ANALISIS DE TRAYECTORIA Y CONSTRUCCION DE MODELOS

Sir Maurice Kendall
C.A. O'Muircheartaigh

PATH ANALYSIS AND MODEL BUILDING

SUMMARY

Path analysis models can be very helpful in disentangling a complex set of relationships and, used with care, can add considerably to our knowledge of the mechanisms at work in the population.

En un sentido amplio, todo análisis científico consiste en la construcción de un modelo. Ya sea en las ciencias físicas o en las ciencias sociales, el científico trata de resumir la complejidad del mundo fenomenal en la forma de aserciones, leyes, hipótesis o modelos simplificados, y por dos razones principales: para comprender y para controlar.

En este sentido, un "modelo" no siempre es una copia física del sistema estudiado, aunque esto es posible, como ocurre cuando una aeronave recién diseñada se prueba parcialmente observando el comportamiento de un modelo en un túnel de viento. En términos más generales, entiéndese por "modelo" una descripción de las relaciones que vinculan entre sí a las variables que interesan - las reglas del juego. El proceso de construcción de un modelo consiste en reunir una serie de expresiones formales de estas relaciones hasta que el comportamiento del modelo reproduce adecuadamente el comportamiento del sistema.

Los modelos se construyen para fines específicos y no tratan necesariamente de describir en detalle cada faceta del sistema. En realidad, la utilidad de algunos modelos reside tanto en lo que omiten por innecesario como en lo que retienen. Así, un mapa carretero cumple su finalidad aunque ignore las calles secundarias y los detalles topográficos sin importancia y un modelo de distribución del ingreso puede, para algunos fines, presentar únicamente la distribución real sin considerar las incontables circunstancias que determinan el ingreso de una persona en particular.

Uno de los problemas básicos en el campo del análisis estadístico es la especificación del modelo que ha de emplearse, esto es, la forma matemática de la población de la cual los datos se consideran una muestra. El problema consiste en inferir de la distribución probable de las variables observadas, la estructura subyacente que genera esta distribución observada. En el campo de las ciencias sociales, los datos observados provienen generalmente de situaciones *no experimentales;* cuando la comprobación experimental no es posible, se puede recurrir a los procedimientos estadísticos.

Los modelos propuestos contienen frecuentemente variables latentes que, aunque no se observan directamente, tienen repercusiones en las relaciones entre las variables observables. En las Guías de Análisis *(Guidelines for Analysis)* se distingue entre variables "explicativas" y variables "intermedias". La idea básica es que las variables explicativas, aunque en general son observables con mayor facilidad, causalmente están más lejos de la fecundidad que las variables intermedias y actúan a través de ellas. Puede considerarse que factores como la frecuencia del coito, la práctica de la anticoncepción y los períodos de lactancia tienen un efecto causal directo en la concepción. En cierto sentido, son variables que explican más que las variables explicativas. Se considera que estas últimas, por ejemplo el nivel educacional, el ingreso y la edad al casarse, influyen en la fecundidad y se las puede describir como causales; pero su influencia se expresa principalmente a través del efecto que ellas mismas producen sobre las variables intermedias.

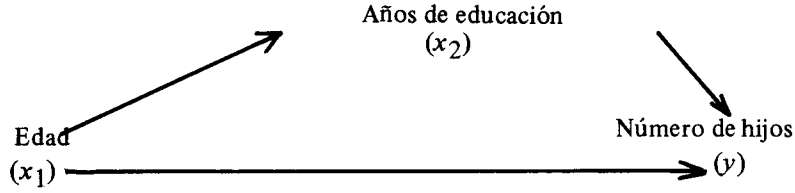
Además, los modelos se construyen generalmente con varias ecuaciones o submodelos que actúan entre sí y deben considerarse simultáneamente. Esta interdependencia de relaciones entre las variables es la fuente de muchas de las dificultades que se presentan cuando se trata de describir adecuadamente un conjunto de datos usando métodos estadísticos convencionales.

En su mayor parte (aunque esto no tiene el carácter de una regla inviolable), los modelos de fecundidad parece que son más útiles si relacionan la fecundidad misma con las variables explicativas, comprueban o refutan las hipótesis referentes a esa relación, y de ser posible,

cuantifican la contribución de cada variable al comportamiento de la fecundidad. Un ejemplo sencillo puede aclarar algunas de las dificultades. Supóngase que nos interesa la relación entre la fecundidad (medida por el número de hijos nacidos) y las dos variables explicativas edad (x_1), y número de años de educación (x_2). Para simplificar la exposición, supóngase que para cada miembro de la muestra se ha determinado un valor satisfactorio de las x y de las y.

Estas son variables "explicativas" y la primera cuestión es saber si necesitamos escribir en el modelo las variables intermedias. La decisión depende de lo que tratemos de hacer con el modelo. Desde cierto punto de vista, las variables explicativas pueden considerarse como estímulos capaces de producir una respuesta y a través de algún mecanismo que no es de interés inmediato. La situación es entonces como la de una "caja negra" que relaciona estímulos y respuestas no preocupándonos de cómo actúan las relaciones dentro de ella. (Puede que en algún momento queramos examinarlas y destapar la caja, pero no es así por el momento). En tal caso los lazos causales entre las x y las y son considerados como directos, aun cuando se reconoce que la causalidad es más tortuosa. Existe un número de relaciones que desearíamos que contuviera el modelo. Primero, esperamos que la edad influya directamente en la fecundidad (quizás a través de algunas variables intermedias no incluídas en el modelo). También podríamos esperar que los años de educación influyeran directamente en la fecundidad. Podemos descartar la posibilidad de que los años de educación influyan en la edad, pero podemos suponer que la edad estará relacionada con la duración de la educación sistemática y que a través de ella influya directamente sobre la fecundidad.

Representamos gráficamente este modelo a continuación, usando flechas en un sentido que llevan de cada variable explicativa a cada variable en que ella influye directamente.



Suponemos que las relaciones son lineales, o que se les ha dado a las variables una forma adecuada para justificar la linealidad. Las rela-

ciones pueden escribirse

$$x_2 = \beta_{21}\ x_1 \qquad \text{(1a)}$$

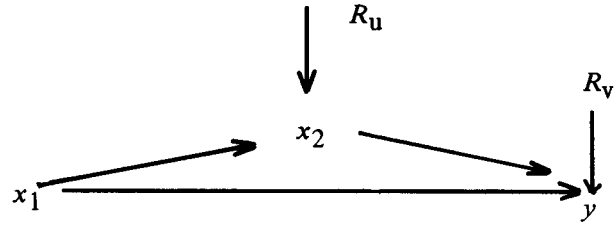
$$y = \beta_{01}\ x_1 + \beta_{02}\ x_2 \qquad \text{(1b)}$$

Sin embargo, estas ecuaciones no son exactas en la práctica y es necesario prever algún margen de inexactitud. Esto se hace generalmente, como en el caso de la regresión, agregando un término a la derecha. Pero ésta no es necesariamente una variable aleatoria. Representa algo que, a propósito o accidentalmente, hemos omitido del modelo, pero que esperamos que no sea tan serio como para menoscabar la representación aproximada que proporciona. Pero para adelantar con la estimación, necesitamos todavía suponer algo acerca de este término residual. Lo que supondremos es que no está correlacionado con ninguno de los determinantes inmediatos de la variable dependiente a que pertenece. Las ecuaciones pueden escribirse ahora

$$x_2 = \beta_{21\,x\,1} + \beta_{2u}\ R_u \qquad \text{(1a)'}$$

$$y = \beta_{01}\ x_1 + \beta_{02}\ x_2 + \beta_{0v}\ R_v \qquad \text{(1b)'}$$

El diagrama puede modificarse en la siguiente forma para incorporar en él los términos residuales:



Sin menoscabo de la generalidad, suponemos por el momento que todas las variables se han estandarizado según una media igual a cero y una variancia igual a la unidad. Convencionalmente, los coeficientes de las ecuaciones con variables estandarizadas se denominan *coeficientes de trayectoria* y se escriben p_{ij}, cuyo primer subíndice representa la variable dependiente y el segundo, la variable cuyo efecto directo sobre la variable se mide mediante el coeficiente de trayectoria. Por lo tanto, el sistema puede escribirse

$$x_2 = p_{21} + p_{2u}\ R_u \qquad \text{(1a)''}$$

$$y = p_{01}\ x_1 + p_{02}\ x_2 + p_{0v}\ R_v \qquad \text{(1b)''}$$

Es un sistema recurrente; en otras palabras, no existen en él canales de retroalimentación por medio de los cuales las x_i pudieran influir en él. En general, no consideraremos modelos que comprendan un canal de retroalimentación directo o indirecto.

El próximo paso es la estimación de los coeficientes de trayectoria. Pueden usarse dos métodos:

(1) *Descomposición de los coeficientes de correlación*

Puesto que las variables están estandarizadas, el coeficiente de correlación r_{ij} puede escribirse

$$r_{ij} = \frac{1}{n} \sum x_i x_j$$

Así, de (1a)"

$$r_{12} = \frac{1}{n} \sum x_1 x_2 = \frac{1}{n} \sum x_1 \ (p_{21} \ x_1 + p_{2u} \ R_u)$$

$$= p_{21} + 0 \text{ puesto que } \frac{1}{n} \sum x_1^2 = 1 \qquad (2a)$$

Y R_u no está correlacionada con x_i.

Del mismo modo

$$r_{01} = \frac{1}{n} \sum x_1 y = \frac{1}{n} \sum x_1 \ (p_{01} \ x_1 + p_{02} \ x_2 + p_{0v} \ R_v)$$

$$= p_{01} + p_{02} \ r_{12} \qquad (2b)$$

y

$$r_{02} = \frac{1}{n} \sum x_2 y = \frac{1}{n} \sum x_2 \ (p_{01} x_1 + p_{02} \ x_2 + p_{0v} \ R_v)$$

$$= p_{01} \ r_{12} + p_{02} \qquad (2c)$$

Las ecuaciones (2b) y (2c) nos permiten despejar p_{01} y p_{02} en función de r_{01}, r_{02} y r_{12}, obteniendo

$$p_{01} = \frac{r_{01} - r_{02} \ r_{12}}{1 - r_{12}^2} \qquad (2c)$$

$$p_{02} = \frac{r_{02} - r_{01}\, r_{12}}{1 - r_{12}^2} \qquad (2d)$$

Así, partiendo de las ecuaciones (2a), (2c) y (2d), los coeficientes de trayectoria p_{01}, p_{02} y p_{21} pueden obtenerse directamente de los coeficientes de correlación.

En este sencillo modelo, los datos provenientes de la encuesta de Fiji, con 4 928 cuestionarios contestados, dan los siguientes valores de los coeficientes de correlación:

$r_{01} = 0{,}64$ (correlación entre la edad y el número de hijos)

$r_{02} = -0{,}34$ (correlación entre los años de educación y el número de hijos)

$r_{12} = -0{,}32$ (correlación entre la edad y los años de educación)

Sustituyendo estos valores en (2a), (2c) y (2d) se obtiene

$$\left.\begin{aligned} p_{21} &= -0{,}34 \\ p_{01} &= 0{,}59 \\ p_{02} &= -0{,}15 \end{aligned}\right\} \qquad (3a)$$

Las trayectorias residuales pueden obtenerse en forma sencilla usando

$$r_{22} = 1 = \frac{1}{n} \sum x_2^2 = \frac{1}{n} \sum x_2 \,(p_{21}\, x_1 + p_{2u}\, R_u)$$

$$= p_{21}^2 + p_{2u}^2$$

De aquí

$$p_{2u}^2 = (1 - p_{21}^2)$$

esto es

$$p_{2u} = \sqrt{(1 - p_{21}^2)} \qquad (2e)$$

y

$$r_{00} = 1 = \frac{1}{n} \sum y^2 = \frac{1}{n} \sum y\,(p_{01}\,x_1 + p_{02}\,x_2 + p_{0v}\,R_v)$$

$$= p_{01}^2 + p_{02}^2 + 2p_{01}\,p_{02}\,p_{21} + p_{0v}^2$$

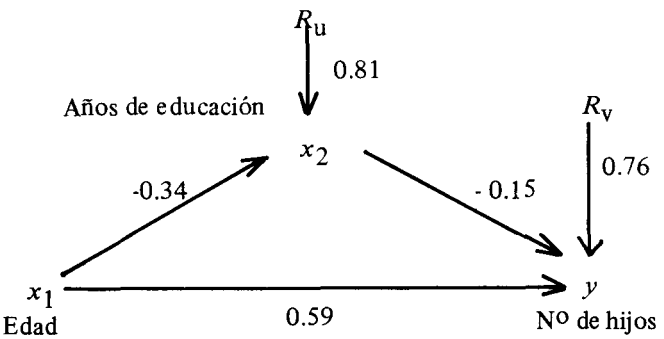
esto es

$$p_{0v} = \sqrt{1 - p_{01}^2 - p_{02}^2 - 2p_{01}\,p_{02}\,p_{21}} \qquad (2f)$$

En este caso

$$\left.\begin{aligned} p_{2u} &= 0{,}81 \\ p_{0v} &= 0{,}76 \end{aligned}\right\} \qquad (3b)$$

Si incorporamos los valores de (3a) y (3b) en el diagrama, obtenemos



(2) *Ecuaciones de regresión*

Este modelo de trayectoria equivale a una serie de análisis ordinarios de regresión y las soluciones de las ecuaciones simultáneas (la)" y (lb)" son simplemente los coeficientes de regresión estandarizados - los "coeficientes beta". Así, la trayectoria p_{21} puede obtenerse mediante la regresión de x_2 sobre x_1, y las trayectorias p_{01} y p_{02} pueden obtenerse mediante la regresión de y sobre x_1 y x_2, usando los mínimos cuadrados ordinarios. Es éste un resultado útil, pues incorpora el método del análisis de trayectorias al marco del análisis estadístico corrien-

te y suministra estimaciones de los errores típicos de los coeficientes que se obtienen.

Como técnica estadística, sin embargo, el análisis de trayectoria no agrega nada al análisis de regresión convencional cuando se aplica por recurrencia a un sistema de ecuaciones, pero aclara la racionalidad del sistema de ecuaciones de regresión y constituye "un método para medir la influencia directa a lo largo de cada trayectoria separada de ese sistema y así, encontrar el grado en que la variación de un efecto dado está determinada por cada causa particular. El método descansa en la combinación del conocimiento del grado de correlación que existe entre las variables de un sistema y del conocimiento que se tenga de las relaciones causales" (Wright 1921).

Un coeficiente de trayectorias convencional da el efecto esperado de un cambio de una desviación estándar en la variable explicativa (manteniéndose constantes las otras variables); este cambio esperado se expresa en función de la desviación estándar de la variable predicha. En este ejemplo queremos repartir la explicación de la variable dependiente entre las dos variables explicativas.

El *efecto total* de la edad puede expresarse mediante la correlación entre la edad y el número de hijos, esto es $r_{01} = 0{,}64$. La ecuación (2b) nos muestra que esto puede expresarse como la suma de dos componentes: p_{01}, el *efecto directo* de la edad, y $p_{02}\, r_{12}$ ($= p_{01}, p_{21}$), el *efecto indirecto* de la edad cuando actúa a través de los años de educación. Numéricamente esto se escribe:

$$0{,}64 = 0{,}59 + (-0{,}34)(-0{,}15)$$

O sea, el efecto directo es + 0,59, y el indirecto + 0,05.

Sin embargo, el efecto total de los años de educación no lo da la correlación de los años de educación con el número de hijos. Parte de esta correlación se debe al efecto de la variable causalmente precedente, la edad, sobre los años de educación. De este modo, en el modelo definido, el efecto total de los años de educación es el efecto directo $p_{02} = -0{,}15$. Volveremos a ocuparnos de este problema en un modelo más complejo más adelante.

Generalización del modelo

Nos ocuparemos aquí de cinco características generales de los modelos de ecuaciones simultáneas. Primero, los modelos se componen de una serie de ecuaciones cada una de las cuales contiene un término de

perturbación que resume la influencia de las variables no medidas o desconocidas sobre la estructura que interesa. O sea, los modelos no son exactos o determinísticos sino estocásticos. Segundo, la mayor parte de las aplicaciones se refieren a variables medidas en encuestas transversales y son por lo tanto modelos estáticos, no dinámicos. Tercero, por lo general los modelos excluyen la causación bipartita y son así recurrentes. Cuarto, se supone que los modelos son lineales en las variables y en las perturbaciones. Y quinto y último, se supone que los errores son independientes.

Nos interesan las relaciones lineales aditivas asimétricas entre una serie de variables que se miden en una escala de intervalos. En el diagrama cualitativo, cada variable incluida está representada o como totalmente determinada por algunas otras o como un factor final (exógeno). (En el ejemplo anterior existe una sola variable exógena - la edad). En un modelo de ecuación estructural, cada ecuación representa una relación causal más que una mera asociación empírica. Esto contrasta con un modelo de regresión, en el que cada ecuación representa la media condicional de la variable dependiente de esa ecuación como una función de las variables explicativas. El rasgo especial más importante del modelo estructural es la simultaneidad de las ecuaciones, es decir, la estimación de los parámetros de una ecuación se realiza al mismo tiempo que los de las otras ecuaciones del sistema. Las propiedades óptimas de la regresión por mínimos cuadrados ordinarios se aplican únicamente a una sola ecuación a la vez. También debemos tener en cuenta que cada ecuación está incluida en una serie de ecuaciones que constituyen nuestro modelo recurrente con perturbaciones independientes. De ahí que necesitemos un método de estimación de mejores resultados que la estimación conjunta de los parámetros del sistema que componen el modelo. El sencillo ejemplo anterior muestra que la solución del sistema de ecuaciones simultáneas (la)" y (lb)" es equivalente a la regresión por mínimos cuadrados ecuación por ecuación. Este resultado vale para todos los modelos recurrentes lineales con perturbaciones independientes (para comprobarlo, véase, por ejemplo, Land (1973) .

El supuesto inicial para el análisis de trayectoria debe ser la especificación del orden causal (o temporal) entre las variables del modelo. Los datos mismos no pueden sernos de ninguna ayuda para esto ni para la elección de las variables que hayan de incluirse en el modelo. La validez de estos supuestos no puede evaluarse a partir de los datos; su base proviene de criterios externos o de la teoría. Independientemente de nuestra ordenación, el método de análisis funcionará y dará resultados. No se captarán señales de error ni los resultados serán incompatibles. Sin embargo, la representación del modelo mediante un diagrama aclara las hipótesis y constituye una base para la apreciación crítica de los resultados.

La otra hipótesis implícita importante es la linealidad de las relaciones. Aunque esto puede no ser exacto en la práctica, la regresión lineal de y sobre x puede interpretarse siempre como la mejor aproximación lineal a la relación cuando ésta no es lineal.

También suponemos que cada ecuación es aditiva. En otros términos, suponemos que un cambio unitario en x_1, por ejemplo, tiene el mismo efecto en x_2 cualquiera que sea el valor de x_2. Y suponemos también que un cambio unitario en x_1 tiene el mismo efecto en x_2, cualesquiera que sean los valores de las otras variables. Puede que estos supuestos no sean realistas, pero afortunadamente, aunque los efectos no lineales y recíprocamente influyentes no se incluyan en el modelo sencillo, pueden introducirse en él, como lo veremos más adelante. Un examen de los términos residuales de las ecuaciones nos suministrará la prueba de si tales modificaciones del modelo son necesarias.

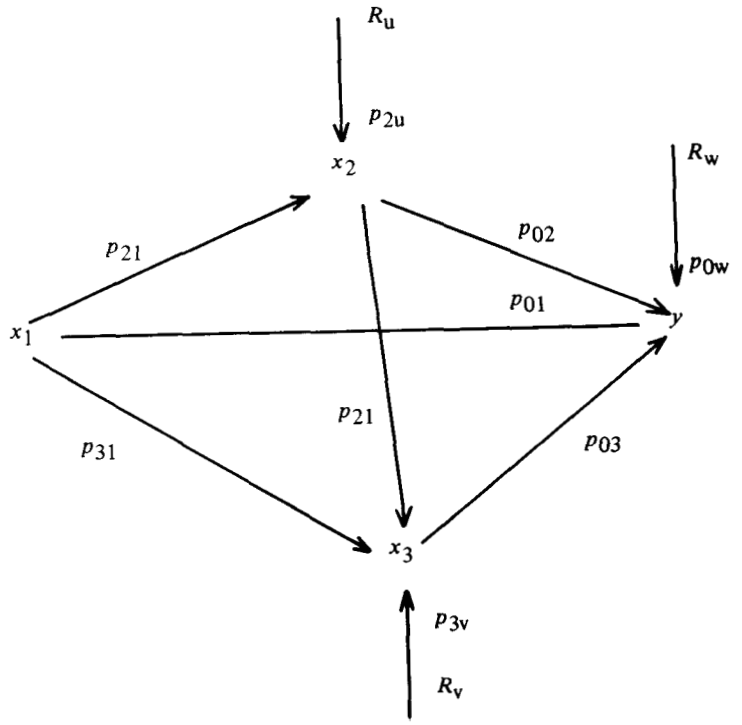
Los términos residuales se supone que no se correlacionan con todas las variables precedentes y, por consiguiente, con ninguna otra. No es necesario sin embargo considerarlos como variables reales sino simplemente como una expresión de la falta de información en el modelo y pueden por lo tanto definirse simplemente como independientes de las variables explicativas de la misma ecuación.

Más arriba se dijo que las variables explicativas se medían en una escala de intervalos. Esta regla tiene una excepción importante. Las variables binarias (dicotomías) pueden tratarse como variables con nivel de intervalo. Los valores *(scores)* dados a las dos categorías no afectarán los resultados y como variables de predicción pueden ser inapreciables. A través de ellas podemos también incorporar polinomías (variables nominativas), aunque pueden presentarse algunas dificultades de interpretación.

Tipos de modelos de trayectoria

(1) *Modelo saturado*

Puede llamarse saturado a un modelo recurrente cada una de cuyas variables se presume dependiente de todos los modelos causales anteriores. He aquí un ejemplo basado en los datos de Fiji:



donde y: número de hijos

x_1: edad en años

x_2: educación en años

x_3: edad al matrimonio

La evidencia práctica sugiere que todas las x_1, x_2 y x_3 se relacionan con la fecundidad. La relación entre x_1 y x_2 expresa el hecho de que mientras más joven es la cohorte de edad más alta es la proporción de personas con educación. La relación de x_1 a x_3 se dará si la edad al matrimonio ha cambiado en el tiempo. La relación entre x_2 y x_3 se basa en la teoría de que la educación retarda el matrimonio directamente o cambiando las alternativas que se le ofrecen a la mujer.

El orden causal implícito en este caso es que la edad es causalmente anterior a la educación, la que es causalmente anterior a la edad al matrimonio, la que a su vez es causalmente anterior al número de hijos. En realidad, este modelo simplemente incluye la variable x_3 (edad al matrimonio) en la secuencia causal entre los años de educación y el número de hijos en el modelo sencillo del primer ejemplo. El modelo puede escribirse bajo la forma de la siguiente serie de ecuaciones:

$$x_2 = p_{21}\, x_1 + p_{2u}\, R_u \tag{4a}$$

$$x_3 = p_{31}\, x_1 + p_{32}\, x_2 + p_{3v}\, R_v \tag{4b}$$

$$y = p_{01}\, x_1 + p_{02}\, x_2 + p_{03}\, x_3 + p_{0w} R_w \tag{4c}$$

Usando el primero método del ejemplo 1 podemos obtener las ecuaciones para cada uno de los seis coeficientes de correlación en función de las trayectorias del modelo. Las ecuaciones son:

$$r_{21} = p_{21} \tag{5a}$$

$$r_{31} = p_{31} + p_{32}\, r_{21} \tag{5b}$$

$$r_{32} = p_{31} + r_{12} + p_{32} \tag{5c}$$

$$r_{01} = p_{01} + p_{02}\, r_{21} + p_{03}\, r_{31} \tag{5d}$$

$$r_{02} = p_{01}\, r_{12} + p_{02} + p_{03}\, r_{32} \tag{5e}$$

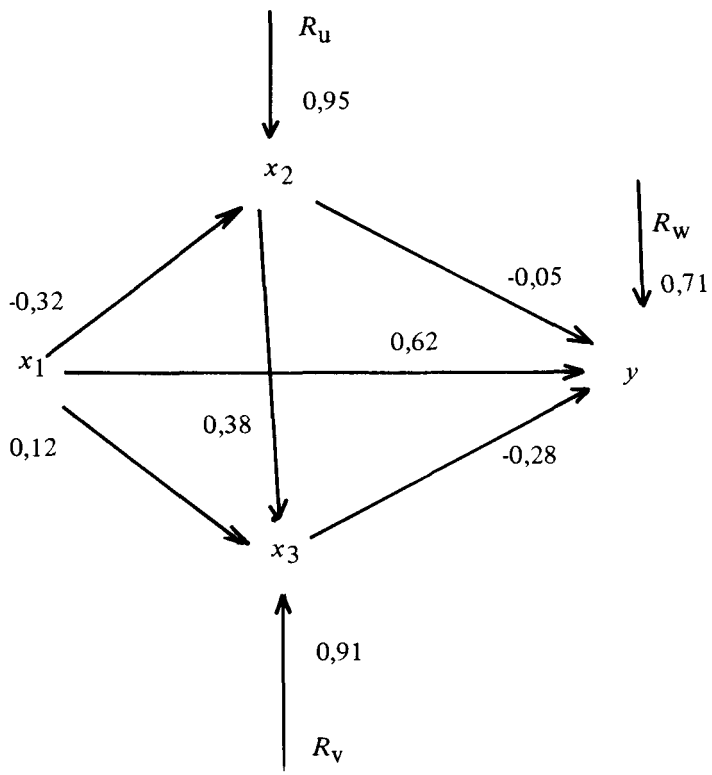
$$r_{03} = p_{01}\, r_{13} + p_{02}\, r_{23} + p_{03} \tag{5f}$$

Todas estas ecuaciones tienen la misma forma general dada por

$$r_{ij} = \sum_{q} p_{iq}\, r_{qj} \tag{5}$$

que es el teorema básico del análisis de trayectoria, donde q abarca todas las variables cuyas trayectorias llevan directamente a x_i.

La ecuación (5a) da una solución para el valor de p_{21}. Las ecuaciones (5b) y (5c) permiten resolver las dos incógnitas p_{31} y p_{32}. Las ecuaciones (5d), (5e) y (5f) aseguran la solución de p_{01}, p_{02} y p_{03}. Sin embargo, el álgebra es engorrosa aun para este modelo y es más fácil llegar a la solución realizando las tres regresiones que indican las ecuaciones (4a), (4b) y (4c). Los coeficientes de regresión tipificados son los valores de los coeficientes de trayectoria. Primero efectuamos la regresión de x_2 (educación) sobre x_1 (edad); esto da p_{21}. En seguida efectuamos la regresión de x_3 sobre x_1 y x_2, que da los coeficientes de trayectoria p_{31} y p_{32}. Por último, procedemos a la regresión de y sobre x_1, x_2 y x_3, que da p_{01}, p_{02} y p_{03}. Los coeficientes de trayectoria residuales p_{2u}, p_{3v} y p_{0w} son las raíces cuadradas de las variancias residuales de las tres regresiones. Al efectuarse estas regresiones con los datos de Fiji se obtuvieron los valores numéricos que se indican en el siguiente diagrama.



El modelo de predicción está representado por la ecuación (4c) y es

$$y = 0{,}62\ x_1 - 0{,}05\ x_2 - 0{,}28\ x_3$$

Esto representa simplemente los efectos directos de las tres variables explicativas y puede obtenerse mediante un análisis de regresión normal. La ventaja principal del modelo estructural consiste en que nos permite avanzar más en el análisis del mecanismo en cuestión.

El efecto total de la edad puede representarse mediante la correlación entre la edad y el número de hijos y es igual a 0,64. No obstante, según la ecuación (5d)

$$r_{01} = p_{01} + p_{02}\, r_{21} + p_{03}\, r_{31}$$

Desarrollando y sustituyendo r_{21} y r_{31} en las ecuaciones (5a) y (5b) tenemos

$$r_{01} = p_{01} + p_{02}\, p_{21} + p_{03}\, p_{31} + p_{03}\, p_{32}\, p_{21} \qquad (5c)$$

Esta es la descomposición de la correlación total de la edad y el número de hijos y puede interpretarse cada uno de los términos de esta fórmula.

p_{01} es el *efecto directo* de la edad = + 0,62

$p_{02}\, p_{21}$ es el *efecto indirecto* de la edad, que actúa a través de su relación con la educación; (-0,32) (-0,05) = + .0,02

$p_{03}\, p_{31}$ es el *efecto indirecto* de la edad, que actúa a través de su relación con la edad al matrimonio; (0,12) (-0,28) = -0,03

$p_{03}\, p_{32}\, p_{21}$ es el *efecto indirecto* de la edad, que actúa a través de la educación, la que a su vez actúa a través de la edad al matrimonio; (-0,32) (0,38) (-0,28) = + 0,03

Los cuatro efectos se suman al efecto total r_{01} = 0,64.

Como se indicó en el ejemplo 1, el efecto total de x_2 (años de educación) no es igual a r_{02} (que es -0,34) sino igual a la suma de las trayectorias directas e indirectas de x_2 a y.

p_{02} es el *efecto directo* de la educación; = - 0,05

$p_{03}\, p_{32}$ es el *efecto indirecto* de la educación, que actúa a través de la edad al matrimonio; (0,38) (-0,28) = - 0,10

El efecto total de la educación es $p_{02} + p_{03}\, p_{32}$ = -0,15

El efecto total de x_3 (edad al matrimonio) es el efecto directo p_{03} puesto que no existen variables entre x_3 e y en el modelo; = -0,28.

La comparación de estos resultados con los resultados obtenidos en el ejemplo 1 muestra el efecto de la introducción de una nueva variable en el modelo estructural. La variable x_3 se introdujo explícitamente en el sistema entre x_2 e y. El efecto total para x_1 y x_2 no cambia por cuanto la nueva variable es causalmente posterior a ambas. Sin embargo, la distribución de este efecto total entre los efectos directos e indirectos cambia tanto para x_1 como para x_2. Los detalles aparecen en el cuadro que se da a continuación.

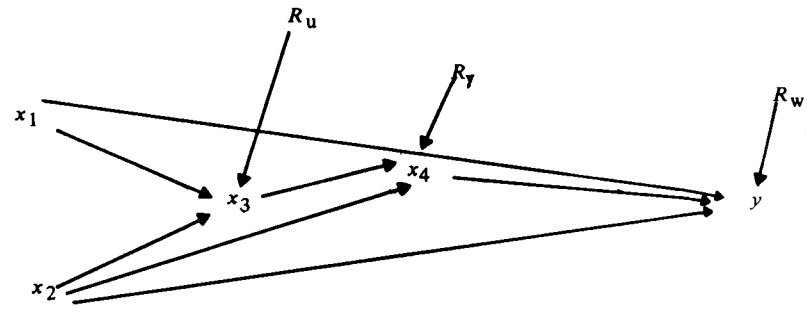
DESCOMPOSICION DEL EFECTO TOTAL PARA LA EDAD (x_1) Y LA EDUCACION (x_2)

Variable	Tipo de efecto	Excluyendo (x_3) del modelo	Incluyendo x_3 en el modelo
Edad (x_1)	*Efecto directo*	+ 0,59	+ 0,62
	Efectos indirectos		
	(i) a través de x_3	No aplicable	- 0,03
	(ii) a través de x_2 y x_3 conjuntamente	No aplicable	+ 0,03
	(iii) a través de x_2 directamente	+ 0,05	+ 0,02
Educación (x_2)	*Efecto directo*	- 0,15	- 0,05
	Efecto indirecto a través de x_3	No aplicable	- 0,10

Así, la omisión de una variable del modelo no invalida los resultados; simplemente reduce la cantidad de información que obtenemos de los datos. La introducción de la variable x_3 en el modelo no reduce el valor explicativo de la educación; en cambio, proporciona una explicación de parte del mecanismo a través del cual la educación influye en la fecundidad.

(2) *Modelo no saturado*

Puede llamarse no saturado un modelo algunas de cuyas variables no están relacionadas entre sí. El ejemplo 3 que sigue representa este sistema. Los datos utilizados también provienen del estudio de Fiji.



donde y: número de hijos

x_1: edad en años

x_2: raza

x_3: años de educación

x_4: tamaño deseado de la familia

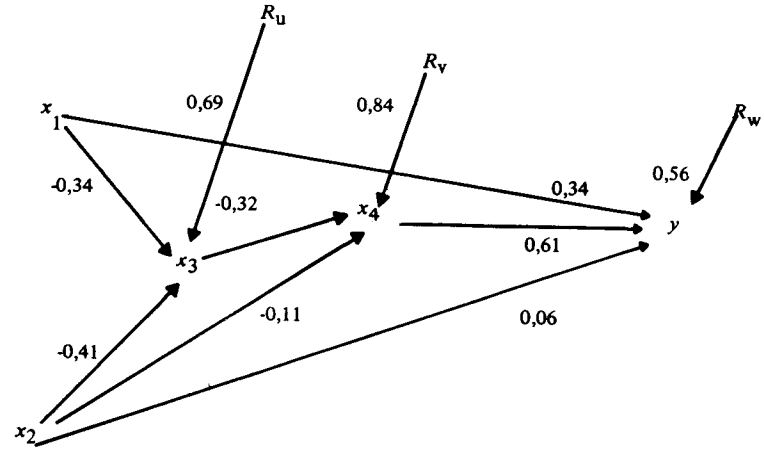
Tres puntos distinguen este ejemplo de los otros. Primero, existen dos variables exógenas (factores finales) x_1 y x_2 que, si bien son anteriores a todas las demás variables, no están ordenadas con respecto a cada una de ellas. Estas no están ligadas en el diagrama puesto que suponemos que no están correlacionadas. Segundo, la variable x_2 es una variable binaria (las dos categorías fijiana e india). Tercero, se omiten del diagrama dos trayectorias: la p_{41} y la p_{03}. El modelo puede escribirse:

$$x_3 = p_{31}\, x_1 + p_{32}\, x_2 + p_{3u}\, R_u \qquad (6a)$$

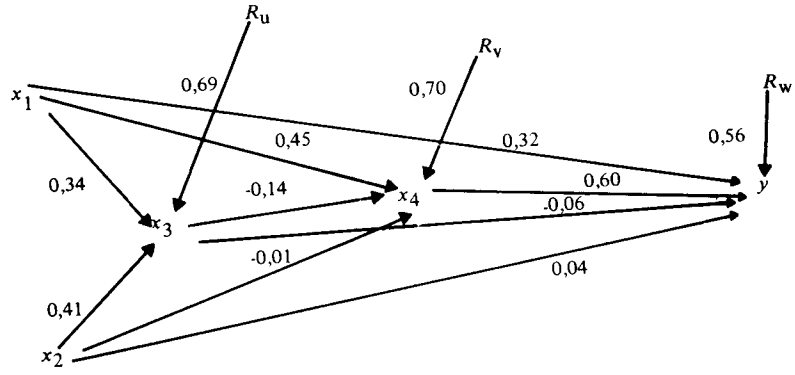
$$x_4 = p_{42}\, x_2 + p_{43}\, x_3 + p_{4v}\, R_v \qquad (6b)$$

$$y = p_{01}\, x_1 + p_{02}\, x_2 + p_{04}\, x_4 + p_{0w}\, R_w \qquad (6c)$$

Las trayectorias pueden estimarse directamente por regresión como antes, la que da los valores numéricos que se insertan a continuación:



Sin embargo, el otro método de estimación indica que pueden existir algunos problemas. Hay nueve ecuaciones simultáneas para estimar las siete trayectorias del modelo. O sea, sin restricciones a priori, el modelo resulta *superidentificado.* El método de regresión nos permite probar estas restricciones. En términos generales, procedemos estimando todos los coeficientes del modelo totalmente saturado y probando la significación de los coeficientes que deseamos eliminar. El resultado del modelo saturado es el siguiente:

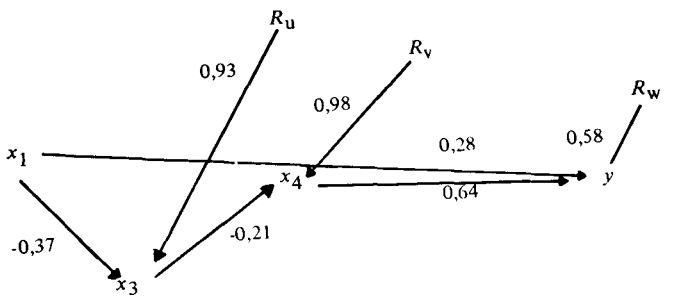


Puesto que las modificaciones del modelo afectan sólo a un coeficiente de cada una de las dos ecuaciones, cada uno de estos coeficien-

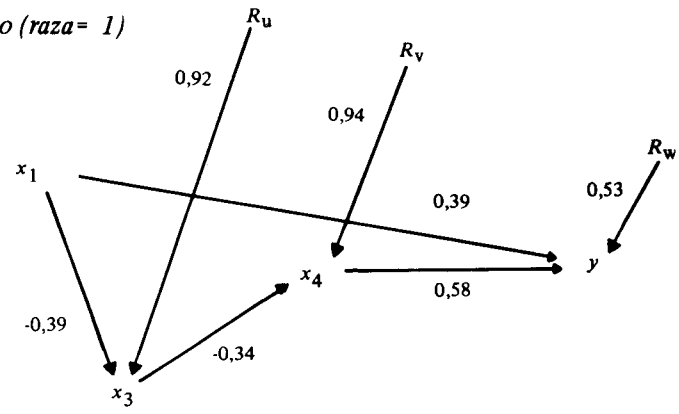
tes puede probarse usando una prueba t con $(n - p - 1)$ grados de libertad, donde p es el número de predictores de la ecuación. En realidad, ambos coeficientes son significativos y no deberían excluirse del modelo. Si queremos probar más de un coeficiente de cualquier ecuación, debemos usar una prueba F con grados de libertad suficiente para probar toda la ecuación comparándola con la ecuación en que se omiten las variables que deseamos eliminar. Como quiera que estamos trabajando con estimaciones de los coeficientes de población, es conveniente comprobar, aun cuando estemos trabajando con un modelo saturado, si los valores obtenidos se deben únicamente al error de muestreo.

La introducción de una variable binaria - la raza (x_2) - en el modelo estructural no crea ninguna dificultad especial. Formalmente, una variable binaria puede tratarse en forma totalmente correcta como una variable de nivel de intervalo. En este caso, puede examinarse con suma facilidad la hipótesis aditiva construyendo modelos separados para cada una de las dos razas y estimando directamente los coeficientes. Volviendo al modelo no saturado anterior, los resultados para los modelos separados son:

Fijiano (raza = 0)

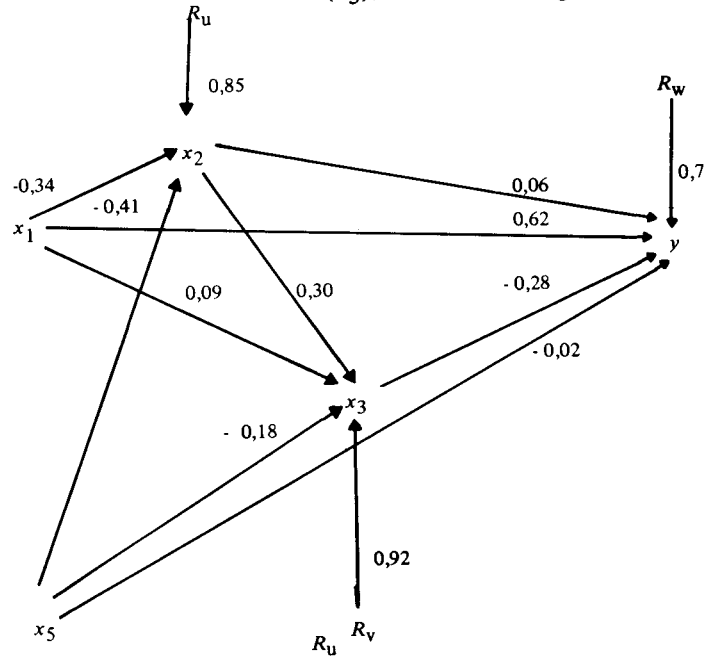


Indio (raza= 1)



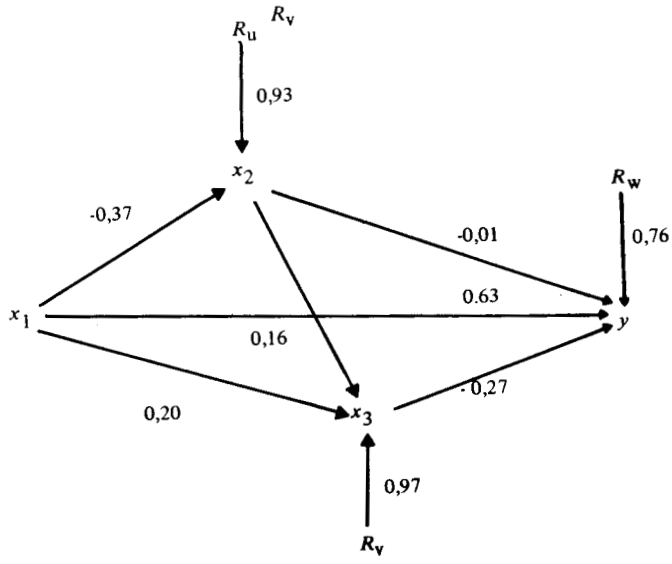
La tendencia de los efectos es la misma en los dos modelos, aunque los valores numéricos de los coeficientes no son iguales. Es posible probar la hipótesis nula de uniformidad de los coeficientes usando una prueba t o una prueba F. En este caso, por observación, las implicaciones de las dos series de coeficientes parecen ser las mismas.

Otro ejemplo subraya la necesidad de actuar con cautela cuando se construye un modelo. Si la raza se incluye en el modelo del ejemplo 2 como una variable adicional (x_5), obtenemos el siguiente resultado.

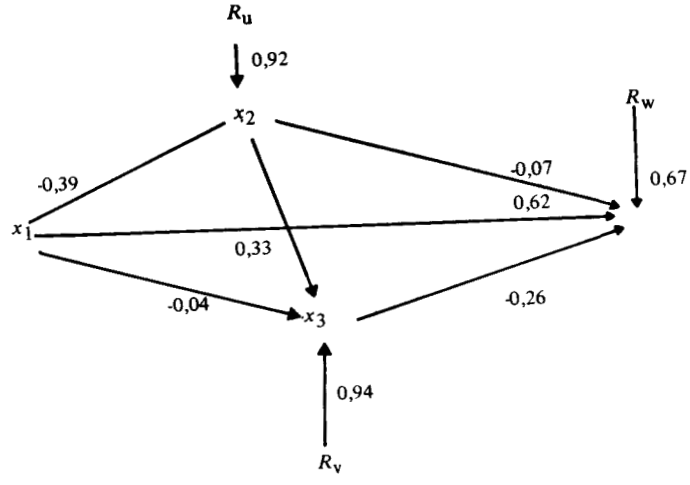


Este modelo da el efecto aditivo de la raza como variable exógena en el modelo estructural. Sin embargo, en este caso el análisis separado de las dos razas arroja resultados sustancialmente diferentes. Los tamaños de las submuestras son: Fijianos, 2045; indios, 2688.

Fijiano (raza = 0)



Indio (raza = 1)



Entre los dos casos existen dos diferencias importantes. Primera, la diferencia en la relación entre la edad y la edad al matrimonio en las dos razas. En el caso de los fijianos, la edad y la edad al matrimonio están positiva y estrechamente correlacionadas; en el caso de los indios, la correlación es negativa. O sea, el efecto indirecto de la edad sobre la fecundidad a través de la edad al matrimonio es negativo en el caso de los primeros y positivo, aunque pequeño, en el caso de los segundos. Segunda, el efecto de la educación tiene una intensidad diferente en los dos casos. Tanto el efecto directo como el indirecto de la educación son considerablemente más amplios en el caso de los indios.

En este modelo existe interacción entre la raza y las otras variables explicativas. O sea, los dos modelos separados dan una representación mucho más valiosa que el modelo conjunto. Esto no invalida el modelo en que se omite la raza. Ese modelo (ejemplo 2) proporciona una descripción media o sumaria de la manera cómo operan las otras variables explicativas.

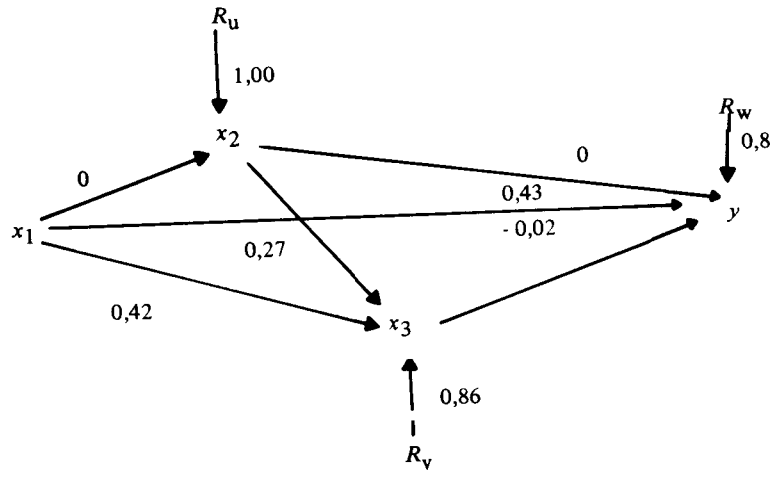
Estos efectos interactivos podemos incorporarlos fácilmente en el modelo construyendo nuevas variables que representen la interacción. También debemos incluir en las ecuaciones la variable binaria. Si construimos un término de interacción para cada predictor, el resultado será equivalente al que se obtendría efectuando dos regresiones separadas. O sea, esta técnica tiene valor sólo si podemos suponer que algunos de los predictores son estables para toda la población.

Uso de la edad como variable explicativa

La asociación entre la edad y el número de hijos es la asociación más fuerte que se da en los datos; la edad es exógena con respecto a todas las variables explicativas del modelo. Convencionalmente, en el análisis de la fecundidad los datos se dividen en cohortes de edad y el análisis se realiza independientemente para cada cohorte. El examen de los datos de la encuesta de Fiji muestra que el efecto de la edad es lineal, pero no proporciona fácilmente una prueba de la interacción entre la edad y las otras variables explicativas. Sin embargo, parece también que la educación y la edad al matrimonio pueden tener diferentes efectos para diferentes grupos de edades. Esto puede investigarse estimando los coeficientes de trayectoria del modelo estructural para cada cohorte por separado y comparando los resultados.

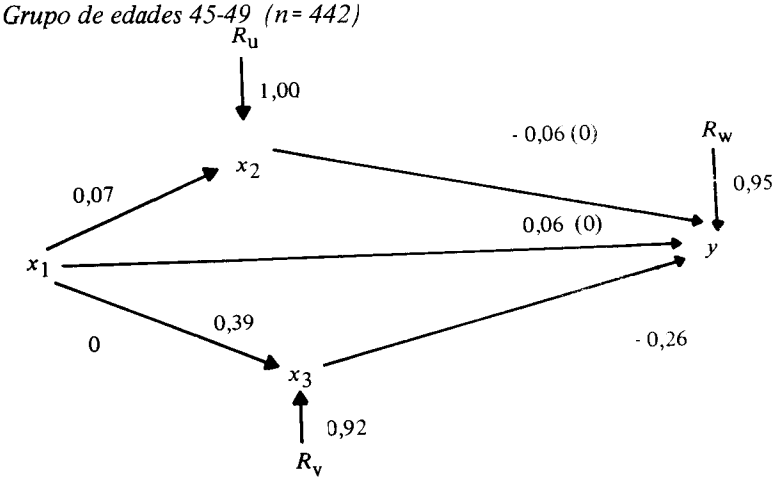
Los dos diagramas que siguen presentan los resultados referentes al grupo de edades más jóvenes y al grupo de edades más avanzadas.

Grupo de edades 15-19 (n= 224)



La edad ya muestra un fuerte efecto directo en este grupo de edades y su omisión del modelo reduciría considerablemente el poder explicativo de éste. El efecto directo de los años de educación desaparece y la explicación la suministra la trayectoria indirecta a través de la edad al casarse. En el grupo de más edad, el efecto directo de la edad casi desaparece y el resto del efecto total se produce principalmente a través de la educación y de la edad al casarse conjuntamente. Los resultados son intuitivamente razonables y abonan la idea de que aun dentro de las cohortes de edad, la edad debería incluirse como una variable explícita en el modelo estructural. No se pierde información por esta inclusión y puede ganarse mucho desde el punto de vista de la fuerza explicativa general y de la descomposición de los efectos.

Grupo de edades 45-49 (n= 442)



Coeficientes estandarizados y coeficientes no estandarizados

Un coeficiente convencional de trayectoria produce en la variable predicha el efecto esperado de un cambio de una desviación estándar en el predictor. Esto abona la hipótesis de que el efecto de una variable es proporcional a la distribución de las variables en la población. El coeficiente *no estandarizado* (el coeficiente de regresión con los datos brutos) da el efecto en términos de un cambio unitario en el predictor. Ambos coeficientes dan una información útil. En el segundo caso consideramos el efecto de un *año* adicional de educación, por ejemplo; en el primero consideramos el efecto de un aumento unitario en años de educación, definida en función de la distribución de la educación en la población. Las dos maneras son compatibles y representan diferentes modos de interpretación, y ambas pueden ser útiles en la identificación de los parámetros estructurales del modelo. En verdad, en el mismo modelo puede usarse una combinación de variables estandarizadas y no estandarizadas.

Conclusión

Cabe aquí formular una advertencia. La estimación de los coeficientes de un modelo estructural nos proporciona ecuaciones que permiten predecir las variables del modelo. Podría considerarse correcto aplicar los resultados en la formulación de una política. Por ejemplo, el análisis muestra claramente que la educación tiene un efecto negativo en la fecundidad. Sin embargo, esto *no* significa que una elevación general del nivel de educación reduzca la fecundidad. La ecuación que hemos obtenido es una ecuación para individuos dentro del sistema actual. Si cambiamos la distribución de la población de la variable (y por consiguiente el sistema), el resultado puede no ser el mismo. Aunque para un individuo en particular un aumento de la educación puede traducirse en una fecundidad menor, tal resultado no se aplica automática-

mente a un cambio en el nivel general de educación de la población. La naturaleza no experimental de los datos impide tales inferencias. Del mismo modo, aunque la edad al casarse se relaciona negativamente con la fecundidad, un cambio en la edad media al casarse en la población puede no tener ningún efecto sobre la fecundidad. La variación de los resultados según el contexto es importante tenerla en mente y se relaciona en parte con la omisión de las variables intermedias del modelo. La educación (o la edad al casarse) puede ser un medio de predicción útil en el sistema debido a una relación con algunas de estas variables intermedias. Pero ambas pueden representar simplemente diferencias culturales o sociales de la población que no estamos midiendo directamente. Si se cambian las distribuciones de la población de las variables, pueden perder su utilidad como elementos de predicción y puede ser necesario otro modelo estructural.

La advertencia señalada no es una crítica al análisis de trayectoria; es una observación acerca de las limitaciones inherentes del análisis transversal de los datos de las ciencias sociales. Los modelos de análisis de trayectoria pueden ser muy útiles para desenredar una serie compleja de relaciones y, utilizados con precaución, pueden contribuir considerablemente a nuestro conocimiento de los mecanismos que actúan en la población.

Referencias bibliográficas

Duncan, O.D. (1966), "Path Analysis: Sociological Examples", *American Journal of Sociology, 72, 1-16.*

Duncan, O.D. (1975), *Introduction to Structural Equation Models,* Nueva York, Academic Press.

Koopmans, T.C., Reiersol, O. (1950), "The identification of Structural characteristics", *Annals of Mathematical Statistics, 21,* 165-181.

Land, K.C. (1973), "Identification, parameter estimation and hypothesis testing in recursive sociological models", A.S. Goldberger, O.D. Duncan (eds.) *Structural Equation Models in the Social Sciences,* (Cap. 2), Nueva York, Seminar Press.

Macdonald, K.I. (1977), "Path Analysis", C.A. O'Muircheartaigh, C.D. Payne (eds.) *Analysis of Survey Data.* (Cap. 3, Vol. 2) Londres, Wiley.

Wright, S. (1934), "The method of path coefficients", *Annals of Mathematical Statistics, 5,* 161-215.

Wright, S. (1960), "Path coefficients and path regressions: alternative or complementary concepts", *Biometrics, 16,* 189-202.