CENTRO LATINOAMERICANO DE DEMOGRAFIA
SECTOR DE FECUNDIDAD

EL USO DEL ANALISIS DE CLASIFICACION MULTIPLE (MCA) EN LA DEMOGRAFIA

Una introducción del método con un ejemplo de su aplicación

Johanna M. de Jong

S.143/26/74.
(Mayo de 1974)
50.

I N D I C E

Indice de cuadros

R E S U M E N

Este trabajo está concebido como una introducción al análisis de clasificación múltiple (MCA) que es parecido al análisis de regresión pero en el que está permitido que las variables predictivas estén medidas con escala nominal y ordinal (además de escala cardinal).

Se hace un análisis con datos del estudio PECFAL-Rural de Costa Rica, comparando la fuerza predictiva en conocimiento de anticonceptivos de un conjunto de variables socio-económicas y otro de variables que miden actitudes. Se concluye que este último conjunto predice mejor el conocimiento y logra especificar cuáles combinaciones de factores se relacionan altamente con el conocimiento.

Como el programa de computación del MCA de que dispone el CELADE en el sistema OSIRIS, produjo muchos problemas, decidimos con el programador Abel Packer, quien hizo todo el trabajo de computación, reescribir el manual, explicando más extensamente las partes en que habíamos cometido errores. Este manual va como apéndice del trabajo.

I. INTRODUCCION

Para relacionar fenómenos sociales representados por variables, hay varios tipos de análisis que se distinguen básicamente en dos grupos por el tratamiento de las variables explicativas. En uno de ellos se especifica la relación con el objeto de estudio (variable dependiente) mediante la sub-división de las categorías de una variable explicativa en cada una de las adicionales. El ejemplo más sencillo de esta manera de analizar es la tabla de entrada doble, triple, etc.

Esta subdivisión ocasiona una disminución grande de casos, por la combinación de categorías cada vez que se agrega una variable explicativa adicional, lo que deriva en una disminución de la confiabilidad de los resultados. Sin embargo, tiene la ventaja de conservar toda la varianza de los datos en forma accesible, lo que facilita el estudio de casos especiales. Estas características hacen que se elija este método de análisis, principalmente con fines explorativos.

En el otro tipo de análisis se considera cada vez el valor predictivo de la variable explicativa en su totalidad y también cada variable, agregada para mejorar la predicción, está tomada en forma entera tanto en su relación con la variable dependiente como con las demás variables explicativas. Este método es la base del análisis de regresión al que nos referiremos.

Para explicar o predecir un cierto fenómeno, es más fructuoso tomar este último camino que posibilita la introducción de muchas variables explicativas sin pérdida de confiabilidad. De gran importancia es esto porque en la realidad social los fenómenos a estudiar casi siempre están interrelacionados con muchos otros, lo que hace necesario usar modelos bastante complicados para llegar a una explicación. El artículo presente se enfocará en uno de ellos derivados del análisis de regresión.

Por lo general este tipo de análisis explicativo tiene la desventaja de usar estadísticas de resumen para indicar las relaciones que, por su característica de cálculo, ocultan mucho de la varianza de los datos. El modelo está basado en los coeficientes de correlación, medida que usa el promedio; para que un promedio tenga sentido, la variable debe poder ser medida en escala de intervalo. (Por ejemplo, no tiene sentido sacar un promedio de "religión" codificado según distintas afiliaciones, tales como 1 católica, 2 protestante, 3 judío, 4 budista. Un promedio general de 1,2 no tiene sentido).

Desafortunadamente, en el campo de las ciencias sociales las variables explicativas no son por lo general cuantitativas, sino cualitativas y su medición se basa en escalas nominales o a lo sumo ordinales. Además, la varianza de los datos es por lo general grande y la relación entre una variable predictiva y la variable

dependiente pocas veces es lineal, hecho que no se ve reflejado en el análisis de regresión.

A pesar de estas discrepancias entre los datos y los supuestos del modelo, se aplica muy a menudo el análisis de regresión, lo que muestra su atractivo que reposa, repetimos, en la posibilidad de usar muchas variables predictivas y la interpretación fácil de los resultados tanto en su relación con la variable dependiente como con las demás variables predictivas, para establecer la fuerza relativa de predicción.

Sin embargo, de existir las fallas descritas, la regresión múltiple muy pocas veces puede dar más que una aproximación global a las relaciones verdaderas entre las variables de índole social, así que para hacer predicciones en este campo había que pensar en adaptaciones de los métodos de análisis existentes o en métodos nuevos. Como la regresión es muy atractiva por la simplicidad de su modelo y la comparabilidad de la fuerza predictiva de las variables usadas, no es de sorprender que se haya buscado adaptar este método de análisis y recientemente se ha presentado el Análisis de Clasificación Múltiple (MCA). Este método está desarrollado específicamente para ser aplicado a datos susceptibles de ser medidos con cualquier tipo de escala, sea cardinal, ordinal o nominal, y para aceptar relaciones no-monótonas.

II. METODOLOGIA

El MCA[1/] es una forma de análisis en que se relacionan varias variables predictivas con una sola variable dependiente en una tentativa de explicar la varianza de ésta. El modelo de predicción resultante permite la respuesta de preguntas del tipo siguiente:

1) ¿Cuán fuerte es la relación entre cada una de las variables predictivas y la variable dependiente?

2) ¿Qué parte de la relación queda vigente cuando las demás variables han sido tomadas en cuenta?

3) ¿Hasta qué punto pueden ser predecidas las variables dependientes usando las variables predictivas en conjunto?

4) ¿Son estas relaciones lo suficientemente fuertes como para ser significativas estadísticamente (con la muestra usada)?

5) ¿Qué puntaje en la variable dependiente puede esperarse para una persona que caiga dentro de ciertas categorías de las variables predictivas?

Como se ve, el tipo de problemas para los que sirve el MCA es parecido a los de la regresión múltiple. La diferencia fundamental se encuentra en el tipo de medición de los datos, cardinal en la regresión vs. cualquier tipo en el MCA, lo cual es reflejo de una base distinta: no se usa en el MCA una estadística resumen de toda la variable, sino que se toman en cuenta todas las categorías de una variable predictiva como si cada categoría fuera una variable independiente.

Por lo tanto, puede permitirse además que los valores que tomen estas categorías varíen en forma no-lineal sin que este hecho importante se oculte tras una medida resumen; cada categoría contribuye con su propio valor independientemente de los demás valores en la predicción y la variable en su totalidad muestra un poder explicativo basado en la contribución ponderada de cada una de sus categorías.

Como lo muestra la pregunta dos antes indicada, el análisis se ocupa de establecer la influencia "neta" de cada una de las variables predictivas, porque casi siempre hay correlación entre ellas. Precisamente se quiere saber lo que queda de la relación original entre las variables independiente y dependiente, después de restar la parte que en realidad está causada por otras variables con que ésta está correlacionada. (Más adelante se discutirá una sola excepción a esta aceptación de correlaciores en el MCA).

El MCA trata de imponer un modelo preconcebido a los datos; por lo tanto es lógico que los datos estén sujetos a limitaciones. Una exigencia es que la variable dependiente sea dicotómica o esté medida en una escala cardinal. En el primer caso el promedio puede considerarse como la proporción de personas que tienen la característica representada por aquella o no la tienen. Otra es que las variables tengan un valor aditivo en cuanto a predicción; la interacción de las variables hace que los resultados del análisis pierdan el sentido. En su forma más sencilla, hay interacción en una variable dependiente cuando la relación entre las variables dependiente e independiente varía según el valor de la variable de control. Lo siguiente sirve como ejemplo. En una muestra de mujeres casadas y convivientes se encuentra una baja relación entre conocimiento de anticonceptivos y su uso. Cuando controlamos por número de hijos, resulta que entre mujeres con cuatro hijos o más la relación entre conocimiento y uso de anticonceptivos es bastante fuerte, mientras que entre las mujeres con menos de cuatro hijos, la relación es casi inexistente. Para evitar que este tipo de problemas dificulten la aplicación del MCA, se aconseja transformar el conjunto de variables causantes de la interacción (número de hijos y conocimiento de anticonceptivos en nuestro ejemplo) en una variable nueva con categorías significativas. Siguiendo nuestro ejemplo, puede pensarse en una variable de cuatro categorías:

1 menos de cuatro hijos y conoce anticonceptivos
2 menos de cuatro hijos y no conoce anticonceptivos
3 cuatro hijos o más y conoce anticonceptivos
4 cuatro hijos o más y no conoce anticonceptivos.

Tal tipo de variable compuesta imposibilita la atribución del valor predictivo de cada uno de sus dos componentes. Sin embargo, tampoco tendría sentido, porque ninguno de los dos actúa por sí mismo.

En tercer lugar, la correlación entre dos categorías de dos variables predictivas no debe ser del ciento por ciento: no es lícito que todas las personas que caigan en una categoría de una variable predictiva también caigan en una sola categoría de otra y viceversa (todas las personas A_1 son también B_1 y todas las B_1 son también A_1). Este traslapo total de dos categorías de dos variables predictivas causa un problema sin solución por la siguiente razón. Entre personas que

tengan indistintamente dos características (en este caso A_1 y B_1) es imposible establecer cuál de las dos es la causante de las diferencias encontradas en la variable dependiente, o cómo está proporcionada la influencia de cada una; ambas categorías tienen iguales derechos. Por eso es aconsejable recodificar o aislar categorías con un traslapo total.

III. UN EJEMPLO DE APLICACION

A continuación se presenta un ejemplo de aplicación de MCA para mostrar más detalladamente su funcionamiento y la interpretación de sus resultados. Para los cálculos, se usó el programa MCA del sistema OSIRIS. Los detalles de la aplicación se presentan en el Apéndice.

Al usar los datos de Costa Rica de la encuesta PECFAL-Rural,[2] nos proponemos hacer una comparación entre grupos de variables predictivas, uno socio-económico y el otro de actitudes hacia el rol de la mujer, para establecer su importancia relativa en la predicción del conocimiento de anticonceptivos (variable dependiente). Como gran parte de las preguntas se hicieron tan solo a las mujeres en unión, usamos este sub-conjunto de la muestra.

1. Variables usadas

En primer lugar, queremos ver cuál es la influencia de las variables socio-económicas (variables 1 hasta 4 del Cuadro 1). Estas son de diferente tipo de escalas de medición. La edad de la mujer corresponde a una escala cardinal; sin embargo, no necesariamente es fácil usar esta variable en un análisis de regresión, porque es probable que su relación con la variable dependiente no sea lineal. En realidad parece en nuestro análisis que la relación con el conocimiento de anticonceptivos es ligeramente curvilínea con un conocimiento mayor en la categoría 20-29 años.

La educación de la mujer corresponde a una escala ordinal: se sabe, por ejemplo, que "primaria completa" es más que "sin educación" pero no exactamente cuánto más. Por lo tanto, una medida basada en el promedio no tiene mucho sentido.

Las variables "lugar de socialización" y "la vivienda tiene luz y/o agua" se miden con una escala nominal, aunque a veces pueden ser ordenadas sus categorías según su comportamiento en el análisis para ser medidas con una escala pseudo ordinal. Para esto ya hay que conocer con anticipación la distribución de las respuestas de esta variable en la variable dependiente, procedimiento que habría que repetir para cada variable dependiente.

En segundo lugar, medimos el impacto de las variables de actitudes (variables 5 hasta 7 del Cuadro 1).[3] A pesar de que estas variables han sido codificadas en cinco categorías que van desde "muy en pro" hasta "muy en contra" o sus equivalentes, la agrupación ha sido hecha a veces en forma arbitraria, por falta de conocimiento, así que la distancia entre las categorías no es conocida.

Cuadro 1

VARIABLES PREDICTIVAS USADAS Y RESULTADOS DEL MCA EN TRES EJEMPLOS SOBRE CONOCIMIENTO DE ANTICONCEPTIVOS

Variables		Número de casos (A)	Promedio (B)	Promedio ajustado Ej.1 a/ (C)	Ej.2 b/ (D)	Ej.3 c/ (E)
1) Lugar socialización:	campo	960	0,63	0,66		0,66
	pueblo	178	0,71	0,66		0,66
	ciudad	192	0,79	0,70		0,68
	no responde	6	0,00	0,25		0,20
2) Luz y/o agua:	nada	461	0,54	0,57		0,57
	luz	76	0,78	0,75		0,74
	agua	322	0,62	0,62		0,64
	luz y agua	469	0,80	0,77		0,75
	no responde	8	0,12	0,34		0,81
3) Educación:	Sin educación	237	0,46	0,49		0,52
	-3 años primaria	552	0,65	0,66		0,67
	4 años prim. incompl.	306	0,72	0,71		0,70
	Primaria completa	170	0,79	0,76		0,72
	Secundaria inc. o más	367	0,92	0,82		0,74
	No responde	4	0,00	0,73		0,53
4) Edad:	15-19	102	0,56	0,53		0,59
	20-24	258	0,70	0,67		0,64
	25-29	263	0,70	0,68		0,66
	30-34	248	0,67	0,68		0,68
	35-39	209	0,63	0,66		0,66
	40-44	146	0,64	0,66		0,69
	45-49	110	0,65	0,68		0,71
5) Aspiraciones para hija:						
	muy altas	242	0,81		0,76	0,71
	altas	541	0,67		0,66	0,66
	neutral	172	0,66		0,67	0,68
	bajas	57	0,56		0,57	0,63
	muy bajas	324	0,55		0,60	0,63
6) Emancipación mujer:	muy en pro	198	0,83		0,78	0,76
	poco en pro	272	0,68		0,65	0,64
	neutral	548	0,64		0,66	0,67
	poco en contra	196	0,61		0,62	0,63
	muy en contra	122	0,52		0,57	0,58
7) Comunicación entre esposos:						
	muy alta	164	0,82		0,79	0,76
	alta	289	0,84		0,82	0,82
	neutral	308	0,67		0,68	0,68
	baja	164	0,72		0,72	0,72
	muy baja	407	0,45		0,46	0,48
Promedio general de conoc. de A.C.			0,66			
Desviación estándar			0,47			
Parte explicada de la suma de los cuadrados				32,32	43,38	60,88
Suma total de los cuadrados			298.75			

a/ Variables socio-económicas.

b/ Variables de actitudes.

c/ Todas las variables.

La variable dependiente, conocimiento de anticonceptivos, se mide dicotómicamente: 0) No conoce; 1) Sí conoce; resultando de esta manera de medir una proporción de mujeres que conoce anticonceptivos.

Con estos predictores hemos hecho tres análisis; siempre con la misma variable dependiente que es conocimiento de anticonceptivos: en el ejemplo 1 entran las variables socio-económicas, en el ejemplo 2 las variables de actitudes y en el ejemplo 3, ambos grupos de variables. Ahora vemos qué tipo de cálculos hace el MCA para llegar a resultados y qué tipo de interpretación podemos dar basada en ellos.

2. Datos no-variables por variable dependiente

Como la predicción está basada en el promedio general de la variable dependiente y los promedios de éste por categoría de cada variable predictiva, se calculan:

a) $\bar{Y}$ promedio general (0,66 en nuestro caso)

$\bar{Y}_{ij}$ promedio de la variable dependiente que obtuvo la categoría j de la variable i (por ejemplo en el ejemplo 1, la categoría "campo" de la variable "lugar de socialización", 0,63. El 63 por ciento de las mujeres en esta categoría tiene conocimiento de anticonceptivos). Estos promedios $\bar{Y}_{ij}$ no son ajustados por la correlación entre variables y muestran, por consiguiente, el efecto bruto de la pertenencia a una categoría dada.

b) La desviación estándar de $\bar{Y}$ (0,47) que en sí da una idea de la varianza de los datos.

c) La suma total de los cuadrados de las desviaciones T (298.75) que constituye la totalidad de la varianza de la variable dependiente a explicar.

Cada vez que se haga un análisis de la misma variable dependiente con el mismo conjunto de casos, aunque con diferentes variables predictivas, el promedio general $\bar{Y}$, los promedios $\bar{Y}_{ij}$, la desviación estándar y la suma total de los cuadrados de las desviaciones, son idénticas porque se refieren a la varianza de la misma variable dependiente y porque no se ha hecho ningún ajuste por correlaciones entre variables. (En el Cuadro 1 se ve reflejada esta similitud: tan solo los datos de los ejemplos 1, 2 y 3 que son los promedios ajustados por la correlación entre variables, son distintos).

3. Datos variables por variable dependiente

A nivel de predicción global, los resultados que dependen del conjunto de variables predictivas usadas son:

a) La parte explicada de la suma de los cuadrados de las desviaciones (E) por el conjunto de categorías de las variables predictivas usadas. (En el ejemplo 1 es 32.32).

b) El residuo de la suma de los cuadrados de las desviaciones (T), lo que representa la parte no explicada de la varianza por el conjunto de categorías de las variables predictivas usadas (en el ejemplo 1: 266.43).

El cuociente $\frac{E}{T}$ $(= R^2)$ es la proporción de la varianza explicada por este conjunto de variables usadas en la predicción. (En el ejemplo 1: $\frac{32,32}{298,75} = 0,108$, este conjunto de variables socio-económicas explica el 10,8 por ciento de la varianza en el conocimiento de anticonceptivos, según los datos de la muestra).

c) La contribución a la varianza total explicada de cada una de las variables predictivas y cada una de sus categorías. Como ya se ha dicho, partiendo de los promedios de cada categoría (Cuadro 1, columna B), se llega a la contribución neta de cada categoría en las columnas C, D y E, tomando en cuenta la interrelación entre las variables. Esta contribución neta puede expresarse en la desviación del promedio global o en el promedio ajustado. (Por ejemplo la desviación neta de la categoría "ciudad" de la variable "lugar de socialización" en el ejemplo 1, es de 0,04; como el promedio general es de 0,66, el promedio "neto" o ajustado correspondiente es de 0,70). Aquí presentamos los promedios ajustados porque nos parece que son más fáciles de interpretar. En el Cuadro 1 se nota que los ajustes están basados en el conjunto de las variables predictivas: por ejemplo la misma categoría "secundaria incompleta o más" de la variable "nivel de educación de la mujer", tiene en el ejemplo 1 un promedio de 0,82 y en el ejemplo 3 de 0,74. Esto quiere decir que del promedio de un 92 por ciento (columna B) de conocimiento de anticonceptivos entre las mujeres con algunos años de enseñanza secundaria, una parte está causada por el hecho de que estas mismas mujeres también se encuentran en ciertas categorías de otras variables socio-económicas que predicen un alto grado de conocimiento de anticonceptivos (baja de 0,92 hasta 0,82). El ejemplo 3 muestra que al tomar en cuenta las variables tanto socio-económicas como de actitudes, el promedio ajustado baja aún más: las mujeres con ciertas características socio-económicas además están ubicadas en algunas categorías de las variables de actitudes que predicen un alto grado de conocimiento.

El promedio bruto original en la categoría de "algunos años de enseñanza secundaria o más" de un 92 por ciento, ocultaba entonces la influencia de otras variables socio-económicas y además de variables de actitudes. Al tomar en cuenta otras variables significativas, la influencia pura de la educación se reduce notablemente. Dicho de otra manera, las mujeres con relativamente mucha educación son más propensas a tener otras características que se relacionen con un alto conocimiento de anticonceptivos: se hace notar un efecto aditivo.

Sabiendo cuáles son los datos invariables y cuáles las variables, veremos la función de cada uno para entender los resultados de un análisis.

4. Resultados no ajustados por categoría

Consideraremos las variables en el orden en que aparecen el el Cuadro 1, excluyendo por el momento las categorías "no responde" que por su bajo número de

casos y su interrelación están sujetos a grandes variaciones. Poniendo en un cuadro la desviación mínima y máxima del promedio general por categorías de cada una de las variables predictivas (acordando que para el MCA no hace ninguna diferencia dónde están ubicadas estas categorías dentro de la variable, porque ellas están consideradas como si no estuvieran relacionadas entre sí), se obtienen los resultados que se muestran en el Cuadro 2.

Cuadro 2

DESVIACION MAXIMA Y MINIMA DEL PROMEDIO GENERAL DE CONOCIMIENTO DE ANTICONCEPTIVOS DE LAS CATEGORIAS DE 7 PREDICTORES ECONOMICO-SOCIALES Y DE ACTITUDES

	Mínima	Máxima	Diferencia
Lugar socialización de la mujer	-0,03	0,13	0,16
Vivienda tiene luz y/o agua	-0,12	0,14	0,26
Nivel de educación de la mujer	-0,20	0,26	0,46
Edad de la mujer	-0,10	0,04	0,14
Aspiraciones para la hija	-0,11	0,15	0,26
Comunicación entre esposos	-0,14	0,15	0,29
Emancipación de la mujer	-0,21	0,13	0,34

El Cuadro 2 muestra que el nivel de educación tiene la mayor dispersión de los promedios por categoría, variando el mínimo y el máximo en 46 puntos, y la edad de la mujer tiene menos poder explicativo; la variación es de tan solo 14 puntos. Sin embargo, en base a estos datos no es posible decir que la educación de la mujer sea la variable con más valor predictivo (haciendo abstracción en este momento de la correlación entre las variables) porque para tal decisión hay que tomar en cuenta además el número de casos en las categorías. Volviendo a los datos del Cuadro 1 se nota que en la variable "educación de la mujer" la categoría "secundaria incompleta o más" contiene 67 casos, lo que constituye una proporción relativamente baja del total de los casos. Además, la categoría más nutrida "hasta 3 años de primaria" da un promedio muy cercano al promedio general (0,65 vs. 0,66). Por el contrario, ambas categorías extremas de la variable "emancipación de la mujer" contienen gran número de casos. Como veremos más adelante en la discusión del valor predictivo de las variables, efectivamente la variable "emancipación de la mujer" es la mejor. Esto quiere decir, generalizando el caso discutido, que el valor predictivo de una variable está constituido por el valor predictivo de cada una de sus categorías, ponderado por el número de casos en cada una de ellas.

5. Ordenación de las categorías por variable predictiva

Al estudiar el comportamiento sin ajuste de las variables se destacan algunas diferencias. El "lugar de socialización de la mujer" diferencia poco y sus categorías estaban a priori ordenadas como si fuera una variable de medición ordinal (con esta variable dependiente). Las categorías de la variable "vivienda tiene

luz y/o agua", estaban ordenadas de la manera presentada para formar una variable seudo-ordinal que midiera el nivel de desarrollo urbano. Por lo general la luz llega primero ya que un sistema de agua potable dentro de la casa (a lo que se llama "agua" en los cuadros) implica inversiones mucho más grandes y por ende un grado de desarrollo urbano más alto. Al relacionar esta variable con conocimiento de anticonceptivos se comporta de otra manera; no es tanto el nivel de desarrollo el que parece estar midiendo sino la accesibilidad al lugar para mensajes que vienen relacionados con la luz: radio y televisión pueden ser las causantes. Así es que las categorías "luz" y "luz y agua" están relacionadas con el mismo nivel de conocimiento y "agua" representa un nivel más bajo. Esta relación irregular que se perdería en una estadística del tipo correlación, se conserva perfectamente en el MCA.

La "educación de la mujer", que es una variable ordinal, muestra una relación monótona, lo que facilitaría su uso en un análisis de regresión "normal" con esta variable dependiente. La "edad de la mujer", en cambio, está relacionada con el conocimiento de anticonceptivos en forma no monótona con un conocimiento más alto en la categoría de 20 a 29 años. Aunque esta variable es la única con escala de medición cardinal, por esta relación curvilínea no sería apta para ser usada en la regresión múltiple. Mirando los promedios de las categorías se nota que el valor predictivo no puede ser alto; si bien la categoría de 15 a 19 años muestra una diferencia notable con el promedio general (0,56 y 0,66 respectivamente), pesa poco porque el número de casos es pequeño.

Las variables de actitudes están basadas en escalas formadas sumando el puntaje que cada individuo obtuvo en varias preguntas relacionadas con la actitud subyacente y dividiéndolas en cinco categorías según el significado de los puntajes. Las categorías no son cardinales porque no se puede cuantificar la diferencia gradual entre, por ejemplo, "muy en pro" y "en pro". Las primeras dos variables de actitudes construidas que son "aspiraciones para la hija" y "emancipación de la mujer", tienen una relación monótona con el conocimiento de anticonceptivos, lo que indica que estas escalas funcionan con esta variable dependiente. La última, por el contrario, da una relación un poco irregular, lo que hace dudar acerca del agrupamiento de sus puntajes para formar sus categorías. Estas se pueden agrupar en tres: alta, neutral-baja, muy baja, para mantener la idea básica de que esta variable es ordinal. Sin embargo, este resultado muestra que el trabajar con variables compuestas (índices) no siempre es lo más apropiado; se ocultan datos y no se sabe a cuál de los componentes atribuir los efectos que perturban los resultados.

6. Resultados con ajuste por categoría

Todo lo antes dicho está basado en los promedios brutos como aparecen en la columna 2 del Cuadro 1, esto es la predicción del conocimiento de anticonceptivos al saber tan solo que la persona pertenece a una categoría dada de una sola variable predictiva. Por ejemplo, si sabemos solamente que la mujer se socializó en el campo, la probabilidad de conocer anticonceptivos es de un 63 por ciento.

Ahora vemos la influencia de pertenecer a una categoría dada, tomando en cuenta que una persona además tiene otras características que influyen a su vez en el conocimiento.

7. Categorías como "no responde"

Hasta este momento no hemos tomado en cuenta las categorías "no responde" que contienen aquí pocos casos y que por lo general se dejan fuera del análisis porque muy a menudo contienen casos especiales. Se nota, comparándolas en las columnas B, C, y E del Cuadro 1, una gran inestabilidad en los promedios (por ejemplo en "vivienda tiene luz y/o agua" respectivamente 0,12, 0,34 y 0,81), que puede deberse a un traslapo casi total entre ellas, lo que quiere decir que las mujeres que cayeron en la categoría "no responde" de una variable también están en la misma categoría de otra. Como se ha dicho, no es posible establecer cuál es el valor predictivo de cada una de ellas. Por ser pequeñas estas categorías, en nuestro caso, sus promedios no perturban significativamente los resultados que dan las demás categorías ni los del predictor total como lo ha mostrado la práctica. Si entran a propósito o por error tales categorías pequeñas, lo aconsejable es no tomarlas en cuenta. Por el contrario, si son grandes, influyen notoriamente en los demás resultados y, por consiguiente, el camino a seguir sería hacer el análisis de nuevo excluyendo los casos que caen en tales categorías.

8. Las categorías de contenido significativo

Por supuesto, en el análisis están solamente las variables predictivas que el investigador hace entrar y por consiguiente el establecer el efecto "neto" de pertenecer a una categoría dada es "neto" en el sentido de haber tomado en cuenta el efecto de las demás variables en este análisis. Así vemos que al tomar en cuenta a las variables predictivas 2, 3 y 4 (vivienda tiene luz y/o agua, educación y edad de la mujer), la diferenciación original entre socialización en el campo o en un pueblo, desaparece; no es el lugar de socialización el responsable de la diferencia original, sino circunstancias vinculadas a éste como, por ejemplo, el hecho de que las mujeres socializadas en el campo tienen menos educación. Esta falta de educación podría ser la causa verdadera. Respecto de la variable "lugar de socialización", se puede decidir que, al usar este conjunto de variables predictivas, el aporte que hace a la predicción es mínimo porque la categoría más grande da el mismo promedio que el promedio general (0,66) y el rango entre las categorías de la variable que originalmente era de 0,63 a 0,79 disminuye considerablemente hasta 0,66 y 0,70.

El hecho de tener luz y/o agua en la casa mantiene casi totalmente su fuerza predictiva en este conjunto, lo que implica que no está muy relacionada con las demás variables. Lo mismo podría concluirse en cuanto a "nivel de educación de la mujer" si su última categoría de "secundaria incompleta o más" no bajara tanto. Entre las personas de esta categoría el tener acceso también a mensajes por radio por ejemplo, podría ser más importante. Por último, la edad de la mujer muestra pequeños cambios que producen el resultado que a partir de los veinte años ya no

es la edad la que cuenta sino la distribución en otras variables. Por ejemplo, se sabe que las mujeres de más edad han tenido menos oportunidad de ir a la escuela. Al tomar constante el nivel de enseñanza, entre otros, se ve que todas las mujeres de 20 años y más casi tienen la misma probabilidad de conocer anticonceptivos. Esta variable representa un ejemplo en que, al considerar otras, la proporción original aumenta (por ejemplo en la categoría de 35 a 39 años aumenta de 0,63 a 0,66). Las mujeres en esta categoría tienen dos o más características relacionadas de las cuales por lo menos una predice un bajo conocimiento. Al tomar constante a esta última, se ve que la otra no era la causa porque sube el porcentaje de mujeres con conocimiento anticonceptivo.

En el ejemplo 2, una comparación de los promedios originales con los netos, muestra que la "comunicación entre esposos" es la variable que mejor mantiene su fuerza predictiva (poco cambio en los promedios por categoría). Los otros dos deben su efecto original en mayor grado a interrelaciones entre las variables y por consiguiente pierden fuerza al considerar al mismo tiempo todas las demás variables predictivas.

La conclusión del análisis de los promedios en los ejemplos 1 y 2 es entonces que, considerados por separado los efectos de índole económico-sociales y de actitudes, las variables "vivienda tiene luz y/o agua", "educación de la mujer" por un lado y "comunicación entre esposos" por otro, son las que tienen más poder explicativo.

9. Comparación de efectos netos por categoría entre ejemplos

Ahora veremos como todas las variables se mantienen al considerarlas juntas (Cuadro 1). Pueden hacerse dos comparaciones con un sentido distinto. Al comparar los promedios brutos con los netos, por ejemplo, las columnas B y E, lo que se estudia es el efecto neto de cada categoría de cada variable predictiva al considerar constantes todas las demás variables. Por otra parte, al comparar los promedios netos de dos análisis distintos, por ejemplo, las columnas C y E, lo que se estudia es el impacto adicional que tienen sobre el promedio neto, la inclusión de variables que no fueron consideradas en uno de los análisis que se están comparando. A este segundo tipo de análisis dedicaremos aquí más atención, al preguntar ¿cuánto cambia el valor predictivo de cada categoría si, además de las variables ya usadas anteriormente, se toma en consideración otro conjunto de variables predictivas?

El lugar de socialización que casi ya no tenía fuerza en el primer ejemplo, ahora (ejemplo 3) quedó sin influencia. Esto quiere decir que el valor predictivo aparente, en realidad se debe a otros factores dentro del conjunto de variables 2 hasta 7.

El hecho de tener luz y/o agua dentro de la casa sigue teniendo fuerza predictiva muy poco relacionada con las actitudes. En cambio el "nivel educacional de la mujer" pierde mucho más de su fuerza distintiva. Por ejemplo, en la categoría secundaria incompleta o más, el promedio de conocimiento llega a ser del 74 por

ciento. Esto puede entenderse como que las personas de esta categoría son en gran parte las mismas que se encuentran en las categorías que predicen un alto grado de conocimiento de otras variables, no sólo socio-económicas sino también de actitudes. La edad de la mujer muestra unos cambios, aunque pequeños, que vale la pena discutir. Se nota que las mujeres de 40 años o más tienen un grado de conocimiento de los anticonceptivos bastante grande si se toman constantes las variables económico-sociales (aumento en la categoría de 45 a 49 de 0,65 a 0,68) y las de actitudes (aumento adicional en la misma categoría de 0,68 a 0,71). Esto quiere decir que además de estar sobrerepresentadas en categorías socio-económicas que están relacionadas con un bajo nivel de conocimiento, también tienen actitudes que por encima de este efecto no favorecen tal conocimiento. Un caso especial se da en la categoría de 15 a 19 años: las mujeres de esta categoría tendrían menos conocimiento, si sus demás características hubieran sido iguales a las de toda la muestra, lo que se entiende por tomar constante a los demás factores. Tal vez, por tener más educación que el promedio de toda la muestra, su nivel de conocimiento bajó en el ejemplo 1 al tomar constante la educación. Ahora en el ejemplo 3 el conocimiento sube, lo que quiere decir que estas mujeres que están al comienzo de su vida conyugal y reproductiva, están sobrepresentadas en las categorías de actitudes en que el conocimiento es pequeño. Tan grande es este efecto negativo, que compensa en exceso los efectos positivos de la posición relativamente buena en los aspectos económico-sociales.

Mirando ahora las variables de actitudes, se nota que la diferenciación en el promedio general disminuye de nuevo respecto de "aspiraciones para la hija". Como las aspiraciones tanto altas como bajas se acercan en el tercer ejemplo más al promedio general (0,66), puede inferirse que las mujeres de ambas categorías están sobrerepresentadas en algunas categorías de las variables socio-económicas; las de altas aspiraciones en categorías que muestran un alto promedio de conocimiento de anticonceptivos y las de bajas aspiraciones en categorías con bajo nivel de tal conocimiento. En cambio "la emancipación de la mujer" no tiene mucha relación con las variables socio-económicas, resultando por eso un cambio pequeño en los promedios por categoría. Lo mismo es válido respecto de "comunicación entre esposos".

10. Resultados por variable predictiva

Los resultados arriba discutidos a base de los promedios por categoría se reducen a estadísticas que facilitan la comparación del poder predictivo de las variables. Las más importantes de ellas son los coeficientes eta y beta, que están basados en el aporte ponderado de todas las categorías de una variable dada; eta representa la correlación entre la variable predictiva y la variable dependiente sin ajuste por la correlación entre variables, y beta representa la correlación después del ajuste por la correlación con las demás variables. Beta, por lo tanto, indica el poder explicativo de la variable. Los cuadrados de eta y beta indican la proporción de la suma total de los cuadrados de las desviaciones que puede ser explicada por la variable (no ajustada y ajustada, respectivamente).

Cuadro 3

COEFICIENTES ETA Y BETA EN TRES EJEMPLOS DE MCA SOBRE CONOCIMIENTO DE ANTICONCEPTIVOS

Variables predictivas	Eta[a/]	Beta Ejemplo 1[b/]	Beta Ejemplo 2[c/]	Beta Ejemplo 3[d/]
Lugar de socialización	0,15	0,07		0,07
Luz/agua	0,25	0,19		0,17
Educación mujer	0,26	0,19		0,14
Edad mujer	0,08	0,08		0,06
Aspiraciones para hija	0,18		0,11	0,06
Emancipación de la mujer	0,18		0,12	0,10
Comunicación entre esposos	0,34		0,31	0,29

a/ Como los coeficientes eta representan la fuerza de la relación sin ajuste, son iguales en cualquiera de los tres ejemplos en que figura la variable correspondiente.
b/ Variables socio-económicas.
c/ Variables de actitudes.
d/ Todas las variables.

Estudiando los coeficientes, se ve reflejado lo antes mencionado y que estaba basado en los datos por categoría. Aquí no se puede decir cuál es la categoría o la parte de la variable que causa una disminución o aumento en el poder explicativo y por lo tanto, el contenido no es tan rico; por otro lado, para comparar la fuerza predictiva de las variables entre sí, es bueno tener una estadística que las haga comparables.

Se ve por ejemplo, en el Cuadro 3, que la baja ya mencionada en el porcentaje de mujeres que conoce anticonceptivos en la categoría "enseñanza secundaria o más", es una de las causas por las cuales el valor predictivo de "educación de la mujer" en el ejemplo 3 en que se toma en cuenta a todas las variables, llega a ser más bajo que el de "vivienda tiene luz y/o agua". La "comunicación entre esposos" continúa teniendo el valor explicativo más alto y junto con las variables "educación de la mujer" y "vivienda tiene luz y/o agua" (por la importancia de tener luz) es responsable de gran parte de la varianza explicada.

Por ser beta y eta tipos de coeficientes de correlación (aunque basados en la desviación del promedio general por categoría de la variable), sus cuadrados tienen el mismo sentido que R^2: la proporción de la varianza total explicada por la variable. En el ejemplo 3 la mejor variable "comunicación entre esposos" da un $beta^2 = 0,292 = 0,08$ (Cuadro 3), lo que indica que explica el 8 por ciento de la varianza total en conocimiento de anticonceptivos. La variable menos fuerte en este mismo ejemplo "aspiraciones para la hija" resulta dar un $beta^2 = 0,06^2 = 0,004$, por lo que esta variable predictiva explica el 0,4 por ciento de la varianza.

11. Resultados por conjunto de variables predictivas

La varianza total explicada y el coeficiente de correlación múltiple, también pueden ser calculados respecto del conjunto de variables dentro de un ejemplo y son dados en R^2 y R, como se muestra en el Cuadro 4.

Cuadro 4

R Y R^2 AJUSTADOS DE TRES EJEMPLOS DE MCA SOBRE CONOCIMIENTO DE ANTICONCEPTIVOS

	R múltiple (ajustado)	R^2 (ajustado)
Ejemplo 1 a/	0,310	0,096
Ejemplo 2 b/	0,370	0,137
Ejemplo 3 c/	0,430	0,185

a/ Variables socio-económicas.
b/ Variables de actitudes.
c/ Todas las variables.

Aquí se ve por ejemplo que la proporción de la varianza total explicada por el conjunto de todas las variables (ejemplo 3) es de 18,5 por ciento, lo que corresponde a un coeficiente de correlación de 0,43 (raíz cuadrada de 0,185). Comparando las estadísticas del Cuadro 2 con las del Cuadro 3, se nota que el coeficiente de correlación múltiple R en el segundo ejemplo, en gran parte está causado por la variable "comunicación entre esposos" (compárese beta 0,31 y R 0,37) mientras que en el primer ejemplo la correlación múltiple de 0,31 está constituida por la contribución de varias variables (luz/agua: beta 0,19 y educación de la mujer: beta 0,19). El R múltiple del tercer ejemplo en que todas las variables actúan, también está constituido por varias variables predictivas.

IV. C O N C L U S I O N E S

Ahora, cuando tratamos de dar una respuesta a la pregunta original de cuál es el mejor conjunto de variables, vemos que ésta tiene muchas graduaciones. Mirando los coeficientes de correlación múltiple, R, se nota que las variables de actitudes tienen más fuerza predictiva que las socio-económicas (0,37 vs. 0,31). Sin embargo, este resultado se basa principalmente en la fuerza predictiva de una sola variable de "comunicación" que está, por su contenido, muy cerca del conocimiento de anticonceptivos. Tomando el ejemplo en que figuran todas las variables y comparando los beta con los eta originales, tan solo hay dos de ellas que conservan más de la mitad de su fuerza explicativa al tomar en cuenta a las demás: la "comunicación" ya mencionada y "luz/agua" (la edad de la mujer se excluye por su bajo nivel

de predicción). La fuerza predictiva de la variable "luz/agua" estaba causada por la presencia de "luz", lo que nos hizo suponer que el acceso a los mensajes difundidos por radio y televisión era el factor principal de este efecto.

Considerando juntas las variables "comunicación" y "luz/agua", podría concluirse que el conocimiento de anticonceptivos está influido primeramente por un proceso de accesibilidad y aceptación. Por lo tanto, las actitudes y opiniones formadas o existentes durante el matrimonio parecen ser las más importantes. De acuerdo con esta hipótesis está el hecho de que la educación de la mujer no tiene una fuerza predictiva tan autónoma, lo que se ve reflejado en la baja en el porcentaje de las que conocen los anticonceptivos entre las mujeres que tienen relativamente mucha educación, cuando se toman en cuenta las demás variables. Estas son las mujeres que tienen relativamente más acceso a "luz" y que hablan más con sus esposos, por ejemplo, y son esos los hechos que importan, no su nivel educativo en sí. Enfocando el problema por otro lado, puede formularse esta conclusión de la manera siguiente: el tener una educación relativamente alta está relacionada con más acceso a medios de comunicación y a más comunicación con el esposo. Por lo tanto, un aumento en la educación puede causar un mejoramiento en la recepción de mensajes y por ende, un aumento en el conocimiento de los anticonceptivos. Los datos vistos de esta manera proporcionan una base para formular una hipótesis de un modelo causal de efectos aditivos que prediga el conocimiento de anticonceptivos.

N O T A S

1/ Para una discusión más elaborada del MCA, vea Frank M. Andrews, James N. Morgan y John A. Sonquist: Multiple Classification Analysis. A report on a Computer Program for Multiple Regression Using Categorical Predictors. Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan, 1971.

2/ PECFAL-Rural, Programa de Encuestas Comparativas de Fecundidad en América Latina, se realizó durante 1968-1969 mediante entrevistas en muestras representativas que fluctuaron entre 2 000 y 2 800 mujeres entre 15 y 49 años de edad y que vivían en las partes rurales y pequeñas ciudades hasta de 20 000 habitantes de Colombia, Costa Rica, México y Perú.

3/ Las variables de actitudes son resúmenes de escalas Likert formadas de los siguientes ítems:

1) Aspiraciones para la hija:
   a) Aspiración para la hija: estudiar o trabajar
   b) Nivel de educación deseado para la hija
   c) Nivel de ocupación deseado para la hija

2) Emancipación de la mujer:
   a) Acuerdo con el uso de la moda urbana
   b) Acuerdo con que jóvenes salgan juntos
   c) Acuerdo con que la mujer vaya a fiestas
   d) Acuerdo con participación política de la mujer.

3) Comunicación entre esposos:
   a) Acuerdo en usar planificación familiar
   b) Frecuencia de conversaciones acerca del número de hijos deseados
   c) Hombre toma en cuenta a la mujer en cuanto a paridez deseada.

A P E N D I C E

MANUAL DE USO DEL PROGRAMA MCA DEL SISTEMA OSIRIS

Johanna de Jong
Abel Packer

El MCA forma parte del sistema de programas llamado OSIRIS. Las notas siguientes están basadas en el USER'S MANUAL de OSIRIS. La idea central es facilitar al máximo la elaboración del programa, razón por la cual hemos eliminado una serie de opciones que tienen que ver con detalles de programación.

I. CALIDAD DE LOS DATOS

El programa acepta datos numéricos. A los casos en blanco deben asignársele números para que no se detenga la ejecución del programa. A estos casos se les puede asignar, por ejemplo, el valor de la categoría "no responde" o "no se aplica" que luego en el análisis puede ser excluida o no.

II. ORGANIZACION DEL PROGRAMA

El programa permite la ejecución continua, en la misma corrida del computador, de varios análisis cada vez con la misma variable dependiente u otra, y las mismas variables predictivas u otras. El único requisito consiste en que todas las variables usadas en la corrida estén declaradas (sin distinguir la calidad en que van a ser usadas) antes de definir el primer análisis. Todas estas variables deben estar en el diccionario (que es una descripción de los datos) o ser agregadas transitoriamente si se trata de variables creadas en la corrida. Se refiere a las variables con una letra V seguida por un número en el caso de variables originales y con R seguida por un número si se trata de variables creadas.

III. RESTRICCIONES Y CAPACIDAD DEL PROGRAMA

1. La variable dependiente es una sola y debe estar medida en escala de intervalo o en forma dicotómica. Como la predicción está basada en el promedio y la desviación estándar de ella, puede tener valores grandes, (hasta seis dígitos) y tanto positivos como negativos. Se aconseja excluir casos extremos porque influyen mucho en el promedio. Para tal propósito, el programa ofrece la posibilidad de asignar un valor máximo a la variable dependiente. Para excluir casos con un valor inferior a un valor mínimo aceptable, hay que recurrir a la posibilidad general de asignarlos a categorías definidas como "casos excluidos".

2. Las variables predictivas. Aunque la capacidad del programa es suficiente como para aceptar de 2 hasta 34 variables, se aconseja usar entre 3 y 10 para facilitar la interpretación de los resultados. La escala de medición puede ser nominal, ordinal o cardinal. El número de categorías por variable puede llegar a 100 (00-99), aunque es preferible usar no más de 6 para que el número de casos por categoría no sea demasiado pequeño y para que quede espacio en el computador para las demás variables predictivas.

3. Filtros. Para el análisis de subgrupos, comparación de mitades de la muestra, etc., pueden seleccionarse conjuntos de casos haciendo uso de un filtro que deje pasar los casos en que se tenga interés. Tales filtros pueden referirse a todos los análisis que se hacen en una sola corrida (filtro global) o a determinados análisis dentro del total de la corrida (filtro local). El filtro también se usa para especificar "casos excluidos".

4. Ponderación. En el análisis pueden aplicarse pesos para equilibrar la muestra. Sin embargo, hay que tener cuidado en la interpretación de algunas estadísticas como la prueba F para las cuales todavía no se ha encontrado un factor de ajuste razonable.

5. Exclusión de categorías como "no responde". Categorías como "no responde", "no se aplica" pueden ser excluidas mediante un filtro. Sin embargo, esto puede resultar en la exclusión de muchos casos al usar muchas variables predictivas. (Por ejemplo, al usar 10 variables, cada una con un 4 por ciento de "no responde", la exclusión puede llegar al 40 por ciento). A veces vale la pena no excluir en cada variable tales categorías sino sólo en las variables en que tendrían un efecto desfavorable por su gran número de casos. Por ejemplo, si algunas variables se refieren tan solo a mujeres, será mejor hacer un análisis de ellas, excluyendo a los hombres en la variable "sexo". Las demás categorías de "no responde" que se supone están distribuidas al azar, podrían quedar incluidas. En la práctica se ha establecido que cuando tales categorías tienen pocos casos, no influyen mucho en las estimaciones de las otras categorías y, por su pequeño número, tampoco en la fuerza predictiva de la variable o del conjunto de variables.

6. Capacidad de almacenaje. En la capacidad de almacenaje del programa, el número de variables explicativas y el número de categorías de ellas entran en forma competitiva. El total de unidades usadas, que no debe superar a los 12.000, se calcula de la manera siguiente:

$$4^* \sum_{i=1} (C_i + 1) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (C_i + 1)(C_j + 1) \leq 12.000$$

en que C_i = la categoría máxima de la variable i

n = número total de variables.

IV. TRANSFORMACION DE VARIABLES

La recodificación o la creación de variables a base de las originales, se hace con una rutina de mucha capacidad del OSIRIS, llamada RECODE. Este tiene un pequeño lenguaje para el usuario que está explicado en el Apéndice K del USER'S MANUAL. Cuando se crea una variable, hay que darle un número que todavía no esté ocupado por otra, sea esta original o recodificada. A pesar de ser precedidas por una letra distinta (V o R), en el diccionario ocuparían el mismo espacio. Las variables creadas son tratadas en el programa con signo negativo para distinguirlas de las originales, de manera que V-6 = R6. En lo que sigue usamos la anotación V-6 y R6 indistinguidamente. Sin embargo, conviene acostumbrarse a la anotación V-6 para limitar las posibilidades de equivocarse. Las nuevas variables son agregadas transitoriamente durante el proceso sin posibilidad de guardarlas. En el diccionario sólo existen en forma permanente las variables originales.

V. EL PROGRAMA

El programa está concebido de forma general y salvo las limitaciones del análisis propiamente dicho, puede procesarse cualquier encuesta mediante el suministro de tarjetas parámetros.

Para preparar estas tarjetas el usuario debe considerar tres puntos: los datos, el diccionario y las tarjetas de control.

1. Los datos pueden estar en cualquier medio de almacenamiento. Cuando no están en tarjetas deben especificarse las características de grabación.

2. El diccionario es una descripción de los datos, en particular de las variables. Cada variable es identificada por un número. Así, todas las veces que es necesario referirse a una variable, se hace por su número. En el diccionario se codifica el nombre de las variables, su ubicación, características y los casos excluidos. Existe una tarjeta de descripción del diccionario y una tarjeta por cada una de las variables que se van a usar.

DICCIONARIO - CODIFICACION

1) Tarjeta de Descripción del Diccionario (TDD)

Columna	Contenido
4	3
5-8	número de la primera variable
9-12	número de la última variable
13-16	N° de tarjetas por cada persona (o cuestionario)
20	BLANCO - se usa notación de columnas para indicar la ubicación de la variable
	1 - se usa la notación de largo de campo para indicar la ubicación de la variable.

2) Tarjetas de Descripción de Variables

Columna	Contenido
1	T
02-05	número de la variable
07-30	nombre de la variable
32-39	ubicación de la variable

Si la columna 20 de la TDD está en blanco, se llenan estas columnas de la manera siguiente:

32-33	número de orden de la tarjeta en que se encuentra la variable
34-35	número de la columna donde empieza la variable.
38-39	número de la columna donde termina la variable

Si en la columna 20 de la TDD hay un 1, se llenan estas columnas de la siguiente manera:

32-35	inicio (columna) donde empieza la variable
36-39	espacio que ocupa la variable

45-58 Se codifican los casos excluidos. Hay dos conceptos:

45-51 MD1 (casos excluidos 1): se perfora un solo valor y sólo este valor va a ser considerado como MD1.

52-58 MD2 (casos excluidos 2): se perfora un solo valor. Si es positivo, los valores iguales o mayores serán considerados como casos excluidos.
Si es negativo, los valores iguales o menores a este valor negativo serán considerados como valores excluidos.

Nota: A diferencia con MD1, en MD2 se puede considerar un rango de valores (consecutivos) como casos excluidos.

3. En las tarjetas de control se perfora la información relacionada con el estudio MCA. Por ejemplo, se define el universo de los datos que van a incluirse en el análisis, las variables a usar distinguidas en variables dependientes, independientes y de ponderación. La forma de estas tarjetas se describe en el orden en que entran en el programa.

1) TARJETA DE FILTRO GLOBAL (opcional): Con esta tarjeta se hace la selección de casos que regirá para toda la corrida. Puede usarse para seleccionar subconjuntos que tengan ciertas características, hacer análisis de las mitades de la muestra y también para excluir casos. La tarjeta empieza con una de las dos palabras claves: INCLUDE o EXCLUDE. Su forma es, por ejemplo:

```
INCLUDE V2=1-5  AND  V70 = 23,27,35 OR R2 = 5*
EXCLUDE V30=00  OR   V27 = 95,99    OR  R5 = 00-02*
```

a) AND se usa cuando todas las expresiones deben cumplirse para la inclusión o exclusión de un caso; OR cuando por lo menos una de las expresiones se

cumpla; AND se ejecuta antes de OR. No se usan paréntesis. El máximo de expresiones es de 15.

b) Pueden ser rangos de categorías (por ejemplo de 1 a 5), o categorías aisladas (por ejemplo 23,27) las que se considerarán.

c) Las categorías deben ser definidas con el mismo espacio con que está definida la variable. Por ejemplo, si la variable V30 tiene 15 categorías, ocupa dos espacios. Al excluir la categoría "0" ésta debe ser indicada como "00" (dos espacios).

d) Las variables pueden aparecer en cualquier orden y más de una vez.

e) Puede usarse cualquier parte de la tarjeta y más de una.

f) La última tarjeta siempre termina con un asterisco.

2) TARJETA DE TITULO. Se da un nombre al programa en cualquier parte de la tarjeta.

3) TARJETA PRINCIPAL DE PARAMETROS. En esta tarjeta se indica qué hacer con los datos no-numéricos como los blancos, signos, etc. Hay cinco opciones: SKIP; MD1: MD2; TERMINATE; STOP. Estas opciones deben precederse con la palabra clave BADDATA seguida del signo = . Por

Ejemplo: BADDATA = SKIP

En una corrida sólo puede ponerse una tarjeta principal de parámetros y en ella sólo puede indicarse una opción. Cuando la opción elegida es STOP, basta con poner un asterisco en la tarjeta (el asterisco significa "fin de la tarjeta"); no hay necesidad de escribir la palabra clave pues el programa procesa automáticamente la opción STOP en este caso. A continuación se explica el significado de cada opción:

SKIP:	Salta el caso
MD1 :	Se convierte el valor al código del primer "caso excluido" (valor mínimo)
MD2 :	Se convierte el valor al código del segundo "caso excluido" (valor máximo)
TERM:	Termina la corrida
STOP:	Termina la corrida

Nota: 1) Si los casos en blanco, con signos, etc. tienen importancia para el análisis, hay que hacer una transformación para convertirlos en datos numéricos, antes de ejecutar el MCA.

2) Las palabras claves pueden abreviarse con las primeras cuatro letras (por ejemplo, TERMINATE=TERM).

3) Si se decide usar SKIP, MD1, MD2, el programa indica el número de casos tratados así.

4) TARJETA DE LA LISTA PRINCIPAL DE VARIABLES. En esta tarjeta aparecen todas las variables usadas en los análisis de esta corrida.

a) Cada número de variable debe ser precedido de una V o una R, si es el resultado de una recodificación. Por ejemplo V35, R46. Los grupos de variables contiguas pueden designarse escribiendo la primera y la última del grupo separándolas por un guión. Ejemplo: V6-V35, que significa desde la variable 6 hasta la variable 35, ambas inclusives. Es importante insistir en que la letra V debe precedir a los dos números. Si se escribiera V6-35 estaría mal. Conviene recordar que en la otra forma de designar a las variables recodificadas se deja la letra V pero se antepone un signo negativo al número. Así, R6-R10 puede escribirse de esta manera: V-6-V-10.

b) Las variables o grupos de variables contiguas se separan con coma (V3,V4-V6). Los grupos de variables contiguas pueden ser ascendentes o descendentes pero no pueden contener la variable 0 (cero); por ejemplo V-6-V5 no es posible.

c) Las variables pueden aparecer en cualquier orden.

d) Se usa cualquier parte de la tarjeta. Si se usa más de una tarjeta, se termina la tarjeta anterior con coma y sigue en la próxima tarjeta en cualquier lugar. La última tarjeta termina con asterisco.

e) Si se requieren residuos, la última variable debe ser una de identificación (ID) que acompañará la salida de residuos (por ejemplo puede usarse el número de la entrevista).

A continuación viene el conjunto de tarjetas que controla cada análisis. Puede haber un número ilimitado de análisis en cada corrida.

5) TARJETA DE FILTRO LOCAL (opcional). Selecciona los casos tan solo para el análisis que le sigue inmediatamente. Su forma y sentido es igual a la del filtro global (Ver 1).

6) TARJETA DE TITULO LOCAL. Se da un nombre al análisis específico en cualquier parte de la tarjeta.

7) TARJETA LOCAL DE PARAMETROS. Se toma una serie de decisiones, entre otras, acerca de la forma y el contenido de la salida, se decide la manera en que el proceso de iteraciones debe ejecutarse, se describe la variable ponderación y la variable dependiente DEPVAR. La descripción de esta variable DEPVAR es obligatoria. Nunca debe omitirse. De esta manera la tarjeta local de parámetros contiene, por lo menos, la palabra clave DEPVAR, con el número de la variable y un asterisco que indica "fin de la tarjeta". Por ejemplo:

DEPVAR = (número de la variable)*

Hay opciones que se presentan de dos maneras. Una de ellas requiere el uso de una palabra clave. A continuación se da la lista con las opciones:

PRINT (NOTABLES/TABLES, NOHISTORY/HISTORY)
TEST = % MEAN/CUTOFF/%RATIO/NONE

MDOPTION = BOTH/MD1/MD2/NONE
CRITERION = n/.005
ITERACIONES = n/25
NOWEIGHT/WEIGHT = n

Otras opciones no están precedidas de una palabra clave. A continuación se da la lista de estas opciones:

NORESIDUAL/RESIDUAL/SRESIDUAL
NOSUPPRS/SUPPRESS

Puede notarse que algunas opciones están subrayadas. Es para indicar que esas opciones serán procesadas automáticamente por el programa, por lo que no es necesario escribir dichas opciones en la tarjeta. Obviamente, no procesa la opción subrayada cuando se optó por la contraria. Ejemplo: si se optó por TABLES no procesa NOTABLES

El proceso de iteraciones necesita una explicación porque es de gran importancia, en esta versión del MCA, para la estimación del efecto neto de cada variable predictiva. El modelo aditivo que se trata de imponer a los datos, necesita la solución de un número de ecuaciones normales, siempre y cuando las variables predictivas estén interrelacionadas, lo que sucede corrientemente en las ciencias sociales. Estas soluciones sirven para estimar cada vez mejor la influencia "neta" de una categoría de una variable, eliminando la parte de la predicción que está causada por la interrelación con otras variables predictivas. Para tal proceso se usa un proceso de repetición de la misma manipulación (proceso iterativo) en que cada vez la estimación se basa en los coeficientes de la estimación anterior, empezando con los promedios "brutos". Este proceso sigue hasta que el cambio en los coeficientes en dos estimaciones consecutivas es tan pequeño que se puede decidir que ellos se han acercado a su valor real. Se dice que el proceso ha llegado a la convergencia, indicando con esto que el valor predictivo estimado de cada variable se ha acercado de tal manera al valor real que el modelo aditivo impuesto coincide de la mejor manera con los datos.

El usuario tiene que decidir el número de iteraciones que el programa ejecutará, para lo cual es útil saber que el número necesario para llegar a la convergencia, depende en forma directa de la correlación entre categorías. Hasta tal punto es este el caso que un traslapo total de dos categorías (correlación del ciento por ciento) hace imposible llegar a una solución porque los coeficientes pueden atribuirse con igual razón tanto a una como a otra categoría, y por lo tanto siguen alternando. Cuando el número de iteraciones establecido por el usuario es demasiado pequeño como para llegar a la convergencia, los ajustes finales son muy toscos e inestables. Por consiguiente, las inferencias hechas a base de ellos son poco confiables. Por esta razón se aconseja no limitar demasiado el número de iteraciones, tomando en cuenta que hay otras maneras más confiables de terminar el proceso iterativo. Estas están especificadas abajo en la palabra clave TEST.

El contenido de la tarjeta es como sigue.

Obligatorio:

DEPVAR = (número de variable, código máximo): Indica cuál es la variable dependiente. DEPVAR debe ser seguido por el signo = y el número de la variable que representa la variable dependiente. Por ejemplo si la variable V6 es la dependiente, se escribe

DEPVAR = (6)

si es R8, se escribe DEPVAR = -(8)

No es necesario que el código máximo sea especificado (el programa presume entonces que es 999999), pero es útil usarlo en casos en que el o los valore(s) más alto(s) que toma la variable, tienen que ser excluidos del análisis. El código máximo es la categoría más alta que va a ser tomada en cuenta.

Opcional:

PRINT (NOTABLES/TABLES, NOHISTORY/HISTORY)

(Nótese que es el único caso en que deben tomarse dos decisiones en una palabra clave)

NOTA/TABL : Se imprimen o no los cruces entre cada combinación de dos variables predictivas.

NOHI/HIST : Se imprimen o no los coeficientes de las iteraciones

NOHI : Se imprimen los coeficientes de las últimas dos iteraciones en caso de no lograr convergencia. Al lograr convergencia no se imprime nada.

HIST : Se imprimen los coeficientes de todas las iteraciones.

NOWEIGHT/WEIGHT=n: Si se usa una ponderación de los datos, WEIGHT debe ser seguido por = y el número de variable que constituye el peso (por ejemplo si la variable V3 es usada como peso se pone WEIGHT=3; si R6 es el peso se escribe WEIGHT=-6).

MD OPTION=BOTH/MD1/MD2/NONE: Esta opción constituye otra posibilidad de excluir casos de la variable dependiente.

BOTH: Elimina casos del análisis en que la variable dependiente tiene el valor mínimo que está definido como caso excluido (MD1) o es igual o mayor que el valor definido como caso excluido en MD2.

MD1: Elimina los casos en que la variable dependiente es igual a MD1.

MD2: Elimina los casos en que la variable dependiente es igual o mayor que MD2.

NONE: No elimina los casos definidos como MD1 o MD2.

ITERACIONES=n/25: n puede ser puesto igual a cualquier valor entre 1 y 99999. Si no se especifica la opción, se hacen 25 iteraciones.

TEST=%MEAN/CUTOFF/%RATIO/NONE: Esta opción constituye el tipo de test de convergencia para las iteraciones que se desea. El valor de este test se determinará en CRITERION.

%MEA: Se averigua si los cambios en los coeficientes en dos iteraciones consecutivas están por debajo de una fracción especificada del promedio general.

CUTO: Los cambios en dos iteraciones consecutivas deben estar por debajo de un valor especificado.

%RAT: Se averigua si el cambio en dos iteraciones consecutivas es menor que una fracción especificada de la razón de la desviación estándar de la variable dependiente por su promedio.

NONE: El programa sigue hasta el total de las iteraciones pedidas.

CRITERION=n/.005: n es el valor numérico que va a ser la tolerancia para el test de convergencia escogida. El rango varía entre 0.0 y 1.0 (hay que perforar el punto decimal). A falta de un valor especificado se toma CRITERION=.005.

NORESIDUAL/RESIDUAL/SRESIDUAL

NORE: No se producen residuos

RESI: Se producen residuos, aplicando el modelo MCA tan solo en los casos:1) que pasen el filtro local; 2) que no sean casos excluidos; y 3) que no superen los códigos máximos estipulados; estos casos no considerados aparecen con valores 9 en los campos del valor predicho y del residuo.

SRES: Se producen residuos para todos los casos que pasen el filtro global.

NOSUPPRS/SUPPRESS: Se suprime o no la impresión de los residuos, si ellos están producidos según RESI o SRES. Si no están pedidos (NORE), el NOSUPPRS o la omisión de esta palabra clave, no tiene ningún efecto.

8) TARJETA DE VARIABLES PREDICTIVAS. De 2 a 34 variables pueden ser especificadas como variables predictivas en cada análisis. Como opción pueden especificarse los valores máximos admisibles de cada variable usada; esta constituye otra manera de excluir casos. Si no se asigna un valor máximo, el programa toma 10 como valor máximo. Sin embargo, los valores mínimos no pueden ser excluidos de esta manera (por ejemplo el valor 0 habría que excluirlo mediante un filtro). Al excluir ciertos valores de una variable, se excluyen de todo el análisis los casos que tengan tales valores.

Ejemplo de la tarjeta de variables predictivas:

P - 10, P4, P30 = 7, P1 = 05, P70*, en que:

P - 10, P4, P30, P1 y P70 son las variables predictivas en este orden.

Para P - 10, P4 y P70, el valor máximo es 10 (por falta de especificación, el programa toma 10 como valor máximo).

P30 tiene como valor máximo 7 y P1, 5.

P - 10 es una variable resultado de una recodificación.

a) Se empieza en cualquier parte de la tarjeta, se separan las variables predictivas por comas. Al tener que usar más tarjetas, la que no sea la última termina con coma, la última con asterisco.