COEFICIENTE DE CORRELACIÓN PARCIAL
(Guía de clase)

FICHA Nº 21

A. El análisis estadístico que utiliza coeficientes de correlación parcial representa un primer paso en la dirección del análisis multivariado.

✎ Esta técnica, al igual que el coeficiente de Pearson, no realiza ningún supuesto sobre cuál es la variable independiente, cuál es la de control y cuál es la variable dependiente.

- Esta distinción es una decisión metodológica del investigador que dependerá de las hipótesis y teorías que se están manejando.
- Por esta razón, se trata de coeficientes SIMÉTRICOS.
- Para los efectos de esta ficha de actividad, se adoptará la convención de anotar con la letra Y a la variable que la teoría identifica como dependiente, en tanto que las restantes variables independientes .

✎ El objetivo de este coeficiente es calcular la magnitud y sentido de la asociación *entre dos* variables (Y, X_1), controlando por el efecto que otra u otras variables (X_2,X_3, X_4.... X_K) tienen sobre dicha relación.

- Las variables que se introducen para controlar la relación originaria se denominarán de control.
- En consecuencia, este análisis parte de una correlación de Pearson entre las dos variables (Y, X_1), y requerirá conocer las otras correlaciones entre las variables de control
- Si se analiza únicamente el efecto de una tercera variable (X_2) para controlar

la magnitud y sentido de la relación original, el coeficiente de correlación parcial se denominará de orden 1 y se lo expresará agregando al subíndice en la notación de correlación de Pearson, un punto y la variable de control:

$$r_{(yx_1.x_2)}$$

- Si se incorporan dos variables de control (X_2 , X_3), entonces el coeficiente calculado se denominará coeficiente de correlación parcial de orden 2 y se presentará:

$$r_{(yx_1.x_2x_3)}$$

✎ El nivel de medición de las variables requerido para el cálculo de la correlación parcial es el mismo que para las correlaciones simples, esto es:

- intervales
- de razón
- dicotómicas
- Excepcionalmente y en el contexto de una flexibilización de la estadística en clave pragmática, se pueden utilizar variables ordinales, siempre que estas tengan un respetable número de categorías, nunca menor que 5. Esta "liberalidad" se debe a que si se aplican coeficientes de Pearson y la Rho de Spearman (de rangos ordenados) a unas mismas variables, si estas tienen muchas categorías, el coeficiente será muy similar.

✎ El análisis de correlaciones parciales tiene como requisito previo el cálculo de una matriz de correlaciones simples (o de Pearson) entre las

variables implicadas en las hipótesis que se están trabajando.

- Las correlaciones simples son también denominadas correlaciones de orden cero, porque en el cálculo de la relación no se han introducido variables de control.

- El resultado de un análisis de correlaciones parciales generalmente desemboca en la construcción de una matriz de correlaciones parciales de orden 1.

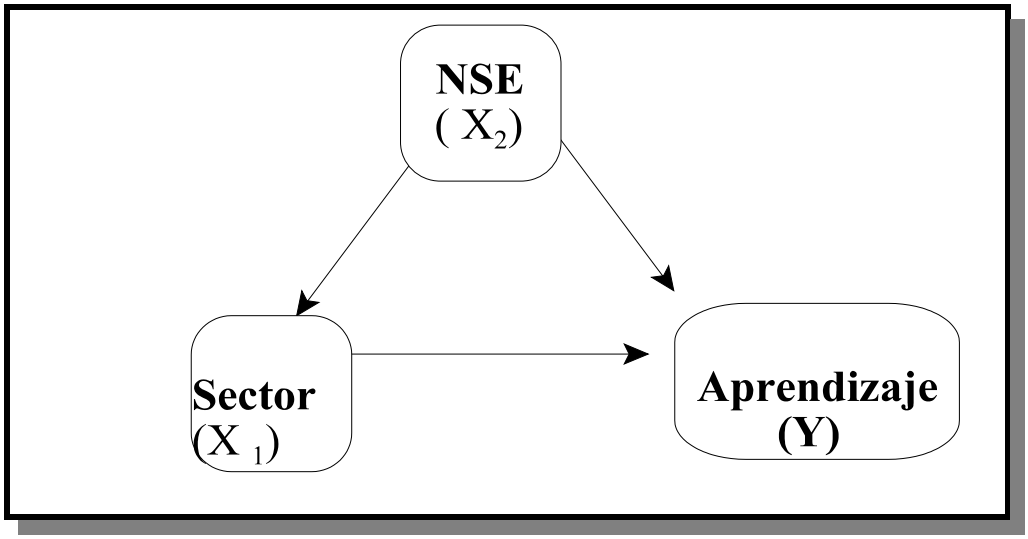
- La matriz de orden 1 es requerida para realizar análisis de correlaciones parciales de orden 2 y así sucesivamente.

B. Los problemas que se aplican son típicos del análisis multivariado Tienen que ver particularmente con la identificación de relaciones espúreas y más en general con sesgos en las magnitudes de las relaciones. En todos los casos por detrás hay una teoría no reduccionista de los fenómenos, o lo que es lo mismo, una teoría unicausal.

✎ Una pregunta propia podría ser como la siguiente: ¿es válida la relación observada entre dos variables o se debe a la relación entre estas dos variables con una tercera? Veremos estos dos problemas como ejemplos.

- En la investigación educativa existe una extensa discusión sobre los efectos que tendría el hecho de que la escuela esté en el sector privado sobre el nivel de aprendizajes de sus alumnos. La hipótesis inicial es que el promedio de los alumnos (en matemática, lengua o ciencias, según se trate) será superior para aquellos que están en las escuelas privadas que en las escuelas públicas. Dicha hipótesis se ha contrastado favorablemente en todos los países (tanto desarrollados como en vías de desarrollo) y se muestra que por

lo general la magnitud de dicho efecto es importante además de estadísticamente significativa. Sin embargo, también se ha encontrado que al controlar la relación original con un indicador de clase social (por ejemplo el nivel socioeconómico del alumnado), la magnitud de la asociación tiende a reducirse o incluso a desaparecer. (Véase la figura 1). Esto se debería a que las escuelas privadas matriculan predominanente un alumnado originado en la clases media-alta y alta, en tanto que las escuelas públicas reciben al resto de la población. Si tal como se sabe desde la teorías de Bourdieu y de Bernstein, la socialización familiar (lenguaje y habitus) incide selectivamente en los aprendizajes, es razonable sostener la hipótesis de que el efecto de las escuelas privadas es en realidad un efecto "espúreo".



- Otro ejemplo proviene de la discusión existente dentro de las políticas educativas sobre los determinantes del efectivo cumplimento de la escolarización universal en los niveles que la sociedad ha definido como obligatoria. Por lo general, esto se informa a través de la tasa escolaridad neta para la cohorte etaria comprendida dentro de las normas constitucionales y legales. La investigación ha mostrado reiteradamente que existen

fenómenos de exclusión educativa que se encuentran presentes simultáneamente en las jurisdicciones y territorios pobres. En las sociedades locales pobres, una mayor proporción de los niños son enviados más tardíamente a escuela y son retirados más tempranamente una vez que los padres consideran que se han cumplido las metas mínimas del aprendizaje de la lectura y la escritura; este retiro está relacionado con la estructura de un mercado laboral donde predominan las tareas agrícolas y el sector informal. Sin embargo, para los países multi-culturales y pluri-étnicos por historia o por inmigración, el factor étnico-lingüístico ha sido identificado como un poderoso determinante ya que la escuela oficial suele ser escuela que enseña una cultura y una lengua y que se impone a las restantes. Para quienes sostienen que la causa profunda de la exclusión educativa se encuentra en la exclusión étnica, la relación entre pobreza y asistencia escolar debería hacerse insignificante y hasta desaparecer cuando se controla por aquel factor.

✎ Alternativamente, podría plantearse un enfoque distinto para las correlaciones parciales y consiste en preguntarse cuál es la magnitud propia y específica en la correlación de dos variables que existe cuando un conjunto de otras variables es controlada simultáneamente.

- Aquí no se hipotetiza que la relación original es espúrea, sino que por el contrario el enfoque teórico es que un conjunto de variables X incide sobre un fenómeno Y. Se sospecha además, que ese conjunto de variables X están intercorrelacionadas entre sí con mayor y menor magnitud. En consecuencia, cada relación X_k Y debe ser examinada tomando en cuenta aquellas intercorrelaciones entre las variables independientes.

- Esta forma de plantearse conceptualmente el problema de análisis se conecta con otros dos temas:

  I. Una interrogante propia del análisis multivariado es cuestionarse por cuánto explican simultáneamente un conjunto de indicadores derivados de una o varias teorías. Así como existe un coeficiente de correlación simple, la investigación demanda un **COEFICIENTE DE CORRELACIÓN MÚLTIPLE.**

II. Un segundo tipo de interrogante consiste en construir una medida del efecto en el fenómeno Y que tiene un cambio en una variable independiente X, cuando se controla el efecto y las intercorrelaciones de las demás variables independientes (X) analizadas. Esta inquietud está respondida a través del

✎ Es importante notar que las hipótesis sobre variables de control no tienen por qué detenerse en una única variable ni tampoco mantener constante la relación original de interés.

- Al primer problema podría agregarse la consideración *simultánea* del ámbito geográfico, en tanto que al segundo podría agregarse un control por analfabetismo. Esto daría lugar a construir *para la misma relación originaria,* coeficientes de correlación parcial de orden 2.

- Podría también ser de interés calcular la correlación entre clase social y aprendizajes controlando por sector institucional de la escuela; o la correlación parcial entre tasa de asistencia y proporción de pobladores que hablan lengua indígena en el municipio controlado por la incidencia de la pobreza.

C. Para responder a las interrogantes y problemas anteriores, la estadística desarrolló un coeficiente de correlación parcial que se calcula para cada par de variables (XY) que definen una relación originaria de interés. La expresión matemática requiere del uso de los coeficientes de correlación simple y de correlación parcial de orden inferior tal como se adelantara más arriba al final del literal A.

✎ Si el objetivo del análisis consiste en computar una matriz de correlaciones parciales de orden 1, para cada relación deberá aplicarse

la siguiente fórmula:

$$r_{YX_1.X_2} = \frac{r_{YX1} - r_{YX2} * r_{X1X2}}{\sqrt{(1 - r^2_{YX2})(1 - r^2_{X1X2})}}$$

- La NOTACIÓN en la fórmula anterior es distinta a la que se encuentra en los libros (por ejemplo, Spiegel 2003: 351). Aquí se ha definido notado con Y a la variable dependiente que representa el fenómeno de interés y la letra X se ha reservado para las variables independientes en el análisis.

- En el numerador de la ecuación anterior aparecen dos términos. En el primero, se indica sencillamente la correlación simple entre las dos variables de interés. El segundo término se resta al primero y consiste en el producto de las otras dos correlaciones lógicamente posibles. Es decir, entre Y y X_2 y entre X_1 y X_2 .

  I. Sustantivamente, esto significa que a la magnitud de la correlación original habrá de descontarse *algebraicamente* la intercorrelación entre las dos variables explicativas introducidas y la correlación que tiene la variable de control sobre Y . En el primer caso, la intercorrelación indicará qué "tan separables" o "independientes entre sí" son los efectos de las dos variables explicativas introducidas. En el segundo caso, indicará la relevancia que la variable de control tiene en el fenómeno estudiado.

  II. Matemáticamente se utiliza el producto y no la suma porque para los efectos del análisis interesará conocer no meramente la intercorrelación o los efectos de la variable de control, sino la combinación de ambos. Podría ocurrir que la variable de control tuviera una fuerte correlación con Y pero que fuera independiente de X_1. Podría ocurrir al contrario, que las dos variables explicativas estuvieran

asociadas pero que la variable X_2 solo tuviera una débil relación con Y. En ambos casos, no representarían cuestionamientos fuertes a la validez de la relación original observada.

- En el denominador se ha introducido una normalización con el fin de encerrrar el resultado del coeficiente entre los valores -1 y + 1.

  I. Es de observarse que existe un producto entre dos términos separados por paréntesis. Estos son análogamente iguales. Se resta la unidad del cuadrado de la correlación simple entre Y y X_2, por un lado, y por el otro lado, se resta la unidad del cuadrado de la correlación simple entre X_1 y X_2 . Es decir, en el denominador se vuelven a incluir las dos medidas de asociación incluidas en el segundo término del numerador.

  II. Entender el significado del cuadrado de una correlación simple conllevaría a introducir nociones nuevas y más complejas, tales como la de "varianza no explicada" y coeficiente de determinación. Ambos serán de importancia en el análisis de regresión.

✎ De la ecuación anterior relativa al coeficiente de correlación de orden 1 se desprenden algunos resultados de interés:

- El coeficiente de correlación parcial de orden 1 siempre será *menor, igual o mayor* al coeficiente de correlación simple de la relación originaria. Formalmente:

$$r_{YX1.X2} \leq r_{YX}$$

- La correlación parcial será igual a la correlación simple siempre y cuando las dos variables explicativas sean independientes:

$$\text{Si } r_{x2x3} = 0 \rightarrow r_{YX1.X2} = r_{YX}$$

- La correlación parcial también será igual a cero si la variable de control

introducida no es estadísticamente relevante para explicar Y.

$$\text{Si } r_{Yx3} = 0 \rightarrow r_{YX1.X2} = r_{YX}$$

✎ De la ecuación de la correlación parcial, también puede hacerse notar alguna similitud con la lógica del análisis que Lazarsfeld propuso para el análisis de asociación entre tres variables categóricas.

- La ecuación de covarianzas de Lazarsfeld expresaba que la relación original se descomponía en dos parciales más el producto de los marginales:

$$(X,Y) = (X,Y;\, T) \oplus (X,Y;T') \oplus [(T,X) \otimes (T,Y)]$$

- La ecuación de la correlación parcial puede ser despejada en forma análoga si sólo se toma el NUMERADOR y se reordenan finalmente los términos y miembros de la igualdad:

$$r_{Yx1.x2} = r_{yx1} - r_{yx2} * r_{x1x2}$$

$$r_{Yx1.x2} + r_{yx2} * r_{x1x2} = r_{yx1}$$

$$r_{yx1} = r_{Yx1.x2} + r_{yx2} * r_{x1x2}$$

- La diferencia en este tipo de análisis es que sólo habrá un resultado parcial "promedio" en lugar de la apertura de los parciales. El segundo término sigue siendo un producto de los marginales, al igual que lo era en Lazarsfeld.

✎ Ahora bien, si el objetivo es avanzar hacia correlaciones de orden 2, la ecuación anterior se puede extender de la siguiente forma:

COEFICIENTE DE CORRELACIÓN PARCIAL
(Guía de clase)

$$r_{YX1.X2X3} = \frac{r_{YX1.X3} - r_{YX2.X3} * r_{X1X2.X3}}{\sqrt{(1 - r^2_{YX2.X3})(1 - r^2_{X1X2.X3})}}$$

- La anterior correlación parte de la base de preguntarse si la correlación parcial obtenida de orden 1 es modificada por la introducción de una segunda variable de control. En el numerador se reitera la misma lógica: al coeficiente de correlación (ahora parcial de orden 1) se resta el producto de dos correlaciones parciales). En el denominador se encuentra otra normalización para que el valor del coeficiente quede encerrado en el recorrido - 1 y + 1.

✎ Siguiendo la misma lógica de formulación de las dos ecuaciones anteriores, se puede seguir avanzando para computar coeficientes de orden 3, 4, ... hasta controlar *k* variables.

- Rápidamente se puede entrever que los cálculos resultarán más engorrosos en cada paso dados los requerimientos de matrices de correlación parcial de orden inferior.

- Por esta razón, el uso de los coeficientes de correlación parcial tienen utilidad práctica mientras que el número de las variables sea reducido. Más allá de 3 variables explicativas, claramente es desplazado por el análisis de regresión.

D. Desarrollemos ahora el primer ejemplo introducido más arriba con el fin de mostrar cómo podría desarrollarse un análisis multivariado con correlaciones parciales.

✎ Utilizando los microdatos de la evaluación internacional de aprendizajes a alumnos de 15 años escolarizados que hiciera la OECD y que se

---

denomina PISA 2000 y PISA +, las relaciones para los países participantes de América Latina (Argentina, Brasil, Chile, México y Perú) indican que las escuelas privadas obtuvieron un promedio en matemática de 404.8 puntos en tanto que las públicas obtuvieron 325.1 puntos. Esto parecería apoyar en una primera instancia la hipótesis *neo-institucionalista.*

✎ La matriz de correlaciones simples presentada a continuación indica que efectivamente existe una muy fuerte correlación entre el carácter privado de la escuela y el promedio en matemática de su alumnado: 0.504. Sin embargo, también se identifica una más fuerte correlación entre el nivel socioeconómico del alumnado de la escuela y el promedio en matemática de la misma: 0.644 . Finalmente, se observa que existe una fuerte correlación entre el sector institucional de la escuela y el nivel socioeconómico del alumnado: 0.611. Estos dos resultados parecerían apoyar en primera instancia la hipótesis reproductivista.

COEFICIENTE DE CORRELACIÓN PARCIAL (Guía de clase)

Matriz de correlaciones simples
para la base latinoamericana de escuelas de PISA 2000

	Prom. Matemática	La Escuela es del Sector Privado (si =1; No =0)	Nivel socioeconómico promedio del alumnado de la escuela	La escuela está en una zona rural (X_3)
Prom. Matemática	1.000			
La Escuela es del Sector Privado (si =1; No =0) (X_1)	0.504	1.000		
Nivel socioeconómico promedio del alumnado de la escuela (X_2)	0.644	0.611	1.000	
La escuela está en una zona rural (X_3)	-0.246	-0.319	-0.459	1.000

Fuente: elaboración propia con base a los microdatos de OECD / PISA 2000, PISA +.

✎ Para avanzar en el análisis y dirimir el conflicto, es necesario calcular correlaciones parciales de orden 1.

- Aplicando la ecuación respectiva de las correlaciones de orden 1 tanto para la relación entre Y y X_1 como para la relación entre Y y X_2, se construyeron las dos siguientes matrices de correlaciones parciales.

- En la primera matriz de correlaciones parciales, se puede apreciar que la magnitud de la relación entre sector y aprendizajes disminuye en forma muy importante al controlar por el nivel socioeconómico del alumnado de la escuela. Originalmente, una relación de 0.504 pasó a 0.183.

- En la segunda matriz de correlaciones parciales, se observa que también la magnitud de la asociación entre el promedio matemático y el nivel socioeconómico se redujo aunque en una magnitud mucho menor a la anteriormente comentada. De un valor de 0.644 pasó a 0.491.

- Se podría interpretar el primer resultado sosteniendo que efectivamente la gran diferencia entre los promedios en matemática según el sector institucional para las escuelas latinoamericanas esconde un efecto de clase social. Una parte de la fuerza de la asociación se debe a que el alumnado de las escuelas privadas tiene un más alto nivel socioeconómico; o lo que es lo mismo, la selección socioeconómica que las escuelas privadas hacen de su alumnado les permiten alcanzar un más alto resultado promedio.

- Sin embargo, la relación entre el sector y el promedio de aprendizajes no resultaría totalmente espúrea de acuerdo a los resultados parciales mostrados. Existe una parte del aprendizaje que parecería estar influido por el sector institucional.

Matriz de correlación parcial de orden 1 controlando por NSE
para la base latinoamericana de escuelas de PISA 2000

	Prom. Matemática	La Escuela es del Sector Privado (si =1; No =0)	La escuela está en una zona rural (si =1; No =0)
Prom. Matemática	1.000		
La Escuela es del Sector Privado (si =1; No =0) (X_1)	0.183	1.000	
La escuela está en una zona rural (si =1; No =0)	0.073	-0.068	1.000

Fuente: elaboración propia con base a los microdatos de OECD / PISA 2000, PISA +.

COEFICIENTE DE CORRELACIÓN PARCIAL (Guía de clase)

Matriz de correlación parcial de orden 1 controlando por SECTOR
para la base latinoamericana de escuelas de PISA 2000

	Prom. Matemática	Nivel socioeconómico promedio del alumnado de la escuela	La escuela está en una zona rural (si =1; No =0)
Prom. Matemática	1.000		
Nivel socioeconómico promedio del alumnado de la escuela (X_2)	0.491	1.000	
La escuela está en una zona rural (si =1; No =0)	-0.104	-0.357	1.000

Fuente: elaboración propia con base a los microdatos de OECD / PISA 2000, PISA +.

✎ Ahora bien, el argumento en discusión podría extenderse para señalar que la ubicación de la escuela incide en la relación original, dado que el sector privado tiene a rehusar localidades rurales (r=-0.319) y que en éstas localidades el nivel socioeconómico del alumnado es muy bajo (r= -0.459). Podría darse el caso de que al introducir la variable rural, la relación entre aprendizajes y sector institucional disminuya aún más. En consecuencia será necesario realizar un análisis de correlaciones de orden 2.

- Para esto será necesario construir una tercera matriz de correlaciones de orden 1, esta vez controlando las correlaciones por la variable rural. En dicha matriz se puede observar que las correlaciones simples originales del sector y del nivel socioeconómico con los aprendizajes han disminuido pero en forma muy pequeña (de 0.504 a 0.452 y de 0.644 a 0.631 respectivamente).

COEFICIENTE DE CORRELACIÓN PARCIAL
(Guía de clase)

Matriz de correlación parcial de orden 1 controlando por RURAL
para la base latinoamericana de escuelas de PISA 2000

	Prom. Matemática	La escuela está en el sector privado (si=1; no =0)	Nivel socioeconómico promedio del alumnado de la escuela (X_2)
Prom. Matemática	1.000		
La escuela está en el sector privado (si=1; no =0)	0.452	1.000	
Nivel socioeconómico promedio del alumnado de la escuela (X_2)	0.631	0.507	1.000

Fuente: elaboración propia con base a los microdatos de OECD / PISA 2000, PISA +.

- Una vez que se cuenta con esta última matriz lógicamente posible, se procederá a calcular todas las matrices de orden 2, aplicando la ecuación introducida oportunamente. Los resultados se presentan a continuación.

Matriz de correlación parcial de orden 1 controlando por NSE y RURAL
para la base latinoamericana de escuelas de PISA 2000

	Prom. Matemática	La Escuela es del Sector Privado (si =1; No =0)
Prom. Matemática	1.000	
La Escuela es del Sector Privado (si =1; No =0) (X_1)	0.197	1.000

Fuente: elaboración propia con base a los microdatos de OECD / PISA 2000, PISA +.

COEFICIENTE DE CORRELACIÓN PARCIAL (Guía de clase)

Matriz de correlación parcial de orden 2 controlando por SECTOR y RURAL
para la base latinoamericana de escuelas de PISA 2000

	Prom. Matemática	Nivel socioeconómico promedio del alumnado de la escuela
Prom. Matemática	1.000	
Nivel socioeconómico promedio del alumnado de la escuela (X_2)	0.523	1.000

Fuente: elaboración propia con base a los microdatos de OECD / PISA 2000, PISA +.

Matriz de correlación parcial de orden 2 controlando por SECTOR y NSE
para la base latinoamericana de escuelas de PISA 2000

	Prom. Matemática	La escuela está en el sector privado (si=1; no =0)
Prom. Matemática	1.000	
La escuela está en zona rural (si=1; no =0)	0.099	1.000

Fuente: elaboración propia con base a los microdatos de OECD / PISA 2000, PISA +.

- La relación originaria, una vez que se ha controlado simultáneamente por el nivel socioeconómico del alumnado y por la ubicación de la escuela en zona rural, parece no sólo mantenerse sino incrementarse levemente respecto del valor obtenido para la correlación de orden 1 (de 0.183 a 0.197).

- La relación entre el aprendizaje y el nivel socioeconómico parecería no sólo estar afectada por el sector institucional sino también por la ubicación geográfica de la escuela. Cuando esta última variable también se incorpora, la magnitud de la correlación disminuye hasta 0.523, cuando inicialmente había sido de 0.644 y controlando por sector fue de 0.491.

- Finalmente, es de observarse que con la ubicación geográfica de la escuela sucede algo bien interesante. A medida que se introducen controles apropiados por nivel socioeconómico y sector institucional se descubre que la asociación negativa con aprendizajes pasa a ser positiva.