independientes, cualquiera sea el nivel de medición (en las variables independientes, la variable dependiente debe ser medida a nivel intervalar).

Antes de desarrollar las alternativas del subprograma de *breakdown* conviene introducir conceptualmente la idea de confiabilidad de la diferencia entre estadísticos.

Para la investigación es importante no solamente estimar los valores poblacionales, sino utilizar el error estándar para interpretar varios resultados, en lo relativo a las diferencias que pueden existir entre ellos.

El tipo de preguntas que nos planteamos aquí es: ¿Cuál es la fiabilidad de la diferencia entre medias proporciones, etc., que hemos registrado en nuestras observaciones? ¿Son los hombres o las mujeres más capaces en comprensión verbal? ¿El rendimiento intelectual en las clases medias es superior, inferior o igual al rendimiento intelectual en las clases bajas?, etcétera.

A) *Error estándar de la diferencia de medias (Subprograma T-Test)*

La magnitud de la oscilación en la diferencia entre medias obtenidas de muestras distintas, dependerá naturalmente de la magnitud de la oscilación que es propia de las medias. La estabilidad de las medias estará representada por sus respectivos errores estándares.

Cuando las N son lo suficientemente grandes, las medias oscilan alrededor de un valor central (parámetro que por lo general no conocemos). Nuestra finalidad es entonces determinar primero si existe diferencia, para luego definir su magnitud.

El problema reside entonces en determinar si la diferencia que se examina entre las dos medias muestrales, implica además una diferencia en la distribución de la población; en otras palabras si la diferencia es la expresión de diferencias reales a niveles poblacionales, o se deben simplemente a los efectos del azar (y por consiguiente del error) en las muestras.

El *test* T de student nos ayuda a establecer cuándo la diferencia entre dos medias es significativa. Para ello se formula una hipótesis nula. Una hipótesis nula supone que las dos muestras provienen de la misma población; consecuentemente las desviaciones son interpretadas como debidas al efecto del azar sobre las muestras. Según la hipótesis nula se supone que la distribución de las diferencias es normal, de donde $M_1 - M_2 = 0$.

El nivel de significación para la aceptación o rechazo de la hipótesis nula es seleccionado por el investigador. Los más comunes son de .05 y .01 aunque esto depende más bien del área que se está investigando (para aceptar o rechazar una vacuna nueva que cure el cáncer puedo elegir un nivel menor; cuando se trata de la introducción de una medicina para suplantar alguna en uso con cierto grado de eficacia, elegiré un nivel de significación mayor).

El valor de t nos informará entonces sobre la probabilidad o improbabilidad para la aceptación o rechazo de la hipótesis nula, o de alguna hipótesis alternativa. Es decir no se afirma que no existe una diferencia en los resultados, sino únicamente que la diferencia es, o no es significativa.

El subprograma *T-Test* computa los valores t y sus niveles de probabilidad para dos tipos de casos: a) *Muestras independientes* o error estándar de la diferencia para medias no correlacionadas, es decir, para situaciones en las que las dos series de observaciones son independientes. Por ejemplo, comparación del rendimiento de hombres y mujeres en una situación de *test*. b) *Muestras apareadas,* o error estándar de la diferencia para medias correlacionadas. El ejemplo típico es el de las mediciones antes-después en diseños experimentales.

Existen casos en los cuales el investigador no plantea la hipótesis nula (la hipótesis de las no-diferencias), sino plantea una hipótesis alternativa, en la que trata de demostrar que la media en un grupo es más grande que la media del otro. En estos casos la interpretación del valor t obtenido se hace a partir de lo que se llama *test* de una sola cola, es decir, se toma en cuenta solamente una mitad de la distribución.

Tablas de contingencia y medidas de asociación (subprograma crosstabs)

Este subprograma contiene tanto tabulaciones cruzadas para tablas de $n \times k$, así como una serie de medidas de correlación, asociación y de confiabilidad de la diferencia entre estadísticos. Comenzaremos por los análisis de tipo más sencillo para continuar luego con los cálculos de medidas más complejas.

Tabulaciones cruzadas

Una tabulación cruzada es simplemente la combinación de dos o más variables discretas o clasificatorias en la forma de tabla de distribuciones de frecuencia. El cuadro resultante puede ser sometido a análisis estadístico, en términos de distribuciones porcentuales, aplicación de *test* de significación, coeficientes de asociación y de correlación, etcétera.

Las tablas cruzadas son muy utilizadas en análisis de encuestas, en tablas de 2×2 o con la introducción de variables de prueba o de control o intervinientes, constituyendo así tablas de $n \times k$. Aquí el investigador debe tener especial cuidado cuando solicita tablas de $n \times k$ que el tamaño de su muestra sea lo suficiente grande para permitir que cada uno de los casilleros contenga las frecuencias esperadas.

De todos modos existe una serie de restricciones al uso de los distintos estadísticos que señalaremos más adelante y que imponen algunas limitaciones en cuanto a la cantidad total de casilleros, ya sea por cantidad de variables o por cortes en cada una de ellas. Por ejemplo: el investigador debe recordar que si combina digamos 4 variables, todas dicotomizadas, la cantidad total de casilleros será de 16; si las variables estuvieran tricotomizadas la cantidad de casilleros ascendería a 81. La fórmula genérica para el cálculo del tamaño final de la matriz es:

$$M = r_1 \cdot r_2 \cdot r_3 \ldots r_n$$

donde:

r = Cantidad de cortes o divisiones en cada una de las variables
M = Tamaño de la matriz de datos

Hay que recordar que en este tipo de cuadros hay que esperar un promedio de 10 a 20 casos en cada uno de los casilleros, lo que hace que las muestras deben tener tamaños considerables cuando se desea cuadros muy complejos.

Normalmente los cuadros imprimen tanto las frecuencias dentro de cada casillero o celda, como los porcentajes con respecto al marginal horizontal y al marginal vertical y al total general (en ese orden), además de todos los coeficientes que incluye la subrutina y que a continuación pasamos a detallar.

Ji cuadrado (χ^2)

Es un modelo matemático o *test* para el cálculo de la confiabilidad o significado de diferencias entre frecuencias esperadas (f_e) y frecuencias observadas (f_o). La utilidad de este *test* no-paramétrico para variables nominales reside en su aplicación para prueba de hipótesis para tres tipos de situaciones:

a) Prueba de hipótesis referidas al grado de discrepancia entre frecuencias observadas y frecuencias esperadas, cuando se trabaja sobre la base de principios apriorísticos;

b) Pruebas de hipótesis referidas a la ausencia de relación entre dos variables. Se trata de pruebas de independencia estadística y son trabajadas en base a cuadros de contingencia, y

c) Pruebas referidas a la bondad de ajuste. En este caso se trata de comprobar si es razonable aceptar que la distribución empírica dada (datos observados), se ajusta a una distribución teórica, por ejemplo, binomial, normal, Poisson, etc. (datos esperados).

Supuestos y requisitos generales

a) Las observaciones deben ser independientes entre sí. *b)* Los sucesos deben ser mutuamente excluyentes. *c)* Las probabilidades que figuran en las tablas de X^2 están basadas en una distribución continua, mientras que el X^2 calculado en la práctica lo está en base a variables discretas. Se supone que esta última puede aproximarse a la primera. *d)* El nivel de medición mínimo es nominal. *e)* Las frecuencias esperadas mínimas por casillero deben ser 5, cuando esto no se cumple es necesario aplicar un factor de corrección (corrección de Yates). *f)* La prueba de X^2 es útil solamente para decidir cuándo las variables son independientes o relacionadas. No nos informa acerca de la intensidad de la relación, debido a que el tamaño de la muestra y el tamaño del cuadro ejercen una influencia muy fuerte sobre los valores del *test*. Existen numerosos estadísticos basados en la distribución de X^2 que son útiles para la determinación de la intensidad de la relación (ver coeficiente Fi, Cramer, *C*, etcétera).

Coeficiente Fi (φ)

Es una medida de asociación (fuerza de la relación) para tablas de 2×2. Toma el valor cero cuando no existe relación, y el valor + 1.00 cuando las variables están perfectamente relacionadas.

Coeficiente V. *de Cramer*

Es una versión ajustada del coeficiente φ para tablas de $r \times k$. El nivel de medición es nominal y el coeficiente varía entre 0 y 1.00.

Coeficiente de contingencia (C)

Basado como los dos anteriores en X^2 se pueden utilizar matrices de cualquier tamaño. Tiene un valor mínimo de 0, y sus valores máximos varían según el tamaño de la matriz (por ejemplo para matrices de 2×2 el valor máximo de C es .707; en tablas de 3×3 es .816, etc., la fórmula genérica: $\sqrt{\frac{k-1}{k}}$) consecuentemente para una interpretación del coeficiente obtenido en cualquier tabla de 2×2 habría que dividir ese valor entre .707.

Limitaciones: *a)* El límite superior del coeficiente está en función del número de categorías; *b)* dos o más coeficientes C no son comparables, a no ser que provengan de matrices de igual tamaño.

El coeficiente Q *de Yule*

También como los anteriores para escalas nominales, se utiliza únicamente en tablas de 2×2. Los valores Q son 0 cuando hay independencia entre las variables, siendo sus límites ± 1.00 cuando cualquiera de las 4 celdas en el cuadro contiene 0 frecuencias: por lo general cuál de los distintos coeficientes es preferible en este caso (φ o Q) depende del tipo de investigación y del tipo de distribución marginal.

Coeficiente lambda (λ)

Es un coeficiente de asociación para tablas de $r \times k$, cuando las dos variables están medidas a nivel nominal.

El coeficiente lambda pertenece a la familia de un grupo de coeficientes (τ_b, λ y otros), que se utilizan para hacer interpretaciones probabilísticas en cuadros de contingencia. El tamaño del coeficiente indica la reducción proporcional en errores de estimación en la variable dependiente cuando los valores en la variable independiente son conocidos.

El valor máximo de λ es 1.00 y ocurre cuando las predicciones pueden ser hechas sin ningún error. Un valor cero significa que no hay posibilidad de mejorar la predicción. Un coeficiente lambda .50 significa que podemos reducir el número de errores a la mitad, etcétera.

Coeficiente τ_b *de Goodman y Kruskal*

Sirve a los mismos propósitos que el coeficiente lambda y debe ser preferido cuando los marginales totales no son de la misma magnitud.

Coeficiente de incertidumbre

También para niveles nominales en cuadros de contingencia de $r \times k$. La computación del coeficiente toma en cuenta simetría y asimetría (el coeficiente lambda toma en cuenta, por ejemplo, solamente la asimetría). El coeficiente asimétrico es la proporción de reducción de la incertidumbre conocido por efecto del conocimiento de la variable independiente. La ventaja de este coeficiente sobre lambda es que considera el total de la distribución y no solamente el modo.

El máximo valor del coeficiente de incertidumbre es 1.00 que denota la eliminación de la incertidumbre, y se alcanza cada vez que cada categoría de la variable independiente está asociada a solamente una de las categorías de la variable dependiente. Cuando no es posible lograr ningún avance en términos de disminución de la incertidumbre el valor del coeficiente es 0. Una versión simétrica del coeficiente mide la reducción proporcional en incertidumbre que se gana conociendo la distribución conjunta de casos.

Coeficiente tau b

Mide asociación entre dos variables ordinales en cuadros de contingencia. Este coeficiente es apropiado para cuadros cuadrados (es decir, donde el número de columnas es idéntico al número de filas). Sus valores varían de 0 a ± 1.00. El valor cero indica que no existe asociación entre pares concordantes y discordantes. El valor ± 1.00 se obtiene cuando todos los casos se ubican a lo largo de la diagonal mayor. En tablas de 2×2 el valor de tau *b* es idéntico al de φ con la ventaja de que el coeficiente tau *b* proporciona información sobre la dirección de la relación a través del signo. Los valores negativos indican que los casos se distribuyen sobre la diagonal menor. Los valores intermedios entre 0 y 1 indican casos que se desvían de las diagonales. A mayor desviación mayor proximidad al valor cero (es decir cuando los pares discordantes son iguales a los pares concordantes).

Coeficiente tau c

Sirve a los mismos propósitos que el coeficiente tau *b*, pero este coeficiente es más apropiado para cuadros rectangulares (cuando el número de columnas difiere del número de líneas). La interpretación de ambos coeficiente es similar.

Coeficiente gamma (γ)

Mide asociación entre dos variables ordinales en cuadros de contingencia de $r \times k$. Mientras que el coeficiente tau *c* depende para su cómputo solamente del número de líneas y de columnas, y no las distribuciones marginales, tomando en cuenta los empates, el coeficiente gamma excluye los empates del denominador de la fórmula de cómputo, siendo además un coeficiente con posibilidades de aplicación en datos no agrupados. Además el coeficiente no requiere de cambios en la forma de la matriz. Los valores numéricos de gamma por lo general son más altos que los valores de tau *b* y de tau *c*.

El coeficiente gamma es simplemente el resultado del número de pares concordantes menos el número de pares discordantes, divididos por el número total de pares unidos. Los valores gamma varían entre 0 y ± 1.00, donde el signo indica la dirección de la relación y los valores la intensidad de la misma.

El SPSS provee valores gamma para cuadros de tres a *n* entradas, en el que se calcula el gamma de orden cero y además gammas parciales. El gamma de orden cero mide la relación entre dos variables, siendo exactamente el mismo al que se discute en los párratos anteriores. Cuando la matriz tiene tres o más dimensiones, el SPSS (subprograma *crosstabs*) computa un coeficiente gamma de orden cero (reduciendo la tabla a variable dependiente e independiente) y además medidas de correlación parcial gamma de la relación entre las dos variables, controladas por una o más variables adicionales. El investigador puede analizar así cómo influyen en la relación de sus variables dependiente e independiente, la introducción de variables adicionales (en la sección correspondiente a correlaciones parciales indicaremos con mayor detalle el uso y significado de las correlaciones parciales).

Coeficiente D *de Sommer*

Para variables ordinales en cuadros de contingencia, este coeficiente toma en cuenta los empates, pero el ajuste es realizado de manera distinta a la utilizada en los coeficientes tau *b* y tau *c*.

Coeficiente eta (η)

Se utiliza cuando la variable independiente es nominal y la variable dependiente intervalar. Este coeficiente indica cuán disimilares son las medias aritméticas en la variable dependiente dentro de las categorías establecidas por la variable independiente. Cuando las medias son idénticas el valor del coeficiente es 0. Si las medias son muy diferentes y sus varianzas son pequeñas, los valores de eta se aproximan a 1.00.

Correlación biserial (r_b)

Para utilizar cuando una de las variables está medida a nivel nominal y la

otra a nivel intervalar, la variable a nivel nominal puede ser una dicotomía forzada. Sus valores oscilan entre 0 y ± 1.00.

Correlación punto-biserial (r_{pb})

Similar al coeficiente biserial, se aplica cuando la variable nominal es una dicotomía real. La interpretación de ambos coeficientes es idéntica y su utilización más común se encuentra en la construcción de pruebas, sobre todo para la determinación de validez.

Coeficiente de correlación Spearman (ρ)

Es un coeficiente de correlación por rangos, cuando las dos variables están medidas a nivel ordinal, e indica el grado en que la variación o cambio en los rangos de una de las variables está relacionado a las variaciones o cambios en los rangos en la otra variable. Tanto el coeficiente ρ (rho) de Spearman como el coeficiente τ (tau) de Kendall, son coeficientes no paramétricos, es decir que no se hacen supuestos acerca de la distribución de los casos sobre las variables. Ambos coeficientes suponen la no existencia de muchos empates, por lo cual los sistemas de organización de los datos y de cómputo, son distintos a los de las tabulaciones cruzadas, y por ello se encuentran en subprogramas diferentes (en este caso el subprograma correspondiente en el SPSS es denominado *nonpar corr*).

Para el cómputo del coeficiente correlación ρ de Spearman (así como para el τ de Kendall), no se toman en consideración los valores absolutos en las variables, sino su orden de rango. El coeficiente rho de Spearman se aproxima más que el coeficiente tau de Kandall al coeficiente de correlación producto-momento de Pearson, cuando los datos son aproximadamente continuos. Los valores del coeficiente varían entre −1.00 y +1.00.

Coeficiente de correlación Tau de Kendall (τ)

Similar al coeficiente rho, se utiliza cuando las dos variantes son ordinales. Por lo general debe preferirse cuando existe abundante número de empates entre rangos, caso que se da especialmente cuando el número total de casos es grande y se los clasifica en un número relativamente pequeño de categorías. El subprograma *nonpar corr* contiene factores de corrección para empates tanto para el coeficiente tau como para el coeficiente rho. Los valores de este coeficiente oscilan entre −1.00 y +1.00.

Coeficiente de correlación producto-momento de Pearson (r)

Para dos variables medidas a nivel intervalar por lo menos, éste es un coeficiente de correlación paramétrico que nos indica con la mayor precisión cuándo dos cosas están correlacionadas; es decir, hasta qué punto una variación en una se corresponde con una variación en otra. Sus valores varían de +1.00

que quiere decir correlación positiva perfecta; a través de 0 que quiere decir independencia completa o ausencia de correlación, hasta −1.00 que significa correlación perfecta negativa. El signo indica por lo tanto la dirección de la covariación y la cifra la intensidad de la misma. Una correlación perfecta de +1.00 indica que cuando una variable se "mueve" en una dirección, la otra se mueve en la misma dirección y con la misma intensidad. La interpretación de la magnitud de r depende en buena medida del uso que se quiera dar del coeficiente, el grado de avance teórico en el área, etc. Guilford [1] sugiere como orientación general, la siguiente interpretación descriptiva de los coeficientes de correlación producto-momento;

r menor que .20 — correlación leve, casi insignificante
r de .20 a .40 — baja correlación, definida, pero baja
r de .40 a .70 — correlación moderada, sustancial
r de .70 a .90 — correlación marcada, alta
r de .90 a 1.00 — correlación altísima, muy significativa

De todos modos la interpretación del coeficiente está además condicionada a su grado de significación (ver significación de los estadísticos).

Premisas o suposiciones fundamentales para el cómputo de r: *a)* Ambas variables deben ser medidas a nivel intervalar al menos. *b)* La dirección de la relación debe ser rectilínea. *c)* La distribución tiene que ser homoscedástica (las dispersiones en las columnas y en las líneas del diagrama de dispersión deben ser similares). Esta condición prevalece cuando las dos distribuciones son simétricas entre ellas.

El programa imprime el valor del coeficiente de correlación, la cantidad de casos, y la significación estadística.

Existen varios coeficientes que se derivan del coeficiente de correlación producto-momento, entre otras, por ejemplo: r^2: mide la proporción de la varianza en una variable que es "explicada" por la otra.

Diagrama de dispersión (Scattergram)

El SPSS puede imprimir además, a través de su subprograma *Scattergram*, el diagrama de dispersión para dos variables, computando además la regresión linear simple. El diagrama de dispersión es un tráfico de puntos donde, basado en los valores en las dos variables, una de las variables define el eje horizontal y la otra el eje vertical. Estos diagramas son de mucha utilidad ya que nos dan una imagen de la relación, que puede ser utilizada para la determinación de la homoscedasticidad, por ejemplo y para decidir si vale o no la pena continuar más adelante.

Para la confección de los diagramas, el usuario tiene que tomar algunas decisiones sobre cómo se va a manejar la falta de datos *(missing data)*, qué clase de escala tiene que ser utilizada y cómo se colocarán las líneas segmentadas.

[1] Guilford, J. P.: *Psychometric Methods;* McGraw-Hill, Nueva York, 1954.

Comúnmente dos líneas verticales y dos líneas horizontales segmentadas dividen cada eje con tres secciones, de manera tal que el gráfico consiste en 9 rectángulos iguales. Si el investigador prefiere, las líneas segmentadas pueden ser diagonales que atraviesan el gráfico.

Los datos (es decir, cada punto sobre el diagrama) están representados por asteriscos (*) cuando un caso cae en alguna intersección, de dos a ocho casos el número es impreso. Nueve o más casos están representados por el número 9. Cuando la escala contiene muy pocas categorías, existe la posibilidad de que los puntos sobre el diagrama se den muy amontonados, lo que limita la utilización del diagrama de dispersión recomendándose para esas situaciones una tabulación cruzada.

Los estadísticos que acompañan al diagrama de dispersión, son aquellos asociados a las regresiones lineares simples: correlación producto-momento, error estándar de la estimación, r^2, significación de la correlación, intersección con el eje vertical, e inclinación.

Es necesario discutir con algún detalle el concepto de regresión, ya que sirve de base para la utilización de predicciones, así como de ayuda para la comprensión del concepto de correlaciones parciales y múltiples.

El concepto de regresión trata de describir no solamente el grado de relación entre dos variables, sino *la naturaleza* misma de la relación, de manera tal que podamos predecir una variable conociendo la otra (por ejemplo, el rendimiento académico a partir del resultado en un *test*, el ingreso a partir de la educación, etc.). Aquí no estamos interesados en explicar por qué las variables se relacionan como se relacionan, sino simplemente a partir de la relación dada, predecir una variable a partir del conocimiento de los valores en la otra. Si la variable X es independiente de la variable Y (es decir, si son estadísticamente independientes), no estamos en condiciones de predecir Y a partir de X o viceversa, es decir nuestro conocimiento de X no mejora nuestra predicción de Y. Por razonamiento inverso, cuando las variables son dependientes —están correlacionadas, co-varían—, el conocimiento de X nos puede ayudar a predecir el comportamiento de Y y viceversa.

Esto se logra mediante lo que se llama ecuación de regresión de Y sobre X, que nos da la forma en cómo las medias aritméticas de los valores de Y se distribuyen según valores dados de X.

La operación de regresión contiene los siguientes supuestos: *a)* Que la forma de la ecuación es linear; *b)* Que la distribución de los valores de Y sobre cada valor de X es normal, y *c)* Que las varianzas de las distribuciones de Y, son similares para cada valor de X; *d)* Que el error es igual a 0.

Cumplidas estas condiciones, la ecuación de la regresión es:

$$Y = \alpha + \beta X$$

donde α y β son constantes y se les da una interpretación geométrica.

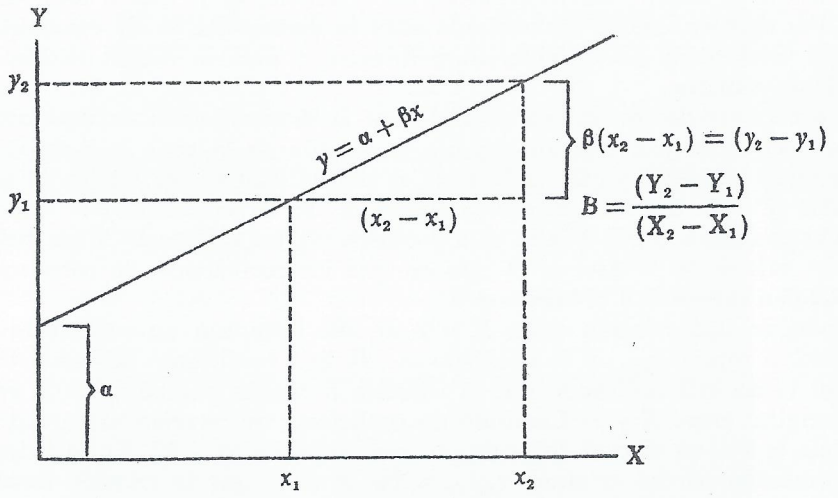
Si X es igual a 0, entonces $Y = \alpha$.

α representa entonces el punto donde la línea de regresión cruza el eje de Y.

La inclinación de la línea de regresión es dada por β, indicando la magnitud en el cambio de Y por cada unidad de cambio en X. Cuando β es igual a 1, y si las unidades de X y Y están indicadas por distancias idénticas a lo largo de sus ejes respectivos, la línea de regresión estará en un ángulo de 45° con respecto al eje de las X. A más grande el tamaño de β, mayor será el declive, es decir más grande el cambio en Y dados determinados valores de cambio en X.

H. Blalock [2] presenta la siguiente figura que aclara la interpretación geométrica del coeficiente de regresión:

FIGURA 1



Es decir que β mide la tangente del ángulo, con lo cual queda identificado el ángulo.

Correlación parcial

Todos los coeficientes de correlación y asociación examinados hasta ahora tomaban en cuenta la relación entre dos variables (con la excepción del coeficiente gamma). La correlación parcial provee medidas del grado de relación entre una variable dependiente Y, y cualquiera de un conjunto de variables independientes, controladas por una o más de esas variables independientes. Es decir, describe la relación entre dos variables, controlando los efectos de una o más variables adicionales.

Es similar a lo que se hace en tabulaciones cruzadas, cuando se introducen variables de control. Sin embargo, ya habíamos visto que para controlar varias variables con varios valores, necesitábamos una muestra demasiado grande, además la inspección del efecto era de tipo literal.

[2] Hubert Blalock: *Social Statistics* (2ª ed.); McGraw-Hill, Kogahusha, Tokio, Japón, 1972.

Con correlaciones parciales, el control no solamente es estadístico, sino además la cantidad de casos no necesita ser muy grande.

$r_{ij \cdot k}$ indica entonces a i y j variable independiente y dependiente (el orden es inmaterial, ya que la correlación entre ij y ji serán idénticas). La variable de control es indicada con k.[3]

Desde la perspectiva de la teoría de la regresión, la correlación parcial entre i y j, controlando por k es la correlación entre los residuales de la regresión de i sobre k y de j sobre k, permitiéndonos establecer predicciones sobre las variables dependientes e independientes a partir del conocimiento del efecto que tiene la variable control sobre ellas.

El coeficiente de correlación parcial puede ser utilizado por el investigador para la comprensión y clarificación de las relaciones entre tres o más variables. Por ejemplo, puede ser utilizado para la determinación de espureidad, para la localización de variables intervinientes, y para la determinación de relaciones causales.

El coeficiente de correlación parcial *para la determinación de espureidad en las relaciones:* una relación espuria es aquella en la cual la correlación entre una variable X y una variable Y, es el resultado de los efectos de otra variable (Z) que es el verdadero predictor de Y. La correlación es espuria cuando, controlando por Z (esto es, a Z constante), los valores de X no varían con los valores de Y. Éste es el caso en que los coeficientes de correlación parcial dan valores 0 o próximos a 0.

Supóngase una relación entre X y Y de .40. Computo un coeficiente de correlación parcial $r_{xy \cdot z}$ y el resultado es .20. Este coeficiente de correlación parcial ya me está indicando que la variable Z explica parcialmente la relación original entre X y Y. Computo un coeficiente de correlación parcial de segundo orden, en el cual controlo por dos variables, Z y W. El coeficiente de correlación parcial es ahora $r_{xy \cdot zw} = .06$, es decir que la relación desaparece, consecuentemente, la relación original era espuria.

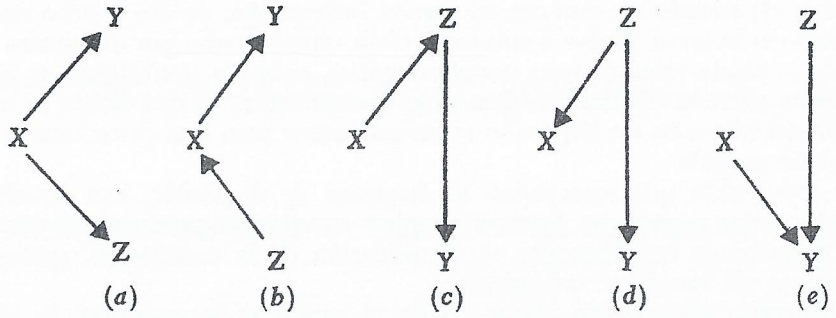
Para la localización de variables intervinientes, así como para la determinación de relaciones causales, el problema es de naturaleza más conceptual, esto es, hay que combinar valores de coeficientes de correlación parcial, con una serie de supuestos sobre las formas de las distribuciones y sobre la intervención de otras variables, además de las que se consideran en el modelo. Los supuestos no pueden ser verificados empíricamente por el análisis estadístico, sino que van a depender del razonamiento teórico.

En cualquiera de los siguientes casos, salvo en (e) la correlación parcial $r_{yz \cdot x}$ debe ser próxima a cero (Y es la variable dependiente, es decir la que va a ocurrir al final en la secuencia temporal); (d) es un caso típico de correlación espuria. La relación entre X y Y se explica en función de las relaciones de X con Z y de Y con Z. (Véase la figura de la página siguiente.)

En el modelo (c) Z actúa como variable interviniente en la relación entre X y Y. La correlación parcial también dará 0. Pero hay que tener mucho

[3] Salvo en el caso de utilizar la correlación parcial para predicciones en la utilización de regresiones en cuyo caso se acostumbra a interpretar $r_{12 \cdot 3}$ denotando 1 la variable dependiente, 2 la variable independiente y 3 la variable de control.

cuidado en no interpretar los modelos (c) y (d) de la misma manera, y la correlación parcial tiene sentido solamente para probar que no hay relación entre X y Y, sino cuando interviene Z. El modelo (b) es similar, aunque ahora X es interviniente.



En el modelo (a) la relación X con Y, y la X con Z son relaciones directas, mientras que no se postula relación entre Y y Z.

En los modelos (a) y (b), la correlación parcial entre X y Z, controlado por Y debe ser 0.

Similarmente, en los modelos (c) y (d), la correlación parcial entre X y Y, controlado por Z, debe ser 0.

Cuando el modelo es (e), la correlación parcial $r_{xy \cdot z}$ dará valores más altos que la correlación entre X y Y. La correlación entre X y Z será 0.

El investigador debe informar a los programadores sobre la lista deseada de correlaciones parciales, en la que se especifiquen las combinaciones de variables (todas las combinaciones posibles o solamente algunas de las combinaciones). Por ejemplo: si se presentan variables: ingreso, educación, actitud frente al cambio y religiosidad, o se especifican las combinaciones deseadas o se deja que se correlacionen todas con todas en n combinaciones.

La palabra *with* especifica en el programa la combinación entre variables cuando la lista incluye solamente algunas combinaciones. Cuando el programa no incluye *with* se calculan todas las combinaciones posibles.

Análisis de regresiones múltiples (subprogram regression)

Este subprograma es considerablemente más complejo que los anteriores, y puede ser utilizado para una variedad bastante grande de análisis de variables múltiples: regresiones polinomiales, regresiones mudas (*dummy*), análisis de la varianza y análisis de la covarianza, predicciones, etcétera.

Por lo general, la regresión múltiple requiere variables medidas a nivel intervalar o racional y que las relaciones sean lineares y aditivas. Sin embargo, hay casos especiales en los cuales regresores mudos, medidos a nivel nominal pueden ser incorporados a la regresión, relaciones no lineares y no aditivas pueden ser manipuladas, etcétera.

Existen algunas diferencias entre análisis de correlaciones múltiples y análisis de regresiones múltiples, que conviene destacar.

Los análisis de correlaciones múltiples se utilizan para: *a)* La evaluación de la medida en que cada variable predictora o subconjunto de variables contribuye a la explicación de los puntajes de un criterio sobre una muestra; o *b)* Para predecir los puntajes de un criterio en una muestra diferente en la cual existe información del mismo grupo de variables predictoras. Aquí no estamos interesados tanto en la relación entre la variable dependiente y cada una de las variables independientes tomadas separadamente, sino en el poder explicativo del conjunto de variables independientes en su totalidad. El coeficiente de correlación múltiple es expresado entonces como $r_{1 \cdot 2345}$ si son 4 las variables predictoras, o $r_{1 \cdot 23}$ si son dos, etcétera.

Los modelos para el análisis de regresiones múltiples, a la vez que son más complejos en términos de cantidad de operaciones o de derivaciones que a través de ellos se puedan realizar, son bastante más simples en términos de los supuestos y condiciones para su utilización.

Por ejemplo, los modelos correlacionales requieren que las variables y los parámetros observables tengan una distribución normal conjunta; los modelos de regresión múltiple requieren solamente que la distribución de las desviaciones de la función de regresión sea normal; no se supone que las variables predictoras provengan de una distribución normal multivariata, o a veces requiere que los datos estén contenidos en códigos binarios.

Nosotros vamos a dar algunos ejemplos de prueba de hipótesis a través de análisis de regresiones múltiples. El análisis de las regresiones múltiples puede ser utilizado ya sea para la descripción de las relaciones entre variables o como instrumento para la inferencia estadística.

Como instrumento descriptivo la regresión múltiple es útil: *a)* Para encontrar la mejor ecuación linear de predicción y para evaluar su eficiencia predictiva; *b)* Para evaluar la contribución de una variable o un conjunto de variables; *c)* Para encontrar relaciones estructurales y proveer explicaciones para relaciones complejas de variables múltiples.

Habíamos visto que el coeficiente de regresión simple se expresaba en la fórmula:

$$Y = \alpha + \beta X$$

Un coeficiente de correlación parcial dijimos era una medida de la cantidad de variación explicada por una variable independiente, después que las otras variables han explicado todo lo que podían. En el coeficiente de correlación múltiple estamos interesados en el poder explicativo de un conjunto de variables independientes sobre la variación dependiente ($r_{1 \cdot 2345}$).

Para ambos casos, la ecuación de la regresión toma ahora la siguiente forma:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Ésta es la ecuación más simple, y parte de los mismos supuestos delineados para la ecuación de regresión simple. En la medida en que nos movemos en espacios multidimensionales, la representación geométrica es imposible.

Los coeficientes β se interpretan de manera distinta que en el caso de regresiones simples, ya que aquí las inclinaciones son varias, y que se obtienen cada una de ellas controlando por cada una de las variables independientes remanentes. Manteniendo X_2 en un valor fijo, β_1 representa la inclinación de la línea de regresión de Y sobre X_1, para el caso en que solamente se estén controlando dos variables. Y así sucesivamente.

Ejemplos

a) *Para encontrar la mejor ecuación linear de predicción y para evaluar la eficiencia predictiva.* Jae-On Kim y Frank Kohout [4] presentan el problema de predecir la tolerancia política, a partir de educación, ocupación e ingreso. A través de técnicas de regresión múltiple, el investigador podría estar interesado en determinar el grado de dependencia linear de la tolerancia política sobre la base de la educación, la ocupación y el ingreso de una persona. Supóngase que la tabla de resultados sea la siguiente:

Correlación múltiple	: .5312	
R^2	: .2822	
Error estándar	: .8604	
Variables independientes	*B*	β *par*
Educación	.1296	.3889
Ocupación	.0089	.1778
Ingreso	.0018	.0556
(Constante A)	2.9889	

La interpretación en este caso podría ser la siguiente: *1)* La cantidad de variación en tolerancia política, explicada por la operación conjunta de educación, ocupación, e ingreso es del 28.22 % de la varianza total. *2)* Si el investigador está interesado en predecir los puntajes que un sujeto va a obtener en tolerancia política a partir de las tres variables independientes, aplicará la ecuación de predicción señalada más arriba.

$$Y = 2.9889 + .1296\ (X_1) + .0089\ (X_2) + .0018\ (X_3)$$

Si el sujeto tiene 10 años de educación formal (X_1), un puntaje de 60 en prestigio ocupacional (X_2) y un ingreso de 100 ($ 10 000) (X_3) entonces

$$Y = 2.9889 + .1296\ (10) + .0089\ (60) + .0018\ (100) = 4.9989$$

El error estándar que figura en la tabla (.8604), predice que los puntajes precedidos en la escala de tolerancia política se van a desviar de los valores parámetros en .8604 unidades.

[4] Jae-On Kim y Frank J. Kohout: "Multiple regression analysis: subprogram regression", en N. Nie, C. H. Hull, J. Jenkins, *et. al.: Statistical Package for the Social Sciences*, 2ª ed.: McGraw-Hill, Nueva York, 1975.

Los valores B en la tabla son coeficientes de regresión parcial, y pueden ser utilizados como medida de la influencia de cada variable independiente sobre la tolerancia política cuando se controlan los efectos de las otras variables.

Obsérvese en el ejemplo que el coeficiente de correlación múltiple (R) es mayor en magnitud que cualquiera de los r, y esto es evidente desde el momento en que es imposible explicar menos variación agregando variables. El máximo valor relativo del coeficiente total ocurre cuando la intercorrelación entre las variables independientes es igual a 0, de manera que si queremos explicar la mayor cantidad de variación en la variable dependiente que sea posible, deberemos buscar por variables independientes que si bien tienen correlaciones moderadas con la variable dependiente, son relativamente independientes unas de las otras.

Relacionado a la intercorrelación entre variables independientes, está el problema de la *multicolinearidad;* esto es, cuando las variables independientes están estrechamente intercorrelacionadas, tanto las correlaciones parciales como la estimación de los β se hacen muy sensitivas a los errores de muestreo y de medición. Cuando la multicolinearidad es extrema (intercorrelaciones del rango de .8 a 1.0) el análisis de regresión no es recomendable.

b) *La regresión múltiple puede ser utilizada también para evaluar la contribución de una variable independiente en particular,* cuando la influencia de otras variables independientes es controlada. Aquí utilizamos coeficientes de regresiones parciales. Hay dos coeficientes designados, la contribución de cada variable a la variación de la variable dependiente: coeficiente de correlación semiparcial *(part-correlation)* y el coeficiente de correlación parcial. El primero se denota como $r_{y(1.2)}$ y el segundo como $r_{(y\ 1).2}$.

El coeficiente semiparcial es la correlación simple entre el Y original y el residual de la variable independiente X^1 a la cual se le extraen los efectos de la variable independiente X^2, es decir que el efecto de X^2 es sacado solamente de la variable X_1, mediante una regresión linear simple de X_2 sobre X_1, entonces ese residual de X_1 es correlacionado con la variable dependiente Y. En el caso de la tolerancia política uno podría estar interesado en determinar de qué manera el ingreso contribuye a la variación de la tolerancia política aparte de lo que es explicado por educación y ocupación. El cuadro siguiente permite calcular los valores del coeficiente semiparcial y el coeficiente parcial:

Regresión con dos variables independientes

A		B	C
Ed. (X_1) y Ocu. (X_2)		Ed. (X_1) e Ing. (X_3)	Ocu. (X_2) e Ing. (X_3)
Regress. mult. (R)	.5292	.5118	.4163
R^2	.2800	.2619	.1733

Regresión con tres variables independientes

Ed. (X_1) y Ocu (X_2), e Ing. (X_3)
Regresión múltiple (R) : .5312
R^2 : .2822

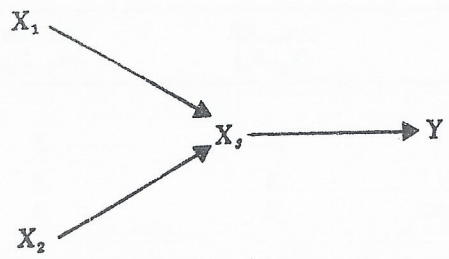
Coeficiente semiparcial. Su cuadrado es igual a la diferencia entre un R^2 que incluye a las tres variables independientes (.2822) y a un R^2 que incluye solamente ocupación y educación (.2800). En nuestro caso entonces $R^2_{y(3.12)}$ es igual a .0022, indicando que ingreso solamente contribuye a un .22 % de incremento en la variación de tolerancia política por educación y ocupación; en otras palabras que el incremento es trivial, y que se puede ignorar ingreso. Para los casos del coeficiente semiparcial para educación y para ocupación los valores respectivos serían .1089 y .0203, es decir que educación explicaría aproximadamente un 11 % de la variación y ocupación un 2 %.

El coeficiente parcial. Es la correlación entre los dos residuales, el residual de Y y el residual de X_1, para los cuales y en ambos se han extraído los efectos de X_2. El cuadrado de una correlación parcial es el incremento proporcional en la variación explicada debido a X_1, expresada como una proporción de la variación que no está explicada por X_2. El coeficiente de correlación parcial indicaría el grado en que una variable da cuenta del remanente de variación del que no dan cuenta las otras variables independientes. En nuestro ejemplo de ingreso la correlación parcial es .0031, es decir, que solamente da cuenta del .31 % de la variable dependiente.

c) *El análisis de regresiones múltiples para la determinación de relaciones estructurales entre variables.* Se trata aquí de una conjunción de la técnica de regresión múltiple con la teoría causal. La teoría causal especificaría un ordenamiento de las variables que refleja una estructura de eslabones causa-efecto, la regresión múltiple determina la magnitud de las influencias directas e indirectas que cada variable tiene sobre las otras variables, de acuerdo al orden causal presumido. El método de *path analysis* es un método para descomponer e interpretar relaciones lineares entre conjuntos de variables, en los que se parte del supuesto que el sistema causal es cerrado, consistente de causas y efectos encadenados. Las relaciones causales *(pathways)* se representan con flechas que conectan la causa al efecto.

Cuando se relacionan tres variables, de las cuales una es dependiente (efecto) existen teóricamente seis maneras a partir de las cuales se puede establecer la relación (ver ejemplos en la sección de correlación parcial), con cuatro variables podemos producir 65 diferentes diagramas, etc. La tarea del investigador es seleccionar de entre los diagramas posibles, aquellos que sean más significativos desde el punto de vista de la teoría sustantiva. Cualquier diagrama, por ejemplo, el de la parte superior de la página siguiente, puede ser representado e interpretado en términos de ecuaciones estructurales: una variable a la que una o más flechas apuntan, es interpretada como una función de solamente aquellas variables desde donde partan las flechas.

Uno de los supuestos principales del *path analysis* es que todas las relacio-



nes son lineares, que las variables son aditivas y que las relaciones son unidireccionales. Cuando se cumplen esas condiciones, la función linear toma la forma:

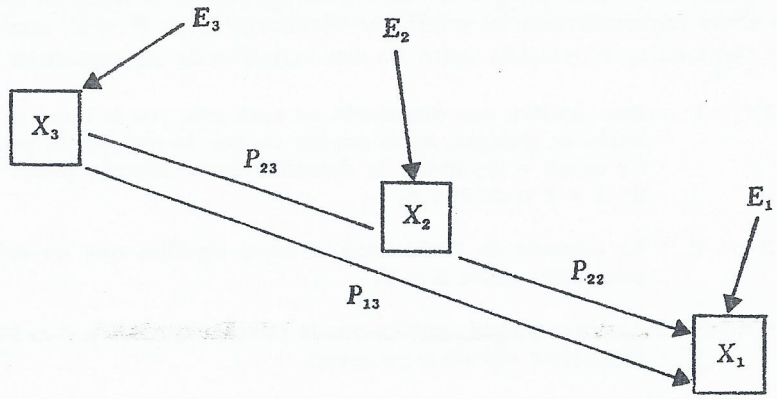
$$X_0 = C_{01}X_1$$

Donde:

X_1 es la variable independiente, o causa,
X_0 es la variable dependiente, o efecto,
C_{01} es una constante que expresa la magnitud del cambio en X_0 para cada unidad de cambio en X_1. Este coeficiente mide el efecto causal linear, o simplemente el coeficiente efecto.

El *paht analysis* no es una técnica para demostrar causalidad. Es un procedimiento para el análisis de las implicaciones de un conjunto de relaciones causales que el investigador impone, a partir de algunos supuestos técnicos, en el sistema de relaciones.

Consideremos ahora un *path analysis* de tres variables, X_3, X_2, X_1. Asumiendo que existe un orden en la relación entre las variables digamos $X_3 \geqslant X_2 \geqslant X_1$, y suponiendo que el sistema sea cerrado podemos representar la relación de la siguiente manera:



o por un sistema de ecuaciones lineares tal como:

$$X_3 = E_3$$
$$X_2 = P_{23}X_3 + E_2$$
$$X_1 = P_{13}X_3 + P_{12}X_2 + E_1$$
$$\text{cov}\ (E_3, E_2) = \text{cov}\ (E_3, E_1) = \text{cov}\ (E_2, E_1) = 0$$

Cada E_i representa todos los efectos residuales en las causas de cada X_i y se denominan errores independientes o perturbaciones independientes, o variables latentes. Cada una de estas variables latentes se estiman a partir de cada R^2 por medio de la fórmula $\sqrt{1 - R^2}$, donde el coeficiente de correlación múltiple R, es la parte de la ecuación de la regresión en la cual X_i es la variable dependiente y todas las variables que la causan son usadas como predictores.

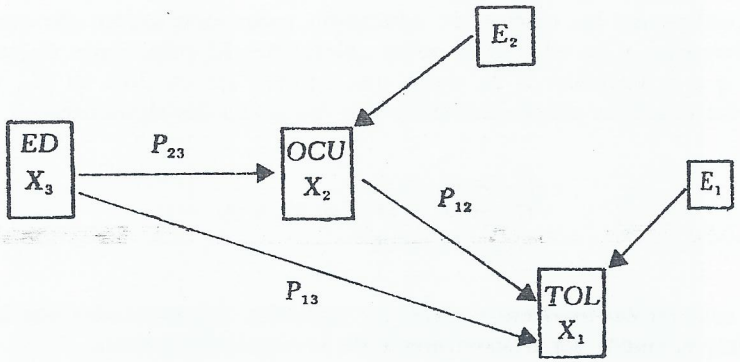
Cada P_{ij} indica un *path* y puede ser estimado a partir de las regresiones de los X_i sobre los X_j. En el *path* que aparece más arriba P_{23} es estimado a partir de la regresión de X_2 sobre X_3, donde $X_2 = B_{23}X_3$.

Y donde P_{13} y P_{12} pueden ser estimados de las regresiones de X_1 sobre X_2 y X_3: $X_1 = B_{13}X_3 + B_{12}X_2$.

Por lo general, dadas n variables en orden $X_n \leqslant \ldots \leqslant X_3, \leqslant X_2, \leqslant X_1$, la estimación de todos los *path* coeficientes requerirá $n - 1$ soluciones de regresión, en las que se toma cada una de las $n - 1$ variables de orden menor en el diagrama como independientes en sucesión y todas las variables de orden mayor como sus predictores.

Sigamos el mismo ejemplo de Kim y Kohout (*op. cit.*): tenemos 3 variables: tolerancia (X_1), *status* ocupacional (X_2) y educación (X_3); si podemos sostener que el grado de tolerancia, *probablemente* va a estar afectado por el nivel educacional y por el nivel del *status* ocupacional, y que el *status* ocupacional del individuo *probablemente* va a estar influido por su nivel educacional, entonces podemos postular un ordenamiento causal débil del tipo $X_2 \geqslant X_2 \geqslant X_1$. Estamos afirmando un juicio de probabilidad en el que no sabemos cómo una variable afecta a la otra. Para continuar con el *path analysis* necesitamos otro supuesto, que el sistema causal es cerrado, que es más difícil de justificar, pero supongamos que esté justificado.

Tenemos entonces un diagrama de la siguiente forma:



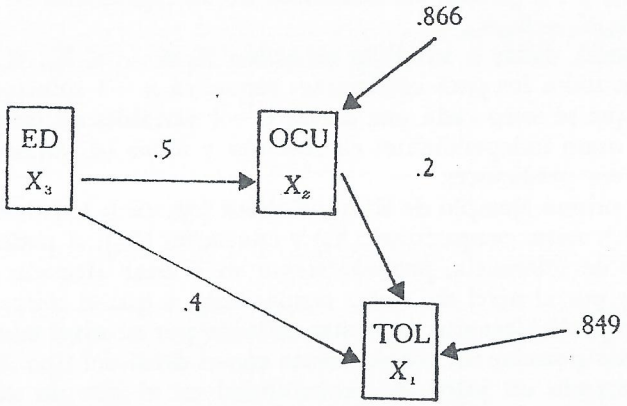
Calculamos los coeficientes de regresión simple para cada uno de los P_{ij} obteniendo los siguientes valores:

$$P_{23} = .5$$
$$P_{12} = .2$$
$$P_{13} = .4$$

Calculamos también los valores E_i que resultan ser:

$$E_2 = .866$$
$$E_1 = .8485$$

El diagrama o *path analysis* tiene entonces la siguiente forma ahora:



¿Cómo interpretamos este *path*? *a)* Primero examinamos cada subsistema, a través de las variables latentes. Y vemos que 75 % de la variación en ocupación y 72 % de la variación en tolerancia, permanece sin explicar por las relaciones causales explicitadas en el modelo.

b) Identificamos los efectos de educación sobre ocupación; de ocupación sobre tolerancia; y de educación sobre tolerancia. El coeficiente C_{ij} mide los cambios que acompañan a X_i dada una unidad de cambio en X_j, estando controladas todas las causas extrañas. Los datos son los siguientes:

$$C_{23} = P_{23} = .5$$
$$C_{13} = (P_{23})\ (P_{12}) + P_{13} = .5$$
$$C_{12} = P_{12} = .2$$

c) La covariación total entre pares de variables, representadas por la correlación simple, puede ser descompuesta de la siguiente manera:

	Ocu., Ed. (X_2, X_3)	*Tol., Ed.* (X_1, X_3)	*Tol., Ocu.* (X_1, X_2)
I) Covariación Original (r_{ij})	.5	.5	.4
II) b_1: Causal-directa	.5	.4	.2
b_2: Causal-indirecta	0	.1	0
Total causal $(b_1) + (b_2) = C_{ij}$	.5	.5	.2
III) No causal $(A) - (B) = r_{ij} - C_{ij}$	0	0	0

Para la relación entre ocupación y educación, el *path analysis* confirma los supuestos, todas las covariaciones entre los dos son tomadas como causales o genuinas.

La covariación entre educación y tolerancia es también tomada como causal, pero la covariación se descompone entre lo que es mediatizado por ocupación y entre lo que no lo es. Aquí parte de la relación entre educación y tolerancia está mediatizada por una variable interviniente.

La relación entre tolerancia y ocupación, esto es la última columna, está descompuesta en componentes causales y componentes espurios.

Casos especiales en el path analysis

Hasta ahora consideramos modelos generales de *path analysis* en los cuales todas las relaciones bivariatas eran asumidas como teniendo una relación causal y el sistema como un todo era cerrado. Es posible introducir en el *path analysis* una cantidad de supuestos diferentes. Sin embargo, siempre hay que recordar que cada vez que incorporamos supuestos ambiguos, producimos como resultado un modelo que da lugar a interpretaciones también ambiguas. Hasta ahora representamos las relaciones bivariatas como $X \rightarrow Y$, también podemos representar la relación entre las dos variables de las siguientes formas:

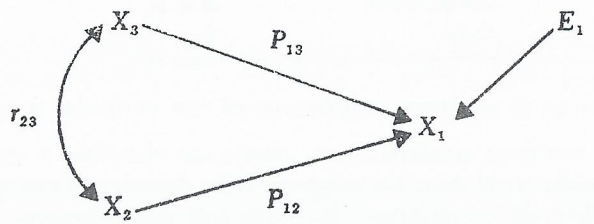
X⌒Y : Esto significa una correlación no analizada, por lo tanto la relación es ambigua, en el sentido en que la covariación puede ser causal o espuria, y la dirección de la relación puede ser de *X* a *Y* o de *Y* a *X*.

X Y : La ausencia de flecha recta o curva significa que no existe covariación entre *X* y *Y*.

X⌒Y : La curva simple significa que la relación entre *X* y *Y* es completamente espuria o no causal.

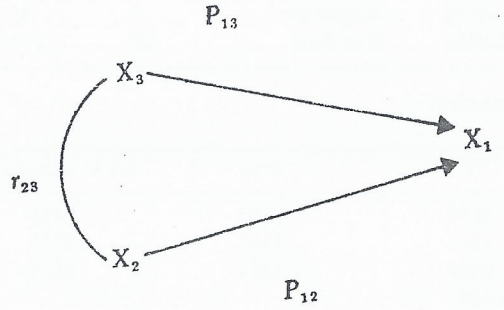
X⇒Y : Representa una relación que es parcialmente causal y parcialmente espuria.

La relación representada hasta ahora en los diagramas anteriores, era del tipo $X \rightarrow Y$, esto es, asumíamos un orden causal entre las variables. Existen situaciones en las cuales no conocemos la verdadera naturaleza de la relación causal, entre algunas de nuestras variables aunque sí conocemos que existe correlación entre ellas. En este caso, el gráfico tendría la siguiente forma:



Es decir, postulamos relación causal entre X_3 y X_1 y entre X_2 y X_1, pero las variables independientes no están conectadas entre sí por una conexión causal, sino simplemente por su correlación. La estimación de P_{13} y de P_{12} se obtiene a partir de las regresiones en las que X_1 es la variable dependiente y X_2 y X_3 variables independientes. Las relaciones entre X_2 y X_3 se expresan por un coeficiente de correlación simple. Nótese que el cambio total en la variable dependiente no está definido en el modelo, lo que dificulta la predicción.

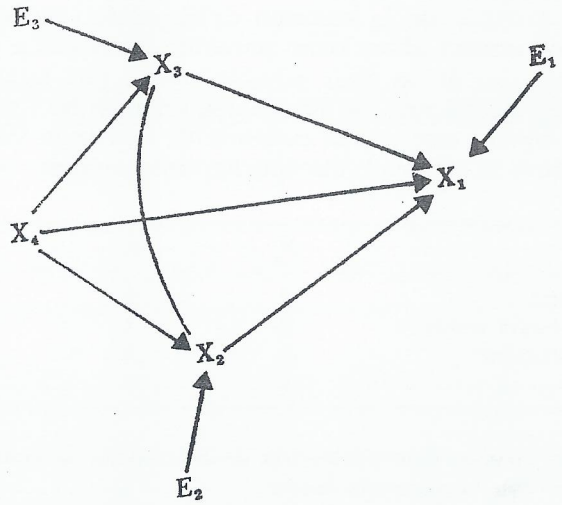
Cuando existen suficientes elementos en la teoría que efectivamente permiten asegurar que la covariación entre las variables exógenas no es de naturaleza causal, el modelo puede ser representado de la siguiente forma:



Ahora sí todas las relaciones entre variables pueden interpretarse de manera causal, simplemente porque partimos del supuesto menos ambiguo que en el del diagrama anterior. Aquí en vez de plantear desconocimiento sobre la naturaleza de la covariación, planteamos que la covariación entre X_3 y X_2 es de naturaleza no causal, es decir que X_3 no causa variación en X_2 y viceversa. De esta manera es posible hacer predicciones en relación a los cambios que una unidad en X_3 o en X_2 producirán en X_1.

Kim y Kohout presentan en otro ejemplo con cuatro variables y un conjunto de supuestos fuertes, lo que da lugar a interpretaciones menos ambiguas: se trata de un esquema en el que se relaciona sexo (X_4) y accidentes de tráfico (X_1), controlando por cantidad de kilómetros conducidos al año (X_3) y frecuencia de conducción en horas de mucho tráfico (X_2).

Los supuestos causales son que el sexo puede afectar tanto la cantidad de kilometraje recorrido como las condiciones en las que se maneja, las cuales a su vez van a determinar las tasas individuales diferenciales de accidentes de tráfico. El investigador no tiene ningún supuesto teórico que le permita relacionar en forma causal el kilometraje recorrido con las condiciones de manejo. Asimismo, ni el total de kilometraje recorrido, ni las condiciones de manejo, se considera sean totalmente explicadas por sexo, sino que a su vez existen una serie de factores que pueden causar ambos. El modelo del *path* adquiere entonces la siguiente forma:



Donde las estimaciones de P_{34} y de P_{24} pueden realizarse a partir de coeficientes de correlación simple, y los P_{13}, P_{12}, y P_{14} por las regresiones de X_1, sobre X_2, X_3 y X_4.

El coeficiente de covariación residual entre X_2 y X_3 se obtiene a partir de: $r_{23} - (P_{34})\ (P_{24})$. Los coeficientes para los E_i se obtienen respectivamente de la siguiente forma:

$$E_1 = \sqrt{1 - R^2_{1.234}}$$

$$E_2 = \sqrt{1 - R_{2.34}}$$

$$E_3 = \sqrt{1 - R^2_{3.24}}$$

Regresiones con variables mudas (dummy variables)

Éste es un caso especial de regresión, en el cual introducimos mediciones a nivel nominal en la ecuación de la regresión. Estas variables mudas se obtienen tratando a cada categoría de la variable nominal como si fuera una variable por separado, asignando puntajes arbitrarios según la presencia o ausencia del atributo en cuestión. Por ejemplo, si en afiliación política tenemos 3 partidos políticos: radical, demócrata cristiano y conservador, cada uno de esos partidos o categoría representa 1 de 3 variables dicotómicas, entonces los puntajes 1 a 0 pueden ser asignados a cada "variable". Si un sujeto tiene afiliación radical, entonces su puntaje en radical será 1, su puntaje en demócrata cristiano será 0, y su puntaje en conservador será 0. Los valores 0 y 1 son tratados como variables intervalares e incluidas así en la ecuación de la regresión. Sin embargo, por un problema de álgebra [5] una de las variables mudas debe ser excluida de la ecuación de regresión. De hecho, la variable muda excluida actuará ahora como punto de referencia a partir del cual los valores en cada una de las otras variables mudas será interpretado. Cada categoría ahora es representada por una combinación de las i variables mudas. Supongamos en nuestro caso que la categoría de referencia sea otro partido; tendríamos entonces la siguiente distribución de puntajes:

	X_1	X_2	X_3
Radical	1	0	0
Demócrata cristiano	0	1	0
Conservador	0	0	1
Otro	0	0	0

Si otro ha sido elegido como categoría de referencia, la ecuación de la regresión puede ser escrita entonces como:

$$Y = A + B_1X_1 + B_2X_2 + B_3X_3$$

donde los casos de la categoría "otros" pueden ser predecidos mediante:

$$Y = A$$

los radicales por:

$$Y = A + B_1X_1$$

y en la medida que el valor de radicales X_1 es 1, entonces:

$$Y = A + B_1$$

[5] La inclusión de todas las variables mudas resultantes de categorías nominales hace que las ecuaciones normales no puedan ser resueltas, ya que la inclusión de las últimas categorías está completamente determinada por los valores de las primeras categorías ya incluidas en la ecuación.

Los valores esperados para cada una de nuestras categorías serán entonces:

	Y
Radical	$A + B_1$
Demócrata cristiano	$A + B_2$
Conservador	$A + B_3$
Otro	A

Análisis de la varianza unidireccional con variables mudas

El análisis de la varianza unidireccional puede ser obtenido a través de diferentes subprogramas en el SPSS: los subprogramas *anova, oneway* y *breakdown* (ver sección análisis de la varianza). En estos tres subprogramas las variables entran como variables nominales, no introduciéndose la creación de variables mudas.

Sin embargo, el investigador puede desear un análisis de la varianza unidireccional con el subprograma *regression*. Para ello debe crear un conjunto de variables mudas según el sistema explicado más arriba e instruir al programador para que incluya las instrucciones pertinentes para la creación de variables mudas en el SPSS.

El *out-put* del subprograma *regression* en su porción referente al análisis de la varianza unidireccional, tiene la siguiente forma, en la cual introducimos cálculos ficticios para las tres variables usadas en nuestro ejemplo con la variable dependiente, actitud frente a la nacionalización del petróleo:

R múltiple	.5844	Análisis de var.	*D. F.*	*Suma cuadr.*	*F*
R^2	.3416	Regresión	3	56.3529	16.5993
Error estándar	1.0638	Residual	96	108.6371	

Variable en la ecuación

Variable	*B*	*Beta*	*Error estándar B*	*F*
D. cristiano	1.3156	.4435	.4135	10.121
Radical	—.3961	—.1497	.3795	1.089
Conservador	—.9444	—.1441	.6393	2.183
(Constante)	2.444			

El valor *F* de 16.5993 tiene una probabilidad mayor que .001, es decir, que las diferencias son muy significativas para el conjunto de partidos. El R^2 es equivalente al coeficiente de correlación múltiple que se derivan del coeficiente de correlación eta (ver correlación simple), y su valor indica que el 34 % de la actitud frente a la nacionalización del petróleo depende o se explica por la afiliación política.

Los promedios para cada categoría pueden ser obtenidos a partir de la columna *B* de "variables en la ecuación" que es el *out-put* de la regresión:

— Actitud frente a la nacionalización del petróleo: $Y = 2.444$
— Radical: $Y = 2.444 + 1.3156 = 3.76$
— D. cristiano: $Y = 2.444 + (-.3961) = 2.05$
— Conservador: $Y = 2.444 + (-.9444) = 1.50$

Regresiones con variables mudas para dos o más variables categorizadas

Las ecuaciones de predicción para dos variables nominales (representadas por dos conjuntos de variables mudas) es la siguiente:

$$Y = A + B_1X_1 + B_2X_2 + B_3X_3 + B_4E_1$$

Donde los X_1 representan una variable nominal con 4 categorías y E_1 una categoría de una variable nominal dicotómica.

El valor predictivo para cada celda de la matriz estará dado por el siguiente cuadro, siguiendo nuestros ejemplos anteriores:

	Varón	*Mujer*
Radical	$A + B_1 + B_4$	$A + B_1$
D. cristiano	$A + B_2 + B_4$	$A + B_2$
Conservador	$A + B_3 + B_4$	$A + B_3$
Otro	$A + B_4$	A

Lo que ocurre ahora es que las categorías mujer y otro actúan como categorías de referencia.

Análisis de la varianza multidireccional con variables mudas

La regresión múltiple con n variables mudas puede ser utilizada para computar análisis de la varianza. Cuando se desea computar análisis de la varianza con las variables nominales sin recurrir a variables mudas se recomienda el subprograma *anova* (ver más adelante en "análisis de la varianza").

Cuando se utilizan variables mudas y queremos realizar análisis de la varianza con el subprograma *regression* es necesario agregar para el caso de dos factores (afiliación política y sexo) los efectos de interacción, es decir necesitamos crear tres nuevas variables mudas en nuestro ejemplo: (X_1E_1), (X_2E_1), (X_3E_1), donde la ecuación de la regresión múltiple tendrá ahora la siguiente forma:

$$Y = A + B_1X_1 + B_2X_2 + B_3X_3 + B_4E_1 + B_5(X_1E_1) + B_6(X_2E_1) + B_7(X_3E_1)$$

Esta regresión representa el *modelo saturado* donde todos los términos de interacción posible están incluidos.

Los valores predictivos para el modelo saturado se obtienen del siguiente cuadro:

	Varón	*Mujer*
Radical	$A + B_1 + B_4 + B_5$	$A + B_1$
D. cristiano	$A + B_2 + B_4 + B_6$	$A + B_2$
Conservador	$A + B_3 + B_4 + B_7$	$A + B_3$
Otro	$A + B_4$	A

Para dos variables nominales A y B, la estrategia del análisis de la varianza sigue lo que se llama modelo clásico de análisis de la varianza en el cual ni los factores A y B (afiliación política y sexo) son ortogonales; esto es, si las frecuencias en las celdas son proporcionales a las frecuencias marginales de afiliación política y sexo la suma de $2a$ y de $2b$ será simplemente la suma de los cuadrados debidos a cada factor y será igual a la suma de los cuadrados debidos a efectos aditivos. Si A y B no son ortogonales, los efectos de A se confundirán con los efectos de B, y la suma de $2(a)$ y de $2(b)$ no será igual a la suma de los efectos aditivos. El siguiente cuadro ilustra el modelo clásico:

Fuente de variación	*Suma de cuadrados*	*Df*	*F*
1) Suma de cuadrados debido a A y B modelo saturado	$SS_y(R^2_{A,B,AB})$	$K = k_1 + k_2 + k_1k_2$	$\frac{(1)/K}{(4)/(N-K-1)}$
2) Suma de cuadrados debidos a A y B modelo aditivo	$SS_y(R^2_{A,B})$	$k_1 + k_2$	$\frac{(2)/(k1+k2)}{(4)/(N-K-1}$
a) Suma de los cuadrados debidos a B ajustados por B	$SS_y(R^2_{A,B} - R^2_B)$	k_1	$\frac{(2a)/k1}{(4)/(N-K-1)}$
b) Suma de los cuadrados debidos a B ajustados por A	$SS_y(R^2_{A,B} - R^2_A)$	k_2	$\frac{(2b)/k2}{(4)/(N-K-1)}$
3) Suma de los cuadrados debida a la interacción	$SS_y(R^2_{A,B,AB} - R^2_{A,B})$	k_1k_2	$\frac{(3)/k1k2}{(4)/(N-K-1)}$
4) Suma de los cuadrados residuales	$SS_y(1 - R^2_{A,B,AB})$	$N-K-1$	

Los significados de este cuadro serán analizados en la sección siguiente que corresponde a análisis de la varianza.