IGNACIO MÉNDEZ RAMÍREZ &
PABLO GONZÁLEZ CASANOVA (ed).

1993 Matemática y ciencias sociales

Miguel A. Porrúa/CIIH/UNAM ediciones

México DF.

371 pp.

Consideraciones sobre el uso de la estadística en las ciencias sociales. Estar a la moda o pensar un poco

FERNANDO CORTÉS*
ROSA MARÍA RUBALCAVA*

INTRODUCCIÓN

"Los análisis (una serie de tabulaciones cruzadas) son relativamente poco sofisticados dada la cantidad de datos disponibles".[1]

La moda no es una buena consejera para la investigación científica; tampoco para seleccionar los instrumentos estadísticos. Sin embargo, pareciera que es uno de los criterios que se han usado y se usan en la investigación social para escoger entre el arsenal de herramientas estadísticas disponibles en las últimas décadas.

Basta una revisión somera de las principales revistas norteamericanas especializadas para corroborar que aparecen ciclos en los que una técnica particular adquiere popularidad. No es claro si este comportamiento se debe a la presencia de algunos problemas sociales relevantes cuya investigación demanda una nueva técnica o la aplicación creativa de instrumentos estadísticos ya existentes; o si obedece a que los problemas a ser investigados se delimitan a partir de los desarrollos instrumentales; o bien a que las técnicas se aplican como recetas de cocina debido a la deficiente formación metodológico-técnica que reciben los científicos sociales (Blalock H. 1989: 450); y aún habría que considerar la presión de los organismos que financian investigaciones

* Investigadores del Centro de Estudios Sociales de El Colegio de México.

[1] Párrafo del dictamen con el que una revista internacional rechazó la publicación de un artículo de un investigador mexicano en 1990.

y los comités de dictamen editorial. Cualquiera que sea la explicación a estos ciclos, se puede decir que en los últimos años la investigación social se ha caracterizado, entre otros rasgos, por la aplicación de técnicas de moda.

Una mirada superficial permite delinear la historia reciente de la investigación social según el uso de los instrumentos estadísticos. En los Estados Unidos, entre los años cincuenta y sesenta, se utilizaban profusamente métodos estadísticos para realizar clasificaciones (análisis factorial, de conglomerados, discriminante) y analizar relaciones entre variables (regresión, correlación, análisis de contingencia y de asociación). Los métodos que demandaban variables métricas[2] (factorial, conglomerados, discriminante, regresión y correlación) tuvieron una aplicación limitada en América Latina, tanto porque los recursos de cómputo eran de difícil acceso como por las debilidades en la formación de los investigadores sociales en técnicas de investigación y estadística.[3] En cambio, los análisis de asociación y de contingencia (que operan sobre variables no métricas)[4] han dominado la investigación social de América Latina hasta la actualidad.

En los años sesenta se popularizó el entonces llamado análisis causal (cuyos orígenes se remontan a un trabajo en genética poblacional escrito por Wright en 1921: 557 a 585), que se percibía como una generalización del análisis multivariado y que tuvo un fuerte impacto sobre la investigación social, a pesar de que sólo se aplicaba sobre variables métricas y variables dicotómicas. Una vez que se apreció la profundidad de la discusión epistemológica en torno a la noción de causalidad, pasó a tomar el nombre con el cual se le conoce hasta hoy: análisis de trayectorias o de senderos (*path analysis*). En este método, las complejidades de cálculo no superaban a las del análisis de regresión.

[2] Son variables métricas las que se miden en escalas de intervalo o de razón y no métricas las medidas en escalas nominal u ordinal.

[3] H. Blalock (1989: 448, 454 y 458) hace mención a la actual deficiencia en el conocimiento de instrumentos de investigación de los estudiantes en ciencias sociales (grado y posgrado) de las universidades norteamericanas. Tradicionalmente sus colegas latinoamericanos adolecen del mismo problema pero con mayor intensidad.

[4] Sólo se requería una clasificadora para realizar los cruces de variables.

La línea del desarrollo estadístico-matemático que se plantea como problema ampliar el acervo de métodos para analizar variables no métricas alcanzó resultados que atenuaron la división tajante entre variables según su nivel de medición. Por una parte, el análisis de regresión se extendió para incorporar variables explicativas y explicadas no métricas (las primeras se conocen con el nombre de variables ficticias o variables mudas, *dummy*, y se tratan como caso particular dentro del análisis de regresión; en tanto que las segundas llevan a modificaciones profundas que cristalizaron en el modelo de regresión logit o logística). Por la misma época se desarrolla una generalización del análisis multivariado (modelo loglineal) que permite modelar un conjunto de relaciones entre variables no métricas. Tanto el modelo logit como el loglineal ganaron popularidad en los años ochenta, en parte, porque sus tediosos procedimientos iterativos de ajuste pudieron realizarse con paquetes de programas estadísticos en computadoras personales.

En las ciencias sociales de América Latina se han hecho aplicaciones esporádicas, en campos temáticos reducidos, de los análisis de trayectorias, loglineal y logit.

En lugar de seguir la moda se puede tomar en cuenta una diversidad de criterios que se han propuesto para orientar la selección de "el mejor" método estadístico. Hay quienes sostienen que uno de los criterios es el nivel de medición de las variables (Siegel 1956), otros ponen el acento sobre los procedimientos que se siguieron para generar las observaciones (Campbell y Stanley 1979), también hay quienes plantean que los instrumentos de registro determinan la viabilidad de análisis estadístico: la información que se obtiene a través de métodos antropológicos (observación participante, observación directa, historia oral, filmación..) no es susceptible, se dice, de análisis cuantitativo (Magrassi, G. E., M. M. Roca y otros 1980: 14). Es cierto que este tipo

de información presenta dificultades particulares pero no son un impedimento insoslayable para aplicar la estadística.

En las páginas que siguen desarrollaremos la idea de que el uso adecuado de los instrumentos estadísticos en una investigación requiere también identificar el isomorfismo entre las estructuras lógicas de la técnica y de las respuestas provisorias (hipótesis) a las preguntas que orientan la investigación. Sostenemos que no basta considerar únicamente el nivel de medición de las variables y los procedimientos de observación, sino que, además, debe examinarse la concordancia entre las preguntas de investigación, las hipótesis de trabajo expresadas en términos de relaciones entre las observaciones o entre las variables y las técnicas que ofrecen diversas posibilidades para el análisis empírico de dichas relaciones.

En ocasiones la pregunta se limita a establecer la presencia o ausencia de relación entre variables, como: ¿hay o no relación entre la salud de los niños [SN] y el trabajo de sus madres fuera del hogar [TM]?, ¿se modifica o no esta relación según la calidad del cuidado [CC] que recibe el hijo en ausencia de la madre?, ¿hay o no relación entre la salud del pequeño y el incremento en el ingreso familiar [IIF] debido al trabajo de su madre? En otras ocasiones, interesa saber la fuerza de esas relaciones o bien cuantificar el efecto neto sobre la salud del menor, originado en el impacto positivo del mayor ingreso familiar y el negativo de la ausencia materna en el hogar, modulado por la atención alternativa al hijo.

El nivel de medición de las variables, los procedimientos que generaron los datos y las relaciones entre variables derivadas de las preguntas de investigación no necesariamente conducen a la selección de *una* técnica estadística en particular, aunque sí delimitan un subconjunto de ellas. Para establecer la relación entre TM y SN se pueden utilizar, por ejemplo, análisis de asociación, análisis de correlación o pruebas de diferencias de medias.[5] La relación TM, SN, en

[5] Se suponen variables dicotómicas.

presencia de CC, podría abordarse con análisis de asociación parcial, mediante la ecuación de covarianzas de Lazarsfeld, con correlaciones parciales o vía análisis de varianza. Para estudiar el efecto de IIF sobre SN, se podría optar entre el análisis de regresión logística, la regresión probit o el análisis discriminante. Si interesa evaluar el efecto neto de estas variables sobre SN, habría que aplicar análisis de regresión múltiple o bien análisis de trayectorias, según la estructura de los vínculos planteados entre IIF, CC, TM y SN.

La coincidencia entre el uso razonado de la estadística (que lleva en sí la correspondencia entre los planteamientos teóricos, las preguntas de investigación, los métodos y técnicas empleadas para recopilar la información y los modelos estadísticos) y la aplicación de moda sólo puede ser un hecho fortuito.

En este trabajo nos proponemos mostrar una gama de técnicas estadísticas que expondremos desde el punto de vista de la correspondencia de sus estructuras con las estructuras de las relaciones entre observaciones y entre variables, propuestas por los esquemas teóricos. Las hemos seleccionado tomando en cuenta su aplicación potencial y su frecuencia de uso en la investigación social en América Latina. Tuvimos presente, sobre todo, aplicaciones a la sociología, la sociodemografía, la antropología, los estudios urbanos y de salud pública, y la ciencia política, cuyos análisis son fundamentalmente de sección cruzada o de estática comparativa; por lo anterior, queda fuera de esta exposición el tratamiento de series de tiempo, procesos estocásticos, estudios en panel y cualquier otro tipo de análisis secuencial.

Por motivos de exposición, consideramos necesario dedicar la primera sección a la matriz de datos, que es el punto de encuentro entre las operaciones teóricas, metodológicas y estadísticas; es la carne que cubre el esqueleto lógico. En la segunda, desarrollamos la idea central de este trabajo: el isomorfismo entre el sistema de relaciones que fluye de la conceptuación y el sistema de relaciones que configura a

cada modelo estadístico. Dada la naturaleza de nuestra tarea, ponemos el énfasis en la presentación de las estructuras de algunas técnicas estadísticas. Esta parte se subdivide en técnicas para construir índices y clasificar, y técnicas para analizar relaciones entre variables. Por último, en la tercera, intentamos abogar por el uso no empirista de la estadística y perfilamos su papel en distintos momentos del proceso de investigación.

La matriz de datos

El análisis estadístico procede sólo si se dispone de un conjunto de mediciones de los atributos de las unidades de observación, unidades que suelen denominarse de diferentes maneras: individuos en sentido genérico (Bouroche, J. M. y G. Saporta 1980: 5), elementos de análisis o unidades de análisis (Galtung, J. 1966: 1), unidades (Castellanos, A. 1977: 35), objetos (Yule, U. y M. Kendall 1959: 15) o bien, casos u observaciones (SPSS, 1988: B-9). Para cada una de ellas se registran, con el fin de caracterizarlas, una serie de rasgos o propiedades que se denominan *variables*. Nosotros preferimos llamar *unidad de registro* a las unidades cuyos atributos se han registrado, con el fin de enfatizar que los datos no surgen de una percepción casi inmediata sino de operaciones que se apoyan en consideraciones teóricas y técnicas. Las unidades de registro y las variables se ordenan en la *matriz de datos*, que no es más que un arreglo rectangular con tantos renglones como unidades haya y con una columna para cada variable. Las casillas de la matriz, definidas por la intersección de renglones y columnas contienen "los valores"[6] de las variables.

La construcción de la matriz de datos es fundamental porque constituye tanto un punto de partida para la aplicación de las herramientas estadísticas, como un punto de llegada que enlaza la esfera de la conceptuación con la del registro empírico, nos parece que el precario tratamiento que han hecho de este tema, tanto la estadística como las ciencias sociales, oculta la complejidad de las operaciones que culminan en la matriz de datos y explica, en buena parte, las dificultades que enfrentan los investigadores para el mejor aprovechamiento de las técnicas en beneficio de su quehacer.

A continuación examinaremos algunos de los temas que involucra la construcción del arreglo rectangular, visto como punto de llegada, de unidades de registro y variables:

i) Las variables de la matriz de datos corresponden a los indicadores de los conceptos teóricos.[7] Hay que consignar la existencia de teorías en que algunos de sus conceptos son inobservables (Blalock, H. 1968: 5 a 27) ya que la correspondencia concepto-indicador puede involucrar varios indicadores para un concepto.

ii) La medición engloba tanto la operacionalización de conceptos teóricos como la confiabilidad y validez de los indicadores. A partir de una definición de cada concepto y de la especificación de sus dimensiones, se llega a establecer uno o más indicadores observables (Lazarsfeld, P. 1973: 35 a 45). La calidad de los indicadores se juzga por su grado de consistencia, estabilidad o precisión (confiabilidad) y por la certidumbre de que miden lo que queremos medir (validez) (Kerlinger, F. 1973: 442 a 473). La estadística auxilia la investigación social, proporcionándole conceptos y medidas (varianzas y correlaciones) que permiten evaluar la calidad de los indicadores.

iii) Cuando se tiene más de un indicador para un concepto se presenta el problema de sintetizar la información:

[6] Entrecomillamos este término porque en las escalas nominal u ordinal no se trata de valores en sentido estricto.

[7] En este trabajo reservaremos al término "variable" para referirnos a su sentido en estadística. En ciencias sociales se usa tanto en este sentido, como para designar a las categorías de mayor abstracción, a los indicadores e índices tal como se definen en la metodología de las ciencias sociales, y también se emplea en relación con un área temática (por ejemplo, cuando se dice "hay que considerar la variable poblacional").

operar sobre el subconjunto de columnas de la matriz de datos (variables) asociadas a un mismo concepto para reemplazarlas por un número menor de variables compuestas (índices). Los índices más utilizados suelen ser aquéllos que se obtienen por operaciones aritméticas elementales, como los índices sumatorios simples (en cada unidad de registro se suman los valores de las variables), o los índices que se construyen a través de cocientes y productos de variables. La estadística también proporciona técnicas que ayudan a la realización de estas operaciones teórico-metodológicas, entre las que se encuentran los análisis de componentes principales y factorial.

iv) La matriz de datos es independiente de las fuentes de información y de los métodos e instrumentos con que ésta se registre. Su forma no se modifica si el investigador obtiene su información de fuentes primarias o secundarias o si utilizó un cuestionario, una entrevista, una grabación, una filmación, un texto, o su propia observación. Lo que sí es esencial, para que sea susceptible de análisis estadístico, es que los datos sean numéricos, entendiendo por ello tanto los números que corresponden a variables métricas como los que pueden asociarse como códigos a variables no métricas.

v) Tampoco importa si la cobertura del estudio es censal o muestral. La aleatoriedad incorporada a la mayoría de las técnicas estadísticas multivariadas, se justifica no sólo por la selección de las unidades de registro a través del muestreo aleatorio, sino también por la imposibilidad de considerar todos los factores asociados a un fenómeno (Hagood, M. 1973: 65 a 78; Johnston, J. 1984: 14) o al argumentar que la aleatoriedad forma parte de los procesos sociales (King, G. 1989: 9 a 37).[8]

[8] La discusión sobre el carácter determinista o aleatorio de la realidad sigue presente en la ciencia. Véase el ríspido debate entre Ilya Prigogine y René Thom, en el encuentro organizado por la Fundación Salvador Dalí, en la Facultad de Física de la Universidad de Barcelona, en noviembre de 1985 (Wagensberg, J. 1986: 187 a 197).

vi) No es trivial, aunque parezca lo contrario, decidir cuál o cuáles son las unidades de registro pertinentes al problema que se investiga. ¿La teoría que orienta la investigación hace alusión a unidades individuales o colectivas? por ejemplo: qué es lo que interesa para un estudio: ¿la condición de "ocupado" o "desocupado" de los individuos o el número (o la proporción) de ocupados dentro de los miembros económicamente activos de un hogar? En el primer caso, las unidades de registro serán los individuos y una de las variables que los caracterice será su condición de ocupación; en el segundo, serán los hogares y una de sus variables el número de ocupados (o su proporción respecto a los económicamente activos del hogar).

vii) Las operaciones que habitualmente se aplican a la matriz de datos no se agotan en la eliminación de los indicadores que no pasaron las pruebas de confiabilidad y validez, ni en la construcción de índices sino que, en ocasiones, la investigación requiere transformar las unidades de registro. En estos casos se aprecia el doble carácter de las unidades de registro: son tales en cuanto sirvieron de base al registro empírico pero sólo constituirán *unidades de análisis* en tanto sean las relevantes para la teoría. Así, una conceptuación que centre la atención en hogares, ante la imposibilidad de acceder directamente a sus rasgos característicos, deberá llegar a la matriz de datos en dos etapas: en la primera se construirá una matriz en que las observaciones sean individuos y una de sus variables indique el hogar al que pertenecen; en la segunda etapa se construirán los hogares como nuevas unidades, unidades de análisis, generadas a partir de la matriz de datos de individuos. Los procedimientos de definición de las variables de los hogares, a partir de las de los individuos, van desde operaciones artiméticas simples como en el caso del índice de ocupación, hasta elaboraciones relacionales complejas para definir variables como el tipo de familia, a partir del parentesco de cada uno de los miembros con el jefe del hogar.

viii) Se presenta una complicación adicional cuando la investigación requiere del manejo estadístico simultáneo de unidades de registro heterogéneas, que deben combinarse en una única matriz de datos que refiera todas las variables a una misma unidad de análisis. Por caso, el problema puede demandar que se combinen las variables del hogar con las de algunos individuos seleccionados (el jefe, el cónyuge, el hijo mayor, el hijo menor) y con características de la vivienda como la zona de residencia, entre otras: a) la edad del cónyuge o la del hijo menor suelen utilizarse como indicadores del ciclo doméstico que, a su vez, se considera como uno de los factores explicativos de la participación femenina en el mercado de trabajo; b) algunas variables de la vivienda y de la zona de residencia pueden usarse en la construción de estratos sociales que, junto con el ciclo doméstico y otras variables, complementan el abanico de factores explicativos de la tasa de participación de la mujer. En los casos en que la estratificación se basa en pocas variables se pueden aplicar procedimientos estadísticos relativamente simples: representación gráfica, comparación de promedios y descomposición de la varianza; pero cuando son muchas, estos métodos no son eficientes y se requiere de algunos más elaborados: análisis de conglomerados, análisis clasificatorio múltiple (*multiple classification analysis*) y análisis discriminante, por citar algunos.

Para finalizar esta sección, debemos hacer notar que la matriz de datos, en la mayoría de las investigaciones (si no es que en todas), experimenta sucesivas transformaciones, las que difícilmente se comprenden con la imagen de un proceso de investigación que avanza, en forma continua, en dirección a su término y cuyo producto se integra "acumulándose en el cuerpo de conocimiento disponible" (Bunge, M. 1979: 19 a 37). Las transformaciones de la matriz de datos aparecen como "casi naturales" si la investigación se conceptúa como un proceso que combina fases de continuidad con rupturas y reordenaciones, es decir,

como un proceso caótico (piaget, J. y R. García,1982: 192 a 194; Prigogine, I. e I. Stengers 1983: 166 a 187; Lazlo, I. 1990: 137 a 149; Balandier, G. 1989: 226 a 237).

Técnicas estadísticas

La matriz de datos, si bien es el punto de llegada de las operaciones teórico-metodológicas, es también el punto de partida del análisis estadístico. En ocasiones, desde el momento en que se selecciona un procedimiento estadístico particular es necesario, para satisfacer sus supuestos, introducir cambios en la matriz de datos. A su vez, las preguntas llevan a seleccionar subconjuntos de variables o de unidades de registro: casi nunca se utiliza la matriz de datos en su totalidad. Más aún, los resultados que arrojan los primeros análisis (supongamos que sea un simple análisis de frecuencias) conllevan, la mayoría de las veces, recodificaciones y redefiniciones de las variables, lo que implica volver al plano de la teoría a la vez que inducen nuevas preguntas de investigación que pueden requerir instrumentos estadísticos diferentes.

Dedicaremos esta sección a presentar la relación entre la estructura del sistema de hipótesis teóricas y la estructura de los instrumentos básicos que proporciona la estadística social. Dados los propósitos de este trabajo se acentuará la exposición de la estructura y características de las técnicas, pero el lector debe retener que, en la práctica, la referencia al campo teórico es permanente.

Para realizar la exposición dividiremos los modelos de análisis estadístico en dos grandes grupos. El criterio de clasificación se inclina en favor del tipo de problema sustantivo que se quiere resolver y responde al uso más frecuente que se hace de las distintas técnicas en las ciencias sociales. En el primer grupo incluimos las que se utilizan con la finalidad principal de construir índices o de clasificar (formar

grupos, estratos, zonas, regiones). En el segundo, están los instrumentos que permiten analizar relaciones conceptuales en la forma de relaciones entre variables. Aunque sea trivial, vale la pena destacar que ésta es sólo una de las varias maneras en que se pueden agrupar las técnicas estadísticas y que responde al tipo de aplicación que habitualmente se hace de ellas, pero de aquí no debe derivarse que no se puedan usar para propósitos distintos; por ejemplo, si bien el análisis factorial se utiliza preferentemente para construir índices, también puede emplearse para contrastar hipótesis (Kim J. O. y Ch. W. Mueller 1978: cap. v; Long, J. S. 1983).

Técnicas para construir índices y clasificar

Análisis de conglomerados (o de cúmulos)

Tiene por propósito agrupar a las unidades de registro cuyas características son las "más parecidas".

La idea básica que subyace a la formación de grupos es que éstos deben ser internamente homogéneos y, a la vez, lo más diferenciados posible entre sí. El caso más simple es el de la formación de grupos a partir de una sola variable, para ello se utilizan procedimientos gráficos, análisis de la distribución de frecuencias y técnicas de descomposición de la varianza. En el método gráfico se toman como referencia los máximos y mínimos de la distribución de frecuencias y se delimitan los valores de variable que marcarán las fronteras entre los grupos. Este recurso se puede complementar analizando el comportamiento de los promedios grupales y de las inter e intravarianzas. También es común examinar los valores de la variable y tomar como puntos de corte aquéllos en que se aprecian discontinuidades.

Sin embargo, cuando el problema involucra a más de una variable, estos criterios pierden operatividad. Las decisiones de agrupación no tienen por qué ser las mismas para

todas las variables: una unidad de registro según una de sus características debería pertenecer a un grupo y, según otra, a uno distinto.

El análisis de conglomerados resuelve este problema considerando simultáneamente *todas las variables*.

El problema consiste en comparar los renglones de la matriz de datos, en las variables seleccionadas, para definir los grupos o conglomerados y decidir cuáles son los más parecidos. Para iniciar un agregado se razona del siguiente modo: dado un renglón se busca entre los restantes al que más se le parezca, se tiene así un grupo de dos elementos; si no hay ninguno similar se comienza la formación de un segundo grupo. El tercer caso se asigna a uno de los grupos ya creados, excepto si es muy distinto. Este procedimiento se repite hasta que todas las unidades de registro hayan sido asignadas a un grupo.

Lo importante para el procedimiento descrito es tener un índice que permita medir el "parecido" de las unidades de registro. El análisis de conglomerado pone varias opciones a disposición del investigador: la distancia euclidiana entre las unidades de registro, la distancia euclidiana al cuadrado, la diferencia entre los valores absolutos, etcétera. (Tryon, R. y D. Bailey 1970: 135 a 181).

Este método se inicia con tantos conglomerados como casos haya y concluye con todos los casos formando un solo grupo. Es decisión del investigador seleccionar cuántos grupos quiere distinguir. Para ello se apoya bien sea en la evolución del valor de la medida resumen utilizada, en gráficas que producen el mismo método o en el análisis de promedios y de varianzas (de las variables) de los conglomerados construidos. Además, puede tomarse en cuenta información externa para afinar la conformación de los grupos, como sería la contigüidad geográfica para una regionalización.

Dependiendo del tamaño de la matriz de datos y de los recursos de cómputo, una aplicación particular puede de-

mandar que el investigador decida de antemano el número de conglomerados. En este caso tendrá que hacer varias pruebas en el entorno del número seleccionado.

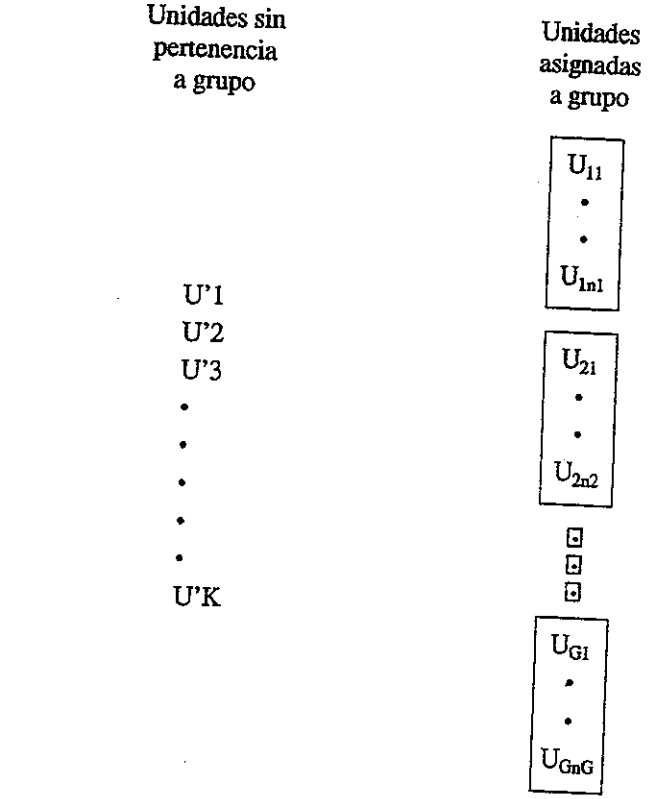
Como se puede apreciar, este procedimiento estadístico no proporciona solución única, sino que provee diferentes conglomeraciones entre las cuales hay que elegir tomando en cuenta no sólo los criterios estadísticos, sino también el conocimiento sobre el tema y la consistencia teórica de la agrupación.

Análisis discriminante

Es útil en los problemas en que hay un subconjunto de unidades de registro asignadas inequívocamente a grupos (cuya existencia se presupone), mientras que las restantes no tienen en ellos una localización precisa. El análisis discriminante entrega elementos para decidir la pertenencia grupal de los casos de este segundo subconjunto.

El modelo presupone que el grupo de pertenencia es una variable cuyo valor en cada unidad de registro particular será asignado por la técnica, a partir de los niveles que esa unidad presente en el conjunto de indicadores con que se caracterizó (llamados variables discriminantes); o también, de manera recíproca, que el grupo de pertenencia determina el nivel de los indicadores. En adición a lo anterior, se considera que el número de grupos es fijo y que cada unidad de registro pertenece a uno y sólo a uno de ellos. Los primeros desarrollos de esta técnica se deben a Fisher (Klecka, W. R. 1984: 12).

Son diversas las situaciones de investigación que comportan un problema como éste. Por ejemplo, aquéllas que se proponen identificar zonas según el grado de urbanización de un conjunto de unidades espaciales. En este caso, se tiene como punto de partida el conjunto de todas las unidades espaciales del país con sus correspondientes indicadores de urbanización.



Con los valores de las variables vinculadas a la urbanización se caracteriza a las unidades que pertenecen a cada grupo y se producen modelos (funciones discriminantes) con los cuales será posible asignar a las unidades con pertenencia grupal indefinida. Los modelos, alimentados con los valores que presenten las variables de urbanización en cada unidad, darán como resultado el grupo al que es más probable que pertenezca esa unidad, garantizando que las unidades de un mismo grupo sean similares y también claramente diferenciadas de las de los otros grupos.

En general, hay más de una función discriminante estadísticamente significativa. Esto quiere decir que cada fun-

ción asigna las unidades en duda de manera diferente y, aunque hay criterios estadísticos para decidir cuál es la función con mayor poder de discriminación, suele ser útil analizar las agrupaciones resultantes de todas las funciones, porque cada una clasifica (discrimina) a las unidades respecto a un rasgo complejo (multidimensional) distinto.

Los paquetes de programas estadísticos para computadoras ponen a disposición del investigador opciones que le facilitan decidir cuál es el mejor modelo, por lo menos desde el punto de vista estadístico. Esto significa identificar cuáles son las variables con mayor poder discriminatorio para el problema de clasificación que se aborda y a través de cuál de las funciones discriminantes se llega a una agrupación estadísticamente válida.

Las aplicaciones de este análisis casi siempre rebasan la finalidad estrictamente clasificatoria, adentrándose el investigador en la interpretación y caracterización de los grupos conformados por los distintos modelos (funciones) discriminantes.

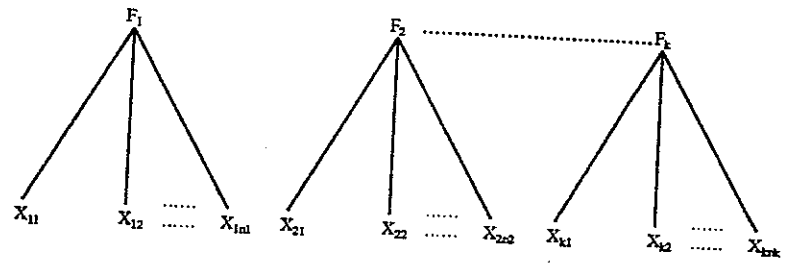
Análisis factorial

Se utiliza las más de las veces para construir índices. El tipo de problemas que demanda su aplicación es aquél en el que se precisa identificar un concepto o conceptos subyacentes (inobservables), a partir de variables de la matriz de datos (indicadores). Al plantearlo de esta forma, el problema refiere la relación entre observables e inobservables que, aunque suele olvidarse, se basa en hipótesis "que deben fundamentarse y confirmarse de algún modo" (Bunge 1979: 737) y cuyo carácter particular se destaca en el término "correlación epistémica".[9]

Este método que reúne una serie de técnicas, el análisis de componentes principales entre otras, permite reducir el

[9] Este término lo atribuye Blalock (1968) a Northrop.

conjunto original de variables a un número más reducido de variables subyacentes llamadas variables latentes o *factores* (por reducir el número de variables, algunos textos llaman al análisis factorial "análisis de conglomerados de variables"). Cada factor representa una combinación lineal de las variables originales que se expresa en un vector cuyos elementos indican el peso de cada una de las variables en el factor. Posteriormente se definen, para cada unidad de registro, índices que resumen cada factor en un sólo número; esto permite caracterizarlas mediante nuevas variables: los índices correspondientes a los factores más significativos desde el punto de vista estadístico. También en este caso, los programas de cómputo ofrecen variadas opciones de definición, cálculo y gráficas.



Con las variables sintetizadas en factores y con los índices correspondientes se pueden formar estratos, zonas, regiones, a través de las técnicas elementales de la estadística descriptiva o mediante los métodos gráficos antes citados.

Una de las primeras aplicaciones del análisis factorial fue hecha en el campo de la psicología de la inteligencia. Las pruebas de inteligencia registraban un conjunto de respuestas que, se hipotetizaba, se vinculaban con dimensiones no observables de la inteligencia (por ejemplo, habilidades y destrezas). Sin embargo, las aplicaciones del análisis fac-

torial se extendieron pronto a varias disciplinas y a múltiples propósitos (Rummel 1967).

Al igual que en el análisis discriminante, el uso de esta técnica busca llegar más allá de una clasificación. Es necesario entender el sentido que tiene cada factor (interpretación) y encontrar justificación tanto para los valores que tienen las unidades de registro en los índices calculados como para el lugar que ocupan en la clasificación multidimensional resultante.

El desarrollo de esta técnica ha extendido sus aplicaciones a la contrastación de hipótesis teóricas, incorporadas por el investigador a través de restricciones en el modelo (Long, J. S. 1983: 12). Se ha propuesto también como un método auxiliar cuando se detecta multicolinealidad en un modelo de regresión (véase el apartado: análisis de regresión), es decir, presencia de combinaciones lineales entre las variables, que viola uno de los supuestos en que se basa la estimación de los parámetros de ese modelo. (Chatterjee, S. y B. Price 1977: 157 a 163).

Técnicas para analizar relaciones entre variables

Análisis de contingencia

Permite dar apoyo empírico a enunciados teóricos que postulan la relación entre dos o más variables no métricas. La articulación entre las hipótesis teóricas y la relación concepto-variable (indicador o índice) conducen al planteamiento de hipótesis estadísticas (formuladas en términos de variables) que se someten a contrastación (Blalock, H. 1968: 5 a 27). Los procedimientos estadísticos se aplican sobre una tabla de contingencia, que se construye a partir del cruce de dos o más variables no métricas de la matriz de datos, en cuyas casillas se registraron las frecuencias absolutas.

El caso más simple de tratar es aquél en el que la hipótesis teórica remite al cruce de dos variables:

	x_1	x_2	. . . . .	x_c
y_1	n_{11}	n_{12}	. . . . .	n_{1c}
y_2	n_{21}	n_{22}	. . . . .	n_{2c}
.	.	.	. . . . .	.
.	.	.	. . . . .	.
.	.	.	. . . . .	.
y_r	n_{r1}	n_{r2}	. . . . .	n_{rc}

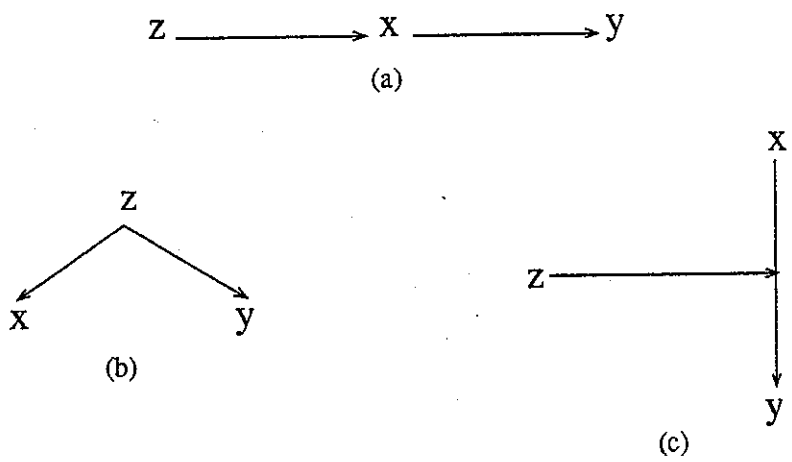
es decir, al análisis de una tabla bidimensional. La relación entre las dos variables se juzga a partir de la distancia respecto a la independencia estadística. Con base en la teoría elemental de probabilidades es posible establecer, para cada casilla, una medida de la discrepancia entre las frecuencias observadas y las esperadas bajo el supuesto de que las dos variables son estadísticamente independientes. Esta idea aplicada a todas las casillas de la tabla, sintetizada de modo conveniente, origina la estadística ji-cuadrada cuyo valor permite juzgar si hay o no relación significativa entre las variables (Cortés, F. y R. M. Rubalcava 1987: 105 a 115).

Cuando las hipótesis teóricas envían a hipótesis estadísticas que involucran la asociación entre más de dos variables aparecen algunas complicaciones técnicas. En efecto, además de la hipótesis: "las tres variables son mutuamente independientes", hay que considerar las pruebas de independencia parcial (la relación entre dos de las variables es

independiente de la tercera), e independencia condicional (dos de las variables son independientes en cada nivel de la tercera) (Everitt, B. 1977: 71 a 77).

Por mucho tiempo gozó de amplia popularidad el modelo de covarianzas de Lazarsfeld, propio para estudiar la asociación entre tres variables dicotómicas (Lazarsfeld, P. 1974(a): 23 a 39). Consiste, en esencia, en analizar la relación entre dos variables una vez que se introduce una tercera (variable "test" o variable control) a partir de una tipología que combina dos formas extremas que puede asumir la ecuación (Lazarsfeld, P. 1974(b): 328 a 352)[10] con la posición temporal (anterior o intermedia) de la tercera variable respecto a las otras dos.

Si representamos por X a la variable antecedente, por Y a la variable consecuente y por Z a la variable control, podemos expresar gráficamente los tres casos que tienen significación teórica según Lazarsfeld:



[10] Una expresión totalmente equivalente había sido desarrollada con bastante anterioridad (Yule, G. U. y Mg. G. Kendall 1911: 36 y 37).

El caso que se representa en (a) se denomina marginal anterior y corresponde a una estructura genuina de relaciones —la variable control es parte de la explicación— a diferencia del (b) en que la relación observada entre X y Y es ilusoria y se explica por las relaciones que mantienen las variables antecedente y consecuente con la variable control. El tercer caso (c) recibe el nombre de relación parcial y simboliza la interacción de Z con las variables X y Y: expresa cómo se modifica la relación entre estas variables en las categorías de Z. En realidad, el propósito principal de esta versión del análisis multivariado es la detección de las relaciones ilusorias o espurias.

A pesar de que el tratamiento matemático sólo considera el caso de tres variables dicotómicas, Lazarsfeld sostiene que la generalización, tanto por el lado del número de categorías como por el del número de variables, es a *fortiori* (Lazarsfeld P. 1974(a): 35).

Las limitaciones del análisis multivariado a la Lazarsfeld se desprenden, por una parte, de la inexistenica de una extensión de la ecuación a más de tres variables y, por otra, en muchos casos, de las dificultades prácticas para decidir si la variable intermedia es anterior a la variable explicativa, si está entre la explicativa y la explicada o bien si interactúa con ellas.

El análisis de asociación

El análisis de contingencia concluye al establecer si dos o más variables son independientes, o bien si están relacionadas; en tanto que el análisis de asociación mide la fuerza de la relación a través de coeficientes. Existe un nutrido número de coeficientes de asociación tanto en la literatura estadística como en los programas de cómputo. El amplio abanico de opciones plantea el problema de la selección adecuada; una solución es la de establecer criterios que ayuden a responder la pregunta ¿qué medida de asociación utilizar?

Hay una serie de coeficientes que, a diferencia de los que son función de ji-cuadrada, definen la asociación no sólo por lejanía respecto a la independencia sino también por la proximidad respecto a un concepto específico de relación. En términos laxos podríamos afirmar que la idea central que subyace a estas medidas es la de comparar la distribución efectiva de los datos y la que se esperaría según la hipótesis teórica (Goodman, L. y H. Kruskal 1954: 732 a 764).[11]

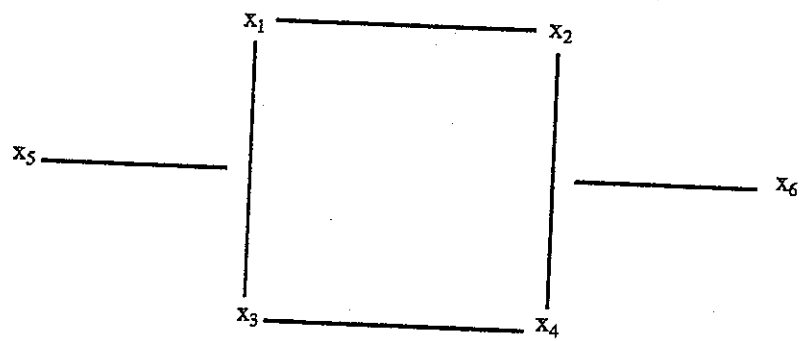
Así, se concluye que la selección del coeficiente adecuado depende más bien de un cierre teórico que de uno de carácter técnico: el investigador debe especificar sus hipótesis y, en función de ellas, generar el coeficiente que mide la proximidad de la distribución teórica respecto a la observada. Esta solución requiere que se coordine la estructura de los enunciados teóricos con la del coeficiente (Cortés, F. y R. M. Rubalcava, 1987).

El análisis loglineal

El trabajo original de Lazarsfeld propone el análisis de covarianzas como un método útil para detectar relaciones entre tres variables dicotómicas, a partir de los datos. Sin embargo, nada impide usarlo desde una óptica inversa, es decir, pronosticar el tipo de relaciones a partir de los enunciados teóricos. Más aún, la ecuación de covarianzas de Lazarsfeld se puede escribir en términos de coeficientes de asociación (Cortés, F. y R. M. Rubalcava 1987: 81 y 82) y, por esta vía, introducir el vínculo entre la hipótesis teórica y la medida de asociación.

Ahora bien, el análisis loglineal puede considerarse como una generalización del análisis de covarianzas de Lazarsfeld y del análisis de asociación. En efecto, es propio aplicarlo en aquellos casos en que el marco teórico conduce a una estructura de relaciones entre un conjunto de variables no métricas. La representación gráfica de una estructura típica de relaciones conceptuales susceptible de análisis loglineal es la siguiente:

[11] De hecho, en la actualidad, se dispone del coeficiente delta-roque que es, en esencia, una función generatriz de coeficientes, cuya forma general se particulariza al especificar el tipo de relación que se deriva de las hipótesis teóricas (Hildebrand, D., J. Laing y H. Rosenthal 1977).



Esta gráfica expresa que la variable X_1 está relacionada con X_2 y X_3; X_2 con X_4; y X_3 con X_4. Además, que la variable X_5 afecta la relación entre X_1 y X_3, y que X_6 impacta la relación entre X_2 y X_4. En la terminología del análisis loglineal las relaciones entre dos variables se denominan interacciones de primer orden. Cuando la relación es entre tres variables se denomina interacción de segundo orden, y así sucesivamente.

A partir de esta estructura se debe especificar la forma del modelo para proceder a su ajuste y obtener estimaciones de los parámetros que representan los efectos de las variables y de sus relaciones. En la especificación del modelo, el investigador supone (basado en su teoría) que las frecuencias de una tabla de contingencia resultan de los efectos de las variables y de sus interacciones.

El ajuste a los datos implica la utilización de procedimientos iterativos de cálculo (que operan sucesivamente de casilla a casilla). Desde el punto de vista práctico es imposible realizar los cálculos manualmente, en particular, cuando se trata de modelos con muchas variables y varias categorías en cada una.

Estos planteamientos conducen a la utilización del modelo loglineal como instrumento que ayuda a decidir respecto al grado de adecuación entre la teorización y las observaciones, lo cual no impide su uso como herramienta para extraer las relaciones que puedan existir en una matriz de datos. Incluso, los paquetes estadísticos propician este uso ya que, por lo común, cuenta con rutinas de cálculo que exploran todas las relaciones posibles (y no tiene costos significativos si se emplean computadoras personales; además, el tiempo que consumen los cálculos en estos equipos se ha reducido de manera considerable con el desarrollo de nuevos microprocesadores de menor costo, mayor rapidez y capacidad).

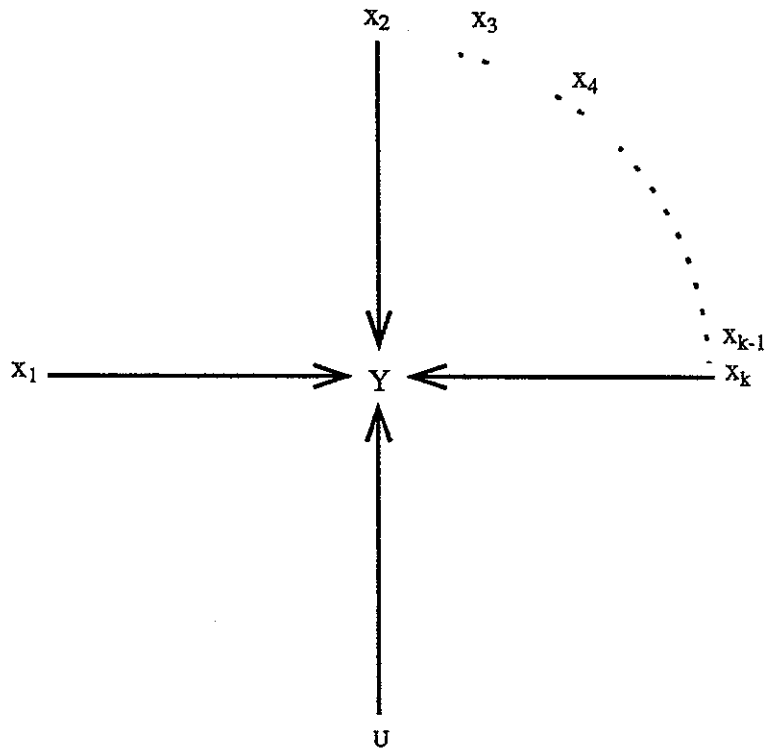
Análisis de regresión

En los casos en que el marco conceptual lleva a la especificación de un modelo que postula la presencia de una variable explicada (Y) a través de un conjunto de variables independientes (explicativas) (X_1, X_2, ..., X_k), la técnica estadística apropiada es el análisis de regresión.

La proposición teórica, que propicia el empleo de esta herramienta, supone que la variabilidad de la variable dependiente se puede explicar a través de las variables independientes. El modelo estadístico correspondiente agrega, además, que opera como variable independiente un factor aleatorio no observable (U), al que se conoce como error estocástico.

La gráfica siguiente muestra K variables explicativas (desde X_1 hasta X_k) y la variable estocástica U. Nótese que

no se han representado relaciones entre las variables explicativas ni entre éstas y el término de error. Las flechas que llegan a la variable explicada señalan que la teorización incluye la idea de direccionalidad de la relación.



El ajuste del modelo estadístico especifica una serie de condiciones que debe satisfacer el término estocástico y que permiten estimar no sólo si el modelo encaja de manera adecuada (para lo cual se tienen pruebas de hipótesis e índices de bondad de ajuste), sino también la significación y magnitud de los efectos de las variables explicativas.

La medición de los impactos de las variables explicativas requiere que no haya relación entre ellas o bien que sea

tenue. En caso contrario se presentan dificultades para estimar la magnitud real de los efectos debido a que el modelo es incapaz de separar cuál corresponde a cada variable.

Los procedimientos de estimación se complican si se detecta que no se cumplen los supuestos relativos al término de error. La construcción de pruebas estadísticas para decidir si se verifican empíricamente y el desarrollo de métodos especiales de ajuste han ocupado por años los esfuerzos de los econometristas.

Durante mucho tiempo se ha sostenido que el modelo de regresión demanda que las variables sean métricas, lo que erigía una fuerte restricción para su aplicación en los problemas típicos de las ciencias sociales. Sin embargo, esta limitación se ha superado al introducir, primero, variables ficticias (dicotómicas) que dieran cuenta de la presencia o ausencia de un suceso (por ejemplo, el impacto de los tiempos de paz o de guerra sobre la función consumo) (Johnston, J. 1960: 221 a 228) y su posterior desarrollo a una o más variables pluricotómicas (Goldberger, A. 1964: 218 a 226; Johnston, J. 1984: 225 a 236).

El ajuste de un modelo de regresión lineal cuando la variable dependiente es dicotómica conlleva una serie de anomalías[12] que se superan a través de una transformación logit de la variable dependiente (Theil, H. 1972: 166 a 196; Aldrich, J. y N. Forrest 1984; Maddala, G. 1989: 13 a 56; Agresti, A. 1990: 79 a 119). El método se ha generalizado para el caso en que la variable dependiente comporta más de dos categorías.

Como se puede apreciar, sucesivos avances han liberado al modelo de regresión de lo que alguna vez se consideró su limitación más importante para aplicarlo al análisis de problemas sociales.

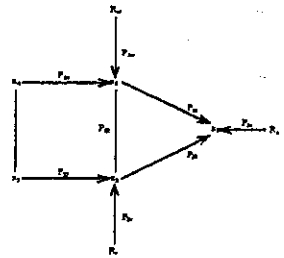
[12] Las estimaciones de la variable dependiente pueden escapar del intervalo [0,1] y las varianzas de los estimadores no son mínimas.

El análisis de trayectoria o de senderos (*Path analysis*)

Otra generalización del modelo de regresión toma lugar cuando la conceptuación desemboca en un sistema de relaciones que conducen a un conjunto de ecuaciones de regresión. Aunque el tema tiene bastante generalidad dentro de la econometría (sistema de ecuaciones simultáneas), en las ciencias sociales se desarrolló el caso particular de los sistemas recursivos.

Para comprender este concepto hay que hacer una especificación terminológica: dado un conjunto de ecuaciones, la distinción entre variables explicativas y explicadas es ambigua porque una misma variable puede ser explicativa en una ecuación y explicada en otra; tomando esto en consideración, cuando se trata de sistemas de ecuaciones se distingue, más bien, entre variables endógenas (las que dependen de otra en algunas de las ecuaciones) y exógenas (que no dependen de ninguna otra del sistema). Se dice que un sistema es recursivo si ninguna variable endógena es a la vez "causa" y "efecto" (Land, K. 1969: 24).[13]

Una representación gráfica típica de la estructura de relaciones que requiere de análisis de trayectorias sería la siguiente:



[13] Este tipo de sistema presenta la no despreciable ventaja de que los estimadores mínimo cuadráticos ordinarios aplicados a cada ecuación son consistentes. eficientes y asintóticamente normales (Johnston, J. 1984: 467 a 469).

Las variables X_4 y X_5 son variables exógenas y están relacionadas (nótese que están unidas por una línea y no por una flecha). Las variables X_1, X_2 y X_3, son variables endógenas y las tres variables R (R_u, R_v y R_w) representan el componente aleatorio de las variables endógenas. Los coeficientes P miden el impacto directo que tiene una variable sobre otra, por ejemplo, P_{14} mide el efecto de la variable X_4 sobre X_1, así como P_{3u}, el de la variable R_u sobre la variable X_3.

Además del efecto directo de una variable sobre otra, este modelo permite evaluar los impactos indirectos. Por caso, a pesar de que la variable X_4 no tiene vínculo directo con X_2, la afecta a través de dos senderos: uno opera a través de X_1 y el otro vía X_5.

El análisis de trayectorias permite tratar estructuras conceptuales bastante más complejas que las del análisis de regresión, con la limitación de que las variables deben ser métricas o dicotómicas.

Por último, hay que señalar que las primeras aplicaciones de esta herramienta a problemas sociales supusieron que por fin se tenía una técnica que permitía medir la fuerza de los vínculos causales (Simon, H. 1957: 37 a 49; Blalock, H. 1964; Boudon, R. 1968). Sin embargo, sin que se haya entrado de lleno a la profundidad epistemológica del tema —sólo limitándose al campo de la medición— se percibió que si bien es posible medir la concomitancia entre dos variables y establecer la precedencia temporal de una de ellas, no es posible garantizar que la relación se esfume una vez que se controlan *todos* los factores que inciden sobre la relación (Asher, H. 1976: 11 y 12). Obsérvese que este último requisito empírico para determinar si una relación es causal o no, toca los límites de lo teórico.[14]

[14] Los temas estadísticos más próximos a la causalidad se limitan a establecer concomitancia entre variables y a evaluar relaciones funcionales; la imputación de causalidad escapa a su dominio. Más aún, las discusiones epistemológicas sobre el tema no están agotadas y todavía existen escuelas cuyas posiciones son diametralmente opuestas (Bunge, M. 1961; Schuster, F.G. 1982; Halbwachs, Fr. 1977; Piaget, J. y R. García 1973).

A MANERA DE CONCLUSIÓN

A lo largo de la exposición hemos insistido en el criterio de selección de la técnica que se basa en las conexiones entre los conceptos y las variables. Por decirlo brevemente: las relaciones entre las variables están supeditadas a las relaciones entre conceptos postuladas por el marco teórico. Sin embargo, hay que reconocer que el estilo de exposición predominante en la estadística es más bien el opuesto. Los textos de análisis de regresión suponen que las relaciones funcionales entre las variables se especifican a partir del diagrama de dispersión (por lo menos cuando se trata de la relación entre dos y hasta tres variables); el análisis de trayectorias se plantea como una herramienta que permite determinar la estructura que mejor se ajusta a los datos; el análisis de conglomerados, como una técnica que se utiliza para reducir casos; el análisis factorial para disminuir el número de variables, y el análisis loglineal como una excelente "caña para pescar relaciones".

Sin embargo, no hay nada intrínseco a estos instrumentos que obligue a usarlos en la dirección que toman los datos como punto de partida.[15] Incluso, dentro de los textos clásicos de regresión se mezclan las dos formas de presentación y algunos se plantean primero la construcción del modelo y luego la estimación de sus parámetros (Intriligator, M. 1978). Lo mismo se puede observar en la exposición de varias de las técnicas que hemos examinado, incluidas las taxonómicas como el análisis factorial (Long, J. S. 1983).

La oposición entre los pares teoría-datos y datos-teoría pierde sentido si conceptuamos el proceso de investigación como "una sucesión de operaciones que pertenecen, sólo preponderantemente, ya sea al campo de lo teórico o al de lo empírico. Sin embargo, la pertenencia a uno u otro cam-

[15] Pablo González Casanova, en un trabajo de 1967 (reimpreso en 1987: 20), distingue cinco tipos de planteamientos de investigación en cuyos extremos están: el que se inicia por las categorías (elementos más abstractos) y el que comienza por los indicadores (elementos más concretos).

po no excluye la relación con el complementario. Las operaciones teóricas contienen siempre referencias empíricas y las operaciones empíricas cobran sentido por su inserción en el proceso de construcción teórica. La imagen que surge así es la de un ir y venir constantemente reiterado entre lo teórico y lo empírico en el cual nunca se vuelve al punto de partida, de modo que el tránsito de uno a otro nunca es circular" (Cortés, F., R. M. Rubalcava y R. Yocelevszky 1987: 13). La transmisión del conocimiento estadístico ha puesto más atención al momento "exploratorio" que al "confirmatorio".[16] Las páginas anteriores muestran que tener en cuenta la estructura de las relaciones teóricas facilita la selección de la técnica estadística adecuada toda vez que se desea examinar su coherencia con los datos (contrastar).

A lo largo del proceso de investigación la matriz de datos experimenta una serie de transformaciones que se originan a veces en el dominio teórico y otras en el empírico. La matriz inicial surge como producto de la objetivación, y la operacionalización de conceptos de la medición y de la construcción de índices. En ocasiones los conceptos teóricos conducen a unidades de análisis que deben construirse a partir de unidades de registro distintas. Además, en la medida que se avanza en la investigación, que se depura y enriquece el marco teórico, aparecen nuevos conceptos que transforman la matriz de datos con la incorporación o eliminación de variables y unidades de registro.

Las alteraciones que sufre la matriz de datos expresan las determinaciones teórico-metodológicas y teórico-estadísticas. Con las primeras nos referimos a las operaciones propias de la "metodología de las ciencias sociales" y con las segundas a las que obedecen a requerimientos de los modelos de análisis para la contrastación estadística de las hipótesis de la investigación.

[16] Detrás de la clasificación de los estudios en "exploratorios" y "confirmatorios" se puede percibir la distinción entre enunciados de observación y enunciados teóricos, una de las piedras angulares del empirismo lógico (Quine, W.V. 1981: 67 a 72; y Hanson, N. R., 1965: cap. 1).

Bibliografía

AGRESTI, Alan. *Categorical Data Analysis*, Nueva York, John Wiley, 1990.

ALDRICH, John y Nelson Forrest. *Linear Probability, Logit, and Probit Models*, California, Sage Publications, 1984.

ASHER, Herbert. *Causal modeling*, California, Sage Publications, 1976.

BALANDIER, Georges. *El desorden. La teoría del caos y las ciencias sociales: elogio a la fecundidad del movimiento*, Barcelona, Gedisa, 1989.

BLALOCK, Hubert. "The Real and Unrealized Contributions of Quantitative Sociology", en *American Sociological Review*, vol. 54, junio de 1989.

——. *Methodology in Social Research*, cap. 1, Nueva York, McGraw Hill, 1968.

——. *Causal Inference in non Experimental Research*, Chapel Hill, The University of North Carolina Press, 1964.

BOUDOND, Raymon. *L' analyse mathematique des faits socieux*, París, Plon, 1968.

BOUROCHE, Jean-Marie y Gilbert Saporta. *L' analyse des données*, París, Presses Universitaires de France, 1980.

BUNGE, Mario. *Causalidad: el principio de causalidad en la ciencia moderna*, Buenos Aires, EUDEBA, 1961.

——. *La investigación científica: su estrategia y su filosofía*, Barcelona, ARIEL, 1979.

CASTELLANOS, Amílcar. *Estadística aplicada a la investigación*, Maracaibo, Fondo Editorial IRFES, 1977.

CAMPBELL, Donald y Julian Stanley. *Diseños experimentales y cuasiexperimentales en la investigación social*, Buenos Aires, Amorrortu, 1979.

CORTÉS, Fernando y Rosa María Rubalcava. *Métodos estadísticos aplicados a la investigación en ciencias sociales: análisis de asociación*, México, El Colegio de México, 1987.

CORTÉS, Fernando, Rosa María Rubalcava y Ricardo Yocelevzky, *Metodología vol. I*, SEP, Universidad de Guadalajara y Comecso, 1987.

EVERITT, B. S. *The Analysis of Contingency Tables*, Londres, Chapman and Hall, 1977.

CHATTERJEE, Samprit and Bertrand Price. *Regression Analysis by Example*, Nueva York, John Wiley and Sons, 1977.

GALTUNG, Johan. *Teoría y métodos de la investigación social*, tomo I, Buenos Aires, EUDEBA, 1966.

GOLDBERGER, Arthur. *Econometric theory*, Nueva York, John Wiley, 1964.

GONZÁLEZ Casanova, Pablo. *La falacia de la investigación en ciencias sociales*, México, Océano, 1987.

GOODMAN, Leo y William Kruskal. "Measures of Association for Cross Classifications", en *Journal of the American Statistical Association*, núm. 49, 1954.

HAGOOD, Margaret. "The Notion of a Hipothetical Universe", en Morrison, Denton y Ramon Henkel. *The Significance Test Controversy*, Chicago, Aldine Publishing Co., 1973.

HALBWACHS, Fr. "Reflexiones sobre la causalidad en física", en Bunge, M., F. Halbwachs, Th. S. Kuhn, *et al. Teorías de la causalidad*, Ediciones Sígueme, 1977.

HANSON Norwood, Russsell. *Patterns of Discovery: an Inquiry into the Conceptual Foundations of Science*, Cambridge University Press, 1965.

HILDEBRAND, David, J. Laing y H. Rossenthal. *Analysis of Ordinal Data*, California, Sage Publications, 1977.

INTRILIGATOR, Michael D. *Econometric Models, Techniques & Applications*, Prentice-Hall, Englewood Cliffs, 1978.

JOHNSTON, J. *Econometric Methods*, Singapur, McGraw-Hill, 1984.

——. *Econometric Methods*, Nueva York, McGraw-Hill, 1960.

KERLINGER, Fred. *Foundations of Behavioral Research*, 2a. ed., Nueva York, Rinehart and Winston, 1973.

KIM, Jae-On y Charles W. Mueller. *Factor Analysis. Statistical Methods and Practical Issues*, núm. 14, Sage Publications, Beverly-Hills, 1978, Serie: Quantitative Applications in the Social Sciences.

KING, Gary. *Unifying Political Methodology*, Cambridge (EUA), Cambridge University Press, 1989.

KLECKA, William R. *Discriminant Analysis*, núm. 19, Los Angeles, Sage Publications, 1984, Serie: Quantitative Applications in the Social Sciences.

LAND, Kenneth. "Principle of Path Analysis", en *Sociological Methodology 1969", San Francisco, Jossey-Bass, 1969.*

LAZARSFELD, Paul. "El álgebra de los sistemas dicotómicos", en Raynond, Boudon y Paul Lazarsfeld. *Metodología de las ciencias sociales II: análisis empírico y causalidad*, Barcelona, Laia, 1974(a).

——. "La interpretación de las relaciones estadísticas como propiedad de investigación", en Raymond, Boudon y Paul Lazarsfeld. *Metodología de las ciencias sociales II: análisis empírico y causalidad*, Barcelona, Laia, 1974(b).

——. "De los conceptos a los índices empíricos", en *Metodología de las ciencias sociales I: conceptos e índices*, Barcelona, Laia, 1974.

LAZLO, Irvin. *La gran bifurcación. Crisis y oportunidad: anticipación del nuevo paradigma que está tomando forma*, Barcelona, Gedisa, 1990.

LONG. J. Scott. *Confirmatory Factor Analysis*, núm. 33, Sage Publications, Beverly-Hills, 1983, Serie: Quantitative Applications in the Social Sciences.

MADDALA, G.S. *Limited-Dependent and Qualitative Variables in Econometrics*, Nueva York, Cambridge University Press, 1989.

MAGRASSI, Guillermo E., Manuel M. Roca, *et al. La "historia de vida"*, Buenos Aires, Centro Editor de América Latina, 1980.

PIAGET, Jean y Rolando García. *Psicogénesis e historia de la ciencia*, México, Siglo XXI, 1982.

——. *Las explicaciones causales*, Barcelona, Barral, 1973.

PRIGOGINE, Ilya e Isabel Stengers. *La nueva alianza: metamorfosis de la ciencia*, Madrid, Alianza Editorial, 1983.

QUINE, W.V. "Five Milestones of Empiricism", cap. 7, en *Theory and Things*, Harvard University Press, 1981, pp. 67 a 72.

RUMMEl, R. J. "Understanding Factor Analysis", en *Journal of Conflict Resolution*, núm. 11, 1967.

SCHUSTER, Félix Gustavo. *Explicación y predicción*, Buenos Aires, Clacso, 1982.

SIEGEL, Sidney. *Nonparametric Statistics for the Behavioral Sciences*, Nueva York, McGraw-Hill, 1956.

SIMON, Hubert. "Spurious Correlation a Causal Interpretation", en *Models of Man*, Nueva York, John Wiley. 1957.

SPSS/PC+V2.0. *Base manual*, Chicago, SPSS Inc., 1988.

THEIL, Henri. *Statistical Decomposition Analysis*, Amsterdam, North Holland, 1972.

TRYON, Robert C. y Daniel Bailey. *Cluster Analysis*, Nueva York, McGraw-Hill, 1970.

WAGENSBERG, Jorge (ed.). *Proceso al azar*, Barcelona, Tusquets Editores, 1986.

WRIGHT, Sewall: "Correlation and Causation", en *Journal of Agricultural Research*, núm. 20, 1921.

YULE, G. Udny y Maurice G. Kendall. *Introducción a la estadística matemática*, Madrid, Aguilar, 1959.

——. *An Introduction to the Theory of Statistics, Nueva York, Hafner, 1911.*

Bibliografía de consulta

Para un lector más interesado en las aplicaciones de la estadística a las ciencias sociales y las humanidades que en el conocimiento estadístico en sí mismo, o en sus bases matemáticas, los textos de supuesta difusión adolecen siempre del defecto de presentar "aplicaciones" a situaciones de investigación ficticias. Sin duda el tratamiento de estos temas dentro de un libro que expone los resultados de una investigación "real" lo harían prácticamente ilegible.

Por eso creemos que el lector debiera comenzar por adentrarse al modelo estadístico que considere más interesante a través de un texto de estadística, ya sea básico o de difusión, y después dirigirse hacia su aplicación en los libros y artículos, producto de investigación en su disciplina, en los que ciertamente no encontrará más que referencias muy generales a la metodología y a los modelos estadísticos utilizados. Esta estrategia permitirá entender las bases de las técnicas y, sobre todo, apreciar sus alcances en tér: :s de la riqueza de las interpretaciones de sus resultados.

A continuación sugerimos un conjunto de lectu e acuerdo con esta propuesta.

Bibliografía básica

1) Análisis de conglomerados

ALDENDERFER, Mark S. y Roger Blashfield. *Cluster Analysis*, núm. 44, Sage Publications, Beverly Hills, 1984, Series: Quantitative Applications in the Social Sciences.

La presentación toma como base dos ejemplos artificiales de investigación en arqueología y psicopatología. Se discuten varias medidas de similitud, se hace una revisión de los métodos de aglomeración, se presentan técnicas de validación y se hace referencia a los programas de cómputo y a la literatura general sobre el análisis de conglomerados.

2) Análisis discriminante

KLECKA, William R. *Discriminant Analysis*, núm. 19, Sage Publications, Beverly Hills, 1980, serie: Quantitative Applications in the Social Sciences.

La exposición de esta técnica aprovecha ejemplos tomados de una investigación en ciencia política. Se discuten los supuestos del modelo y la derivación e interpretación de las funciones discriminantes. También se presentan los procedimientos de clasificación y de inclusión selectiva de variables, con mención de los principales criterios de selección.

3) Análisis factorial

KIM, Jae-On y Charles W. Mueller, *Introduction to Factor Analysis. What It Is and How to Do It*, núm. 13. Sage Publications, Beverly Hills, 1978, serie: Quantitative Applications in the Social Sciences.

En este texto se presentan los fundamentos lógicos de esta técnica y la correspondencia entre modelos factoriales y estructuras de covarianzas. También se exponen las etapas en la aplicación del método, el uso de paquetes estadísticos y las complicaciones que surgen del análisis de datos reales.

KIM, Jae-On y Charles W. Mueller. *Factor Analysis. Statistical Methods and Practical Issues*, núm. 14, Sage Publications, Beverly

Hills, 1978, serie: Quantitative Applications in the Social Sciences.

Este volumen profundiza en los métodos para extraer los factores iniciales y en los métodos de rotación. Trata el problema del número de factores, presenta una introducción al análisis factorial "confirmatorio" y discute el problema de la construcción de escalas. Ofrece también una sección con respuestas a las preguntas más frecuentes acerca de este método.

4) Análisis de regresión

CHATTERJEE, Samprit y Bertram Prince. *Regression Analysis by Example*, Nueva York, John Wiley, 1977.

Este libro presenta de manera resumida y clara los principales problemas de estimación en el modelo de regresión. La exposición es conceptualmente profunda con requerimientos matemáticos mínimos. Hace uso extenso de representaciones gráficas y de ejercicios numéricos, con los cuales muestra de manera simple las consecuencias si no se cumplen los supuestos en que se basa la estimación mínimo cuadrática ordinaria. Pueden acudir a este libro todas aquellas personas que quieran informarse de manera rápida y precisa de temas tales como: heterocedasticidad, autocorrelación, regresiones con variables independientes cualitativas, multicolinealidad, etcétera.

Aquellos que necesiten un conocimiento más completo y profundo del modelo de regresión, pueden recurrir al libro ya clásico y cuya primera edición data de 1960: JOHNSTON, J. *Econometric Methods*, Nueva York, McGraw-Hill, 1984.

5) Regresión logística

ALDRICH, John y Nelson Forrest. *Linear Probability, Logit, and Probit Models*, núm. 45, California, Sage Publications, serie: Quantitative Applications in the Social Sciences 1984.

Es un excelente texto para comprender por qué es inadecuado utilizar el modelo de regresión lineal cuando la variable dependiente es dicotómica, o bien, su recorrido está limitado a un intervalo. La exposición se basa en ejemplos de investigación y demanda conocimientos mínimos de matemáticas, aunque requiere un manejo fluido de los conceptos estadísticos básicos.

Las personas que pretendan aplicar el modelo logístico a sus propias investigaciones deberían consultar: Hosmer, David W. y Lemeshow Stanley. *Applied Logistic Regression*, Nueva York, John Wiley, 1989; obra fuertemente orientada hacia la construcción de modelos, interpretación de los parámetros y hacia la evaluación del grado de bondad de ajuste. Los ejemplos se refieren al campo de la salud.

Un conocimiento más profundo se puede obtener en el libro de: Agresti, Alan. *Categorical Data Analysis*, Nueva York, John Wiley, 1990; que muestra con detalle la matemática de este modelo y sus problemas de estimación y bondad de ajuste; también dedica un capítulo entero a modelos multinomiales y al tratamiento de variables ordinales.

En el libro de: Maddala, G. S. *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, 1989, se presenta un tratamiento matemático riguroso.

6) Análisis de trayectorias

ASHER, Herbert. *Causal Modeling*, núm. 3, California, Sage Publications, 1976, serie: Quantitative Applications in the Social Sciences.

Después de hacer una revisión histórica y analítica de las principales contribuciones al tema, presenta los conceptos básicos del análisis de trayectorias. Utiliza ejemplos tomados de investigaciones para exponer los procedimientos que permiten pasar de un sistema de hipótesis expresadas en lenguaje natural a una gráfica y, a partir de ésta, escribir sistemas de ecuaciones recursivas y lineales. Enseña de manera didáctica las reglas de Wright para calcular qué valor toman las correlaciones calculadas a partir del modelo y evaluar su bondad de ajuste, comparándolas con las correlaciones observadas. Muestra, además, las reglas que se usan para calcular los impactos directos e indirectos entre las variables.

Una presentación, en el mismo estilo que la obra de Asher, pero con una utilización un poco más intensa de las matemáticas se encuentra en: Land Kenneth, C. "Principles of Path Analysis", en *Sociological Methodology, 1969*, San Francisco, Josey-Bass, 1969.

En los casos en que fuera necesario manejar niveles de complejidad mayores que los expuestos en las referencias anteriores, se recomienda recurrir al capítulo "Sistema de ecuaciones" de un

buen libro de econometría, por ejemplo, al de Johnston citado en análisis de regresión.

7) Análisis de contingencia y de asociación

CORTÉS, Fernando y Rosa María Rubalcava. *Métodos estadísticos aplicados a la investigación en ciencias sociales: análisis de asociación*, México, El Colegio de México, 1987.

Este libro presenta las nociones básicas de los análisis de contingencia y de asociación a partir de ejemplos seleccionados de la investigación social realizada en América Latina, evitando, en lo posible, las complejidades matemáticas. Permite introducirse al tema ligando los problemas conceptuales con las tablas de contingencia y los coeficientes de asociación.

Un tratamiento simple para el caso particular de variables nominales (aunque exige algún manejo de matemáticas) se encuentra en: Reynolds, H. T. *Analysis of Nominal Data*, núm. 7, California, Sage Publications, 1984, serie: Quantitative Applications in the Social Sciences. Para variables ordinales ver: Hildebrand, D. J. Laing y H. Rosenthal. *Analysis of Ordinal Data*, núm. 8, California, Sage Publications, 1977, serie: Quantitative Applications in the Social Sciences.

El texto clásico para examinar en profundidad tanto el análisis de contingencia como el de asociación es el de: Kendall, M. G. y A. Stuart. *The Advanced Theory of Statistics, vol. 2*, Londres, Charles Griffin, 1961.

8) Análisis loglineal

KNOKE, David y Peter Burke: *Log-Linear Models*, núm. 20, California, Sage Publications, 1980, serie: Quantitative Applications in the Social Sciences.

Este libro proporciona la manera más simple de introducirse al estudio del modelo loglineal. Desarrolla los conceptos básicos de este modelo apoyándose en un nutrido número de ejemplos de investigación. Introduce paulatina y cuidadosamente tanto la construcción de modelos, como la interpretación de los coeficientes y la contrastación.

En los capítulos 4, 5 y 6 del libro de : Everitt, B. S. *The Analysis of Contingency Tables*, Londres, Chapman and Hall, 1977, se muestra un desarrollo que vincula las tablas de contingencia multidimensionales con el análisis loglineal. Este libro exige que el lector tenga una buena formación en inferencia estadística. El libro clásico para este tema es el de: Bishop, Y., S. Fienberg y P. Holland, *Discrete Multivariate Analysis: Theory and Practice*, Cambridge Massachusset, MIT, 1975, que lo expone tanto en su complejidad conceptual como matemática.

9) Paquetes de programas estadísticos

Para paquetes de programas estadísticos para computadora, sugerimos consultar las secciones correspondientes a estos modelos en: SPSS/PC. *Statistical Package for the Social Sciences. Advanced Statistics V2.0*, Chicago, SPSS, 1988.

Bibliografía de aplicaciones

UNIKEL, Luis, con la colaboración de Crescencio Ruiz Ch. y Gustavo Garza. *El desarrollo urbano de México*, cap. IV, México, El Colegio de México, 1976.

En el capítulo IV, "El proceso de metropolización en México", se aplica el análisis discriminante para delimitar áreas de influencia de centros urbanos. El interés focaliza las doce zonas metropolitanas más importantes y los ciento quince municipios que podrían conformarlas. A través del modelo estadístico fue posible identificar los municipios centrales, periféricos y en transición de las zonas metropolitanas.

Rubalcava, Rosa María y Marta Schteingart. "Estructura urbana y diferenciación socioespacial en la zona metropolitana de la Ciudad de México (1970-1980)", en *Atlas de la Ciudad de México*, México, Departamento del Distrito Federal y El Colegio de México, 1987.

Este trabajo muestra una aplicación del análisis factorial para identificar los factores más importantes en la diferenciación de las delegaciones y municipios que conforman la zona metropolitana, caracterizados por variables físicas, espaciales y demográficas de

los censos de población y vivienda. La construcción de índices para los factores resultantes permitió diferenciar áreas según sus niveles de consolidación urbana y socio-económico, y comparar resultados de los años 1970 y 1980.

JEONG-HWA, Lee y Adam Przeworski. *Cui Bono? Corporatims and Welfare*, 1990, (mimeo).

Utiliza análisis de regresión múltiple aplicado a catorce países para examinar el impacto del "poder organizacional de los trabajadores" y del "control partidario de la izquierda" sobre el ingreso esperado. Uno de los hallazgos más interesantes es que los ingresos promedio por adulto, por persona ocupada y por trabajador en la manufactura, son menos en aquellos países en que la izquierda tiene mayor poder político.

SMITH, Peter H. *Los laberintos del poder: el reclutamiento de las élites políticas en México, 1900-1971*, cap. 4, México, El Colegio de México, 1981.

Este autor emplea el análisis de trayectorias para encontrar las determinantes del cargo más elevado que alcanzan los miembros de las élites políticas, distinguiendo entre la cohorte prerrevolucionaria (1900-1911), la revolucionaria (1917-1940) y la postrevolucionaria (1946-1971). Las conclusiones más generales que obtiene del análisis le permiten sostener que, a lo largo del siglo XX, el origen social ha tenido impacto diferencial en el reclutamiento político. Las probabilidades de acceso a la élite de la política nacional han dependido de la educación, la ocupación, la ocupación del padre y, hasta cierto punto, el lugar de nacimiento. Sin embargo, una vez incorporado un miembro a la élite, el origen social pierde relevancia en la determinación del puesto más alto que puede ocupar en la escala política.

GARCÍA, Brígida. *Desarrollo económico y absorción de fuerza de trabajo en México: 1950-1980*, cap. V. México, El Colegio de México, 1988.

Ajusta un modelo loglineal a la evolución de la población económicamente activa, según rama de actividad y posición en la ocupación, con datos censales comprendidos entre los años 1950 y 1980. Los resultados de la investigación mostraron que en la agricultura y el comercio prevalecen los trabajadores por cuenta propia, mientras que en la industria, servicios y construcción tienen mayor importancia los asalariados. El modelo permite afirmar, además, que esta situación no varió a lo largo del periodo.

CHRISTENSON, B., Brígida García y Orlandina de Oliveira. "Los múltiples condicionantes del trabajo femenino en México", en *Estudios Sociológicos*, núm. 20, vol. VII, México, El Colegio de México, 1989.

Utilizan el modelo logit para identificar el impacto de las determinantes del trabajo femenino en México. Identificaron las condicionantes contextuales (regiones y diferenciación rural-urbana), y las condicionantes familiares e individuales. Aun cuando una mujer se desenvuelva en un contexto favorable a su participación económica, operan factores individuales y familiares que contrarrestan su efecto: la carencia de instrucción, la presencia de muchos hijos entre las casadas y la de un padre o esposo para afrontar las responsabilidades económicas.

Las aplicaciones del análisis de contingencia y el de asociación son tan frecuentes que no consideramos necesario citar alguna en particular: En el libro básico sugerido para estas técnicas (Cortés, F. y R. M. Rubalcava 1987), se mencionan varias investigaciones desarrolladas en América Latina que utilizaron estos modelos.