociación negativa se espera que si $i_1 > i_1'$ entonces $i_2 < i_2'$, o si $i_1 < i_1'$ entonces $i_2 > i_2'$, lo que se conoce como órdenes diferentes. Hay empate cuando $i_1 = i_1'$ ó $i_2 = i_2'$ en este caso no se puede decir que tipo de relación hay.

Sean:

$$\Pi_s = Pr \{ i_1 > i_1' \text{ e } i_2 > i_2' \text{ ó } i_1 < i_1' \text{ e } i_2 < i_2' \}$$

$$\Pi_d = Pr \{ i_1 > i_1' \text{ e } i_2 < i_2' \text{ ó } i_1 < i_1' \text{ e } i_2 > i_2' \}$$

$$\Pi_t = Pr \{ i_1 = i_1' \text{ ó } i_2 = i_2 \}$$

entonces la probabilidad condicional de órdenes iguales dado que no hay empates es $\Pi_s / (1 - \Pi_t)$ y la probabilidad condicional de órdenes diferentes dado que no hay empates es $\Pi_d /(1 - \Pi_t)$.

La gamma de Goodman - Kruskal se define como:

$$\gamma = \frac{\Pi_s - \Pi_d}{1 - \Pi_t}$$

que es la diferencia entre las probabilidades condicionales de órdenes iguales y diferentes. Mide qué tanto más probable es obtener órdenes iguales que diferentes en la tabla bivariada.

Como $\Pi_s + \Pi_d + \Pi_t = 1$ entonces se pueden simplificar los cálculos para incluir sólo a Π_s y Π_t . Haciendo:

$$\gamma = \frac{\Pi_s - \Pi_d}{1 - \Pi_t} = \frac{\Pi_s - [1 - \Pi_s - \Pi_t]}{1 - \Pi_t} = \frac{2\Pi_s - 1 - \Pi_t}{1 - \Pi_t}$$

con:

$$\Pi_s = 2 \sum_{i_1} \sum_{i_2} p_{i_1 i_2} \left\{ \sum_{i_1' > i_1} \sum_{i_2' > i_2} p_{i_1' i_2'} \right\} .$$

y

$$\Pi_t = \sum_{i_1} p_{i_1 \cdot}^2 + \sum_{i_2} p_{\cdot i_2} - \sum_{i_1} \sum_{i_2} p_{i_1 i_2} .$$

Las propiedades de esta medida son las siguientes:

i) γ es indeterminada si la población cae enteramente en un sólo renglón o columna de la tabla de clasificación cruzada.

ii) γ es 1 si la población está concentrada en la diagonal arriba - izquierda a abajo - derecha de la tabla . γ es -1 si la población está concentrada en la diagonal abajo - izquierda a arriba - derecha de la tabla.

iii) γ es cero en el caso de independencia, pero lo contrario no necesariamente se cumple excepto para el caso 2 x 2 . Para poder establecer propiedades estadísticas de γ es necesaria una expresión para la varianza.

Esta medida sólo se aplica a tablas bivariadas de contingencia, cuando se estudian tablas multivariadas hay una generalización debida a Kawasaki (1980) para medir asociación parcial, es decir, se mide la relación controlando la influencia de variables adicionales. A esta medida se le ha llamado la gama-componente γ_c.

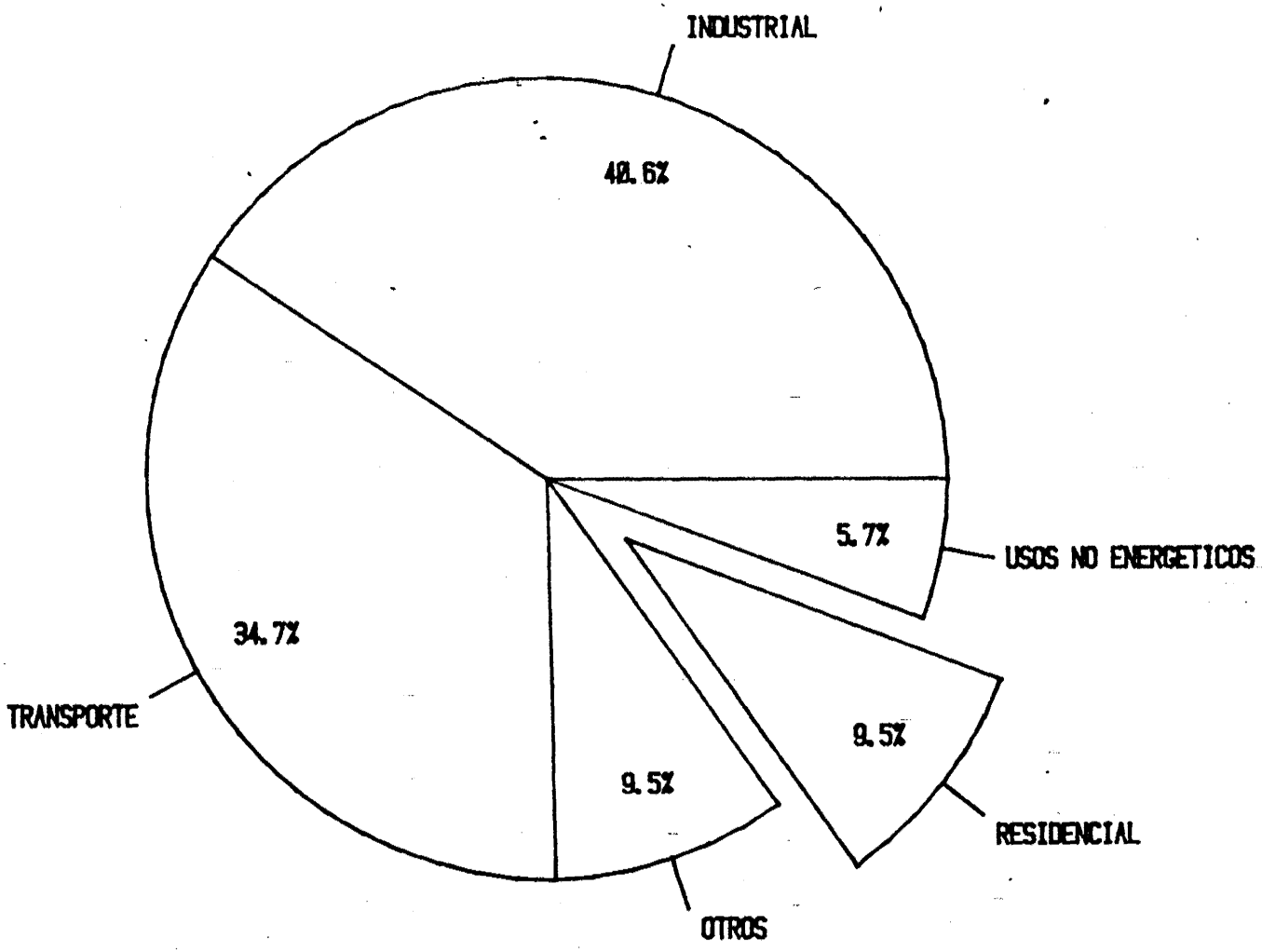
V. UNA APLICACION DEL MODELO

Introducción.

Se ha presentado el modelo log-lineal para analizar una tabla de contingencia. El objetivo de esta sección es plantear aquel modelo que resulte más adecuado para los datos bajo estudio, que presente las principales interrelaciones que existen entre las variables consideradas.

Dada la importancia que ha cobrado recientemente el estudio del consumo de energía, y como el sector residencial representa aproximadamente el 9% del gasto total en energía, (ver gráfica 1), el objetivo del análisis es ver qué relación guardan entre sí variables como el ingreso de la familia y su gasto en energía, si las familias con mayor ingreso gastan más energía, si los habitantes de zonas rurales se las arreglan con menos energéticos, si existe alguna relación entre el consumo de energía y el tamaño que tiene la vivienda, etc.

PARTICIPACION DE LA DEMANDA RESIDENCIAL EN LA DEMANDA TOTAL DE ENERGIA (1976)



A partir de estas interrogantes se pueden formular hipótesis de trabajo las cuales deberán probarse a partir de los datos. El objeto no es presentar al modelo log-lineal como la única posibilidad de análisis, sino señalar sus posibilidades y sus limitaciones. Evidentemente existen otras metodologías que, dependiendo del acceso a la información original, podrán arrojar mejores resultados o resultados complementarios; se pueden mencionar el análisis de varianza, modelos de regresión, detección automática de interacciones (AID), Analisis de Conglomerados (Cluster Analysis), etc.

Otro punto que conviene señalar es el relativo a la consideración del diseño muestral utilizado para obtener los datos. Existe una controversia entre dos grandes escuelas de estadísticos. Una señala que el diseño específico utilizado (estratificado, por conglomerados, etc.) debe incorporarse explícitamente en el análisis; la otra sostiene que esto no es necesario y que los análisis pueden llevarse a cabo como si se tratara de una muestra aleatoria. Para un comentario más a fondo, véase Tomberlin (1979 y 1980) Imrey y Sobel (1980) y Fay (1982). Evidentemente este problema existe tanto en relación con la aplicación del modelo log-lineal a datos de encuesta, como con la aplicación de cualquier otro método de análisis: ANOVA, etc.. Si se tienen los datos originales y una descripción detallada del diseño muestral es posible intentar

formas de análisis más completas, sin embargo es bien sabido que con mucha frecuencia lo único que tiene es un cuadro y una idea vaga de cómo se obtuvieron los datos.

Así pues, el análisis que a continuación se presenta, supone una muestra aleatoria y como tal debe tomarse como un primer análisis de los datos.

V.1 Datos Utilizados.

Los datos utilizados corresponden a la Encuesta de Ingreso-Gasto de los Hogares (S.P.P. 1976). El paquete utilizado para dividir las variables en categorías fue el S.P.S.S. (Statistical Package for the Social Science, versión 6, 1º abril 1975), considerándose las siguientes variables con sus respectivas categorías:

1) Zona
- Rural ($i_1 = 1$)
- Urbana ($i_2 = 2$)

La zona rural corresponde a municipios con menos de 100,000 habitantes y son las unidades primarias no autorepresentados (UPNAR) en la encuesta.

2) Número de cuartos
- 1 - 2 ($i_2 = 1$)
- 3 - más ($i_2 = 2$)

Para distinguir viviendas chicas y grandes. En esta variable, al igual que en las tres siguientes, se fijaron categorías hasta cierto punto arbitrarias, con el fin de aplicar los modelos log-lineales. Es bien conocido el hecho de que la elección de categorías influye en los resultados, Agresti (1976), por lo que nuestros resultados son condicionales en las categorías utilizadas.

3) Tamaño de la familia
- 0 - 3 hijos ($i_3 = 1$)
- 4 - más ($i_3 = 2$)

Estas categorías reflejan familias pequeñas y numerosas, ya que el número promedio de hijos es de aproximadamente 3.5 por pareja.

4) Ingreso (pesos/mes)
- 0 - 2400 ($i_4 = 1$)
- 2400 - 5725 ($i_4 = 2$)
- 5725 - más ($i_4 = 3$)

El ingreso así clasificado sirve para diferenciar ingresos bajos, medios y altos; estos niveles se consideran tomando en cuenta el salario mínimo vigente en esa época.

5)	Gasto en energéticos (pesos/mes)	0 - 50	(i_5 = 1)
		50 - 150	(i_5 = 2)
		150 - más	(i_5 = 3)

Esta categorización es un poco arbitraria pero intenta reflejar gastos bajos, medios y altos en energéticos. Para justificar esta clasificación puede observarse la tabla 1 que presenta para cada estrato de ingreso cual es su promedio de gasto en energéticos: para un nivel bajo de ingreso (0 - 2400) el gasto en energéticos fluctua entre 10.29 y 72.87, por lo tanto puede tomarse como referencia 50.00 que refleja de alguna manera un gasto medio aproximadamente. De esta manera pueden justificarse las otras divisiones.

TABLA I

PORCENTAJE DE GASTO EN ENERGETICOS
CON RESPECTO AL INGRESO

Estrato de Ingreso mensual	total de energéticos	gasto en energéticos *
0.1 - 700	2.94	10.29
700 - 1000	2.86	24.31
100 - 1350	2.80	32.9
1350 - 1800	3.22	50.71
1800 - 2400	3.47	72.87
2400 - 3150	3.89	107.95
3150 - 4300	3.66	136.34
4300 - 5725	3.89	194.99
5725 - 7500	4.35	287.64
7500 - 10150	4.57	403.30
10150 - 13400	5.32	626.43
13400 - 18000	4.64	728.48
18000 ó más	4.03	

Fuente: Encuesta Nacional de Ingresos y Gastos de los Hogares. 1977, S.P.P.

* Tomando el valor medio del intervalo como referencia.

TABLA 2

	i_1	1	1	1	1	2	2	2	2
	i_2	1	1	2	2	1	1	2	2
	i_3	1	2	1	2	1	2	1	2
i_4	i_5								
1	1	602	1178	170	425	151	136	61	37
1	2	103	228	83	190	105	132	92	83
1	3	15	25	21	38	15	19	23	26
2	1	47	207	18	142	115	168	52	69
2	2	54	251	67	360	170	507	211	601
2	3	11	80	36	194	30	177	95	395
3	1	9	19	8	22	18	27	39	40
3	2	16	52	35	150	55	155	195	60(
3	3	11	31	42	277	21	122	237	133

Variables consideradas

1) Zona
2) Número de cuartos
3) Tamaño de la familia
4) Ingreso
5) Gasto en energéticos

Los datos utilizados en este capítulo se presentan en la Tabla 2, en donde aparecen los valores en las celdas elementales. La tabla consta de 72 celdas que resultan de considerar todas las posibles combinaciones de las categorías (2x2x2x3x3). El número total de familias en la muestra es de 11539. No existe el problema de celdas vacias debido a la magnitud de la muestra y gracias a esto fue posible estimar cualquier tipo de modelo.

En las secciones que siguen se presentan los resultados obtenidos al ajustar modelos log-lineales, mediante el programa CALM (Conditional Anova Log-Linear Models, desarrollado por John P. Link, Northwestern University. Noviembre, 1980), que permite ajustar modelos jerárquicos conjuntos y que presenta entre los resultados que imprime:

a) Algunas medidas de bondad de ajuste: X^2 , G^2 , etc.

b) Un análisis paramétrico: para cada configuración incluída en el modelo se presentan las estimaciones de los parámetros y su correspondiente t , así como la gama componente de Kawasaki para cada configuración bivariada.

Mediante estos resultados se analizan los modelos y el objetivo de esta sección será encontrar aquel modelo que presente un buen ajuste y que comparado con los demás sea el más adecuado.

V.2 Resultados obtenidos.

V.2A) Algunos modelos jerárquicos.

Para facilitar la comprensión se utilizarán las siguientes abreviaturas al referirse a las variables:

1) ZONA (Zona)
2) NCUAR (Número de cuartos)
3) TFAM (Tamaño de la familia
4) INGR (Ingreso)
5) GASTO (Gasto en energéticos)

Con objeto de examinar la relación que guardan entre sí estas variables, se consideraron los siguientes modelos:

1) Modelo saturado: que incluye a todos los efec-

tos posibles entre las variables, involucra a la configuración C_{12345} como estadística suficiente.

2) Modelo con efectos de orden 4 : incluye todos los efectos hasta los de orden 4; las estadísticas suficientes son C_{1234} , C_{1235} , C_{1245} y C_{2345} .

3) Modelo con efectos trivariados: cuyas configuraciones mínimas suficientes son: C_{123} , C_{124} , C_{125} , C_{134} , C_{135} , C_{145} , C_{234} C_{235} , C_{245} , donde los efectos incluídos involucran relaciones entre las variables indicadas en cada configuración.

4) Modelo con efectos bivariados: que tiene como configuraciones mínimas C_{12}, C_{13} , C_{14} , C_{15} , C_{23} , C_{24} , C_{25} ,C_{34} , C_{35} , C_{45} .

5) Modelo con efectos univariados: solamente incluye como configuraciones: C_1 , C_2 , C_3 , C_4 y C_5 . Este supone que no existe re-

lación alguna entre las variables.

En la Tabla 3 se presentan los valores de la X^2 , G^2 , los grados de libertad y el nivel de significancia asociado al ajuste de cada modelo (en paréntesis). Como la hipótesis nula a probar es lo de que el modelo no ajusta bien, una G^2 significativa, la cual corresponde a un valor menor a .1 entre paréntesis, implica que el modelo correspondiente no es bueno. Se observa que en los modelos H_4 y H_5 la G^2 es altamente significativa. Esto es una indicación de que no basta con incluír sólo efectos principales e interacciones de orden dos, síno que habrá que incluir al menos algunas interacciones de orden tres.

Formalmente esto se puede hacer mediante el uso de las G^2, ya que éstas nos permiten decidir si los parámetros incluídos en un modelo y no en el otro hacen una contribución significativa al ajuste del modelo. Así, entre el modelo H_1 y el H_2 la diferencia está en que el primero incluye todas las interacciones, hasta las de orden cinco (saturado) mientras que el segundo sólo incluye hasta interacciones de orden cuatro; entonces G^2 (1/2)= G^2 (1) - G^2(2) = 1.62 tiene una distribución con 4 grados de libertad. La comparación con el valor de tablas, χ^2_4 (.05) = 7.779 indica que la contribución no es significativa. La Tabla 4 presenta

TABLA 3

MODELO	Configuraciones mínimas suficientes	X^2	G^2	g.l.
H_1	saturado	0.0 * (1.0)	0.0 * (1.0)	0
H_2	de orden 4	1.63 (.804)	1.62 (.805)	4
H_3	Trivariados	16.7 (.671)	16.8 (.668)	20
H_4	Bivariados	216 (0.0)	201 (0.0)	45
H_5	Univariados	18700 (0.0)	12500 (0.0)	66

* Valor obtenido 0.0

las comparaciones entre los modelos H_2 y H_3 y la de H_3 con H_4. El valor entre paréntesis corresponde al nivel de significancia.

TABLA 4

	G^2	g.l.
H_3 - H_4	184.2 (.005)	25
H_2 - H_3	15.18 (.5115)	16

Se tiene pues que la diferencia entre H_2 y H_3 no es significativa, es decir, las interacciones de orden 4 no son relevantes. Por el contrario, la diferencia de H_3 con H_4 es altamente significativa lo que implica que por lo menos algunas interacciones de orden tres sí son relevantes. Esto significa que se intentará encontrar un modelo entre H_3 y H_4.

V.2B) Selección del Modelo.

A partir de un modelo que incluya los efectos trivariados se probará excluyendo aquellos parámetros menos significativos, hasta llegar á aquél que posea un buen ajuste y un número menor de parámetros.

Teniendo como base el modelo con efectos trivariados se prueba excluyendo algunos de éstos parámetros que presentan menor significancia, de acuerdo a los valores χ^2 correspondientes. Las interacciones de orden 3 se denotan con la letra griega delta, δ.

En la Tabla 5 se presentan los resultados de excluir algunas configuraciones trivariadas. En ella se señalan varios modelos en los que se han omitido diversas configuraciones de orden tres. Estas configuraciones corresponden a los parámetros de las interacciones de orden tres. Así, en el modelo H_A se ha eliminado la interacción entre las variables 1, 2 y 5. La estadística G^2 permite analizar la bondad de ajuste de cada uno de ellos, al compararla con un valor de χ^2 obtenido de tablas, para los grados de libertad correspondientes, y que aparecen en la última columna de la tabla. Nuevamente una G^2 significativa

TABLA 5

Modelo	Configuraciones excluidas H_3	G^2	g. l.
H_3		16.8 (.668)	20
H_A	C_{125}	17.5 (.733)	22
H_B	C_{125} C_{235}	19.1 (.748)	24
H_C	C_{125}, C_{134}, C_{235}	27.3 (.395)	26
H_D	C_{125}, C_{234} C_{235}	22.4 (.668)	26
H_E	C_{125}, C_{134}, C_{235}, C_{234}	30.5 (.340)	28
H_F	C_{123}, C_{125}, C_{134}, C_{235}, C_{234}	39.9 (.108)	30

TABLA 6

Prueba	Parámetro Extra a Probar	G^2	g.l.
H_3 - H_A	δ_{125}	1.04 (.5945)	2
H_A - H_B	δ_{235}	1.6 (.4492)	2
H_B - H_C	δ_{134}	8.2 (.0166)	2
H_B - H_D	δ_{234}	3.3 (.1921)	2
H_D - H_E	δ_{134}	8.1 (.0174)	2
H_E - H_F	δ_{123}	9.4 (.0091)	2

implica que el modelo correspondiente no es bueno. El nivel de significancia aparece entre paréntesis. Observamos que todos los modelos ajustan bien con excepción del H_F , en el cual se ha eliminado el mayor número de interacciones. Como nos interesa obtener un modelo con el menor número de parámetros, parece ser que cualquiera en que se han eliminado 3 ó 4 interacciones podría ser adecuado.

Ahora veamos cuales de los parámetros específicos hacen una contribución significativa al ajuste del modelo. Para esto pasamos a la Tabla 6. En ella utilizamos la propiedad de que G^2 (A/B) = G^2 (A) - G^2 (B) como se hizo al comparar modelos en la Tabla 3. Nótese que al referirse en la Tabla 6 que se está probando el parámetro δ_{125} en realidad se están probando los 12 parámetros correspondientes a las combinaciones de los niveles de las variables 1, 2 y 5: δ_{125} (i, j, k), i = 1,2 j = 1,2 y k = 1,2,3. Los dos grados de libertad se deben a que, con las restricciones tipo ANOVA, al fijar dos cualesquiera de los parámetros, los restantes (diez) quedan fijos. Entonces se tiene que:

1) H_A es preferible a H_3, o sea que la eliminación de los parámetros δ_{125} no reduce en forma significativa el ajuste del modelo.

2) H_A es preferible a H_B, o sea que la interacción δ_{235} sí contribuye significativamente al ajuste del modelo.

3) H_B es preferible a H_C; no se puede excluir la configuración (interacciones de orden tres) C_{134} del modelo B sin afectar su ajuste.

4) H_D es preferible a H_B; la combinación de interacciones C_{234} se puede eliminar del modelo B.

5) Con base en lo anterior, se eliminan simultáneamente las interacciones que se indica en el modelo H_E y se compara con el H_D resultando que H_D es preferible a H_E. No se puede eliminar la configuración C_{134}.

6) El modelo H_D es preferible al H_3 ya que G^2 (D/3) = G^2 (D) - G^2 (3) = 22.4 - 16.8 = 5.6 y tiene

6 g.l.

7) Se hace luego el intento de eliminar más interacciones de orden 3; por ejemplo la C_{123}, y esto resulta en el modelo F. Aquí la conclusión es que el modelo E es preferible al F.

8) El modelo finalmente seleccionado fue el H_D, del cual se excluyen las siguientes interacciones:

i) ZONA x NCUAL x GASTO

ii) NCUAR x TFAM x INGR

iii) NCUAR x TFAM x GASTO

No se consideró seguir eliminando parámetros ya que los demás efectos trivariados son significativos, y aunque el efecto bivariado β_{23} no es significativo, no es posible eliminarlo dada la estructura jerárquica implícita. Un intento de eliminarlo, es mediante el modelo H_F que excluía a la configuración C_{123}, pero este modelo se rechazó en favor de H_E.

V.2C) Interpretación de los resultados.

El modelo seleccionado en función de los parámetros que contiene es:

$$\ell_{m\, i_1 i_2 i_3 i_4 i_5} = \mu + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5$$

$$+ \beta_{12} + \beta_{13} + \beta_{14} + \beta_{15} + \beta_{23} + \beta_{24} + \beta_{25} + \beta_{34} + \beta_{35} + \beta_{45}$$

$$+ \delta_{123} + \delta_{124} + \delta_{135} + \delta_{145} + \delta_{245} + \delta_{345} + \delta_{134}$$

El cual contiene 8 efectos menos que el modelo saturado. Puede utilizarse

$$G^2(1) - G^2(D) / G^2(1) = 12\,500 - 22.4 / 12\,500 = .9986$$

o sea que el 99.9% de la variación se explica mediante este modelo, lo cual representa un buen ajuste.

Se puede ver de la Tabla 7 que en general los valores observados y los estimados son muy semejantes, lo cual no es de sorprender dado el buen ajuste del modelo. Por otra parte los valores de la estadística Freeman-Tukey no muestran ningún comportamiento anormal; el valor más alto, 1.40, está dentro de los límites razonables para una normal estándar.

T A B L A 7

i_4	i_5	i_1 i_2 i_3	1 1 1	1 1 2	1 2 1	1 2 2	2 1 1	2 1 2	2 2 1	2 2 2
1	1		602.0	1 478.0	170.0	425.0	151.0	[illegible]	[illegible]	[illegible]
			591.06	1 185.38	172.21	426.34	164.45	[illegible]	58.25	[illegible]
			(.45)	(- .21)	(- .15)	(- .05)	(-1 .05)	(.59)	(.05)	(-.15)
1	2		103.0	224.0	53.0	190.0	105.0	182.0	92.0	[illegible]
			114.08	220.63	79.5	189.79	104.17	129.21	85.25	93.47
			(-1.04)	(.51)	(.42)	(.03)	(.11)	(.27)	(.74)	(4.05)
1	3		15.0	25.0	21.0	38.0	15.0	19.0	23.0	25.0
			16.13	23.71	21.02	38.14	14.56	19.59	22.9	26.55
			(- .22)	(.31)	(-.04)	(.02)	(.18)	(-.08)	(.20)	(-.04)
2	1		47.0	207.0	16.0	142.0	115.0	168.0	52.0	[illegible]
			50.11	213.94	23.91	126.04	104.14	166.82	63.84	77.2
			(-.41)	(-. 46)	(-1.23)	(1.40)	(1.06)	(-.04)	(-.22)	(-.95)
2	2		54.0	251.0	67.0	360.0	170.0	[illegible]	231.0	[illegible]
			52.08	248.9	62.47	368.55	168.51	512.51	248.84	564. 1
			(.30)	(.15)	(.59)	(-.43)	(.13)	(-.25)	(-.52)	(-.50)
2	3		11.0	80.0	36.0	194.0	30.0	[illegible]	[illegible]	[illegible]
			13.38	71.59	31.04	204.98	36.52	[illegible]	97.25	382.72
			(-.60)	(.99)	(.89)	(-.76)	(-1.05)	(.59)	(.42)	(.46)
3	1		9.0	19.0	8.0	22.0	18.0	27.0	[illegible]	[illegible]
			8.48	18.76	8.23	22.51	20.26	25.49	37.04	41.22
			(.25)	(.11)	(0.0)	(-.06)	(-.46)	(.04)	(-.65)	(-.15)
3	2		16.0	52.0	35.0	150.0	55.0	[illegible]	[illegible]	[illegible]
			15.39	53.93	34.48	149.2	48.11	100.57	248.02	599.8
			(.21)	(-.23)	(.13)	(.09)	(.99)	(-.42	(-.55)	(.25)
3	3		11.0	31.0	42.0	277.0	21.0	122.0	[illegible]	1 [illegible]
			7.29	34.13	47.13	272.44	19.45	124.10	257.10	1 336.52
			(1 .29)	(- .50)	(-.73)	(.29)	(.39)	(-.17)	(.00)	(.64)

Valor Observado
Valor Ajustado
(componente Freeman-Tukey)

T A B L A 7

i_4	i_5	i_1 / i_2 / i_3	1 / 1 / 1	1 / 1 / 2	1 / 2 / 1	1 / 2 / 2	2 / 1 / 1	2 / 1 / 2	2 / 2 / 1	2 / 2 / 2
1	1		602.0	1 178.0	170.0	425.0	151.0	136.0	61.0	37.0
			591.06	1 185.38	172.21	426.34	164.45	126.10	56.28	38.17
			(.45)	(- .21)	(- .15)	(- .05)	(-1 .05)	(.89)	(.65)	(-.15)
1	2		103.0	228.0	83.0	190.0	105.0	132.0	92.0	83.0
			114.08	220.63	79.5	189.79	104.17	129.11	85.25	93.47
			(-1.04)	(.51)	(.42)	(.03)	(.11)	(.27)	(.74)	(-1.09)
1	3		15.0	25.0	21.0	38.0	15.0	19.0	23.0	26.0
			16.13	23.71	21.02	38.14	14.56	19.66	22.9	26.55
			(- .22)	(.31)	(-.04)	(.02)	(.18)	(-.08)	(.20)	(-.06)
2	1		47.0	207.0	18.0	142.0	115.0	168.0	52.0	69.0
			50.11	213.94	23.91	126.04	104.14	168.82	53.84	77.2
			(-.41)	(-. 46)	(-1.23)	(1.40)	(1.06)	(-.04)	(-.22)	(-.93)
2	2		54.0	251.0	67.0	360.0	170.0	507.0	211.0	601.0
			52.08	248.9	62.47	368.55	168.51	512.51	218.94	589.04
			(.30)	(.15)	(.59)	(-.43)	(.13)	(-.23)	(-.52)	(.50)
2	3		11.0	80.0	36.0	194.0	30.0	177.0	95.0	395.0
			13.38	71.59	31.04	204.98	36.32	176.71	91.25	392.72
			(-.60)	(.99)	(.89)	(-.76)	(-1.05)	(.59)	(.41)	(.13)
3	1		9.0	19.0	8.0	22.0	18.0	27.0	39.0	40.0
			8.48	18.78	8.23	22.51	20.26	25.49	37.04	41.22
			(.25)	(.11)	(0.0)	(-.06)	(-.46)	(.34)	(.36)	(-.15)
3	2		16.0	52.0	35.0	150.0	55.0	155.0	195.0	606.0
			15.39	53.93	34.48	149.2	48.11	160.57	203.02	599.3
			(.21)	(-.23)	(.13)	(.09)	(.99)	(-.42	(-.55)	(.28)
3	3		11.0	31.0	42.0	277.0	21.0	122.0	237.0	1 337.0
			7.29	34.13	47.13	272.44	19.48	124.10	237.10	1 336.32
			(1 .29)	(- .50)	(-.73)	(.29)	(.39)	(-.17)	(.01)	(.03)

Valor Observado
Valor Ajustado
(componente Freeman-Tukey)

La Tabla 8 presenta los valores de los parámetros estimados para el modelo, correspondientes a los efectos principales y a las interacciones de orden dos. Por la dificultad en interpretar las de orden mayor éstas sólo se presentan en forma de listado en un apéndice al final del artículo. Junto con los parámetros correspondientes a las interacciones se ha incluído la gama-componente que mide el grado de asociación entre las dos variables a que se refiere la interacción, eliminando el efecto de los demás. Conviene señalar que la gama-componente sólo tiene sentido en el caso de que ambas variables involucradas sean ordinales, cosa que es válida en todas las variables, pues aún la distinción entre zona rural y urbana implica un orden en cuanto al tamaño de la localidad.

A partir de esta tabla se observa que todos los efectos principales son significativos; esto se determina a partir de los valores t que aparecen entre paréntesis. Esto quiere decir que cada una de ellas contribuye a determinar el número de personas que cae en cada celda. Los resultados para las interacciones indican que con excepción de ZONA x NCUAR y NCUAR x TFAM todos son significativos. La gama-componente de los restantes son significativos y positivos lo cual es de esperarse. La interacción ZONA x TFAM es negativa, indicando que hay más familias pequeñas en la zona urbana y más familias numerosas en la zona rural.

TABLA 8

ECTOS PRINCIPALES

Rural	Urbana
-.1849	.1849
(10.63)	(10.63)

DE CUARTOS (NCUAR):

Chica (de 1 a 2)	Grande (3 y más)
-.1721	.1721
(10.153)	(10.153)

DE FAMILIA (TFAM):

Pequeña (de 0 a 3 hijos)	Numerosa (4 o más hijos)
-.4759	.4759
(28.446)	(28.446)

INGRESO (INGR):

Bajo	Medio	Alto
.0509	.3531	-.3840
(1.230)	(15.739)	(13.820)

GASTO (GASTO):

Bajo	Medio	Alto
-.1210	.5155	-.3944
(4.648)	(24.171)	(14.669)

INTERACCIONES BIVARIADAS

ZONA x NCUAR:

	Chica (1-2)	Grande (>2)
Rural	.0311 (2.161)	-.0311 (2.161)
Urbana	-.0311 (2.161)	+.0311 (2.161)

GAMA COMP. = .0622 (2.167)

ZONA x TFAM:

	Pequeña (0-3)	Numerosa (>3)
Rural	-.1338 (9.508)	.1338 (9.508)
Urbana	.1338 (9.508)	-.1338 (9.508)

GAMA COMP. = -.2613 (9.968)

ZONA x INGRESO:

	Bajo	Medio	Alto
Rural	.5353 (6.975)	-.1512 (6.975)	-.4051 (14.680)
Urbana	-.5563 (23.024)	.1512 (6.975)	.4051 (14.680)

GAMA COMP. = .5646 (26.209)

ZONA x GASTO:

	Bajo	Medio	Alto
Rural	.3078 (12.219)	-.1380 (6.834)	-.1698 (6.614)
Urbana	-.3078 (12.219)	.1380 (6.834)	.1698 (6.614)

GAMA COMP. = .3072 (10.915)

NCUAR x TFAM:

	Pequeña (0-3)	Numerosa (>3)
Chica (1-2)	.0110 (.888)	-.0110 (.888)
Grande (>2)	-.0110 (.888)	.0110 (.888)

GAMA COMP. = .0220 (.888)

NCUAR x INGR:

	Bajo	Medio	Alto
Chica	.3430 (14.621)	.0943 (4.846)	-.4373 (17.3[illegible])
Grande	-.3430 (14.621)	-.0943 (4.846)	.4373 (17.3[illegible])

GAMA COMP. = .4777 (20.799)

NCUAR x GASTO:

	Bajo	Medio	Alto
Chica	.4225 (18.346)	-.0158 (.872)	-.4067 (17.[illegible])
Grande	-.4225 (18.346)	.0158 (.872)	.4067 (17.[illegible])

GAMA COMP. = .5032 (23.5[illegible])

TFAM x INGRESO:

	Bajo	Medio	Alto
Pequeña	.2993 (12.499)	-.1409 (6.945)	-.1584 (4.125)
Numerosa	-.2993 (12.499)	.1409 (6.945)	.1584 (4.1[illegible])

GAMA COMP. = .2637 (9.390)

TFAM x GASTO:

	Bajo	Medio	Alto
Pequeña	.1824 (7.660)	-.0359 (1.861)	-.1465 (5.575)
Numerosa	-.1824 (7.660)	+.0359 (1.861)	.1465 (5.575)

GAMA COMP. = .2158 (7.577)

INGR x GASTO:

	Bajo	Medio	Alto
Bajo	1.0226 (31.465)	-.1179 (4.033)	-.9047 (22.2[illegible])
Medio	-.1598 (5.032)	.1103 (4.22[illegible])	.0495 (1.4[illegible])
Alto	-.8632 (20.150)	.0075 (.232)	.[illegible] (23.1[illegible])

Se pueden resumir los hallazgos a partir de las interacciones y los coeficientes gama - componente como sigue:

i) el gasto en energía está áltamente relacionado con el nivel de ingreso.

ii) el ingreso varía de acuerdo a la zona.

iii) el gasto en energéticos y el número de cuartos están relacionados entre sí.

iv) el número de cuartos que posee la vivienda y el nivel de ingreso están relacionados.

v) el tamaño de la familia guarda poca relación con el ingreso y con el gasto, pues las gamas son pequeñas aunque significativas.

vi) el gasto depende de la zona.

vii) las familias de las zonas rurales tienen más hijos, mientras que las de las zonas urbanas menos.

viii) el número de cuartos y el tamaño de la familia no tienen ninguna relación.

Otro tipo de análisis que se puede hacer a partir del modelo es con base en los "momios" para variables con dos categorías. Estos son el cociente de la probabilidad de caer en una ca-

tegoría entre la de caer en la otra. Con base en el modelo se representa de la siguiente manera para la variable ZONA:

$$\frac{m_{1\,i_2\,i_3\,i_4\,i_5}}{m_{2\,i_2\,i_3\,i_4\,i_5}} = \frac{e^{\alpha_1(1)}}{e^{\alpha_2(2)}} \cdot \frac{e^{\beta_{12}(1,i_2)}}{e^{\beta_{12}(2,i_2)}} \cdot$$

$$\frac{e^{\beta_{13}(1,i_3)}}{e^{\beta_{13}(2,i_3)}} \cdot \frac{e^{\beta_{14}(1,i_4)}}{e^{\beta_{14}(2,i_4)}} \cdot \frac{e^{\beta_{15}(1,i_5)}}{e^{\beta_{15}(2,i_5)}} \cdot \frac{e^{\delta_{123}(1,i_2,i_3)}}{e^{\delta_{123}(2,i_2,i_3)}} \cdot$$

$$\frac{e^{\delta_{124}(1,i_2,i_4)}}{e^{\delta_{124}(2,i_2,i_4)}} \cdot \frac{e^{\delta_{135}(1,i_3,i_5)}}{e^{\delta_{135}(2,i_3,i_5)}} \cdot \frac{e^{\delta_{145}(1,i_4,i_5)}}{e^{\delta_{145}(2,i_4,i_5)}}$$

$$\frac{e^{\delta_{134}(1,i_3,i_4)}}{e^{\delta_{134}(2,i_3,i_4)}}$$

que se interpreta como el "chance" relativo de que una familia caiga en la categoría rural en lugar de en la urbana, dependiendo de los niveles de las otras variables.

El valor de los momios correspondientes se presentan en la Tabla 9 para cada uno de los valores de los índices para las variables TFAM, NCUAR, INGR, GASTO.

Esta tabla se interpreta de la siguiente manera, por ejemplo si se observa en la Tabla el valor de la celda para i_2=1 i_3 = 1, i_4 = 1, i_5 = 1 se ve que hay aproximadamente 4 familias de la zona rural por una de la zona urbana que caen en esta celda (ya que el valor del momio es de 3.75).

De acuerdo a esta tabla si la familia presenta un ingreso medio superior es más probable pertenecer a la zona urbana que a la rural y tener un consumo de energía medio o superior. También se observa que es más probable que para una familia de la zona rural su patrón de gasto de energía sea menor (los momios más altos se observan en gasto bajo en energéticos y nivel de ingreso bajo). Otra característica que hacen resaltar los momios es que las familias de bajos ingresos se concentran en las zonas rurales, así como que en general el tamaño de la familia sea mayor en las zonas rurales (que en la tabla se observa ya que los momios de estar en la categoría 4 hijos ó más son en general mayores que los de estar en la categoría de 1 a 3 hijos para un mismo tamaño de vivienda, ingreso y gasto en energía).

En la Tabla 10 se presenta un análisis similar utilizando los momios de NCUAR. Se observa que el tamaño de la vivienda aumenta con el nivel de ingreso.

TABLA 9

MOMIOS PARA RURAL/URBANA

Gasto en Energéticos	Ingreso	NCUAR			
		$i_2 = 1$		$i_2 = 2$	
		TFAM			
		$i_3 = 1$	$i_3 = 2$	$i_3 = 1$	$i_3 = 2$
$i_5 = 1$	$i_4 = 1$	3.59	9.407	3.059	11.182
	$i_4 = 2$	0.481	1.268	0.444	1.635
	$i_4 = 3$	0.467	0.690	0.222	0.547
$i_5 = 2$	$i_4 = 1$	1.094	1.709	0.884	2.032
	$i_4 = 2$	0.528	0.486	0.271	0.626
	$i_4 = 3$	0.546	0.336	0.170	0.249
$i_5 = 3$	$i_4 = 1$	1.108	1.211	0.943	1.439
	$i_4 = 2$	0.369	0.406	0.340	0.734
	$i_4 = 3$	0.374	0.275	0.199	0.204

TABLA 10

MOMIOS DE TAMAÑO DE LA VIVIENDA CHICA /GRANDE

Gasto en Energéticos	Ingreso	ZONA			
		$i_1 = 1$		$i_1 = 2$	
		TFAM			
		$i_3 = 1$	$i_3 = 2$	$i_3 = 1$	$i_3 = 2$
$i_5 = 1$	$i_4 = 1$	3.433	2.781	2.923	3.305
	$i_4 = 2$	2.183	1.697	1.934	2.186
	$i_4 = 3$	1.072	0.833	0.547	0.618
$i_5 = 2$	$i_4 = 1$	1.435	1.162	1.222	1.382
	$i_4 = 2$	0.869	0.676	0.770	0.870
	$i_4 = 3$	0.446	0.361	0.237	0.268
$i_5 = 3$	$i_4 = 1$	0.766	0.621	0.653	0.738
	$i_4 = 2$	0.431	0.349	0.398	0.450
	$i_4 = 3$	0.154	0.125	0.082	0.093

TABLA 11

MOMIOS DE TAMAÑO DE LA FAMILIA PEQUEÑA/NUMEROSA

Gasto en Energéticos	Ingreso	ZONA			
		$i_1 = 1$		$i_1 = 2$	
		NCUAR			
		$i_2 = 1$	$i_2 = 2$	$i_2 = 1$	$i_2 = 2$
$i_5 = 1$	$i_4 = 1$	0.498	0.404	1.306	1.475
	$i_4 = 2$	0.236	0.191	0.617	0.698
	$i_4 = 3$	0.451	0.365	0.794	0.898
$i_5 = 2$	$i_4 = 1$	0.517	0.418	0.807	0.912
	$i_4 = 2$	0.209	0.170	0.329	0.372
	$i_4 = 3$	0.285	0.231	0.300	0.339
$i_5 = 3$	$i_4 = 1$	0.680	0.551	0.743	0.840
	$i_4 = 2$	0.187	0.152	0.206	0.233
	$i_4 = 3$	0.214	0.173	0.157	0.178

Al comparar los momios entre las zonas rural y urbana estos aumentan para las familias pequeñas o sea que en la zona urbana hay una tendencia al hacinamiento.

La Tabla 11 contiene los momios para el tamaño de la familia. En ella se nota que en general predominan las familias numerosas, con excepción de los niveles de ingreso bajo en la zona urbana. Los que más cerca le siguen son los niveles de ingreso alto en la misma zona.

En la zona urbana el número de familias numerosas aumenta con el ingreso, mientras que en la rural el aumento se da al pasar de niveles de ingreso bajos al nivel medio, pero disminuye nuevamente al pasar a ingresos altos. El número de cuartos no parece tener ningún efecto. Lo cual es consistente con lo señalado a partir de la Tabla 10.

V.2D) Conclusiones

Hemos presentado una aplicación del modelo log-lineal para el análisis de los datos sobre ingresos y gastos familiares, con especial interés en el gasto en energéticos. Hemos confirmado

algunas hipótesis como son el que el gasto depende del nivel de ingreso, del número de cuartos y de la zona, mas no del número de miembros en la familia. Esto es congruente con el hecho de que a ciertos niveles de ingreso existe hacinamiento en viviendas de uno y dos cuartos.

Algunas de las variables utilizadas son de tipo "intervalo", y se categorizaron para este estudio. Un análisis más profundo, y ya con la información que hemos obtenido con el presente podría llevarse a cabo modificando los criterios de clasificación o bien, si estuviera disponible la información original, utilizando alguna de las técnicas alternativas.

Hemos querido ilustrar la aplicación del modelo log-lineal a datos económicos y contribuir a su difusión como un método más dentro del análisis econométrico, concientes de que tiene limitaciones, pero que en el caso de variables que por su naturaleza son categóricas tal vez sea lo más adecuado.

APENDICE I

TABLA 8 (Continuación)

Parámetros Estimados del Modelo H_D
Efectos Trivariados

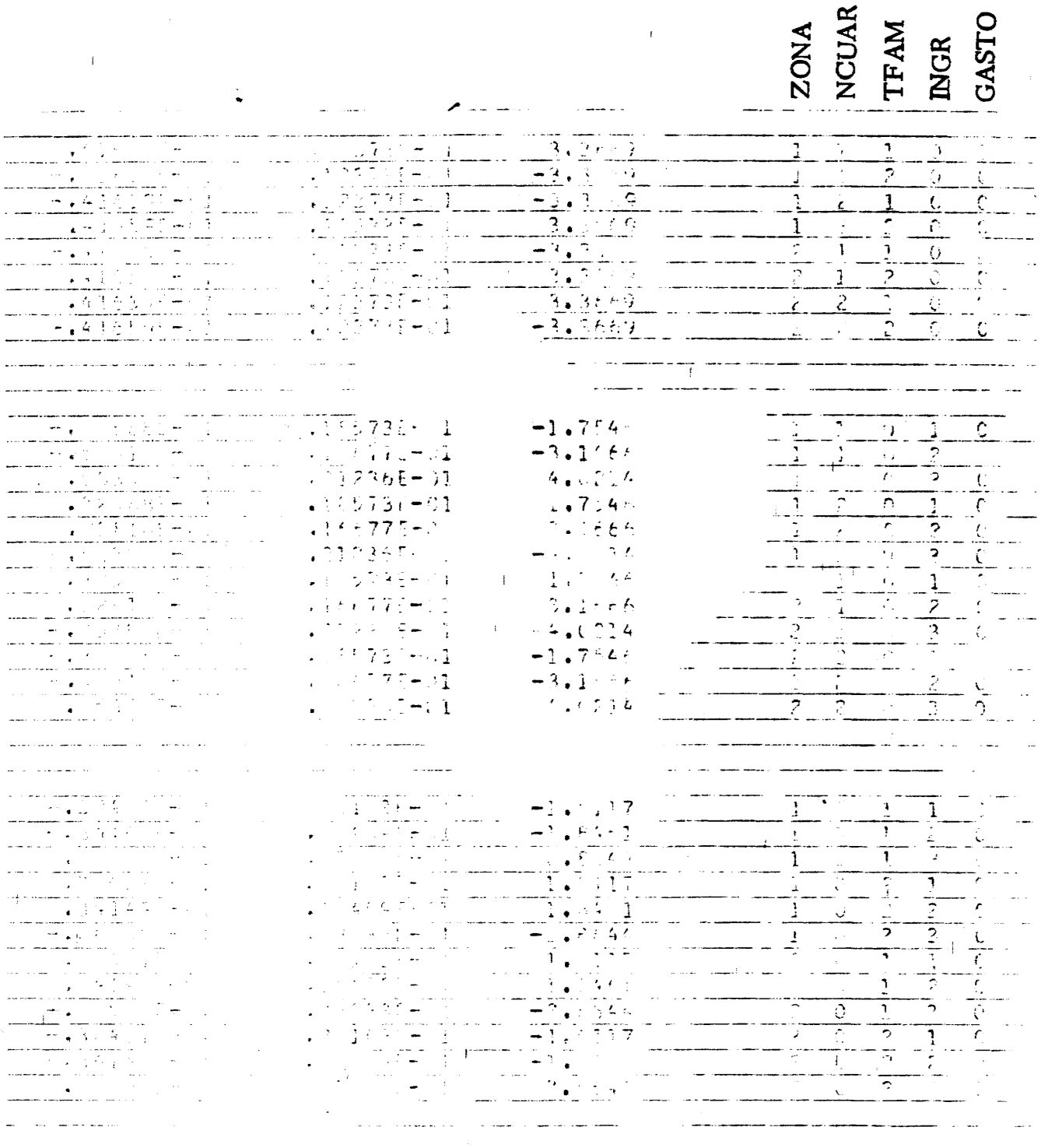


TABLA 8 (Continuación)
Parámetros Estimados del Modelo H_D
Efectos Trivariados

			ZONA	NCUAR	TFAM	INGR	GASTO
[illegible]	[illegible]	-[illegible]	1	0	[illegible]	[illegible]	1
[illegible]	[illegible]	.77[illegible]7[illegible]	1	0	1	0	[illegible]
[illegible]	[illegible]	[illegible].3213	1	0	1	0	3
[illegible]	[illegible]	[illegible]	1		2		1
[illegible]	[illegible]	-.7[illegible]7[illegible]	1		2		[illegible]
[illegible]	[illegible]	-4.[illegible]	1	0	2	0	3
[illegible]	[illegible]	[illegible]	2	0	1	0	1
[illegible]	[illegible]	-.[illegible]			1		
[illegible]	[illegible]	-6.[illegible]			1	0	2
[illegible]	[illegible]	-4.4[illegible]1	2	2	2	0	1
[illegible]	[illegible]	.[illegible]72	2	0	2		
[illegible]	[illegible]	[illegible]	2		2	0	2
[illegible]	[illegible]	[illegible]	1		[illegible]		1
[illegible]	[illegible]	-.7[illegible]	1	0	[illegible]	1	
[illegible]	[illegible]	-3.[illegible]	1	0	0	1	3
[illegible]	[illegible]	-2.46[illegible]	1	0	0	2	1
[illegible]	[illegible]	.[illegible]	1			2	2
[illegible]	[illegible]	1.[illegible]42	1			2	3
[illegible]	[illegible]	-2.1[illegible]	1	0	0	3	1
[illegible]	[illegible]	1.8[illegible]	1	0	0	3	2
[illegible]	[illegible]	[illegible].37[illegible]	1			3	3
[illegible]	[illegible]	-2.[illegible]7[illegible]	2	2	2	1	1
[illegible]	[illegible]	2.72[illegible]7	2		0	1	2
[illegible]	[illegible]	3.[illegible]	2			1	[illegible]
[illegible]	[illegible]	2.[illegible]	2	0	0	2	1
[illegible]	[illegible]	-.2[illegible]	2		0	2	2
[illegible]	[illegible]	-1.[illegible]4[illegible]	2	0	0	2	3
[illegible]	[illegible]	2.1295	2		0	3	1
[illegible]	[illegible]	-1.[illegible]	2			3	2
[illegible]	[illegible]	-2.[illegible]	2	0		3	[illegible]

TABLA 8 (Continuación)

Parámetros Estimados del Modelo H_D
Efectos Trivariados

			ZONA	NCUAR	TFAM	INGR	GASTO.
-.27[illegible]	.[illegible]504E-01	-.93102	[illegible]	1	0	1	1
-.25[illegible]	.[illegible]6610E-01	-.94507	[illegible]	1	0	1	2
.52[illegible]	.[illegible]81518E-01	1.3796[illegible]	0	1	0	1	3
-.23[illegible]	.[illegible]8173E-01	-.10153	[illegible]	1	0	2	1
-.27[illegible]	.[illegible]21695E- 1	-1.[illegible]	[illegible]	1	0	2	2
.33[illegible]	.[illegible]8115E-01	1.1833	1	1	0	2	3
.32[illegible]	.[illegible]85495E-01	.85213	0	1	0	3	1
.2[illegible]	.27740E-01	1.9114	0	1	0	3	2
-.[illegible]	.[illegible]2[illegible]E-01	-2.[illegible]	[illegible]	1	0	3	3
.274[illegible]	.[illegible]9504E-01	.93102	0	2	0	1	1
.25[illegible]	.[illegible]6610E-01	.94507	0	2	0	1	2
-.[illegible]	.[illegible]8114E- 1	-1.[illegible]	[illegible]	2	0	1	3
.[illegible]	.[illegible]1173E- 1	.19153	0	2	0	2	1
.27[illegible]	.[illegible]8169E-01	1.[illegible]	0	2	0	2	2
-.[illegible]	.[illegible]8115E-01	-1.1[illegible]	0	2	0	2	3
-.[illegible]	.[illegible]4[illegible]E-01	-.[illegible]	0	[illegible]	0	3	1
-.[illegible]	.[illegible]7405E-01	-1.[illegible]	[illegible]	2	0	3	2
.[illegible]	.[illegible]3[illegible]E-01	[illegible]	0	2	0	3	3
-.1[illegible]	.[illegible]1[illegible]E-[illegible]	-[illegible]	[illegible]	[illegible]	1	1	1
-.17[illegible]	.[illegible]4[illegible]E-01	-.[illegible]	[illegible]	0	1	1	2
.14[illegible]	.[illegible]2[illegible]E-01	3.[illegible]	[illegible]	[illegible]	1	1	3
-.[illegible]	.[illegible]5[illegible]E-01	-.[illegible]	[illegible]	[illegible]	1	2	1
.22[illegible]	.[illegible]3[illegible]E-01	.94[illegible]	[illegible]	0	1	2	2
-.1[illegible]	.[illegible]6521E-01	-.41[illegible]	[illegible]	[illegible]	1	2	3
.1[illegible]	.[illegible]E- 1	[illegible].4[illegible]	[illegible]	[illegible]	[illegible]	3	1
-.[illegible]	.[illegible]- 1	-.1[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2
-.1[illegible]	.[illegible]1[illegible]7E-01	-4.[illegible]	[illegible]	[illegible]	1	3	3
.1[illegible]	.[illegible]1[illegible]E-01	[illegible].[illegible]	[illegible]	[illegible]	2	1	1
.[illegible]	.[illegible]1[illegible]E-01	.[illegible]	[illegible]	[illegible]	2	1	2
-.[illegible]	[illegible]	-[illegible].[illegible]	[illegible]	[illegible]	2	1	3
.1[illegible]	.[illegible]E- 1	.[illegible]	0	0	2	2	1
-.[illegible]	.[illegible]E-01	-.[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2
.[illegible]	.[illegible]E- 1	[illegible].[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3
-.1[illegible]	.[illegible]E- 1	-[illegible].[illegible]	[illegible]	[illegible]	[illegible]	3	1
.[illegible]	.[illegible]E- 1	.[illegible]	[illegible]	[illegible]	2	3	2
.1[illegible]	.[illegible]E- 1	[illegible].[illegible]	[illegible]	[illegible]	[illegible]	3	3

BIBLIOGRAFIA

1) Agresti, A., The Effect of Category Choice on Some Ordinal Measures of Association, Journal of the American Statistical Asociation, Vol. 71, pp 49 - 53.

2) Bishop, Y.M.M., Fienberg, E.S., Holland, P.W. Discrete Multivariate Analysis: Theory and Practice. Ed. MIT Press. Sexta Edición, Cambridge, Massachusetts (1980).

3) Fay, R., Contingency Table Analysis for Complex Sample Designs: CPLX, Proceedings of the Survey Research Section ASA. (1982).

4) Haberman, Shelby. Analysis of Qualitative Data. Vol. I Introductory Topics. Ed. Academic Press. Nueva York. (1978).

5) Imrey, P.B., y E. Sobel, Modeling Contingency Tables From Complex Surveys, Proceedings of the Survey Research Section ASA (1978).

6) Kawasaki, Seiichi. Application of Log-Linear Probability Models in Econometrics. Tesis doctoral presentada en la Universidad de Northwestern. Departamento de Economía (1979).

7) Link, John P. CALM. Conditional ANOVA Log-Linear Models: User's Guide. Universidad de Northwestern (1980).

8) Maxwell, A.E. Análisis Estadístico de Datos Cualitativos. Ed. UTEHA (Manuales 222) traducción al español por Ramón Galhd Garza. México (1965).

9) Mendoza, Y., El Modelo Log-Lineal y su Aplicación. Tesis Profesional para obtener el Título de Licenciado en Matemáticas Aplicadas (1984). (ITAM)

10) Nerlove, M., Press, S.J. Univariate and Multivariate Log-Linear and Logistic Models. Preparado para la Administración del Desarrollo Económico y el Instituto Nacional de Salud. RAND CORPORATION - R - 1306 - EDA&NIH. (diciembre, 1973).

11) Nerlove, M., Press S.J. Multivariate Log-Linear Probability Models for the Analysis of Qualitative Data. Artículo para discusiónNum. 1. Universidad de Northwestern. Centro de Estadística y Probabilidad (1976).

12) Nerlove, Marc. Expectations, Plans and Realizations in Theory and Practice. Econométrica, Vol. 51. Num 5, (septiembre, 1983).

13) Payne, C. "The Log-Linear Model for Contingency Tables". Tomado de The Analysis of Survey Data: Exploring Data Structures. Ed. John Wiley and Sons. Nueva York (1977) pags. 106.

14) Samaniego, R., Berndt, E. Residential Energy Demand in México. publicado por el Instituto de Tecnología de Massachussetts. Laboratorio de Energía. (Agosto, 1982).

15) Tomberlin, T.J., The Analysis of Contingency Tables of Data From Complex Surveys, Proceedings of the Survey Research Section, ASA (1979).

16) Tomberlin, T.J., A Model - Based Approach to the Analysis of Contingency Tables of Data From Complex Surveys, Proceedings of the Survey Research Section, ASA (1980).

Serie Documentos de Trabajo 1984

No. I Alberro, José Luis, "Introduction and Benefit of Technological Change under Oligopoly".

No. II Serra-Puche, Jaime y Ortíz, Guillermo, "A Note on the Burden of the Mexican Foreign Debt".

No. III Bhaduri, Amit, "The Indebted Growth Process".

No. IV Easterly, William, " Devaluation in a Dollarized Economy".

No. V Unger, Kurt, "Las Empresas Extranjeras en el Comercio Exterior de Manufacturas Modernas en México".

No. VI De Alba, Enrique y Yolanda Mendoza, "El Uso de Modelos Log-Lineales para el Análisis del Consumo Residencial de Energía".

Serie Documentos de Trabajo 1983

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

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

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

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

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

No. VI Lustig, Nora "Políticas de Consumo Alimentario: Una Comparación de los Efectos en Equilibrio Parcial y Equilibrio General".

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

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

No. IX Nadal, Alejandro "La Construcción del Concepto de Mercancía en la Teoría Económica".

No. X Cárdenas, Enrique "Some Issues on Mexico's Nineteenth Century Depression".

No. XI Nadal, Alejandro "Dinero y Valor de Uso: La Noción de Riqueza en la Génesis de la Economía Política".

No. XII Blanco, Herminio y Garber, Peter M. "Recurrent Devaluation and Speculative Attacks on the Mexican Peso".

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

Editor: José Luis Alberro

Serie Documentos de Trabajo 1982

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

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

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

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

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

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

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

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

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

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