UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

ANALISIS DE TRAYECTORIAS

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A

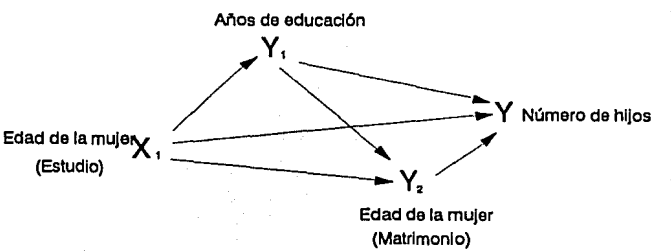
JACINTA CELIA IGLESIAS MOLINA

ASESOR DE TESIS:
DR. GUSTAVO J. VALENCIA RAMIREZ

MEXICO, D. F. 1992.

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

ANALISIS DE TRAYECTORIAS



INDICE

INTRODUCCION

El Análisis de Trayectorias es una forma de análisis de regresión lineal con respecto a variables estandarizadas (media cero y varianza uno). Los resultados que se desprenden del Análisis deben ser consistentes con la estructura y compatibles con los datos observados de las variables involucradas.

El Análisis de Trayectorias fué desarrollado por Sewall Wright en 1918, para explicar relaciones causales en poblaciones genéticas. Más tarde (1960) se extendió su utilización en el campo de las ciencias sociales y la sociología.

Debido a que el Análisis de Trayectorias es un tipo de Análisis de Regresión. En el capítulo 1 se hace notar la diferencia entre los dos tipos de análisis, desde el punto de vista de la predicción o de la explicación de un suceso.

Aunque el tema tiene bastante generalidad, se desarrollará el caso particular de los sistemas recursivos. Se dice que un sistema es recursivo, si dos variables no pueden ser recíprocamente "Causa" y "efecto". Los términos causa y efecto, se presentarán explícitamente en el capítulo 2 al igual que los diferentes tipos de relaciones causales y la construcción de un modelo causal.

El capítulo 3 trata del Análisis de regresión y correlación. Esto como una introducción a lo que es el análisis de trayectorias por ser un análisis de regresión estandarizado (cabe mencionar que existe dicho análisis para variables no estandarizadas) de tipo causal.

En el capítulo 4 se desarrolla el análisis de trayectorias donde se plantea como una herramienta que permite determinar la estructura que mejor se ajusta a los datos. En el capítulo 5 se muestran varias aplicaciones del Análisis de Trayectorias, en estudios realizados en la Biología y en diferentes áreas de las Ciencias Sociales.

CAPITULO 1

EL PROBLEMA

Los modelos estadísticos lineales son la base de los métodos estadísticos más usuales en áreas como regresión y diseño de experimentos. En estas áreas, el planteamiento del modelo no contempla la relación de causa-efecto que puede existir entre las variables. Esto se debe a que el investigador está consciente de que tal relación puede no existir y por lo tanto no la contempla. Pero si él aplica un modelo de regresión y sucede que debe optar por un modelo que considere la relación causal (por ejemplo el modelo de trayectorias), entonces existe un problema para el cual los resultados y conclusiones obtenidas via regresión por lo general no son suficientemente útiles.

Si de antemano el investigador sabe o infiere que existe una relación de causa-efecto, como en el caso de una población

genética, un estudio médico, factores que influyen en la explosión demográfica o en el aumento o disminución del precio de un producto etc. el investigador no se puede apoyar simplemente en un modelo de regresión en el cual se observa la asociación entre las variables explicativas y la variable respuesta, pero no se puede asegurar nada en el sentido de que las variables explicativas causen directamente la variación de la variable respuesta (Méndez, 1987). En tal caso,es necesario plantear algún modelo que considere la relación de causa efecto, como por ejemplo el modelo de trayectorias. Esto no quiere decir que se utilice el modelo de trayectorias en lugar del modelo de regresión, más bien se utiliza el primero como complemento del segundo, debido a que teóricamente se establecen relaciones de causa y efecto entre las variables.

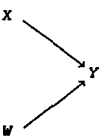
La estadística tradicional derivada del método de mínimos cuadrados está dirigida a la predicción y explicación de las variables involucradas. Mientras que el método de trayectorias no se ocupa de la predicción, más bien, procura dar una interpretación plausible de las relaciones entre las variables. En otras palabras se preocupa de elegir una estructura causal compatible con los datos observados.

Para ilustrar la distinción entre predicción e interpretación plausible se muestra el siguiente ejemplo (Li, 1975, pp. 2-4):
-en un estudio de la aceptación o rechazo de anticonceptivos de las amas de casa jóvenes en Taiwan, el equipo de investigación (compuesto por científicos sociales, bioestadísticos y médicos) observó las siguientes variables (causales) para esposo y esposa: edad, educación, ocupación, ingreso, vivienda, salud, número de hijos vivos, etc. Ajustando una regresión para todas estas variables encontrarón que el factor simple más importante, fué el número de artefactos eléctricos en el hogar (ventiladores, tostadores etc.) poseídos por cada familia.

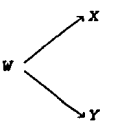
Si la predicción es el propósito, entonces, el número de artefactos eléctricos sería probablemente el mejor indicador o predictor para la aceptación de anticonceptivos. Si se considera que los artefactos eléctricos constituyen una causa importante para la aceptación de anticonceptivos entonces, la planeación familiar estaría asegurada mediante una distribución apropiada de aparatos eléctricos, lo cual resultaría más barato que la atención y recursos que requiere un niño, en términos económicos.

Si se aplica un Análisis de trayectorias al mismo conjunto de datos, lo primero sería intentar una explicación plausible del fenómeno mediante la construcción de un diagrama que especifique la naturaleza de las relaciones de causalidad, y de acuerdo a ese diagrama se efectuaría el análisis. En este caso particular, probablemente se observaría que la aceptación de anticonceptivos (*Y*) y el número de artefactos eléctricos (*X*), son consecuencia de las mismas variables causales, tales como la educación, ingreso, etc.. Esto es, que la gente con más educación y más altos ingresos tienen mayor aceptación a los anticonceptivos y también puede adquirir más artefactos eléctricos.

La diferencia entre análisis de regresión múltiple y el de trayectorias en este caso muy simple, se refleja en los siguientes dos diagramas, tomando a *W* como el conjunto de las variables educación, ocupación, ingresos etc, a *X* como número de artefactos eléctricos y a *Y* como aceptación de anticonceptivos.



Regresión Múltiple



Explicación Plausible

El método de lectura de tales diagramas se comentará en detalle en el capítulo 4. El diagrama de la izquierda, representa la aceptación de anticonceptivos (Y), como una combinación lineal de las variables X y W. En el diagrama de la derecha, la variable X (el número de artefactos eléctricos) y Y (la aceptación de anticonceptivos), se observan como dos consecuencias del mismo conjunto de variables W.

Note que podemos aplicar cualquiera de los dos métodos para resolver este problema en particular, sin embargo debemos estar conscientes, y saber cuál es nuestro próposito (si predicción o explicación), al momento de decidir cuál aplicar.

CAPITULO 2

SISTEMAS CAUSALES

El concepto de causalidad es de suma importancia debido a que para la mayoría de los sucesos que ocurren, podrían establecerse relaciones de causa-efecto.

Ejemplo.- En la teoría clásica del precio: El decremento del precio de un bien puede ser causado por el aumento de la oferta de esa mercancía, por la disminución de su demanda o por la combinación de ambos elementos. Si lo que se desea es explicar la baja del precio del trigo en un año concreto, lo más factible será buscar el volumen de la cosecha de aquel año, esto a su vez está en función de la superficie sembrada, de la cantidad de lluvia y de la cantidad de abono aplicado. Como se observa en este ejemplo, el efecto es la baja del precio y las causas son, el aumento en la oferta o la disminución de su demanda.

Se puede generalizar este tipo de problemas: Sea $A(\alpha)$ una proposición relativa a la situación α (p. ej., la oferta aumenta en el mercado α), y sea $B(\alpha)$ una proposición más acerca de α (p. ej., el precio desciende en el mercado α). Entonces la proposición de relación entre $A(\alpha)$ y $B(\alpha)$ es de la forma:

$$A(\alpha) \longrightarrow B(\alpha)$$

que se interpreta: $A(\alpha)$ ocasiona o tiene por efecto $B(\alpha)$.

DEFINICÍON DE EFECTO CAUSAL

Para dar una definición de efecto causal es necesario dar una definición de causalidad. Según Norman y Hadlai(1975).

Definición "operacional" de causalidad: α_1 es una causa de α_0 si y sólo si α_0 puede ser cambiada por manipular α_1.

Primero, la noción de causalidad implica predicción, pero predicción de una clase particular. Predicciones matemáticas o estadísticas que no implican la capacidad de producir cambios son excluidas de esta definición. Segundo, para juzgar que se entiende por "sólo" en la definición, se debe entender la noción de jerarquía causal y control relevante. Por ahora, sólo note que al manipular α_1 no implica que todas las otras causas de α_0 estén controladas, o se mantengan constantes, ya que al cambiar α_1, esto traerá cambios en muchas otras variables que sean afectadas por α_1.

Efecto causal.- Tomando a α_0 como la variable respuesta y a α_1 como la variable explicativa en un experimento ideal y suponiendo que el sistema causal es lineal, aditivo y unidireccional, entonces la relación entre α_0 y α_1 es de la forma:

$$\alpha_0 = c_0\alpha_1 \qquad (2.1)$$

donde c_0 es una constante fija para la magnitud de cambio en x_0 por una unidad de cambio en x_1.

El coeficiente c_0 así medido será llamado el coeficiente lineal del efecto causal, o simplemente el coeficiente del efecto. Si ésto se relaciona con el modelo de regresión lineal sin ordenada al origen, se observa, que dada una regresión entre x y y el modelo ajustado es,

$$y = bx$$

Aquí, b no puede ser interpretada como el coeficiente del efecto si no como el coeficiente de regresión, debido a que regresión lineal no considera la relación de causa y efecto.

Análogamente, si se interpretan los coeficientes de regresión como coeficientes de efecto (es decir, considerando el orden causal) se está haciendo una interpretación analítica de trayectorias.

En la ecuación 2.1 se observa una relación de causalidad lineal entre x_0 y x_1. Para una discusión amplia sobre este punto, ver Blalock(1971).

RELACION DE CAUSALIDAD

Una vez definido lo que es efecto causal, se pueden ver relaciones de causalidad.

Desde el punto de vista filosófico, se puede entender como causa aquello que es capaz de producir o modificar la manera de ser de lo que ya es.

Esta noción de causa implica una distinción importante, sin duda en la investigación científica, entre la causa que produce algo nuevo de aquella que soló modifica lo ya existente.

Se dice que dos variables, presentan una relación de causalidad, cuando una variable influye en la otra, en el sentido de que una modificación en la primera produce o dá lugar a una modificación en la segunda.

Como menciona Ph. Van Parijs en su trabajo "La syntase de I'explicatión dans les sciences sociales" existe una relación causal cuando modificando A se puede producir una modificación en B.

De aquí que, científicamente, las definiciones de causa y relación causal se centren en esta idea de sucesión necesaria, constante e irreversible.

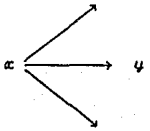
J. A. Davis escribe (1972, p. 108) que "dos variables tienen una relación causal cuando una reordenación de casos de una va seguida de una reordenación de casos de la otra". Y Stefan Novak (1975, p. 82) afirma que "decir que S es la causa de B significa que donde quiera que y siempre que, dentro de los límites de la generalización causal de que se trate, S ocurre (o debería ocurrir) es (o debería ser) seguido por B, independientemente de si S ocurre o (debería ocurrir) espontáneamente o fue traido a la existencia por alguna acción voluntaria o por cualquier "Actor" o "producto"".

Como menciona Sierra(1981, p. 268). Una consecuencia importante de esta idea científica de causa, es que la causalidad no se puede comprobar de manera inductiva, empírica, sino que supone una inferencia deductiva, es decir un apoyo o sustento teórico.

Una nota importante al respecto es como dice Parijs(p.228) "incluso en la experimentación más cuidada, uno no está nunca seguro de que todas las variables susceptibles de influir en la variable efecto han sido controladas".

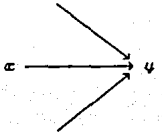
Tipos de relaciones causales.-Dentro de las relaciones causales existen varios tipos, por la forma en como se presentan. Hoy en día como menciona Méndez (1987) en la ciencia moderna se reconocen los siguientes tipos: las causas necesarias, las suficientes, las contribuyentes, y las necesarias y suficientes.

Causa necesaria.-Una condición x es una causa necesaria para el efecto y, si y siempre es precedida por x, pero puede ocurrir x sin que ocurra y. Esquematizando:



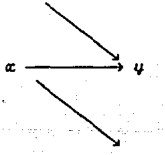
Por ejemplo: x puede ser introducción de amibas en el organismo y y la enfermedad amibiasis.

Causa suficiente.-Una condición x es una causa suficiente para el efecto y, si x siempre causa a y, pero puede ocurrir y sin que ocurra x.



Ejemplo: x falta de hierro en la dieta y y la anemia.

Causa contribuyente.-Una condición x es contribuyente para y, si se asocian frecuentemente x y y. Pero puede ocurrir x sin que ocurra y, y y sin que ocurra x.



Ejemplo: x fumar, y cáncer pulmonar.

Causa suficiente y necesaria (determinista).-Una condición α es necesaria y suficiente para ψ, si ψ siempre es precedida por α y α siempre conduce a ψ.

$$\alpha \longrightarrow \psi$$

Ejemplo: Considere un gas a volumen constante, entonces α presión y ψ temperatura del gas.

MODELOS CAUSALES

Los modelos causales lineales constan de una o más ecuaciones. Cada ecuación permite representar las relaciones causales que existen entre las variables. Al conjunto de estas ecuaciones se le llama estructura lineal.

Estos modelos se suelen considerar también como estructurales, en cuanto que estudian las relaciones de influencia entre las variables de un sistema y pretenden determinar la estructura de dichas relaciones. Por ejemplo, el modelo de trayectorias. A este respecto, cabe recordar la importante distinción entre la investigación experimental y la no experimental.

Investigación experimental.- se basa, en la reducción de los complejos sistemas de causalidad reales a sistemas o modelos sencillos, mediante el aislamiento y control experimental de las variables, es decir las condiciones causales son modificadas a voluntad del investigador.

Investigación no experimental.- En las investigaciones no experimentales, no existe el aislamiento y el control experi-

mental. En este tipo de investigación no hay modificación producida por el investigador de las condiciones causales.

Los modelos causales, se basan en la idea de causalidad, por lo que es conveniente hacer referencia a el problema de identificación, que se presenta en relación a estos modelos, debido a que desembocan en un sistema de relaciones que conducen a un conjunto de ecuaciones (Sistema de ecuaciones simultáneas).

EL PROBLEMA DE IDENTIFICACION

Para un modelo dado y unos datos determinados, se dice que una estructura está identificada, si existe únicamente una estructura que sea a la vez admisible respecto a los datos y al modelo.

En relación a este punto existen para un modelo determinado, únicamente tres alternativas:

1.- La subidentificación.
2.- La identificación exacta.
3.- la sobreidentificación.

Subidentificación.- El sistema de ecuaciones que determina el modelo, no ofrece información suficiente para encontrar el valor de todas y cada una de las incógnitas

Identificación exacta.- Aquí, por el contrario, sí existe dicha información y el sistema tiene solución única.

Sobreidentificación.-Por último, en este caso se da un exceso de información, por lo que no es posible hallar una solución que satisfaga a la vez a todas las ecuaciones del sistema. Sin embargo, pueden elegirse subsistemas determinados que proporcionen

soluciones al sistema, si bien éstas serán diversas y tantas como subsistemas se puedan construir.

La identificabilidad de un modelo depende en esencia de que el sistema de parámetros sea, en términos matemáticos:

Determinado.-Igual número de ecuaciones independientes que de incógnitas.

Indeterminado o subidentificado.-Mayor número de incógnitas que de ecuaciones.

Incompatible o sobreidentificado.-Mayor número de ecuaciones que de incógnitas.

Una vez visto los posibles problemas que se pueden presentar, se procedera a construir un modelo.

CONSTRUCCIÓN DEL MODELO

Los científicos intentan resumir la complejidad de los fenómenos en leyes, hipótesis o modelos, para el mejor entendimiento y control de los mismos.

Un modelo en este sentido no es una réplica física de el sistema bajo estudio, más bien intenta describir las principales relaciones entre las variables de interés.

El proceso de construcción del modelo consiste en proponer un conjunto de expresiones para estas relaciones, y ver que tanto la conducta del modelo imita adecuadamente la conducta del sistema o fenómeno. Los modelos son construidos para propósitos específicos y no necesariamente intentan describir en detalle cada faceta del fenómeno.

Uno de los problemas básicos en estadística es la especificación del modelo a ser usado.

ESTRUCTURA DEL MODELO

Considerando el sistema en el cual las cantidades $y_1, y_2, y_3, \ldots, y_k$ y $x_a, x_b, x_c, \ldots, x_q$ son medibles, donde las K variables quedan determinadas por las x's. Es decir, cada variable endógena y_i queda absoluta y únicamente causada por el grupo de variables exógenas x's.

El modelo causal.- Sean las ecuaciones siguientes con las cuales se expresan estas relaciones causales.

$$\begin{aligned} y_1 &= F_1\ (x_a, x_b, \ldots x_q;\ \alpha_{11}, \alpha_{12}, \ldots \alpha_{1m}) \\ y_2 &= F_2\ (x_a, x_b, \ldots x_q;\ \alpha_{21}, \alpha_{22}, \ldots \alpha_{2m}) \\ &\vdots \qquad\qquad\qquad \vdots \\ y_k &= F_k\ (x_a, x_b, \ldots x_q\ \ \alpha_{k1}, \alpha_{k2}, \ldots \alpha_{km}) \end{aligned} \qquad (2.2)$$

Estas ecuaciones son llamadas estructurales, en donde los parámetros denotados por α, pueden ser constantes físicas, químicas, biológicas o psicológicas, económicas, tecnológicas, etc. las y_i son los efectos y las x_i las causas.

Introduciendo la variación aleatoria en el sistema de ecuaciones se puede describir la situación en la que existen: variaciones debidas al conjunto de factores no observados explícitamente, o a posibles errores de medición.

En los modelos causales, 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; Cuando se trata de sistemas de ecuaciones se distingue, más bien, entre variables endógenas, que son las que dependen de otra en alguna de las ecuaciones; y exógenas, que no dependen de ninguna otra de las consideradas en el sistema.

La idea básica es que las variables pueden ser dispuestas jerarquicamente en términos de sus prioridades causales.

ANALISIS CAUSAL

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, pp. 557-585), 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 a variables métricas (son variables métricas las que se miden en escalas de intervalo o de razón) y a variables dicotómicas. Una vez que se apreció la profundidad de la discusión en torno a la noción de causalidad, toma el nombre con el cual se le conoce hasta hoy:Análisis de Trayectorias o de senderos (Path Analysis).

Ejemplo.- Suponiendo que para realizar la explicación del rendimiento escolar (y) se dispone de las variables, estatus socioeconómico del padre (x_4), grado de educación del padre (y_1) y prestigio del hijo en la clase (y_2). Suponiendo además que se recurre a un modelo de trayectorias (tomando en cuenta la relación de causa y efecto que existe entre las variables).

En seguida se construyen dos posibles diagramas de las relaciones de causa-efecto.

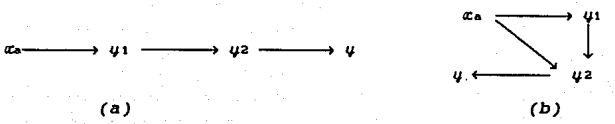


FIG.2.1 Sistemas causales.

Las ecuaciones correspondientes a cada diagrama son construidas directamente del mismo. Para el caso del diagrama (*b*) las ecuaciones son escritas como,

$$y_1 = P_{1a}x_a$$
$$y_2 = P_{21}y_1 + P_{2a}x_a$$
$$y = P_{02}y_2$$

las cuales no contienen variables de error porque el diagrama no las contempla. Sin embargo, puede considerarse que estos errores existen por las variables (aleatorias) que han sido excluidas del modelo.

Note que estos diagramas no son todos los que se pueden construir con cuatro variables, existen todas las combinaciones posibles; pero la construcción de diagramas y la decisión de cual tomar para el análisis se discute en el capítulo 4 (Análisis de trayectoria), con apoyo en el capítulo 3 (Análisis de regresión).

CAPITULO 3

REGRESION Y CORRELACION

En algunos problemas interesa a menudo no sólo las pruebas de significancia y las medidas de asociación, sino que también, se quiere describir la naturaleza de la relación entre dos variables, de modo que conociendo una de ellas se pueda anticipar la otra.

Cuando el interés se centra ante todo en la tarea exploradora de encontrar cuáles variables se relacionan con una variable determinada, interesa por lo regular el grados o fuerza de las relaciones. En particular los coeficientes de correlación, pero si lo que se intenta es predecir el valor de una variable apartir de otra, además de conocer la *forma* o la naturaleza de la relación entre las variables, entonces se debe hacer una Regresión.

REGRESION LINEAL SIMPLE

El Análisis de Regresión simple, es un instrumento estadístico importante y general que sirve para estudiar la naturaleza y forma de asociación lineal entre dos variables.

En la relación de regresión sólo se considera aleatoria una de las variables, llamada variable respuesta (y), la otra, variable explicativa (x) esta fija. La variable respuesta es aquella cuyo cambio se considera en función de la variable explicativa.

Al plantear el modelo, podrían esperarse problemas de estimación muy difíciles. Sin embargo, los métodos de estimación se componen de técnicas básicas simples. El conocimiento detallado del principio de mínimos cuadrados, hace falta para presentar el material que sigue. De esta manera se explicará regresión lineal simple,

$$y = f(x) \qquad (3.1)$$

este paso simplemente identifica a la variable x, la cual se considera que influye sobre la variable y.

Un supuesto más consiste en especificar la forma de la relación entre y y x. El conocimiento del interesado puede sugerir la forma funcional precisa que debe usarse, o ciertas condiciones secundarias que debe cumplir con respecto a la ordenada en el origen, la pendiente y la curvatura de la función. Una gran variedad de funciones pueden cumplir estos requerimientos, por lo que, el Análisis Estadístico proporciona la ayuda para elegir entre ellas. Si bien la idea de regresión es general, la mayoría de los desarrollos estadísticos se han realizado con los modelos más simples. En particular, se va a suponer que la forma de la ecuación de regresión es lineal.

Ecuación de la línea recta.- Si la regresión de α y ψ es lineal, o sea una relación en linea recta, se puede escribir la ecuación como sigue:

$$\psi = \beta_0 + \beta_1 \alpha \qquad (3.2)$$

Donde β_0 se conoce como la ordenada al origen y β_1 como la pendiente de la recta. β_1 es la razón de cambio en ψ por una unidad de cambio en α; como lo muestra la figura 3.1.

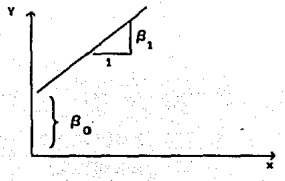


FIG. 3.1 La linea recta.

Los números β_0 y β_1 son llamados parámetros. En aplicaciones estadísticas, estos parámetros no son conocidos y deben ser estimados usando los datos de la muestra.

En este caso la forma de la relación entre α y ψ esta dada por la ecuación de la línea recta. Sin embargo, existen otras relaciones entre dos variables, como pueden ser:

$$\psi = \alpha e^{bx} \qquad \psi = \alpha\alpha^{b} \qquad \psi = \alpha + \beta(1/\alpha)$$

La tercera relación es lineal en las variables y y $1/x$, y las otras dos son lineales si se toman logaritmos en ambos miembros, con lo cual se tiene:

$$\log_e y = \log_e \alpha + bx \qquad \log_e y = \log_e \alpha + b \log_e x$$

respectivamente. La primera es lineal en x y en logaritmo de y, y la segunda en los logaritmos de ambas variables.

Errores " ε ".- Devido a que existe dispersión alrededor de la ecuación de regresión, se representa el valor real de y mediante una ecuación lineal que contiene un término de error, denotado por ε_i. Este término de perturbación puede originarse por errores de medición en y. El efecto de variables explicitamente no incluidas en el modelo tambien contribuye a los errores.

Si la mayor parte de estas variables omitidas tienen individualmente un efecto menor, y si además son independientes entre ellas, es razonable suponer que el valor esperado correspondiente al factor de perturbación o error $E(\varepsilon_i)$ es igual a cero. Los errores ε son variables aleatorias no conocidas.

En la práctica estas suposiciones generalmente requieren ser verificadas. Para hacer los errores de ajuste más pequeños se utilizan las transformaciones de las variables originales, además para que algunos supuestos sobre las características aleatorias de esos errores se cumplan. Por ejemplo la normalidad y la homogeneidad de varianzas.

En resumen, se supone que los errores ε_i tienen media cero,

$$E(\varepsilon_i) = 0, \qquad i = 1,2,3,\ldots, n$$

varianza desconocida constante,

$$Var(\varepsilon_i) = \sigma^2, \qquad i = 1,2,3,\ldots, n$$

y que los errores no están mutuamente correlacionados,

$$Cov(\varepsilon_i,\varepsilon_j) = 0, \qquad \forall\ i \neq j,$$

En algunas aplicaciones es necesario hacer suposiciones sobre la distribución de los errores, en este caso se supone Normalidad (los errores siguen una distribución probabilística Normal). Esto se hace ya que es usada principalmente para realizar pruebas de hipótesis y construir intervalos de confianza.

Tomando en cuenta las suposiciones anteriores, se dice que los ε_i son Normales e independientemente distribuidos con media cero y varianza común σ^2.

$$\varepsilon_i \sim N\ (0,\ \sigma^2) \qquad i = 1,2,3,\ldots,n$$

El modelo de regresión lineal simple.- En los casos en los que el marco conceptual lleva a la especificación de un modelo que postula la presencia de una variable explicada (y) determinada linealmente a través de una variable explicativa (x), el modelo apropiado es el de regresión simple,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1,2,3,\ldots,n \qquad (3.3)$$

con

$$E(\varepsilon_i) = 0$$

$$Var(\varepsilon_i) = \sigma^2$$

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j$$

Literalmente se tiene que los valores de y_i son determinados por los valores de x_i y ε_i, a través de la ecuación especificada. Las ε_i's son cantidades aleatorias no conocidas. β_0, β_1, y σ^2, son no conocidas, pero pueden ser estimadas mediante los valores observados x's y y's, por cualquiera de los siguientes dos métodos:

El método de mínimos cuadrados.

El método de máxima verosimilitud.

Estimación por el método de mínimos cuadrados.- Existen problemas donde un conjunto de pares de datos indican que la regresión es lineal, pero la distribución de las variables involucradas no es conocida y se quieren estimar los coeficientes de regresión β_0 y β_1. Problemas de este tipo son usualmente resueltos con el método de mínimos cuadrados, un método de ajuste de curvas sugerido por el matemático francés **Adrien Legendre** a principios del siglo XIX, que no considera la distribución de las ε_i's o de las y_i's, Freund y Walpole(1980, pp. 436-440).

Denotando a los errores de ajuste observados o residuos, por e_i, donde cada e_i esta dado por la ecuación

$$e_i = y_i - \hat{y}_i \qquad i = 1,2,3,\ldots,n \qquad (3.4)$$

como lo muestra la figura siguiente:

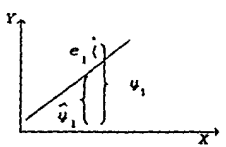


FIG. 3.2 Criterio de mínimos cuadrados.

donde $\hat{\psi}_i$ es el estimador del valor de ψ_i. Así que los valores ajustados o estimados $\hat{\psi}_i$ están dados por la ecuación

$$\hat{\psi}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad i = 1,2,3,\ldots,n \tag{3.5}$$

La diferencia entre los ε_i's y e_i's es importante, los residuales e_i son observables y serán usados para verificar suposiciones, mientras que los errores ε_i no son observables.

Este método está basado en minimizar la suma de cuadrados de los residuales (*SCR*). Dado un conjunto de pares de datos $\{(x_i,\psi_i);\ i=1,2,\ldots n\}$, los estimadores por mínimos cuadrados de los coeficientes de regresión son los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ para los cuales la cantidad

$$SCR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [\psi_i - \hat{\psi}_i]^2 \tag{3.6}$$

es mínima. Derivando parcialmente con respecto a β_0 y β_1 e igualando a cero, se obtiene

$$\begin{aligned} \sum_{i=1}^{n} \psi_i &= \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i \psi_i &= \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \end{aligned} \tag{3.7}$$

Resolviendo este sistema de ecuaciones llamadas normales para $\hat{\beta}_0$ y $\hat{\beta}_1$. Los estimadores por mínimos cuadrados de β_1 y β_0 son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\ (\psi_i - \bar{\psi})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3.8}$$

donde

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

son el promedio de la muestra para x_i y y_i respectivamente. Y

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

sea diferente de cero.

Propiedades de los estimadores por mínimos cuadrados.-Los estimadores por mínimos cuadrados son función lineal de las y_i's y estas a su vez de las ε_i's. Si todas las ε_i's tienen media cero, y el modelo es correcto, entonces como lo muestra el apendice A, los estimadores por mínimos cuadrados son insesgados,

$$\begin{aligned} E(\hat{\beta}_0) &= \beta_0 \\ E(\hat{\beta}_1) &= \beta_1 \end{aligned} \tag{3.9}$$

Para la varianza de los estimadores, tomando en cuenta los supuestos anteriores, $Var(e_i) = \sigma^2$ y $Cov(e_i,e_j) = 0 \quad \forall\, i \neq j$, entonces, del apendice A se tiene que,

$$Var(\hat{\beta}_1) = \sigma^2 \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.10}$$

$$Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \qquad (3.11)$$

Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ están correlacionados, porque su covarianza es, en general, diferente de cero:

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (3.12)$$

Estimación por el método de máxima verosimilitud.- Cuando se analiza un conjunto de pares de datos $\{(x_i, y_i);\ i=1,2,\ldots n\}$ por análisis de regresión se supone que las x_i son constantes mientras que los valores correspondientes a las y_i son variables aleatorias independientes. Esta sección se dedicará a problemas básicos del Análisis de Regresión Normal, donde se supone que para cada x_i fija, la densidad condicional de la correspondiente variable aleatoria y_i es *Normal*

$$f(y_i/x_i) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-1/2\left[\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma}\right]^2} \qquad -\infty < y_i < +\infty \quad i=1,2,\ldots n \qquad (3.13)$$

donde β_0, β_1 y σ son constantes para cada i.

Dada una muestra aleatoria de pares de datos, el Análisis de Regresión Normal consiste en estimar σ y los coeficientes de regresión β_0 y β_1. Con este método es posible realizar tanto

pruebas de hipótesis concernientes a los parámetros, como predicciones basadas en la ecuación de regresión estimada,

$$\hat{\psi} = \hat{\beta}_0 + \hat{\beta}_1 \alpha,$$

donde $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores de β_0 y β_1.

Para obtener los estimadores de máxima verosimilitud de los parámetros, se obtienen las derivadas parciales de la función de verosimilitud de β_0, β_1 y σ; ya que se busca maximizar la verosimilitud y es posible utilizar el criterio de la segunda derivada. Como el logaritmo es una función monotona no decreciente entonces el máximo de L (la función de verosimilitud) y el de $ln\ L$ coinciden, de donde

$$L(\beta_0,\beta_1,\sigma/\psi_1,\alpha_1) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-1/2\left[\frac{\psi_1-(\beta_0+\beta_1\alpha_1)}{\sigma}\right]^2}$$

$$ln\ L = -n\ ln\ \sigma - \frac{n}{2}\ ln\ 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[\psi_i-(\beta_0+\beta_1\alpha_i)]^2 \qquad (3.14)$$

derivando parcialmente con respecto a los parámetros e igualando las expresiones a cero se obtiene,

$$\frac{\partial\ lnL}{\partial\beta_0} = \frac{-2}{\sigma^2}\sum_{i=1}^{n}[\psi_i - (\beta_0+\beta_1\alpha_i)] = 0$$

$$\frac{\partial\ lnL}{\partial\beta_1} = \frac{-2}{\sigma^2}\sum_{i=1}^{n}\alpha_i[\psi_i - (\beta_0+\beta_1\alpha_i)] = 0 \qquad (3.15)$$

$$\frac{\partial\ lnL}{\partial\sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}[\psi_i - (\beta_0+\beta_1\alpha_i)]^2 = 0$$

de las primeras dos ecuaciones se obtienen los estimadores de máxima verosimilitud de β_0 y β_1,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\ (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (3.16)$$

los cuales son idénticos a los de mínimos cuadrados en 3.8. Por lo tanto tienen las mismas propiedades.

Sustituyendo $\hat{\beta}_0$ y $\hat{\beta}_1$ en la tercera ecuación, se obtiene inmediatamente el estimador de máxima verosimilitud de σ dado por

$$\hat{\sigma} = \sqrt{(1/n) \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2} \qquad (3.17)$$

Una vez obtenidos los estimadores por máxima verosimilitud se pueden contrastar hipótesisis y construir intervalos de confianza para β_0 y β_1. Para esto es necesario emplear el siguiente teorema.[1]

Teorema 3.1. Bajo las suposiciones de Análisis de Regresión Normal. $\frac{n\hat{\sigma}^2}{\sigma^2}$ tiene distribución Ji-cuadrada con $n-2$ grados de libertad y $\hat{\beta}_1$ tiene distribución normal con media β_1 y varianza $\frac{\sigma^2}{\Sigma(x - \bar{x})^2}$ además $\frac{n\hat{\sigma}^2}{\sigma^2}$ y $\hat{\beta}_1$ son independientes.

[1] La demostración del teorema 3.1 y otros detalles matemáticos pueden ser encontrados en textos como los de Wilks(1962) o . Searle(1971).

En este trabajo no se discute más sobre intervalos de confianza y pruebas de hipótesis. Si desea más información consultar Freund y Walpole(1980), Weisberg(1985, pp. 20-23), MendenHall y Reinmuth (1981, pp. 327-340) entre otros.

ANALISIS DE CORRELACION SIMPLE

En ocasiones no sólo se desea conocer la naturaleza de la relación entre x y y, sino que también es necesario conocer el grado o fuerza con la que dos variables x y y se encuentran linealmente relacionadas.

Si la relación es muy débil, no tiene objeto tratar de predecir y a partir de x. En ocasiones los investigadores tienen interés en descubrir cuales de entre un gran número de variables se relaciona linealmente con una variable dependiente determinada. Para este tipo de situación es necesario contar con una medida de correlación o de asociación. Comúnmente la medida de correlación lineal usada es el coeficiente de correlación r, que se define como:

$$r = r(x,y) = r_{xy} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i-\bar{x})^2 \sum_{i=1}^{n} (y_i-\bar{y})^2}} \qquad (3.18)$$

Se observa que los denominadores, tanto de r como de $\hat{\beta}_1$ en (3.8) son siempre positivos por ser sumas de cuadrados. También se observa que el numerador de r y de $\hat{\beta}_1$ es el mismo. De lo anterior que el coeficiente de correlación r y $\hat{\beta}_1$ tengan el mismo signo y además r será cero sólo si $\hat{\beta}_1 = 0$. Así que $r = 0$ implica la ausencia de correlación lineal entre x y y. Un valor

positivo de r implica que la pendiente de la recta es positiva (la recta crece a la derecha), un valor de r negativo indica que la recta decrece a la derecha (pendiente negativa) como lo muestra la figura 3.4.

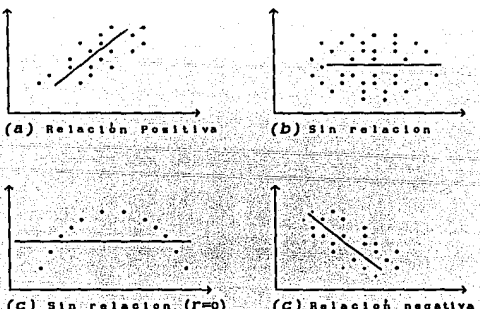


FIG 3.3 Diagrama de dispersión que muestra las diferentes fuerzas y direcciones de las relaciones entre X y Y.

PROPIEDADES DEL COEFICIENTE DE CORRELACION

Es simétrico.-se observa que la expresión (3.18) es simétrica con respecto a y y x, por lo tanto $r(x,y) = r(y,x)$.

Se encuentra entre menos uno y uno.- El coeficiente de correlación en valor absoluto no excede a la unidad; esto es,

$$-1 < r < +1 \ , \ r^2 < 1$$

Cuando r=1 o r=-1, todos los puntos deben pertenecer a una linea recta, cuando r=0, se encuentran dispersos sin tendencia a crecer o decrecer, o tienen la forma de una curva, es decir, no muestran evidencia alguna de relación lineal; cualquier otro valor de r simplemente sugiere el grado de dependencia lineal.

Es independiente del origen y de las unidades.- Considerando una transformación de x y y en;

$$x' = K_1 + b_1 x \qquad y' = K_2 + b_2 y \tag{3.19}$$

con b's y K's constantes, suponiendo que las K's pueden tomar cualquier valor diferente de cero y las b's valores positivos. En la transformación (3.20), al agregar una constante a x y a y la desviación $(x-\bar{x})$ y $(y-\bar{y})$ no cambian. El multiplicador constante b_1 y b_2 se elimina, es decir,

$$r(x',y') = \frac{\sum_{i=1}^{n} b_1(x_i-\bar{x})b_2(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n} b_1^2(x_i-\bar{x})^2 \sum_{i=1}^{n} b_2^2(y_i-\bar{y})^2}} =$$

$$= \frac{b_1 b_2 \sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y})}{b_1 b_2 \sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} = r(x,y)$$

donde

$$\bar{x}' = K_1 + b_1\bar{x} \qquad \text{y} \qquad \bar{y}' = K_2 + b_2\bar{y}$$

por lo tanto, las cuatro correlaciones siguientes son iguales:

$$r(x',y') = r(x',y) = r(x,y') = r(x,y) \tag{3.20}$$

Transformaciones del tipo (3.19) no tienen efecto sobre la correlación, en otras palabras, el coeficiente de correlación es independiente del origen y de las unidades en que se midan las variables x y y.

Las correlaciones no son transitivas.- considerando tres variables denotadas por x_1, x_2, x_3 y suponiendo que el coeficiente de correlación entre x_1 y x_2 es r_{12}, y entre x_2 y x_3 es r_{23}. Conociendo estos coeficientes no es posible conocer el valor de r_{13}, en general $r_{13} \neq r_{12} r_{23}$, por lo tanto, en este sentido la correlación no es transitiva.

Variables independientes y no correlacionadas.- cuando el coeficiente de correlación es igual a cero, simplemente se dice que x y y no están linealmente correlacionadas, más no que x y y son independientes. Si $f_i(x)$ es la distribución marginal de x_i y $f_i(y)$ es la distribución marginal de y_j. Una distribución independiente significa que la distribución conjunta de (x_i, y_j) es igual al producto de sus distribuciones marginales,

$$f_{ij}(x_i, y_j) = f_i(x_i) . f_j(y_j)$$

entonces se dice que x y y son independientemente distribuidas, en tal situación la correlación $r(x,y)$ es siempre cero. Por lo tanto, distribución independiente implica no correlación, pero $r(x,y) = 0$ no implica distribución independiente, sólo se dice que x y y no están correlacionadas linealmente. En el caso particular de que $f_{ij}(x_i, y_j)$ sea la distribución normal bivariada, $r(x,y) = 0$ si implica la independencia.

REGRESION MULTIPLE

En Regresión múltiple se trata de predecir una sola variable respuesta y, apartir de cierto número de variables explicativas

$x_1, x_2, \ldots, x_p$; siguiendo la idea de regresión simple el modelo es especificado por la ecuación lineal:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \qquad (3.21)$$

Como se dijo anteriormente en regresión simple, las β's son parámetros no conocidos, las ε's son los términos de error, la y es la variable respuesta y las x's son los predictores.

La ecuación (3.21) puede ser reescrita para n casos como sigue:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad i=1 \ldots n \qquad (3.22)$$

Notación Matricial[2].- Se denota a Y y e como vectores columna de dimensión $n \times 1$, de la siguiente forma :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \qquad (3.23)$$

β como el vector de párametros de dimensión $(p+1)\times 1$, incluyendo β_0.

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad (3.24)$$

(2) En esta sección se supone que el lector está familiarizado con álgebra de matrices.

X es la matriz del "diseño" de $n\times(p+1)$ dada por :

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \qquad (3.25)$$

La matriz X dada por los valores observados de los predictores, la columna de unos corresponde a β_0, la próxima columna corresponde al primer predictor x_1 y así sucesivamente.

Usando esta notación, la ecuación (3.22) puede ser escrita en términos matriciales, como:

$$Y = X\beta + \varepsilon \qquad (3.26)$$

El error ε .- El término de error es un vector de variables aleatorias. Las suposiciones concernientes a las ε_i's son las mismas que están dadas en regresión simple, por lo tanto :

$$E(\varepsilon) = 0, \quad Var(\varepsilon) = \sigma^2 I_n$$

Donde $Var(\varepsilon)$ significa la matriz de varianzas - covarianzas (ε), I_n es la matriz identica o identidad de dimensión nxn, y 0 es un vector de ceros de dimensión nx1. Agregando la suposición de que cada ε_i está Normalmente distribuida entonces:

$$\varepsilon \sim N(0, \sigma^2 I_n) \qquad (3.27)$$

Estimación por mínimos cuadrados.- El estimador por mínimos cuadrados $\hat{\beta}$, de β es aquel que minimiza la suma de cuadrados de los residuales (SCR).

$$SCR = e^t e = (Y - X\hat{\beta})^t (Y - X\hat{\beta}) \qquad (3.28)$$

$$= Y^tY - 2\hat{\beta}^tX^tY + \hat{\beta}^tX^tX\hat{\beta}$$

Donde e denota el vector columna de residuales, e^t el vector de residuales transpuesto de dimensión n y $\hat{\beta}$ el vector columna de valores estimados de β.

derivando (3.28)

$$\frac{\partial (e^te)}{\partial \beta} = -2X^t Y + 2X^t X\beta$$

igualando a cero, se obtiene,

$$X^tX\hat{\beta} = X^tY$$

suponiendo que X tiene rango $p<n$ $(X^tX)^{-1}$ existe, entonces el estimador por mínimos cuadrados es

$$\hat{\beta} = (X^tX)^{-1} X^tY \qquad (3.29)$$

El estimador depende solamente de (X^tX) y (X^tY); las cuales son matrices de suma de cuadrados y productos cruzados.

Cuando se calcula $\hat{\beta}$, se encuentran varias cantidades; el vector de valores ajustados $\hat{Y} = X\hat{\beta}$ tiene el i-ésimo elemento igual a $\hat{y}_i = x_i^t \hat{\beta}$. El vector de residuales es $e = Y-\hat{Y}$ con el i-ésimo elemento $e_i = y_i-\hat{y}_i$.

Propiedades de los estimadores por mínimos cuadrados.- Las propiedades de los estimadores están desglosadas en el apéndice *B* y solamente se resumen aquí. Suponiendo que $E(\varepsilon)=0$ y $Var(\varepsilon)=\sigma^2 I_n$, entonces $\hat{\beta}$ es insesgado,

$$E(\hat{\beta})=\beta \qquad y \qquad Var(\hat{\beta})=\sigma^2(X^tX)^{-1} \qquad (3.30)$$

Estimación por máxima verosimilitud.- suponiendo que Y es un vector Normal multivariado con vector de medias XB y matriz de varianzas-covarianzas $\sigma^2 I$, donde X es una matriz conocida $nx(p+1)$, B es un vector de constantes no conocidas $(p+1)x1$ y σ^2 es un escalar no conocido. La función de verosimilitud de la muestra es:

$$L\,(B,\sigma^2/Y,X) \quad = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma)^2 (Y-XB)^t (Y-XB)}$$

$$log\ L = -(n/2) log\ 2\pi - (n/2) log\sigma^2 - (1/2\sigma^2)(Y-XB)^t (Y-XB)$$

entonces L (y $log\ L$) es una función diferenciable de parámetros no conocidos, esperando encontrar aquellos valores de B y σ^2 que maximicen L. Obteniendo las derivadas parciales de $log\ L$ con respecto a σ se tiene,

$$\frac{\partial log\ L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + (1/2\sigma^4)(Y-XB)^t (Y-XB)$$

Suponiendo que σ^2 es constante, $log\ L = k - (1/2\sigma^2)Q$ donde $Q = (Y-XB)^t (Y-XB)$, todas las derivadas parciales de $log\ L$ con respecto a B se obtienen

$$\frac{\partial log\ L}{\partial B} = -\frac{1}{2\sigma^2} (-2X^t Y + 2X^t XB)$$

haciéndolas simultáneamente iguales a cero y despejando, se tiene que:

$$\hat{B} = (X^t X)^{-1} X^t Y$$

$$\hat{\sigma}^2 = (1/n)\ (Y-X\hat{B})^t (Y-X\hat{B})$$

$\hat{\beta}$ es el mismo estimador que el encontrado por mínimos cuadrados (3.29) por lo tanto tiene las mismas propiedades, además que $\hat{\beta}$ se distribuye como una $N(\beta,\sigma^2(XX)^{-1})$ y $n\hat{\sigma}^2/\sigma$ se distribuye como $\chi_{(n-p)}$. $\hat{\beta}$ y $\hat{\sigma}^2$ independientes.

ANALISIS DE VARIANZA

El análisis de varianza tiene una aplicación mucho más amplia que un problema general de comparar varias medias. Para llevar a cabo el análisis, es necesario formular un modelo matemático en términos de los parámetros desconocidos y las variables aleatorias asociadas.

Como su nombre lo indica, el propósito del procedimiento de análisis de varianza es analizar la variabilidad de la variable respuesta y asignar componentes de esa variablilidad a cada uno de los conjuntos de variables explicativas. La idea detrás del procedimiento es que las variables de respuesta cambian debido a la variación de un conjunto de variables explicativas desconocidas.

El objetivo del análisis de varianza es determinar cuáles son las variables explicativas de importancia en un estudio, y en qué forma interactúan y afectan la respuesta.

En esta sección se tratará el problema de probar la hipótesis de que los parámetros β asociados a las variables x's son cero, suponiendo que se tiene el modelo completo,

$$MC: Y = X\beta + \varepsilon \qquad (1)$$

y el modelo reducido (sin considerar ninguna x's)

$$MR: Y = \beta_0 1 + \varepsilon$$

donde 1 es el vector de unos de dimensión nx1. Estos corresponden a:

$$MC: \ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$MR: \ y_i = \beta_0 + \varepsilon_i$$

de análisis de varianzas para regresión lineal simple.

El vector de β's para el modelo completo en (1) es representado como

$$\mathcal{B} = \begin{pmatrix} \beta_0 \\ \mathcal{B}^* \end{pmatrix} \qquad \text{con} \qquad \mathcal{B}^* = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

como se desea probar que los parámetros β asociados a las variables x's son cero, entonces las hipótesis a probar son:

$$HN: \ \mathcal{B}^* = 0$$

$$HA: \ \mathcal{B}^* \neq 0$$

La prueba se describe intuitivamente. Se utiliza el método de mínimos cuadrados para ajustar el modelo reducido, entonces $\hat{\beta}_0 = \bar{y}$ y la suma de cuadrados total (SCT) es $\sum_{i=1}^{n}(y_i - \bar{y})^2$. Para el modelo completo el estimador de $\mathcal{B}$ está dado en (3.29) y la SCR en (3.28). Se tiene que $SCR < SCT$, la diferencia entre estos dos es,

$$SCreg = SCT - SCR \qquad (3.31)$$

mientras más grande resulte la diferencia de $SCT - SCR$, más grande será la evidencia de que los términos deben incluirse.

Para probar la hipótesis antes mencionada se utiliza la estadística

$$f = \frac{(SCreg)/(p)}{SCR/(n-p-1)}$$

cuando se satisfacen las suposiciones de que los valores de y se distribuyen Normal e independientemente, con media $E(y)$ y varianza σ^2 entonces esta estadística f tiene una distribución F con (p -, n-p-1) grados de libertad.

Mientras más grande sea la reducción en *SCR*, más evidencia se tendrá para rechazar la hipótesis nula y aceptar la hipótesis alternativa.

Los resultados de Análisis de varianza se resumen en la siguiente tabla:

FUENTE	*ANALISIS DE VARIANZA TOTAL* GRADOS DE LIBERTAD	SUMA DE CUADRADOS	SUMA DE CUAD. MEDIOS	F
Regresión sobre $x_1, \ldots, x_p$	p	$SCreg$	$SCreg/p=CMreg$	$\frac{CMreg}{CMR}$
Resiudual	$n-p'$	SCR	$SCR/n-p'=CMR$	
Total	$n-1$	SCT	$SCT/n-1$	

Una prueba adicional es la de probar la hipótesis de que uno o varios de los parámetros β son cero. Suponiendo que se tiene el modelo,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

o equivalentemente

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

y se desea saber si ciertas variables contribuyen con información para la predicción de y. Si un conjunto de variables x no contribuyen con información alguna para la predicción de y, entonces sus parámetros β debieran ser igual a cero. En consecuencia, el probar si ciertas variables x deben incluirse en el modelo es equivalente a probar la hipótesis de que ciertos parámetros β son cero.

Suponiendo que se tienen dos modelos para $E(y)$, uno que es referido como "modelo completo" (MC) y otro como "modelo reducido" (MR). El modelo reducido incluye sólo parte de los términos del modelo completo. El propósito de la prueba es el probar la hipótesis de que los parámetros β asociados a estos términos adicionales son cero. En otras palabras, se prueba si los términos adicionales contribuyen con información para la predicción de y.

Representando al modelo completo y reducido por:

$$MC: E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_g x_g + \ldots + \beta_p x_p$$

$$MR: E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_g x_g \qquad \text{con } g<p$$

si el vector de β's para el modelo completo se representa como

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_g \\ \mathbf{B}_p \end{pmatrix}$$

con

$$\mathbf{B}_g = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_g \end{pmatrix} \qquad \mathbf{B}_p = \begin{pmatrix} \beta_{g+1} \\ \vdots \\ \beta_p \end{pmatrix}$$

entonces las hipótesis a probar son:

$$HN: \beta_p = 0$$

$$HA: \beta_p \neq 0$$

La prueba se describe intuitivamente. Se utiliza el método de mínimos cuadrados para ajustar el modelo reducido y obtener la suma de cuadrados de las desviaciones para dicho modelo (SCE_R). Posteriormente se obtiene la suma de cuadrados de la desviaciones para el modelo completo (SCE_C).

se comparan las dos sumas de cuadrados; si las variables $x_{q+1}...x_p$ de verdad contribuyen con información para la predicción de y, entonces la SCE_C debiera ser significativamente menor que SCE_R. El incorporar estas variables al modelo produce una reducción en la suma de cuadrados de la regresión. En consecuencia, mientras más grande resulte la diferencia (SCE_R - SCE_C), más grande será la evidencia de que los términos deben incluirse, es decir habrá más evidencia que indique que almenos uno de los parámetros del vector β_p difiera de cero.

Para probar la hipótesis antes mencionada se desarrollará el estadístico de prueba que esta en función de (SCE_R - SCE_C), cuya distribución se conoce cuando HN es verdadera.

Suponiendo que HN es verdadera, se examinarán las cantidades que se han calculado.

$$SCE_R = SCE_C + (SCE_R - SCE_C)$$

entonces $S_1^2 = \frac{SCE_R}{n-(q+1)}$ es un estimador insesgado de σ^2.

$$S_2^2 = \frac{SCE_C}{n-(p+1)} \qquad y \qquad S_3^2 = \frac{SCE_R - SCE_C}{(p-q)}$$

son estimadores insesgado de σ^2 y son estadísticamente independientes.

$$f = \frac{S_3^2}{S_2^2}$$

tiene una distribución F con $(p-g)$ y $(n-(p+1))$ grados de libertad.

CORRELACION MÚLTIPLE

Así como el Análisis de Regresión Múltiple consiste en la extensión del Análisis de Regresión Simple, así también, los coeficientes de correlación múltiple sirven para extender los coeficientes de correlación simple.

El concepto de correlación se generaliza en dos formas: la correlación parcial para designar la correlación entre dos variables cualquiera, cuando los efectos de otras variables se han controlado y la correlación múltiple, para indicar la variación de la variable respuesta que puede explicarse linealmente por medio de un conjunto de variables explicativas.

Correlación parcial .- en los modelos en los que se relacionan más de dos variables, ocurre que si se encuentra la correlación simple entre las variables del modelo (dos a dos), esta correlación expresa un grado de asociación entre dichas variables sin considerar a las otras. Si alguna de las otras se fija, la correlación puede verse afectada. Se emplea el término de correlación parcial para designar la correlación entre dos variables cualesquiera cuando los efectos de otras variables se han controlado o mantenido fijas.

Ejemplo.- Sean tres variables x_1, x_j, y x_k. La correlación entre las variables x_i y x_j, controlando a x_k se representa mendiante $r_{ij.k}$, en forma análoga se denota a la correlación entre las variables j y k, controlando a i por medio de $r_{jk.i}$.

Esta relación puede extenderse a cualquier número de variables de control, añadiendo más índices a la derecha del punto central, el cual indica además el orden de la correlación. Así, un primer orden parcial tendrá un control (un subíndice a la derecha); un segundo orden, dos controles y así sucesivamente. En relación con esta terminología, la correlación sin controles se designa a menudo como correlación de orden cero. El término correlación *total* se emplea también para designar una correlación entre dos variables sin controles.

Se puede dar ahora la fórmula del coeficiente de correlación parcial de primer orden para las variables x_1, x_j, y x_k,

$$r_{ij.k} = \frac{(r_{ij} - r_{ik}\, r_{jk})}{(1-r_{ik}^2)\ (1-r_{ik}^2)} \qquad (3.32)$$

Observe que la primera correlación del numerador es la correlación total entre las dos variables a relacionar i y j. La variable de control figura en la segunda expresión del numerador, en donde se relaciona con cada una de las otras variables, así como en ambos términos del denominador.

Examinando la ecuación (3.32) para ver cómo la correlación parcial se comporta en relación con las tres correlaciones totales. Suponiendo primero que r_{ij} es positiva. Si r_{ik} y r_{jk} tienen ambas el mismo signo, su producto será positivo, y el numerador será o bien un número positivo menor que r_{ij}, o será

cero o negativo. Por otra parte, el denominador será siempre menor que la unidad, a menos que $r_{ik} = r_{jk} = 0$. Por consiguiente, la fracción resultante puede ser cualquier número entre -1 y +1, según sea la magnitud de las tres correlaciones totales.

Suponga ahora que las correlaciones con la variable de control son de signos opuestos. Se obtiene en tal caso un producto negativo a sustraer de un número positivo, y el resultado será un número positivo mayor. Esto significa que si se empieza con dos variables relacionadas positivamente y si se encuentra una variable de control relacionada negativamente con una de ellas pero positivamente con la otra, la parcial resultante será mayor que la correlación de orden cero. Si la variable de control esta correlacionada ya sea positiva o negativamente con las variables, el denominador será menor que la unidad, y la correlación parcial volvera a ser mayor que la correlación total.

Si se empieza con una correlación total negativa, una variable de control relacionada con cada una de las otras dos en la misma dirección (ya sea positiva o negativa) producira una correlación negativa mayor. Si la variable de control no se relaciona con ninguna de las otras variables, la correlación parcial sería igual a la correlación total.

Si se eleva al cuadrado el coeficiente de correlación parcial, el número resultante representará la proporción de variación de la variable i, no explicada por k, pero que puede explicarse por los valores ajustados de x_j.

El Análisis de correlación no se puede emplear directamente para establecer causalidad debido al hecho de que las correlaciones sólo miden el grado en que diversas variables cambian juntas. Sin embargo, uno de los intereses más comunes en una investigación es el de establecer relaciones causales.

Correlación múltiple.- Al igual que el cuadrado del coeficiente de correlación de orden cero indica el porcentaje de variación explicada por la recta de mejor ajuste, el cuadrado del coeficiente de correlación múltiple puede interpretarse compo el porcentaje de variación explicada por las variables explicativas através de una relación lineal.

La correlación múltiple representa la correlación de orden cero entre los valores reales obtenidos para la variable dependiente y los valores ajustados a partir de la ecuación de mínimos cuadrados. Si todos los puntos se encuentran exactamente en la superficie de mínimos cuadrados, los valores real y ajustado coincidirán, y la correlación múltiple será la unidad. Y cuanto mayor sea la dispersión alrededor de la ecuación de mínimos cuadrados será menor la correlación entre los valores real y ajustado.

Ejemplo.- para el caso de tres variables, tomando a dos como variables independientes y a la primera variable como dependiente. El coeficiente de correlación múltiple se puede escribir como $R_{1.23}$ y el cuadrado será,

$$R^2_{1.23} = r^2_{12} + r^2_{13.2}\,(1-r^2_{12}) \qquad (3.33)$$

$$\begin{pmatrix}\text{Proporcion}\\ \text{explicada}\\ \text{por 2 y 3}\end{pmatrix} = \begin{pmatrix}\text{Proporcion}\\ \text{explicada}\\ \text{por 2}\end{pmatrix} + \begin{pmatrix}\text{Proporcion}\\ \text{adicional}\\ \text{explicada}\\ \text{por 3}\end{pmatrix} \begin{pmatrix}\text{Proporcion}\\ \text{no explicada}\\ \text{por 2}\end{pmatrix}$$

Las correlaciones múltiples sólo tienen una cifra a la izquierda del punto, cifra que indica la variable dependiente. Los números de la derecha, en cambio, indican aquellas variable independientes que se están empleando para explicar la variación de la variable dependiente.

Se opera con los cuadrados tanto de la correlación total como de las correlaciones parciales, ya que se obtienen los porcentajes de la variación explicada. Por lo que no se tiene que preocupar por los signos de estas correlaciones.

Resolviendo la ecuación (3.33) en relación con la parcial $r^2_{13.2}$ se obtiene:

$$r^2_{13.2} = \frac{R^2_{1.23} - r^2_{12}}{1 - r^2_{12}} \qquad (3.34)$$

Esto permite ver la relación entre los coeficientes de las correlaciones múltiples y parciales. En el numerador se sustrae la proporción de la variación de 1 explicada por 2, de la proporción explicada por 2 y 3 actuando juntas ($R^2_{1.23}$). El resultado es el incremento explicado por 3, después de haber permitido actuar a 2. Si dicho incremento de divide entre la proporción de variación dejada sin explicar por 2, se obtiene la parcial entre 1 y 3 controlando a 2. Esto concuerda con la interpretación del coeficiente de correlación parcial.

La fórmula de la correlación múltiple puede escribirse también totalmente en términos de correlaciones de orden cero. Tomando la ecuación (3.32) de $r_{ij.k}$ en términos de coeficientes de orden cero y simplificando la expresión algebraica resultante, se tiene que,

$$R^2_{i.jk} = \frac{r^2_{ij} - r^2_{ik} - 2r_{ij}\, r_{ik}\, r_{jk}}{1 - r^2_{jk}} \qquad (3.35)$$

En particular, si la correlación entre las dos variables independientes j y k es cero, se obtiene:

$$R^2_{i.jk} = r^2_{ij} + r^2_{ik} \qquad (3.36)$$

Puede observarse que R no puede ser menor en magnitud que cualquiera de las correlaciones totales, ya que es imposible explicar menos variación agregando más variables. Normalmente R será mayor que una de las r totales. El cuadrado de la correlación múltiple será en este caso igual a la suma de los cuadrados de las demás correlaciones.

CAPITULO 4

ANALISIS DE TRAYECTORIAS

El Análisis de Trayectorias, es un análisis de tipo cuantitativo; inicialmente desarrollado por el genetista Sewall Wrigth en 1918; para explicar relaciones causales en poblaciones genéticas, y ha sido popularizado por Duncan(1966) en las llamadas Ciencias Sociales, como menciona Wichern(Cap 7, pp. 345-346).

En 1925 este tipo de análisis se aplicó a los precios de grano de cereales y a la carne de cerdo; fue desarrollado extensamente en econometría, pero generalmente sin el uso de los coeficientes de regresión estandarizados.

Wright usó en 1921 los coeficientes no estandarizados y el término de regresión de trayectorias, fué hasta 1954 cuando favoreció la forma estandarizada. Puntualizó que los coeficientes de trayectoria los cuales son medidas estandarizadas, tienen ciertas ventajas sobre los coeficientes no estandarizados, a los cuales Wright denominó "Coeficientes de regresión de trayectorias"

Wright declara en su estudio "Correlation y Causation", el Análisis de Trayectorias es " un método de medida de la influencia directa en un sistema a lo largo de cada trayectoria y de hallar de este modo el grado con que la variación de un efecto dado es determinado por cada causa particular. El método depende de la combinación del conocimiento de los grados de correlación entre las variables de un sistema con el conocimiento que pueda tenerse de sus correlaciones causales".

Como indican Leik y Meeker(1975, 110) el Análisis de Trayectorias permite estimar los parámetros de un modelo causal; mismos que son interpretados como indicadores de la cantidad de cambio (estandarizados) en una variable dependiente, que es atribuible al cambio en una variable anterior.

El objetivo del Análisis de Trayectorias o análisis de ecuaciones estructurales es proporcionar una explicación plausible de correlaciones observadas entre variables mediante la construcción de diagramas de relación causa-efecto. La importancia de los resultados, no se deriva de la simple aplicación de las técnicas estadísticas y matemáticas, sino que depende y está condicionada por el grado en que los datos utilizados, cumplan con los supuestos exigidos (mismos que se verán más adelante).

Un hecho muy importante es que un coeficiente de correlación "significativo" no implica una relación causal. Ciertamente, una correlación observada nunca puede ser usada como prueba de una relación causal. La relación causal debe basarse en argumentos

convincentes construidos o establecidos por una teoría no necesariamente estadística.

El Análisis de Trayectorias puede considerarse como un procedimiento para expresar problemas de regresión mediante un simple diagrama. El diagrama representa el flujo de causa y efecto. Las expresiones resultantes pueden parecer diferentes de aquellas utilizadas en los procedimientos ordinarios de mínimos cuadrados; pero sin embargo, puede demostrarse que los dos sistemas (Análisis de trayectorias y Análisis de regresión) son matemáticamente equivalentes a condición de que se empiece con los mismos modelos y suposiciones.

Ejemplo.- Sea $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ el modelo de regresión de y sobre x_1 y x_2 y $\bar{y} = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \varepsilon$, restando $\bar{y}$ de y,

$$y - \bar{y} = \beta_1 (x_1 - \bar{x}_1) + \beta_2 (x_2 - \bar{x}_2) + \varepsilon - \bar{\varepsilon}$$

se cancela β_0, pero aparece $\bar{\varepsilon}$ que usualmente se considera negligible, sin embargo puede considerarse. Dividiendo la ecuación por $S_0 = S_y$ (se hace esta igualdad sólo por notación).

$$\frac{y - \bar{y}}{S_0} = \frac{\beta_1 (x_1 - \bar{x}_1)}{S_0} + \frac{\beta_2 (x_2 - \bar{x}_2)}{S_0} + \frac{\varepsilon - \bar{\varepsilon}}{S_0}$$

multiplicando el primer término de la derecha por S_1/S_1 y el segundo por S_2/S_2

$$\frac{y - \bar{y}}{S_0} = \beta_1 \frac{S_1}{S_0} \left(\frac{(x_1 - \bar{x}_1)}{S_1} \right) + \beta_2 \frac{S_2}{S_0} \left(\frac{(x_2 - \bar{x}_2)}{S_2} \right) + \frac{\varepsilon - \bar{\varepsilon}}{S_0}$$

$$Y = P_{01}X_1 + P_{02}X_2 + U \qquad (4.1)$$

donde $U = \frac{\varepsilon - \bar{\varepsilon}}{S_0}$

El modelo (4.1) es el modelo de trayectorias, donde todas las variables, Y, X_1 y X_2 estan estandarizadas a media cero y varianza 1. El error U expresa que la variable Y no queda totalmente determinada. Los coeficientes de regresión estandarizados son los coeficientes de trayectorias P_{01} y P_{02}, donde

$$P_{01} = \beta_1 \left(\frac{S_1}{S_0} \right) \qquad P_{02} = \beta_2 \left(\frac{S_2}{S_0} \right)$$

El significado es el mismo que en regresión múltiple, es decir P_{01} es el cambio en desviaciones estándar que experimenta Y al aumentar una desviación estándar X_1, manteniendo constante a X_2.

SUPUESTOS PARA PODER APLICAR EL ANALISIS DE TRAYECTORIAS

Las condiciones que debe reunir un modelo para que sea aplicable este tipo de análisis son las siguientes:

Modelos o sistemas de variables cerradas, o completas.-. En este tipo de modelos como señala Duncan(1974) cada variable endógena debe estar completamente determinada por alguna combinación de variables exógenas en el sistema. En los casos en que no se mantiene la determinación completa por las variables medidas, debe introducirse una variable residual que no esté correlacionada con otras variables determinantes del modelo.

El modelo debe ser recursivo.- Un modelo es recursivo cuando las relaciones entre las variables que la forman son o se suponen que son asimétricas. Este tipo de modelo implica que dos variables no pueden ser recíprocamente causa y efecto una de otra.

Esta condición se exige por la misma idea de causalidad; según la cual, el efecto de una causa no puede ser a su vez causa de su causa al mismo tiempo y en referencia al mismo aspecto.

El modelo debe ser lineal.- Esto quiere decir que las relaciones entre las variables que forman el modelo se pueden representar por ecuaciones lineales, y en el caso en que estas ecuaciones sean de otro tipo, se deben transformar en lineales.

Relaciones de causa-efecto entre las variables del modelo.- El Análisis de Trayectorias por decirlo así es una técnica para el análisis de las estructuras causales, y por tanto exige como condición que todas las variables se relacionen causalmente. Como se menciona en el capítulo 2, el requisito empírico para determinar si una relación es causal toca con los límites de lo teórico.

Nivel de medida de intervalo o de razón,- El Análisis de Trayectorias requiere que las variables del modelo sean de tipo cuantitativo, continuas, y que sus valores tengan por lo tanto escalas de intervalo o de razón. También se pueden emplear variables cualitativas, pero a condición de clasificarlas. Esta clasificación no es tan inmediata cuando son más de dos categorías.

residuales.- Las variables residuales, también llamadas errores, que representan los errores de medición de las variables observadas, o las variables que pueden influir en el sistema, pero que no están incluidas en él; se supone que no están correlacionadas entre sí y que ejercen una influencia aleatoria.

FORMULACIÓN DE ANÁLISIS DE TRAYECTORIAS

Suponga que se tiene un sistema causal cerrado que consiste de q factores primarios o causas X's, y p efectos resultantes Y's. Estas $p+q$ variables deben estar asociadas unas con otras por una red de trayectorias causales. Es conveniente hacer un diagrama de ésta red para representar causalidad utilizando flechas de una sola cabeza, conectando la causa (al final) al correspondiente efecto (cabeza).

Tomando tres variables existen tantos diagramas como combinaciones se puedan formar, conectando la causa al correspondiente efecto. la selección del más significativo y prometedor diagrama, estará basado en el juicio del investigador.

Ejemplo.- Blalock(1971, pp. 78-81) considerando un sistema en el cual dos causas primarias juntas determinan un efecto, el cual a su vez determina un efecto diferente.

El diagrama de trayectorias para éste sistema es:

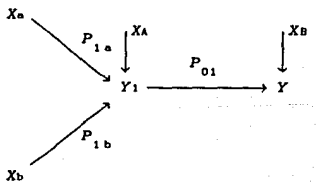


FIG. 4.1 Diagrama de trayectorias con 2 causas X_a y X_b. Y_1 como efecto, y como causa del efecto Y.

Las causas como los efectos son representados con letras mayúsculas. Para los factores primarios o causa (X's) tienen letras del alfabeto minúsculas en subíndices, mientras que para los efectos (Y's) son números. El efecto final sin subíndice, y los errores representados por la letra X con subíndice alfabético en mayúsculas.

La variable en la cabeza de una o más flechas, se puede expresar como una función de las variables en el extremo de estas flechas; así en el ejemplo, Y_1 y Y se pueden expresar como función de las variables X_a y X_b, y Y_1 respectivamente.

Suponiendo linealidad como en el capítulo 2 y 3, se pueden escribir las ecuaciones estructurales con su correspondiente residual. Para el diagrama de la figura 4.1, la estructura lineal es,

$$Y_1 = P_{1a}X_a + P_{1b}X_b + P_{1A}X_A \qquad (4.2)$$

$$Y = P_{01}Y_1 + P_{0B}X_B$$

Donde las P's contienen doble subíndice, los cuales denotan los coeficientes de trayectoria, el primer subíndice indica a la variable en la cabeza de la flecha y el segundo a la variable en el extremo. Note que para el efecto final el subíndice para P es cero. X_A y X_B son los errores correspondientes a Y_1 y Y respectivamente.

Siempre debe haber p ecuaciones estructurales, una ecuación por cada efecto.

Se ha hecho referencia a diagramas de trayectorias (figura 4.1), sin embargo, no se ha dado una explicación explícita de lo que es un diagrama, como se construye y para que sirve; por lo que, acontinuación se da un tratamiento especial a este punto.

DIAGRAMAS DE TRAYECTORIAS

Para el mejor entendimiento del Análisis de Trayectorias es conveniente hacer una representación mediante la cual se expresan gráficamente las relaciones de causalidad, que se supone existen en un conjunto de variables, a esta representación se le llama Diagrama de Trayectorias.

Dentro de un diagrama se pueden distinguir tres tipos de variables del modelo: Endógenas, Exógenas y Residuales. Las primeras son las variables que dependen de otras variables, o son influidas; las variables exógenas son aquellas que no dependen de otras varibles; las residuales son las variables que representan a los factores implícitos no observados, así como, a los posibles errores de medición de las variables del modelo.

Construcción de diagramas.- Un diagrama se construye en base a la experiencia y conocimiento del investigador; fija cuales son las variables endógenas y cuales las exógenas; además de ver cuales son las más relevantes para su estudio.

Pueden construirse tantos diagramas como posibles combinaciones existan de estas variables, sin embargo, como ya se mencionó antes, es el investigador quien decide cuál es el diagrama representativo.

Al construirse el diagrama, las variables se representan por letras mayúsculas, las relaciones entre las variables mediante flechas unidireccionales rectas, que empiezan en la variable independiente (o que influye) y cuya punta termina en la variable dependiente (o influida).

La posible correlación entre las variables exógenas, o no dependientes de otras variables del modelo, se representan por flechas de doble punta (o cabeza) y con la línea de unión curva

en lugar de recta. Es una práctica heurística para añadir realidad a un sistema causal.

Ejemplo.- El siguiente diagrama muestra los diferentes tipos de variable y las relaciones que existen entre ellas.

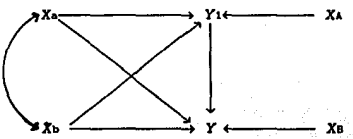


FIG 4.2 Diagrama de trayectorias.

Como se observa en la figura 4.2 las variables se representan por letras mayúsculas; las relaciones entre las variables mediante flechas unidireccionales, si las causas no están correlacionadas, y por flechas de doble cabeza si las causas están correlacionadas.

Las variables X_a y X_b son variables exógenas, Las variables Y_1 y Y son variables endógenas y las dos variables X_A y X_B representan el componemte aleatorio de las variables endógenas.

COEFICIENTES DE TRAYECTORIAS

Una trayectoria no solamente tiene una dirección sino tambien una cantidad o valor que mide por decirlo así la importancia de la trayectoria. A este valor asignado se le llama coeficiente de trayectoria y se denota con la letra *P*.

En el Análisis de Trayectorias los coeficientes *P*, son las incognitas cuyo valor se halla mediante la aplicación de mínimos cuadrados a cada ecuación del sistema estructural del modelo.

ESTIMACIÓN E INTERPRETACION DE LOS COEFICIENTES DE TRAYECTORIAS

Existe un método para estimar los coeficientes de trayectorias, que consiste en la aplicación de regresión de mínimos cuadrados, para cada una de las ecuaciones en el sistema.

Por ejemplo en modelos de fertilidad, se observa que para medir la fertilidad de una mujer es necesario relacionar el número de hijos (Y), con las dos variables; edad (X_2) y número de años de educación (Y_1). Para simplificar la exposición, supongase que para cada miembro de la muestra los valores de las X's y de las Y's están determinados.

Cuando la variable dependiente es una medida de fertilidad, la edad es el mejor ejemplo de una variable control (en el sentido de que un cambio en el valor de esta variables tiene como efecto un cambio en la variable fertilidad); puesto que la manera en la cual fertilidad varía con la edad de una mujer es conocida. Alguna de las siguientes estrategias puede ser adoptada para el análisis:

(*a*) La población podría ser subdividida en cohortes, o grupos de edades y cada una analizada separadamente.

(*b*) La variable dependiente podría ser definida de tal forma que el impacto de edad no afecte.

(c) La edad debe ser incluida explícitamente como una variable de control.

La decisión de como plantear el modelo y que variables tomar depende de lo que se quiera probar con el modelo, desde el punto de vista particular del investigador.

Existe un número de relaciones las cuales se desea que el modelo contenga, para poder hacer un análisis de trayectorias; suponiendo que primero se espera que la edad influya directamente sobre fertilidad al igual que los años de educación, y segundo, que a través de los años de educación la edad influya indirectamente sobre fertilidad. Estas relaciones se muestran en el diagrama 4.3, usando flechas que salen de la variable explicativa y llegan a la variable influida con su correspondiente residual.

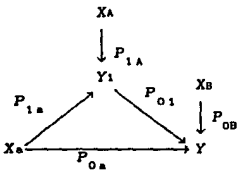


FIG 4.3 Diagrama de trayectoria que representa la fertilidad de una mujer mediante Y=número de hijos, Xa=edad y Y1= años de educación, xA y xB los residuales.

El modelo o sistema de ecuaciones estructurales que se obtiene directamente del diagrma es:

$$(1) \quad Y_1 = P_{1a}X_a + P_{1A}X_A$$

$$(2) \quad Y = P_{0a}X_a + P_{01}Y_1 + P_{0B}X_B \qquad (4.3)$$

Ecuaciones de regresión.- este modelo de trayectorias sigue una secuencia similar a la del Análisis de Regresión convencional y a la solución de ecuaciones simultáneas. Este método consiste en resolver el sistema de ecuaciones por mínimos cuadrados para los parámetros *P*.

Así del sistema (4.3), aplicando mínimos cuadrados a cada ecuación, considerando a X_A y X_B como errores, ya que son variables no observables, los dos sistemas de ecuaciones normales son:

(1) $$\Sigma Y_1 X_a = P_{1a}\Sigma X_a^2$$

(2) $$\Sigma Y X_a = P_{0a}\Sigma X_a^2 + P_{01}\Sigma Y_1 X_a$$
$$\Sigma Y Y_1 = P_{0a}\Sigma X_a Y_1 + P_{01}\Sigma Y_1^2$$

por estar estandarizadas las variables, los coeficientes de correlación son iguales a la suma de los productos de las variables.

$$r_{0a} = \frac{\Sigma(Y - \bar{Y})(\bar{X}_a - \bar{X}_a)}{\sqrt{\Sigma(Y - \bar{Y})^2\Sigma(X_a - \bar{X}_a)^2}} = \frac{\frac{\Sigma(Y - \bar{Y})(X_a - \bar{X}_a)}{n-1}}{\frac{1}{n-1}\sqrt{\Sigma(Y - \bar{Y})^2\Sigma(X_a - \bar{X}_a)^2}}$$

$$= \Sigma Y X_a$$

y $\Sigma Y_1^2 = 1$, $\Sigma X_a^2 = 1$. Bajo esta consideración las ecuaciones normales son las siguientes:

(1) $$r_{1a} = P_{1a}$$

(2) $$r_{0a} = P_{0a} + P_{01} r_{1a}$$

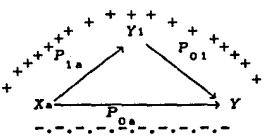
$$r_{01} = P_{0a} r_{1a} + P_{01}$$

La correlación r_{1a} es P_{1a} en (1), las otras correlaciones pueden ser expresadas, a partir de las ecuaciones en (2) en términos de los coeficientes de trayectorias. Como menciona Méndez(1991), las correlaciones entre variables resultan de las trayectorias o caminos. por los cuales se "pueden comunicar" cada pareja de variables. Una correlación se descompone en la suma de los productos de los coeficientes por cada posible trayectoria que conecta las dos variables de esa correlación.

La primera ecuación del sistema (2) es,

$$r_{0a} = P_{0a} + P_{01} r_{1a}$$

sustituyendo r_{1a} de la ecuación (1) entonces $r_{0a} = P_{0a} + P_{01} P_{1a}$, esta ecuación representa las dos trayectorias que conectan Y con X_a. Cada una de ellas esta asociada con uno de los términos P_{0a} y $P_{01} P_{1a}$



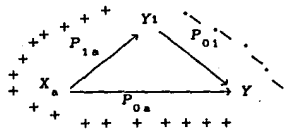
Se dice que r_{0a} es el efecto total (*EF*) entre *Y* y X_a, P_{0a} es el efecto directo (*ED*) de X_a a *Y*, la diferencia entre *ET* y *ED*, es *EI*, efecto indirecto originado por un camino o trayectoria, que conectan a *Y* con X_a de forma indirecta.

$$EI = ET - ED = P_{01}P_{1a}$$

El efecto indirecto mide la influencia de una variable en otra, que se origina por la correlación con otras variables. Para el ejemplo, el *EI* de X_a sobre *Y* es el que se origina por la asociación de ambas con Y_1.

La segunda ecuación del sistema (2),

$$r_{01} = P_{0a}r_{1a} + P_{01} = P_{01} + P_{0a}P_{1a}$$



existen dos trayectoria entre *Y* y Y_1. El $ET = r_{01}$, $ED = P_{01}$, y el $EI = P_{0a}P_{1a}$.

El coeficiente de trayectorias de los errores para cada variable endógena, mide el grado de indeterminación de dicha variable. Como R^2 es el coeficiente de determinación, entonces la

indeterminación es $1 - R^2$, y el coeficiente de trayectorias del error a la variable será,

$$P_{OB} = \sqrt{1 - R^2}$$

para el diagrama 4.3 se tiene

$$P_{OB} = \sqrt{1 - R^2_{Y.Y1Xa}} \qquad P_{1A} = \sqrt{1 - R^2_{Y1.Xa}}$$

Si las variables son estandarizadas (transformadas a media cero y varianza la unidad), se puede utilizar como un segundo método, el uso de las correlaciones observadas de orden cero entre las variables en el sistema. Esta manera de estimar los coeficientes de trayectorias para variables estandarizadas es descrito completamente en Duncan(1966), se le conoce como el método de descomposición de los coeficientes de correlación. los estimadores que se obtienen son idénticos a aquellos obtenidos por el método de mínimos cuadrados.

Descomposición de los coeficientes de correlación.- Tomando las variables estandarizadas, los coeficientes de correlación pueden ser escritos como,

$$r_{ij} = \frac{1}{n} \sum X_i Y_j \qquad \text{con} \qquad i=a,b,\ldots \quad j=1,2,\ldots$$

Sustituyendo la primera ecuación de (4.3) se tiene que

$$r_{a1} = \frac{1}{n} \sum X_a Y_1 = \frac{1}{n} \sum X_a (P_{1a} X_a + P_{1A} X_A) = P_{1a} \qquad (4.4)$$

con $\frac{1}{n} \sum X_a^2 = 1$ y X_A no esta correlacionada con X_1.

Similarmente,

$$r_{0a} = \frac{1}{n}\sum X_a Y = \frac{1}{n}\sum X_a(P_{0a}X_a + P_{01}Y_1 + P_{0B}X_B)$$

$$= P_{0a} + P_{01}r_{a1} \qquad (4.5)$$

$$r_{01} = \frac{1}{n}\sum Y_1 Y = \frac{1}{n}\sum Y_1(P_{0a}X_a + P_{01}Y_1 + P_{0B}X_B)$$

$$= P_{0a}r_{a1} + P_{01} \qquad (4.6)$$

Resolviendo las ecuaciones (4.5) y (4.6) para encontrar P_{0a} y P_{01} en términos de r_{0a}, r_{01} y r_{a1}, se obtiene,

$$P_{0a} = \frac{r_{0a} \quad r_{01}r_{a1}}{1 \quad - \quad r_{a1}^2}$$

$$P_{01} = \frac{r_{01} \quad - \quad r_{0a}r_{a1}}{1 \quad - \quad r_{a1}^2}$$

Así de (4.4), (4.5) y (4.6) los coeficientes P_{0a}, P_{01} y P_{1a} pueden ser obtenidos directamente de los coeficientes de correlación.

Las ecuaciones anteriores, son un caso particular del teorema básico del Análisis de Trayectorias según Tukey(1954),

Teorema:

$$r_{ij} = \sum P_{iK}r_{Kj}$$

donde K denota todos las variables en el correspondiente diagrama de las cuales sus trayectorias conducen directamente a X_j.

Conociendo los valores de las correlaciones se obtienen los coeficientes de trayectorias.

tomando los datos de la Encuesta Mundial de Fertilidad aplicada a 4928 mujeres. Las correlaciones obtenidas de los datos son las siguientes (Kendal, 1977):

$r_{Oa} = 0.64$ Correlación entre edad y número de hijos

$r_{O1} = -0.34$ " entre años de educ. y # de hijos

$r_{a1} = -0.32$ " entre edad y años de educación

sustituyendo estos valores en (4.4), (4.5) y (4.6),

$P_{1a} = -0.32$ $P_{Oa} = 0.59$ $P_{O1} = -0.15$

Los residuales se obtienen usando;

$$r_{11} = 1 = \frac{1}{n}\sum Y_1^2 = \frac{1}{n}\sum Y_1(P_{1a}X_a + P_{1A}X_A)$$

$$= P_{1a}^2 + P_{1A}^2$$

entonces

$$P_{1A} = \sqrt{(1 - P_{1a}^2)} \qquad (4.7)$$

$$r_{OO} = 1 = \frac{1}{n}\sum Y^2 = \frac{1}{n}\sum Y(P_{Oa}X_a + P_{O1}Y_1 + P_{OB}X_B)$$

$$= P_{Oa}^2 + P_{O1}^2 + 2P_{Oa}P_{O1}P_{1a} + P_{OB}^2$$

$$P_{OB} = \sqrt{1 - P_{Oa}^2 - P_{O1}^2 - 2P_{Oa}P_{O1}P_{1a}} \qquad (4.8)$$

En este caso

$$P_{1A} = 0.94 \qquad P_{0B} = 0.76$$

Ejemplo para ilustrar los métodos de solución de los coeficientes de trayectorias: Usando el modelo aditivo saturado (es un modelo recursivo en el cual cada variable se supone dependiente de todas las variables causales, se dice que es o está saturado). Continuando con el ejemplo anterior de fertilidad, agregándole una variable, el diagrama queda como sigue:

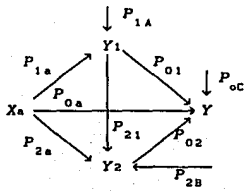


FIG 4.4 Diagrama de trayectorias, donde Y= número de hijos, Xa=edad en años, Y1=años de educación y Y2=edad al momento del matrimonio.

En este caso el ordenamiento causal implica que edad es causa de educación, de edad al matrimonio y de número de hijos. La educación es causa de edad al momento del matrimonio y número de hijos. Finalmente la edad al momento del matrimonio es causa del número de hijos, por lo tanto el modelo puede ser escrito como el siguiente conjunto de ecuaciones:

$$Y_1 = P_{1a}X_a + P_{1A}X_A \qquad (4.9)$$

$$Y_2 = P_{2a}X_a + P_{21}Y_1 + P_{2B}X_B \qquad (4.10)$$

$$Y = P_{0a}X_a + P_{01}Y_1 + P_{02}Y_2 + P_{oC}X_C \qquad (4.11)$$

Solución:

Método de mínimos cuadrados: utilizando las ecuaciones de regresión (4.9), (4.10) y (4.11). Resolviendo primero, la regresión de Y_1 sobre X_a; ésto da P_{1a}, después la regresión de Y_2 sobre X_a y Y_1 de lo cual se obtiene P_{2a} y P_{21}, finalmente, la regresión de Y sobre X_a, Y_1 y Y_2 obteniendo P_{0A}, P_{01} y P_{02}. Los coeficientes de trayectoria de los residuales se obtienen de la raíz cuadrada de las varianzas de los residuales en las tres regresiones.

Obteniendo con los datos de la muestra los valores numéricos siguientes (Kendall Y Muirchertaingh, 1977):

$P_{1a} = -0.32$	$P_{01} = -0.05$	$P_{02} = -0.28$
$P_{2a} = 0.12$	$P_{0a} = 0.62$	$P_{21} = 0.38$
$P_{1A} = 0.95$	$P_{oC} = 0.71$	$P_{2B} = 0.91$

El modelo final esta representado por la ecuación,

$$Y = 0.62X_a - 0,05Y_1 - 0.28Y_2$$

este simplemente representa el efecto directo de las tres variables explicativas. La ventaja principal del modelo estructural es que permite avanzar más en el análisis de los parámetros involucrados.

El efecto total de edad es representado por la correlación entre edad y número de hijos que es igual a 0.64. Sin embargo de (4.15), se tiene que

$$r_{0a} = P_{0a} + P_{01}r_{1a} + P_{02}r_{2a}$$

sustituyendo r_{1a} y r_{2a} de (4.12) y (4.13) se obtienen,

$$r_{0a} = P_{0a} + P_{01}P_{1a} + P_{02}P_{2a} + P_{02}P_{21}P_{1a}$$

que es la descomposición de todas las correlaciones de edad y número de hijos y por tanto, cada uno de los términos puede ser interpretado como sigue:

P_{0a} es el efecto directo de edad; = 0.62

$P_{0b}P_{1a}$ es el efecto indirecto de edad en relación con educación; (-0.32)(-0.28)= -0.03.

$P_{0c}P_{ca}$ es el efecto indirecto de edad en relación con edad al matrimonio; (0.12)(-0.28)=0.03.

$P_{0c}P_{cb}P_{1a}$ es el efecto indirecto de edad a través de educación y de edad al matrimonio; (0.32)(0.38)(-0.28)=0.03.
Los cuatro efectos sumados al efecto total son r_{0a}= 0.64

Se puede observar que el efecto directo es el más importante.

Usando el segundo método (sólo si las variables estan estandarizadas), se obtienen las ecuaciones para cada una de las seis correlaciones en términos de los coeficientes de Trayectorias en el modelo,

$$r_{1a} = P_{1a} \quad (4.12)$$

$$r_{2a} = P_{2a} + P_{21}r_{1a} \quad (4.13)$$

$$r_{21} = P_{2a}r_{1a} + P_{21} \quad (4.14)$$

$$r_{0a} = P_{0a} + P_{01}r_{1a} + P_{02}r_{2a} \quad (4.15)$$

$$r_{01} = P_{0a}r_{1a} + P_{01} + P_{02}r_{21} \quad (4.16)$$

$$r_{02} = P_{0a}r_{2a} + P_{01}r_{21} + P_{02} \quad (4.17)$$

La ecuación (4.12) proporciona la solución para los valores de P_{1a}. Las ecuaciones (4.13) y (4.14) para P_{2a} y P_{21}, y (4.15), (4.16) y (4.17) proporcionan la solución para P_{0a}, P_{01} y P_{02}.

COEFICIENTES DE TRAYECTORIAS EN VARIABLES CAUSA

Dentro del análisis de trayectorias, al momento de plantear el diagrama se pueden distinguir dos tipos de causas o variable exógenas (Li, 1975): causas no correlacionadas, causas correlacionadas.

Causas no correlacionadas.- Se dice que una causa (X_i) es no correlacionada, cuando no esta correlacionada con ninguna otra causa (X_j), es decir que en el diagrama no existe una trayectoria curva de doble cabeza entre X_i y alguna otra causa.

Tomando un primer caso (sólo limitandose a tres variables). Sean X_a y X_b variables exógenas o causas y Y variable endógena o efecto. Suponiendo que no existe correlación entre X_a y X_b, y que además Y queda completamente determinada por dichas causas, el diagrama es el siguiente:

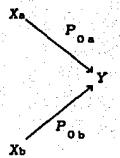


FIG. 4.5 Diagrama de trayectorias con 2 causas Xa y Xb no correlacionadas, Y como efecto.

La ecuación que se obtiene directamente del diagrama es,

$$Y = P_{0a}X_a + P_{0b}X_b \qquad (4.18)$$

Aplicando mínimos cuadrados a la ecuación para los parámetros P's, las ecuaciones normales son:

$$r_{0a} = P_{0a} + P_{0b}r_{ba}$$

$$r_{0b} = P_{0b}r_{ba} + P_{0b}$$

como no existe correlación entre las variables exógenas entonces $r_{ab} = 0$, por lo tanto,

$$r_{0a} = P_{0a}$$

$$r_{0b} = P_{0b}$$

es decir, <u>los coeficientes de correlacipn son igual a los coeficientes de trayectorias cuando las causas son no correlacionadas.</u>

Considerando un segundo caso, suponiendo que se tiene un diagrama de la forma,

$$X_a \xrightarrow{P_{1a}} Y_1 \xrightarrow{P_{01}} Y$$

Para este diagrama las ecuaciones son:

$$Y_1 = P_{1a}X_a \qquad (4.19)$$

$$Y = P_{01}Y_1$$

si se desea encontrar el coeficiente de trayectorias P_{0a}, apartir de P_{1a} y P_{01}. Retomando el hecho de que los coeficientes de trayectorias son iguales a los coeficientes de correlación para causas no correlacionas entonces,

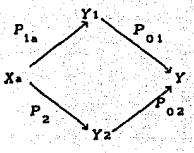
$$r_{01} = P_{01} \qquad r_{1a} = P_{1a}$$

por lo tanto

$$P_{0a} = r_{0a} = \Sigma YX_a = \Sigma(P_{01}Y_1)X_a$$
$$= \Sigma P_{01}P_{1a}X_a^2$$
$$= P_{01}P_{1a}$$

luego entonces, el coeficiente de trayectorias de Xa directamente a Y es igual al producto de las dos componentes de trayectorias.

Considerando un último caso, suponiendo que se tiene un diagrama de la forma,



Las ecuaciones del modelo son,

$$Y_1 = P_{1a}X_a$$

$$Y_2 = P_{2a}X_a \qquad (4.20)$$

$$Y = P_{01}Y_1 + P_{02}Y_2$$

suponiendo que se desea encontrar La trayectoria total P_{0a},

$$\begin{aligned} P_{0a} &= r_{0a} = \Sigma YX_a \\ &= \Sigma(P_{01}P_{1a}X_a^2 + P_{02}P_{2a}X_a^2 \\ &= P_{01}P_{1a} + P_{02}P_{2a} \end{aligned}$$

entonces la trayectoria total que involucra variables intermedias es la suma de todas las trayectorias .

Cuando las causas no están correlacionadas la cadena puede ser extendida a cualquier número de pasos.

Causas correlacionadas.- En este tipo de causas, por el contrario si existe una flecha curva de doble cabeza entre una causas X_i y alguna otra causa X_j, que representa a la correlación que existe entre dichas causas.

Ejemplo.- retomando el diagrama de la figura 4.5, y suponiendo además que existe una flecha curva de doble cabeza entre X_a y X_b, es decir, que las causas están correlacionadas por medio de r_{ab}, el nuevo diagrama es,

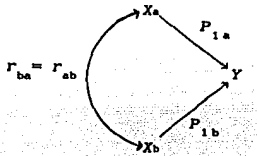


FIG. 4.6 Diagrama de trayectorias con 2 causas correlacionadas X_a y X_b. Y como efecto.

La ecuación es la misma que (4.18), por lo tanto las ecuaciones normales son iguales, solo que ahora la correlación r_{ba} es diferente de cero por lo cual,

$$r_{0a} = P_{0a} + P_{0b} r_{ba}$$

$$r_{0b} = P_{0b} r_{ba} + P_{0b}$$

Esto significa que <u>los coeficientes de correlación son diferentes a los coeficientes de trayectirias para causas correlacionadas.</u> En este caso particular, r_{0a} y r_{0b} son diferentes de P_{0a} y P_{0b} respectivamente.

Una de las características de los diagramas de trayectorias (con causas correlacionadas o no correlacionadas) es que cada paso de la cadena se puede tomar como un diagrama aparte y puede ser analizado separadamente sin afectar las partes restantes del diagrama.

REGLAS DE LECTURA PARA DIAGRAMAS DE TRAYECTORIAS

Las siguientes reglas son necesarias para poder examinar un diagrama. El siguiente ejemplo muestra varias características que ilustran estas reglas.

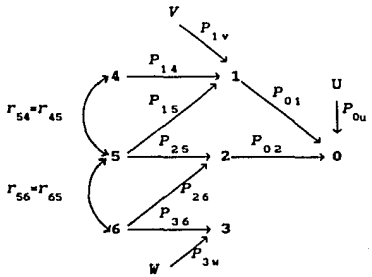


FIG 4.6 Diagrama de trayectorias que ilustra las reglas de lectura de diagramas. Las variables observadas son 0,1,2,3,4,5,6, y los residuales son U, V y W.

(a) Note que no existe trayectoria directa entre 4 y 6. La existencia de r_{45} y r_{56} no conectan 4 con 6.

(*b*) El coeficiente de trayectoria de una causa no correlacionada a un efecto es el coeficiente de correlación respectivo, en el ejemplo $P_{36} = r_{36}$, se debe a que todas las otras variables que determinan a 3 no están correlacionadas con 6, como lo muestra la flecha de *W* a 3, similarmente $r_{0u} = P_{ou}$.

(c) El coeficiente de trayectoria de una causa correlacionada a un efecto es diferente del coeficiente de correlación .

P_{14} no es igual a r_{14} porque 4 está correlacionado con otras causas de 1, así que el coeficiente de correlación en este caso es:

$$r_{14} = P_{14} + P_{15}r_{54}$$

(*d*) Una trayectoria compuesta por una variable intermedia es el producto de dos trayectorias elementales a lo largo de la ruta de conexión. Así el coeficiente de trayectoria de 4 a 0 es:

$$P_{04} = P_{014} = P_{01}P_{14}$$

Note que el coeficiente de trayectoria P_{04}, no es la correlación r_{04}, porque la causa 4 está correlacionada con otras causas que conducen a 0.

(*e*) La cadena ejemplificada arriba puede ser extendida a más de dos variables, permitiendo un cambio en la dirección, es decir

$$P_{015} = P_{01}P_{15}$$

$$P_{0152} = P_{01}P_{15}P_{25}$$

La cadena puede incluir una correlación en medio, o al final, dependiendo de las causas que estén correlacionadas, por ejemplo:

$$P_{0145} = P_{01}P_{14}r_{45}$$

$$P_{01452} = P_{01}P_{14}r_{45}P_{25}$$

Una conexión de trayectoria no puede pasar por una misma variable más de una vez.

(f) La trayectoria total que involucra variables intermedias es la suma de todas las trayectorias

El coeficiente de trayectoria de 5 a 0 involucra dos variables intermedias 1 y 2, por lo que su trayectoria total es la suma de las dos conexiones de trayectoria:

$$P_{05} = P_{015} + P_{025} = P_{01}P_{15} + P_{02}P_{25}$$

En este caso el coeficiente de trayectoria no es el mismo que el de correlación.

g) Una vez familiarizado con el trazado y lecturas de diagramas, al lector no se le dificultara ver los siguientes resultados; cada correlaciónes es la suma de dos, tres o más conexiones de trayectoria.

$$r_{23} = P_{26}P_{36} + P_{25}r_{56}P_{36}$$

$$r_{12} = P_{15}P_{25} + P_{14}r_{45}P_{25} + P_{15}r_{56}P_{26}$$

$$r_{02} = P_{02} + P_{0152} + P_{01452} + P_{01562}$$

$$r_{05} = P_{015} + P_{025} + P_{0145} + P_{0265}$$

h) Para cada porción semicontenida del Diagrama 4.3 podemos calcular la fracción de determinación de los efectos bajo consideración, Las siguientes expresiones dadas para la determinación de las variables 3,2,1, y 0 respectivamente.

$$P_{3u}^2 + P_{36}^2 = 1$$

$$P_{25}^2 + P_{26}^2 + 2P_{25}P_{26}P_{56} = 1$$

$$P_{1v} + P_{14}^2 + P_{15}^2 + 2P_{14}P_{15}r_{45} = 1$$

$$P_{0u}^2 + P_{01}^2 + P_{02}^2 + 2P_{01}P_{02}r_{12} = 1$$

Donde r_{12} esta dada en *(g)*. En la práctica los valores de los coeficientes de trayectoria pueden ser calculados a partir de estas ecuaciones.

Todas las expresiones en esta sección son verdaderas solamente con respecto al punto de vista expresado por la figura 4.3. Si las siete variables (0,1,...,6) son rearregladas de alguna otra forma, las medidas absolutas (Las correlaciones entre las variables) permanecerían iguales, Pero las expresiones involucradas para los coeficientes de trayectorias ya no serían verdaderas. El valor de r_{36} permanecería igual mientras que P_{36} ya no sería igual, dependería del nuevo arreglo(diagrama). Por lo tanto todas las expresiones que involucran coeficientes de trayectoria deben ser siempre leidas con respecto al diagrama de trayectorias especificado.

Por el primer método el álgebra llega a ser incomoda, por otro lado resolviendo por el segundo método es fácil, sin tanta álgebra, y más fácil aún mediante el uso de una computadora.

En el paquete SPSS existe explícitamente una rutina para resolver el Análisis de trayectorias mediante las ecuaciones de regresión como en el segundo método, una vez que ya se haya planteado el diagrama de trayectorias.

Sin embargo existen otros paquetes de tipo estadistico (como NUMBER CHRUNCHER; STATGRAPHICS por mencionar alguno) por medio de los cuales también se pueden obtener los coeficientes de trayectorias, mediante Análisis de Regresión múltiple, ya que los coeficientes de trayectorias son coeficientes de regresión estandarizados. Esto implica que con cualquier paquete que contenga Análisis de Regresión Múltiple se puede trabajar Análisis de Trayectorias.

Primero se construye el diagrama de trayectorias del cual se obtiene el modelo estructural (las ecuaciones de regresión) y una vez obtenidas estas meterlas al paquete y obtener los valores correspondientes a los coeficientes *P's*.

Dentro del Análisis de Trayectorias existe una gran variedad de temas, los cuales no se tratan en este trabajo, por ejemplo existe el análisis de trayectorias con variables no estandarizadas (Blalock, 1971 pp. 406-408 y Kendal y Muirchertaingh, 1977). El análisis de trayectorias para variables binarias Kendal y Muirchertaingh(1977, p. 20), para variables dummy y datos ordinales (Blalock, 1971 pp. 432-437).

CAPITULO 5

AREAS DE APLICACION DEL ANALISIS DE TRAYECTORIAS

Existe un gran número de areas en las cuales se puede aplicar el análisis de trayectorias, a continuación se describirá brevemente algunas de estas.

BIOLOGÍA

Las aplicaciones del análisis de trayectorias en el campo de poblaciones genéticas, son especialmente en aquellos problemas que conciernen a parejas. La razón principal para el uso del método en poblaciones genéticas es que las "causas" y "efectos" son conocidos através de las leyes hereditarias de Mendel, por medio de las cuales se pueden hacer directamente los digramas de trayectorias y de estos las ecuaciones estructurales concernientes al análisis. El mismo Sewall Wright de 1918 a 1921 desarrolló el método para explicar relaciones causales en poblaciones genéticas, (Wichern, 1982). Si desea mayor información sobre este tipo de aplicaciones, consultar Li(1975, Cap. 7,8,9).

EPIDEMIOLOGÍA Y PROBLEMAS MEDICOS

Podemos citar algunos ejemplos de aplicaciones de coeficientes de trayectorias en epidemiología y problemas médicos. La epidemiología involucra varios factores cuyas relaciones causales no son claras y pueden variar de una región (geográfica) a otra. Los epidemiologos exploran la posibilidad de usar análisis de trayectorias en sus estudios. Goldsmith y Berglund(1974) presentan diagramas tentativos de trayectorias para (*i*) Asma y bronquitis infantil, (*ii*) Efisema y bronquitis en los adultos, (*iii*) El deterioro del aparato respiratorio en niños y adolecentes y (*iv*) Cáncer pulmonar; sus metas son identificar los factores etiológicos de las enfermedades humanas através de métodos numéricos.

Un estudio hecho por Kalimo y Bice(1973) es el de las siguientes cuatro variables: X_1=edad, clasificada en tres categorias (15-44, 45-64 y 65 y más); X_2=presencia o ausencia de una enfermedad crónica; X_3=número de dias enfermo(sin trabajar) en las dos últimas semanas(0=ninguno, 1=uno o dos dias, 2=tres o

más dias); X_4=visitas al médico(0 ó 1). El diagrama de trayectorias propuesto es mostrado en la figura 5.1. Las siguientes séis correlaciones observadas entre las cuatro variables estan basadas en las respuestas de casi 35,000 individuos, Li(1975, pp. 331-332).

		Problema crónico X_2	*Dias enfermo* X_3	*Uso méd.* X_4
Edad	X_1	0.25	0.11	0.04
Problema crónico	X_2		0.21	0.08
Dias enfermo	X_3			0.30

Las cinco ecuaciones para los cinco coeficientes de trayectoria se pueden descomponer en tres grupos. Hay solamente una trayectoria entre X_1 y X_2, y hay dos trayectorias entre pares de variables, como se observa en el diagrama de la figura 5.1.

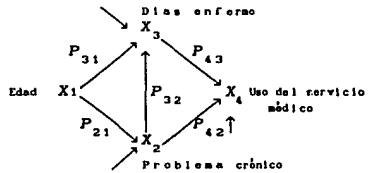


FIG. 5.1 Modelo causal para el uso del servicio médico. X_1 = edad X_2 = problema crónico, X_3 = dias enfermo, X_4 = visitas al médico. Las flechas pequeñas no etiquetadas son los residuales (Kalimo y Bice, 1973).

obteniendo directamente del diagrama, las siguientes ecuaciones:

$$r_{21} = P_{21} = 0.25$$

$$r_{31} = P_{31} + P_{32}r_{21} = 0.11$$

$$r_{32} = P_{31}r_{12} + P_{32} = 0.21$$

$$r_{42} = P_{42} + P_{43}r_{32} = 0.08$$

$$r_{43} = P_{42}r_{23} + P_{43} = 0.30$$

Despejando para obtener los valores de P, se tiene que las soluciones son :

$P_{21} = 0.25$ $\quad$ $P_{31} = 0.061$ $\quad$ $P_{32} = 0.195$

$P_{42} = 0.018$ $\quad$ $P_{43} = 0.29$

Hay cinco coeficientes de trayectoria y séis correlaciones observadas, por lo tanto el sistema es sobre-identificado por una ecuación. Suponga que deseamos conocer si los valores de P son consistentes con la ecuación no usada concerniente a r_{41}. El valor calculado de la correlación entre X_1 y X_4 esta indicado por una r':

$$r'_{41} = P_{42}r_{21} + P_{43}r_{31} = 0.037$$

el cual coincide con el observado $r = 0.04$ de la figura 5.1, dos conclusiones son obvias. La primera es que el número de dias enfermo es la causa más importante para el uso del servicio médico que un problema crónico. El segundo es que el uso del servicio

médico es en gran parte debido a factores no indicados en el diagrama. Sacando el grado de determinación de X_4 para obtener mayor información,

$$P_{43}^2 + P_{42}^2 + 2P_{43}P_{42}r_{23}$$

$$=(.296) + (.018) + 2(.296)(.018)(.21) = 0.09$$

entonces el grado de indeterminación esta dado por $\sqrt{1 - R^2}$, es decir, se deja el 91% de la varianza del uso médico sin contar, esto es debido a los factores residuales. Las características del diagrama es que no existe una conexión directa entre edad y uso del servicio médico; aunque sería razonable incluirla.

URBANIZACIÓN E INGRESOS

Blalock(1961) considera varias posibles relaciones entre cinco variables en 150 ciudades, seleccionadas aleatoriamente, tomadas del censo de 1950. las cinco variables escogidas para el estudio son:

X_1 = Indice de urbanización.

X_2 = Porcentaje de negros en la ciudad.

X_3 = Ingresos de la gente blanca.

X_4 = Indice de el nivel de educación de los negros.

X_5 = Ingresos de la gente negra.

Las correlaciones observadas entre las variables son:

	% *Negros* X_2	*Ingreso de Los blancos* X_3	*Educación* X_4	*Ingreso* X_5
Urbanización X_1	-0.389	0.670	0.264	0.736
% Negros X_2		0.067	-0.531	-0.440
Ingresos de los blancos X_3			0.042	0.599
Educación de los negros X_4				0.386

El diagrama de trayectorias de la figura 5.2 de las variables anteriores fué hecho por Boudon(1965), omitiendo los residuales por simplicidad. Este mueastra esencialmente un estudio a cerca de urbanización e ingresos. El diagrama consiste de dos subsistemas; 1) las variables X_1, X_2 y X_3 forman un sistema semi-contenido concerniente al ingreso de los blancos y 2) Las variables X_1, X_2, X_4 y X_5 forman otro sistema concerniente al ingreso de los negros. La existencia de los subsistemas implica que las ecuaciones involucran coeficientes de trayectoria que están comprendidos en grupos y pueden ser resueltos en cada grupo. Primero, note que hay solamente una trayectoria entre X_1 y X_2, y entre X_2 y X_4. Entonces,

$$r_{21} = P_{21} = -0.389$$

$$r_{42} = P_{42} = -0.531$$

las siguientes dos ecuaciones forman otro grupo

$$r_{31} = P_{31} + P_{32} r_{21} = 0.670 \qquad P_{31} = 0.820$$

$$r_{32} = P_{31} r_{12} + P_{32} = 0.067 \qquad P_{32} = 0.386$$

obteniendo asi, los valores correspondientes a las P's involucradas,

$$P_{31}=0.820 \qquad P_{32}=0.386$$

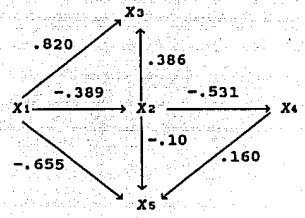


FIG. 5.2 Un modelo de relación causal entre urbanización($X1$) e ingreso de la gente blanca ($X3$) e ingreso de la gente negra ($X5$). Las variables intermedias son: $X2$= Porcentaje de población negra en una ciudad; $X4$= Nivel de educación de los negros (Blalock, 1961 y Boudon, 1965)

Similarmente, las siguientes tres ecuaciones forman otro grupo:

$$r_{51}= P_{51}+ P_{52}r_{21}+ P_{54}r_{41}= 0.736$$

$$r_{52}= P_{51}r_{12}+ P_{52}+ P_{54}r_{42}= -0.440$$

$$r_{54}= P_{51}r_{14}+ P_{52}r_{24}+ P_{54}= 0.386$$

sustituyendo los valores observados de r y resolviendo, se obtiene,

$$P_{51}=0.655 \qquad P_{52}=-0.100 \qquad P_{54}=0.160$$

Así se han obtenido los valores de los siete coeficientes de trayectoria. Las diez correlaciones entre las cinco variables han sido observadas y solamente hay siete coeficientes de trayectoria en el diagrama, por lo tanto el sistema esta sobre-determinado. Las tres correlaciones restantes pueden ser usadas para ver si son consistentes con los coeficientes de trayectoria o no. Una r' es usada para indicar los valores cálculados de las correlaciones.

CALCULADOS	*OBSERVADOS*
$r'_{41}= P_{42}r_{21}= 0.207$	$r_{41}=0.264$
$r'_{34}= P_{42}r_{23}= -0.036$	$r_{34}=0.042$
$r'_{35}= P_{31}r_{15}+ P_{32}r_{25}= 0.434$	$r_{35}=0.599$

Note que las correlaciones cálculadas son menores que las observadas. Al agregar una trayectoria directa de urbanización (X_1) al nivel de educación de los negros (X_4) en el diagrama, daría una nueva ecuación:

$$r_{41}= P_{41}+ P_{42}r_{21}= P_{41}+ (-0.531)(-0.389)=0.264$$

obteniendo

$$P_{41}=0.264 - 0.207 = 0.057$$

Pero esto no cuenta comparativamente a la correlación alta entre el ingreso de los blancos y de los negros, r_{35}=0.599. El esquema causal tiene solamente un factor común X_1 para X_3 y X_5. Este mismo sugiere algún otro factor común para el ingreso de blancos y negros que no ha sido incluido.

FERTILIDAD

En un estudio sobre fertilidad, una encuesta es aplicada a 4700 mujeres, midiendo para cada mujer las siguientes variables, Wonacott(1981).

X_A = Edad de la mujer al momento del estudio.
Y_1 = Años de educación (estudio) de la mujer.
Y = Número de hijos que tiene.
Y_2 = Edad de la mujer al momento del matrimonio.

De acuerdo con las variables anteriores se construye el siguiente diagrama de fertilidad de la mujer. Presentando las variables de izquierda a derecha, ordenadas de acuerdo a sus causas y efectos.

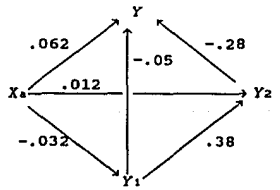


FIG. 5.3 Diagrama de trayectorias para XA= Edad al momento del estudio Y1= Años de educación, Y2= Edad de la mujer al momento del matrimonio y Y= Número de hijos. Es el diagrama para la fertilidad de una mujer.

obteniendo las ecuaciones,

$$Y = P_{0a}X_a + P_{01}Y_1 + P_{02}Y_2$$

$$Y_2 = P_{2a}X_a + P_{21}Y_1$$

$$Y_1 = P_{1a}X_a$$

siguiendo los pasos descritos en el capítulo 4 para encontrar los coeficientes de trayectoria, y de acuerdo a la matriz de correlaciones (Wonnacott, 1981); las ecuaciones de regresión son las siguientes:

$$Y = 0.062X_a - 0.05Y_1 - 0.028Y_2$$

$$Y_2 = 0.012X_a + 0.38Y_1$$

$$Y_1 = -0.032X_a$$

cálculando el efecto total indirecto de las otras variables sobre el número de niños:

a. Educación (Y_1).
b. Edad presente (X_a).
c. Edad al matrimonio (Y_2).

Resolviendo para *a*: Tomando el diagrama semi-contenido entre Y_1, Y_2 y Y. De Y_1 a Y existen dos trayectorias, una directa de Y_1 a Y y la otra indirecta via Y_2. Entonces

efecto directo de Y_1 *sobre* Y		$= -0.05$
efecto indirecto vía Y_2	$(.38)(-.28)$	$= -0.11$
efecto total		$= -0.16$

Esto es, una mujer que tiene mayor educación tiene en promedio menor número de hijos (es decir 0.16 menos hijos por cada año de educación que ha recibido). Esto es en parte, porque a mayor educación menos hijos (relación directa), y en parte, porque a mayor educación produce que se casen a mayor edad, el cual da como resultado pocos hijos (relación indirecta).

b. De X_a a Y hay varias trayectorias, una directa y varias indirectas a través de las variables Y_1 y Y_2; considerando a cada una en su momento:

efecto directo de X_a *sobre* Y		= -0.062
efecto indirecto vía Y_2	(.012)(-.28)	= -0.003
efecto indirecto vía Y_1	(-.032)(-.05)	= 0.002
vía Y_1 *y* Y_2	(-.032)(.38)(-.28)	= 0.003
efecto total		= 0.064

c. De Y_2 a Y hay justamente una trayectoria, por lo tanto la respuesta es inmediatamente obtenida: *efecto total de* Y_2 = -.28

En resumen, el efecto total de una variable (X_a) sobre otra variable (Y) esta definido como el cambio en desviaciones estándar que ocurre en Y cuando X_a cambia en una desviación estándar -hablando de todos los cambios con la intervención de variables entre X_a y Y. El efecto total puede ser cálculado de la cadena de efectos directos usando el siguiente hecho: Cada trayectoria de X_a a Y, multiplicando juntos todos los coeficientes se encuentra que:

El efecto total de X_a sobre Y = *la suma de todas las trayectorias (siguiendo las flechas de* X_a *a* Y*).*

Existen otros ejemplos de aplicación del análisis de trayectorias en las Ciencias Sociales, de los cuales se mencionan algunos. Existe un interesante estudio de Sewell Y Shah(1967) sobre estatus socioeconómico, inteligencia y educación; Hauser(1973) en estudios socioeconómicos y desempeño educacional. Análogamente a Duncan(1966) en su artículo "Psychological examples", Wert y Linn(1970) aplican el análisis de trayectorias a un conjunto de problemas en el seno familir. El economista Gintis(1971) reporta un Análisis de Trayectorias sobre educación y productividad en el trabajo, Griliches y Mason(1972) emplean este tipo de análisis en un estudio sobre educación e ingreso, haciendo una prueba de aptitudes como indicador de una variable no observada representando habilidad y capacidad.

Sewall Wright uso el método para estimar un modelo de oferta y demanda para la carne de puerco, uno más para las variables X_1=papas y X_2=precio rezagado de la carne; también desarrolló la versión dinámica del modelo, LI(1975, pp. 335-337). Las aplicaciones que se hicieron en 1925 fuerón a los precios de granos de cereales y a la carne de cerdo, Wichern(1982).

El Análisis de Trayectorias puede observarse como una forma especial de "ecuaciones estructurales" empleados por varios autores en diferentes campos. En el volumen editado por Blalock(1971) se encuentran una serie de aplicaciones, ya que es una colección de artículos de modelos causales y una fuente de referencias para estos métodos.

CONCLUSION

Se trato en este trabajo de dar a conocer el análisis de trayectorias como una técnica estadística de tipo causal.

El análisis de regresión y de trayectorias son matemáticamente equivalentes a condición de que uno empiece con los mismos modelos y suposiciones. La diferencia estaría en la forma de las ecuaciones, es decir que mientras el análisis de trayectorias requiere de un sistema de ecuaciones recursivo (ésto viene exigido por la misma idea de causalidad), el análisis de regresión no.

El análisis de trayectorias por trabajar con sistemas recursivos, presenta la ventaja de que los estimadores mínimo cuadraticos ordinarios aplicados a cada ecuación son consistentes, eficientes y asintóticamente normales. Además tiene ventajas sobre el análisis de regresión, ya que mientras la regresión sirve para explicar y predecir sin considerar causalidad, el análisis de trayectorias sirve para explicar considerando causalidad. El requisito de causalidad en determinado momento se puede pensar como una limitante del método, ya que para concluir que entre dos variables existe relación de causa efecto es preciso examinar cuidadosamente la naturaleza del fenómeno que se estudia (este requisito caé en los límites de lo teórico). Por lo anterior, no se quiere decir que el análisis de trayectorias se deba usar en lugar del análisis de regresión o de las técnicas multivariadas, si no, más bien como complemento de las mismas.

En algún tiempo (años 60's) se penso que por fin se tenía una técnica que permitía medir la fuerza de los vínculos causales, sin embargo sólo limitandose al campo de la medición.- Se percibió que si bien es posible medir la relación entre dos variables y establecer la precedencia temporal de una de ellas, no es posible garantizar que la relación se mantenga una vez que se controlan "todos" los factores que inciden sobre la relación.

Por otra parte existen técnicas para probar que tan confiables son los estimadores de los coeficientes de trayectorias, como son, intervalos de confianza y pruebas de hipótesis, estos resultados quedarón fuera del presente estudio.

Para la solución de los coeficientes de trayectorias, existen paquetes estadísticos para microcomputadoras. Uno de los paquetes que contiene explicitamente subrutinas para la solución del análisis de trayectorias es el SPSS (Statistical Package for the Social Sciences), sin embargo en el StatGraphics, NumberCroncher (por mencionar algunos) no existe. En estos casos lo que se debe hacer, es tomar las subrutinas de análisis de regresión múltiple y obtener los coeficientes de regresión estandarizados. Ya que como se mencionó en el capítulo de análisis de trayectorias son los coeficientes de regresión estandarizados. Tomando ese hecho, se obtienen finalmente los coeficientes de trayectorias.

APENDICE A

PROPIEDADES DE LOS ESTIMADORES POR MINIMOS CUADRADOS

Los estimadores mínimo cuadrados son función lineal de las y_i, y éstas a su vez de las ε_i's. De estos estimadores se obtiene la media, la varianza y la covarianza.

En particular se supone que el modelo de regresión simple es

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1,2,3,\ldots, n$$

con $E(y_i) = \beta_0 + \beta_1 x_i$ y $Var(y_i) = Var(\varepsilon_i) = \sigma^2$. Ahora considere el estimador $\hat{\beta}_1$, dado en (3.8). Suponga que se definen las constantes $c_1, c_2, \ldots c_n$ por la ecuación,

$$c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

entonces las x_i son consideradas como números fijos, estos son las c_i. El estimador $\hat{\beta}_1$ es igual a una combinación lineal de las y_i's. Aplicando el desarrollo que dice: Si a_i es una constante y u_i una variable aleatoria entonces $E(a_0 + \Sigma a_i u_i) = a_0 + \Sigma a_i E(u_i)$, entonces la media de β_1 se encuetra como sigue:

$$E(\hat{\beta}_1) = E(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} c_i E(y_i)$$

$$= \Sigma c_i (\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \Sigma c_i + \beta_1 \Sigma c_i x_i$$

$$= \beta_1$$

porque

$$\sum_{i=1}^{n} c_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\bar{x} - n\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = 1$$

lo cual muestra que $\hat{\beta}_1$ es un estimador insesgado de β_1. También se puede demostrar facilmente que $E(\hat{\beta}_0) = \beta_0$.

La varianza de $\hat{\beta}_1$ se encuetra aplicando el siguiente desarrollo para la varianza de una suma de variables correlacionadas:

$$Var\left(\sum_{i=1}^{n} a_i u_i\right) = \sum_{i=1}^{n} a_i^2 Var(u_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j Cov(u_i, u_j)$$

por lo tanto

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} c_i \psi_i\right)$$

$$= \sum_{i=1}^{n} c_i^2 Var(\psi_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_i c_j Cov(\psi_i, \psi_j)$$

pero $Cov(\psi_i, \psi_j) = Cov(\beta_0+\beta_1 x_j+\varepsilon_j, \beta_0+\beta_1 x_i+\varepsilon_i) = Cov(\varepsilon_j, \varepsilon_i) = 0$ por suposición. También $Var(\psi_i) = Var(\varepsilon_i) = \sigma^2$,

entonces

$$Var(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (A.1)$$

porque

$$\sum_{i=1}^{n} c_i^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)} = \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

para encontrar la variaza de $\hat{\beta}_0$, se tiene que,

$$Var(\hat{\beta}_0) = Var(\bar{y} - \hat{\beta}_1 \bar{x})$$

$$= Var(\bar{y}) + \bar{x}^2 \, Var(\hat{\beta}_1) - 2\bar{x} \, Cov(\bar{y}, \hat{\beta}_1)$$

la $Var(\bar{y}) = \frac{\sigma^2}{n}$, la $Var(\hat{\beta}_1)$ está dada en (A.1) y $Cov(\bar{y}, \hat{\beta}_1) = 0$, este último resultado puede ser demostrado, sin embargo, es intuitivamente claro porque el valor promedio $\bar{y}$ no depende en algún momento del valor ajustado de $\hat{\beta}_1$. Así que

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

Finalmente aplicando la regla para la covarianza, $Cov(a_0 + a_1 u_1, a_3 + a_2 u_2) = a_1 a_2 Cov(u_1, u_2)$ se tiene que,

$$\begin{aligned} Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) \\ &= Cov(\bar{y}, \hat{\beta}_1) - \bar{x} Cov(\hat{\beta}_1, \hat{\beta}_1) \\ &= -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned}$$

APENDICE B

Un vector cuyos elementos son variables aleatorias, es llamado un vector aleatorio. En regresión, el vector de parámetros estimados $\hat{\beta}$ es un vector aleatorio al igual que el vector de errores ε, el vector de valores observados Y, valores ajustados $\hat{Y}$ y el residual $\hat{e}$.

La media o el valor esperado de un vector aleatorio es el vector de medias de las variables aleatorias en ese vector. Así por ejemplo, la media de ε es un vector de ceros, escrito como $E(\varepsilon)=0$.

Para encontrar la media de $\hat{\beta}$ usaremos el hecho de que la media es lineal, el cual dice que si z es un vector aleatorio de longitud $nx1$, C es una matriz de qxn, y d cualquier vector fijo, entonces la media de la variable aleatoria Cz + d es:

$$E(Cz + d) = CE(z) + d$$

entonces

$$E(\hat{\beta}) = E((X^tX)^{-1}X^tY) = E((X^tX)^{-1}X^t(X\beta+\varepsilon))$$

de acuerdo al modelo $Y = X\beta + \varepsilon$.

$$E(\hat{\beta}) = (X^tX)^{-1}X^tX\beta + (X^tX)^{-1}X^tE(\varepsilon)$$

pero $E(\varepsilon) = 0$,

$$E(\hat{\beta}) = (X^tX)^{-1}X^tX\beta = \beta$$

por lo tanto $\hat{\beta}$ es un estimador insesgado de β.

Un vector aleatorio tiene asociada una matriz de varianzas-covarianzas, las entradas de la diagonal son las varianzas de los elementos del vector aleatorio, mientras que las entradas fuera de la diagonal son las covarianzas entre los elementos; el elemento (i,j) de la matriz es la covarianza entre el elemento i-ésimo del vector aleatorio y el elemento j-ésimo. Usamos el simbolo $Var(z)$ para denotar la matriz de varianzas-covarianzas del vector z.

El vector de errores ε, se ha supuesto que tiene elementos con varianza común σ^2 y covarianza 0 (cero) para toda i. En resumen $Var(\varepsilon)=\sigma^2 I_n$.

Aplicando el hecho de que la $Var(Cz + d) = C[Var(z)]C^t$ para encontrar la varianza de $\hat{\beta}$. se tiene que

$$Var(\hat{\beta}) = Var((X^tX)^{-1}X^tY)$$

$$= Var[(X^tX)^{-1}X^tX\beta + (X^tX)^{-1}X^t\varepsilon]$$

$$= [(X^tX)^{-1}X^t[Var(\varepsilon)][(X^tX)^{-1}X^t]^t$$

$$= (X^tX)^{-1}X^t(\sigma^2 I_n)X(X^tX)^{-1}$$

$$= \sigma^2(X^tX)^{-1}X^tX(X^tX)^{-1}$$

$$= \sigma^2(X^tX)^{-1}$$

BIBLIOGRAFIA

BLALOCK, H.M.Jr. Causal models in the science social. *Londres McMacmillan, Aldine atherton/chicago/1971.*

BLALOCK, HUBERT M. Estadistica social. *Fondo de cultura económica. Segunda reimpresión 1983.*

BLALOCK, Hubert M.Jr. Causal Inferences in nonexperimental Research. *Chapel Hill:Univ. of Nort Carolina Press. 1961.*

BOUDON, R. A method of linear Causal Analysis: dependence analysis. *Am Social Rev. 30:365-374. 1965.*

BRAITHWAITE, Richard B. Scientific Explanation. *Cambridge Univ. Press, 1953.*

CORTES, F. RUBALCAVA, R.M. La moda: consideraciones sobre el uso de la estadistica en ciencias sociales. *Colegio de México.*

CUTHBERT Daniel, Freud S. Wood. Fitting Equations to data. *JohWiley and Sorr, 1980.*

DAVIS, J.A. Elementary survey analysis. *Printice. HALL, 1971.*

DUNCAN, Otis Dudley. Path Analysis: Sociological Examples. *American Journal of Sociology 72:1-16. 1966.*

FREUND E. Jonh, WALPOLE, E. Ronald. Mathematical statistics. *Prentice Hall 1980.*

GINTIS, H. Education, technology, and the characteristics of worker productivity. *Amer. Economic Rev. 61:226-279, 1971.*

GOLDSMITH,J. R. and BERGLUND, K. Epidemiological approach to multiple factor interactions in pulmonary diseases: the potential usefulness of path analysis. *Annals N. Y. Acad. Sci 221:361-375, 1974*

GRILICHES, Z., and HAZON, W. Education, income, and ability. *J. Political Economy 80:S74-103. 1972.*

HAUSER, R. M. Socioeconomic background and education performance. *Amer. Sociological Assoc., Washington, D. C. 1973.*

HEISE, D. R. Problems in path analysis and causal inference. *In Sociological methodology, ed. E.F. Borgatta, pp. 38-73. San Francisco: Jossey Bass. 1969.*

HOGG, Robert v. Allen T. Introduction to mathematical statistics. *Co., Inc. New York Collier Macmillan Publishers London. Four Edition 1978.*

JOHNSTON, J. Econometric Methods. *McGraw-Hill, New York 1972.*

KALIMO, E, and BICE, T, W. Causal Analysis and ecological Fallacy in crossnational epidemiological research. *Scand. J. Soc. Med. 1:17-24.1973.*

KENDALL M. G., MUIRCHERTAINGH C. A. O. Path analisis and model building. *Technical bulletins, March 1977. International statistical institute. Prinses Beatrixloon 428. Netherlands.*

KOTZ, Samuel, JOHNSON L. Norman. Encyclopedia of statistical sciences. *Volumen 6. USA 1985.*

KRUSKAL, I. WILLIAM H. International encyclopedia of statistics. *New York:Free Press, 1978.*

LAND, K.C. Principles of path analysis. *In Sociological methodology, ed. E.F. Borgatta. pp. 3-37. San Francisco: Jossey-Bass. 1969.*

LARSON, Harold J. Introduction to the theory of Statistics. *United State of America. 1973.*

LEIK, R. K. MEEKER, B. Mathematical sociology. *Printice Hall, 1975*

LI, Ching chun, Ph. D. Path Analysis -A primer-. *the boxwood press Second printing with corrections, 1975.*

MENDEZ, Ramires I. La ubicación de la estadística en la metodología científica. *Comunicaciones Técnicas, IIMAS-UNAM. Octubre 1987.*

MENDEZ, Ramires I. Análisis de Senderos en tablas de contingencia. *Matemáticas Aplicadas Estadistica y Computación. No. 2 pp. 227-254. Colegio de postgraduados.*

MENDEZ, Ramírez I. Modelos estadisticos lineales, interpretación y aplicaciones. *Consejo Nacional de Ciencia y Tecnología. 1981.*

MENDENHALL W. REINMUTH J. E. Estadistica para administración y economia. *Duxbury Press. 1978.*

NANNY, Wermuth. Linear Recursive ecuation, Covariance Selection, and Path Analisis. *Journal of the American Statistical Association. Teory and Methods Section. December 1980, Volume 75, Number 372.*

NORMAN, H. Nie, HADLAI, Hull C. Statistical Package for the Social Science (SPSS). *USA, McGraw-hill 1975.*

NOVAK, S. Causal interpretation of statistical relationship in social research in quantitative Sociologie. *Nueva York, 1975.*

PARIJS, PH. Van. La systase de l'explication dans les sciences sociales. *Recherches Sociologiques, 1977.*

SEARLE S. Linear Models. *Wiley, New York. 1971.*

SEWELL, W. H., aand, SHAH, V.P. Socioeconomic status, intelligence, and the attainment of higher education. *Sociology of Education 40:1-23. 1967.*

SIERRA, Bravo R. Modelos Mátematicos en las ciencias sociales. *Madrid. Paraninfo 1981.*

SILLS, David L. Enciclopedia internacional de las ciencias sociales. *Volumen 1. Madrid: Aguilar, 1974.*

SIMON, Herbert A. Models of man, Social and Rational: Mathematical Essays on rational human behavior in a social setting. *New York: Wiley, 1957.*

TUKEY, J. W. Causation, Regression and Path analysis. *ed. Kempthorne, Boncroft, Grwen, and Iush, Ames: The Iowa State College Press. 1954.*

WEISBERG, Sanford. Applied Linear Regression. *Johnwiley University of Minnesota. Second Edition. April 1985.*

WERTS, C. E., and LINN, R. L. Path analysis: psychological examples. *Psychological Bulletin 74:193-212. 1970.*

WICHERN Dean W, JOHNSON, Richard Arnold. Applied Multivariate statistical Analysis. *Englewood Cliffs, New Jersey: Printice Hall, 1982.*

WILKS S, S, Mathematical Statistics. *New York: John Wiley &Sons, Inc., 1962.*

WONNACOTT, H. Thomas. WONNACOTT J. Ronald. Regresion:A second course in statistics. *John Wiley &Sons, Inc. New York, 1981.*

WRIGHT, S Correlation and Causation. *F. Agri. Res. 20, 1921, pp. 85-557*