25 ANIVERSARIO | UBA Sociales
1988 – 2013 FACULTAD DE CIENCIAS SOCIALES

CARRERA DE SOCIOLOGÍA – UBA
METODOLOGÍA DE LA INVESTIGACIÓN
CÁTEDRA: COHEN

CUADERNO DE CÁTEDRA Nº 1

“La aplicación de técnicas multivariadas en Ciencias Sociales”

Autores: Patricia Fernández; Guadalupe López y Jontef Enrique

Cuaderno de cátedra Nº 1
"La aplicación de técnicas multivariadas en ciencias sociales"
Autores: Patricia Fernández, Guadalupe López y Enrique Jontef

1. Una aproximación a la explicación

Dentro del proceso de investigación, cuando comenzamos la etapa de análisis, solemos realizar un primer entendimiento de las características de la muestra y la relación que se establece entre las variables construidas teóricamente. En esta primera instancia, asociado al análisis univariado y bivariado, describimos el comportamiento de las variables y su variación conjunta. Sin embargo, cuando nos proponemos además objetivos explicativos de investigación, es necesario avanzar hacia análisis multivariados. De esta forma, el análisis multivariado se enmarca en una estrategia de análisis cuantitativa que responde a objetivos explicativos de una investigación.

"La existencia de relación entre dos acontecimientos no permite suponer que uno de ellos explique o de cuenta de por qué existe el otro. En todo caso, coinciden, se vinculan, hasta puede haber relación de dependencia entre uno y otro, pero no necesariamente esto determina que el suceso dependiente es explicado por el independiente". (Cohen y Gómez Rojas, 2003: 118).

Siguiendo a Padua (1979), la explicación requiere de condiciones de verificación tanto lógicas como empíricas. No basta preguntarse el "por qué" de determinado fenómeno, además debemos resolver la necesidad de analizar la relación en principio planteada entre dos variables, controlando la influencia de terceras variables.

Siguiendo a los objetivos explicativos, el **diseño experimental** responde a la necesidad de poner a prueba una hipótesis original entre dos variables y con ello profundizar y enriquecer la explicación de los fenómenos que nos proponemos estudiar, como menciona Archenti, "*dando lugar a mejores aproximaciones a la realidad social donde resulta muy difícil encontrar la explicación al comportamiento de una variable solamente a partir de otra"* (Archenti, 2007: 279). Es decir, se parte de un supuesto en el cual esa relación aparente que observamos a partir de la convergencia entre dos variables puede deberse al efecto de otra u otras variables que están jugando algún tipo de influencia y se encuentran "ocultas" o ignoradas en ese primer análisis descriptivo de la relación. Dicho en otras palabras, uno de los objetivos del diseño experimental es explicar un fenómeno social, de la forma más completa posible, dando cuenta de las variables que podrían explicar la ocurrencia de dicho fenómeno.

En su corriente clásica, el **experimento** necesita de por lo menos dos grupos, que pueden ser individuos, grupos sociales o comunidades. A uno de ellos se le aplica un estímulo (**grupo experimental**) y al otro un *placebo* (**grupo de control**). Este procedimiento cuenta como mínimo de los siguientes pasos: la medición inicial de ambos grupos, la incorporación del estímulo al grupo experimental, la medición posterior de ambos grupos y finalmente la comparación de resultados. El grupo que brinda el parámetro de cambios o modificaciones es el grupo de control.

A diferencia de las posibilidades de experimentar en el sentido anteriormente descripto que permiten algunas ciencias, como la biología, química, entre otras, en las ciencias sociales estos diseños presentan dificultades en su implementación. Mientras que en el diseño experimental se interviene empíricamente en el momento en que ocurre el fenómeno (manipulando el objeto de estudio y realizando mediciones previas y posteriores a la aplicación del estímulo), en el estudio de los fenómenos sociales este tipo de intervención que implica aislar determinados fenómenos de su situación real, no es posible (no podemos aislar el fenómeno de la pobreza, violencia social u otro objeto de estudio). Por el contrario, sólo puede realizarse post facto, reproduciendo determinadas condiciones de experimentación a posteriori, en el momento del análisis.

Por esta razón en ciencias sociales más que hablar de experimentación en el sentido clásico, nos referimos al concepto de **explicación.** Hablar de análisis explicativo implica la presencia de tres condiciones necesarias:

- **Covariación entre variables**. Describe la concomitancia o relación conjunta entre las variables. No es exclusivo del análisis explicativo pero lo requiere.
- **Orden temporal de las variables**. En el análisis explicativo el orden temporal tiene un rol importante en esta secuencia. El tiempo actúa así como una variable que interviene o afecta la explicación del fenómeno social.
- **Control de variables**. Confronta con la idea de covariación a partir del supuesto de existencia de terceras variables ocultas - no controladas - que podrían modificar o estar afectando la relación entre las variables originales en donde se detectó dicha covariación.

En síntesis, tomando las tres condiciones anteriormente mencionadas podemos afirmar que en el análisis multivariado se presenta una explicación de la relación entre dos variables, donde además de demostrar la covariación que existe entre ambas, debemos explicitar la secuencia temporal que establecen entre sí y **garantizar que esa relación esté controlada, a la luz de otras variables que podrían estar afectando esa relación.** Es importante señalar

que la incorporación al análisis de estas variables surge del marco teórico, ya que la decisión de controlar una relación surge de considerar aquellas variables que son relevantes teóricamente para el estudio de ese fenómeno social.

Ahora bien, ¿cómo controlamos las variables en ciencias sociales? El control de variables **consiste en transformar las terceras variables en constantes** (es decir, quitarles su variación) **y luego analizar cuáles son sus efectos comparando qué ocurre ante su presencia y ante su ausencia.** Por ejemplo, si quisiéramos analizar la relación entre intención de voto y zona de residencia controlándola por la variable sexo, deberíamos transformar en constante esta última variable. Transformarla en constante implica quitarle su variabilidad. Por lo tanto, deberíamos reproducir la relación entre intención de voto y zona de residencia entre los hombres por un lado y entre las mujeres por otro. Para cada una de esas situaciones la variable "sexo" se transforma en una constante. En un grupo tenemos sólo hombres y en el otro sólo mujeres.

Tras el proceso de control de variable, nos aseguramos en avanzar en la posible explicación de un fenómeno social. Siempre para poder explicar tengo que poder controlar esa relación por otras variables, es decir evidenciar aquellas variables que en un primer momento aparecen ocultas al primer análisis de la situación.

Antes de avanzar con la ejemplificación del análisis multivariado mencionaremos algunos autores que han desarrollado dicha temática en ciencias sociales. Entre los **antecedentes** que han desarrollado esta técnica, podemos encontrar a Emile Durkheim en su estudio sobre el suicidio (1897). Durkheim plantea a partir de una relación inicial diferentes y sucesivos controles, donde está presente el objetivo de poner en juego la influencia de otras variables y observar qué relaciones se producen. En un primer momento, Durkheim parte de una hipótesis original, en la cual la confesión religiosa causaría algún efecto sobre la tasa de suicidio, de esta forma pone en relación dos variables, la confesión religiosa, como variable independiente y la tasa social de suicidio en el papel de variable dependiente.
En esta lógica del diseño experimental, lo que se propone es controlar esta relación por terceras variables que podrían estar incidiendo también en la tasa de suicidio como la nacionalidad, el lenguaje, la región, la instrucción de los cónyuges, distintos grupos de provincias. De esta forma, comienza a plantear nuevas hipótesis–alternativas que van surgiendo a medida que aparecen nuevas variables externas que podrían estar influyendo en la tasa social de suicidio. Finalmente, en su estudio encuentra un elemento común de mayor abstracción que lo lleva a pensar que no es la religión la que impulsa o genera tendencia a una mayor incidencia de suicidios, sino la expresión de un cierto libre examen, una cierta posibilidad de desarrollo de la individualidad y con ello la idea de que la tasa de suicidios está expresando una situación de anomia.

Lo interesante de esa obra es que permite ver que la relación entre variables no empieza ni termina entre dos variables por sí, sino que necesita tener un examen atento de terceros, cuartos, quintos o más aspectos que pueden estar contaminando, cuestionando, invalidando esa relación.
Recién luego de esos controles se puede decir que se está en condiciones de explicar un fenómeno: en este caso las causas de la incidencia en la tasa social de suicidios.

Contemporáneamente, Hyman (1968) y Lazarsfeld (1966) también han abordado el problema de las relaciones espurias o aparentes a través de la incorporación de variables de control.

Como síntesis de todo lo descripto anteriormente, el **aporte de la lógica experimental** a la comprensión de la explicación de la sociología permite dar cuenta de distintas situaciones en las cuales terceras variables pueden estar influyendo en una relación original entre lo que llamamos una variable dependiente e independiente:
1- La posible existencia de una **relación espuria**, es decir una relación que en principio parecía existir entre dos variables pero que sólo se manifiesta por la existencia de una tercera variable que produce la relación.
2- La **explicación** de la variable dependiente por la independiente, demostrando la no influencia de la variable de control en la relación.
3- La existencia de determinadas condiciones bajo las cuales una relación se manifiesta, esto es **especificando** las situaciones en las que dicha relación se presenta.

2. Un ejemplo de análisis multivariado para variables cualitativas

Supongamos que nos encontramos realizando un estudio entre jóvenes de 20 a 28 años que residen en AMBA en el año 2013, con el objetivo de analizar el tipo de vínculo que poseen actualmente con la lectura de diarios en versión impresa, en un contexto de fuerte incremento de otros medios como fuentes de información, entretenimiento y comunicación.
En este contexto, la variable sobre la que nos interesa profundizar el análisis es la *lectoría habitual de diarios en formato impreso,* entendiendo por habitual una frecuencia de lectura igual o mayor a una vez por semana.
En una primera instancia, se decide analizar la relación entre *lectoría habitual de diarios impresos* y *la lectura habitual de diarios en Internet.* Cuando hablamos de la lectura habitual de diarios online nos referimos también a una frecuencia de una vez por semana o superior, y a la versión digital de diarios impresos o diarios que se emiten solamente por Internet.
La hipótesis que acompaña la decisión de realizar este análisis es que los jóvenes que acceden habitualmente a la lectura de diarios online abandonan el

hábito de lectura de diarios impresos, encontrando en este medio el tipo de acceso a la información que buscan.
Para poner a prueba la hipótesis anteriormente mencionada, se construye el siguiente cuadro:

Cuadro 1: Lectura habitual de diarios impresos según lectura habitual de diarios en Internet. Jóvenes de 20 a 28 años que residen en AMBA - año 2013.

		Lectura de diarios en internet		Total
		Lee diarios en internet	No lee diarios en internet	
Lectura de diarios impresos	Lee diarios impresos	30%	81%	57%
	No lee diarios impresos	70%	19%	43%
%		100%	100%	100%
n		270	250	520

Fuente: Datos ficticios

Como podemos observar en el cuadro anterior, más de la mitad de los entrevistados (57%) lee habitualmente diarios impresos. Sin embargo, si lo analizamos en función de la lectura de diarios en Internet, observamos que entre aquellos que sí lo hacen, la lectura de diarios impresos disminuye al 30%. Por el contrario, entre quienes no leen diarios en Internet, la lectura de diarios impresos aumenta al 81%, presentándose una diferencia porcentual del 51% entre una y otra categoría.
Al tratarse de dos variables dicotómicas, esa diferencia porcentual se mantiene para aquellos que no leen habitualmente diarios impresos según lectoría de diarios por Internet. Entre aquellos que no tienen experiencia de lectura en el hogar de origen el 70% no lee habitualmente diarios, comparado con un 19% entre los que sí tienen experiencia de lectura.
De la lectura anterior, podemos concluir que la hipótesis planteada tiende a corroborarse y que efectivamente existe una asociación entre la lectura de diarios impresos y la lectura de diarios por Internet, en el sentido que a mayor tendencia a la lectura de diarios por Internet disminuye la lectura de diarios impresos.

Ahora bien, ¿podríamos afirmar con este análisis que el hábito de lectura de diarios por Internet es una variable que explica el leer o no leer habitualmente diarios impresos en la actualidad? La respuesta es no, ya que pueden existir otras variables que de alguna u otra manera intervengan en la relación, ya sea para fortalecerla o para demostrar que esa relación inicial se ve modificada al incorporar en el análisis terceras variables. Como vimos anteriormente, para

poder hablar de explicación tres condiciones son necesarias: la covariación, el control de variables así como la secuencia temporal que teóricamente definimos en la relación entre las variables involucradas.

Para esto ponemos a prueba la relación inicial u original (en el ejemplo anterior, la relación entre *lectoría habitual de diarios impresos y la lectoría de diarios por Internet),* controlándola por terceras variables que consideremos teóricamente relevantes y con capacidad de influencia en dicha relación.

Volviendo al ejemplo, decidimos controlar la relación original por un conjunto de variables que consideramos podrían estar influyendo en la relación. A continuación presentaremos el análisis de la relación original controlada por tres variables: sexo, nivel educativo y experiencia de lectura de diarios impresos en el hogar de origen. A través de cada uno de estos análisis se podrán observar distintas situaciones que permiten comprender mejor la relación entre la lectura actual de diarios impresos y diarios online.

Para comenzar, controlamos la relación por sexo ya que consideramos que podría estar ejerciendo algún tipo de influencia en la relación. Esta nueva hipótesis de trabajo confronta así con nuestra hipótesis original, poniendo a prueba la capacidad explicativa que tiene la variable *lectura de diarios en Internet* sobre la lectura habitual de *diarios impresos.* Operativamente, reproducimos la relación original para cada una de las categorías de la variable de control (sexo). Quedan así conformados dos grupos diferenciados, uno de hombres y otro de mujeres. Para cada uno de los cuadros el sexo se transforma ahora en una constante.

Cuadro 2: Lectura habitual de diarios impresos según lectura habitual de diarios en Internet por sexo. **Hombres de 20 a 28 años** que residen en AMBA - año 2013.

		Lectura de diarios en internet		**Total**
		Lee diarios en internet	No lee diarios en internet	
Lectura de diarios impresos	Lee diarios impresos	31%	83%	59%
	No lee diarios impresos	69%	17%	41%
%		100%	100%	100%
n		146	134	280

Fuente: Datos ficticios

Cuadro 3: Lectura habitual de diarios impresos según lectura habitual de diarios en Internet por sexo. **Mujeres de 20 a 28 años** que residen en AMBA - año 2013.

		Lectura de diarios en internet		**Total**
		Lee diarios en internet	No lee diarios en internet	
Lectura de diarios impresos	Lee diarios impresos	29%	81%	56%
	No lee diarios impresos	71%	19%	44%
%		100%	100%	100%
n		124	116	240

Fuente: Datos ficticios

Nótese que al introducir la variable de control el número total de casos queda reducido en cada cuadro a la cantidad de jóvenes hombres en un cuadro y mujeres en otro.

Yendo al análisis del Cuadro 2, podemos observar que dentro de los jóvenes hombres casi 6 de cada 10 (59%) lee habitualmente diarios impresos. Sin embargo, cuando lo analizamos en función de la lectura de diarios online, entre quienes leen diarios por Internet el 69% no lee diarios impresos. Este porcentaje disminuye al 17% entre quienes no leen diarios en Internet. La diferencia porcentual es del 52% entre una y otra categoría, valor similar a lo que sucede en la relación original, sin controlarlo por la variable sexo.

En el caso de las mujeres (cuadro 3), la relación se presenta muy similar a los hombres. El 56% lee habitualmente diarios impresos. Sin embargo, cuando lo analizamos en función de la lectura de diarios online, entre quienes leen diarios por Internet el 71% no lee diarios impresos. La diferencia porcentual con quienes no leen diarios online es en este caso del 52%.

Luego de analizar al relación original controlada por la variable sexo concluimos que ésta no produce ningún tipo de influencia en la relación. Dicho en otros términos podemos afirmar que sexo es una variable ajena a la relación original.

Decir que la variable sexo es ajena a la relación no implica necesariamente afirmar que no puedan existir otras variables que afecten la relación, ya que entendemos la complejidad del mundo social desde una perspectiva multidimensional.

Controlamos en segundo lugar la relación, esta vez por la variable nivel educativo del entrevistado. La variable se presenta dicotomizada en dos categorías: hasta secundario incompleto y secundario completo y más.

Cuadro 4: Lectura habitual de diarios impresos según lectura habitual de diarios en Internet por nivel educativo. Jóvenes de 20 a 28 años **con hasta secundario incompleto** que residen en AMBA - año 2013.

		Lectura de diarios en internet		**Total**
		Lee diarios en internet	No lee diarios en internet	
Lectura de diarios impresos	Lee diarios impresos	45%	65%	57%
	No lee diarios impresos	55%	35%	43%
%		100%	100%	100%
n		138	128	266

Fuente: Datos ficticios

Al analizar la relación entre hábito de lectura de diario impresos y online, sólo para aquellos entrevistados que poseen hasta secundario incompleto observamos que la relación se mantiene pero debilitándose, presentando una diferencia del 20% entre aquellos que leen habitualmente diarios en Internet y aquellos que no lo hacen.

Veamos qué sucede al controlar la relación por aquellos que tienen un nivel educativo superior.

Cuadro 5: Lectura habitual de diarios impresos según lectura habitual de diarios en Internet por nivel educativo. Jóvenes de 20 a 28 años **con secundario completo o más** que residen en AMBA - año 2013.

		Lectura de diarios en internet		**Total**
		Lee diarios en internet	No lee diarios en internet	
Lectura de diarios impresos	Lee diarios impresos	12%	90%	70%
	No lee diarios impresos	88%	10%	30%
%		100%	100%	100%
n		132	122	254

Fuente: Datos ficticios

En el cuadro precedente (cuadro 5) podemos observar que dentro de los jóvenes con secundario completo y más el 70% lee habitualmente diarios impresos, pero que este porcentaje aumenta entre quienes no leen diarios en Internet. En este caso la diferencia porcentual en relación a los que leen diarios online es del 78%, 27 puntos más que lo observado para la relación original.

El anterior análisis nos permitió avanzar en la comprensión de la relación entre las dos variables, que se ve fortalecida entre aquellos entrevistados que poseen nivel educativo superior. De esta forma podemos afirmar que el nivel educativo influye en la relación, incrementándola entre los jóvenes que tienen mayor nivel educativo (secundario completo o más).

Supongamos ahora que controlamos la relación por la variable *experiencia de lectura de diarios impresos en el hogar de origen* (es decir, hogar donde crecieron durante la infancia).

Cuadro 6: Lectura habitual de diarios impresos según lectura habitual de diarios en Internet por experiencia de lectura de diarios impresos en hogar de origen. Jóvenes de 20 a 28 años **con experiencia de lectura en el hogar** que residen en AMBA - año 2013.

		Lectura de diarios en internet		Total
		Lee diarios en internet	No lee diarios en internet	
Lectura de diarios impresos	Lee diarios impresos	96%	98%	97%
	No lee diarios impresos	4%	2%	3%
%		100%	100%	100%
n		100	120	220

Fuente: Datos ficticios

Al analizar la relación entre hábito de lectura del diario de diarios impresos según la lectura de diarios en Internet bajo el control de la variable experiencia de lectura en el hogar, observamos que la relación original tiende a desaparecer. Dentro del grupo de jóvenes que tienen experiencia de lectura en el hogar, el 97% lee diarios impresos, independientemente de leer o no leer diarios en Internet.

Cuadro 7: Lectura habitual de diarios impresos según lectura habitual de diarios en Internet por experiencia de lectura de diarios impresos en hogar de origen. Jóvenes de 20 a 28 años **sin experiencia de lectura en el hogar** que residen en AMBA - año 2013.

		Lectura de diarios en internet		Total
		Lee diarios en internet	No lee diarios en internet	
Lectura de diarios impresos	Lee diarios impresos	4%	8%	5%
	No lee diarios impresos	96%	92%	95%
%		100%	100%	100%
n		170	130	300

Fuente: Datos ficticios

Al analizar la relación entre hábito de lectura de diarios impresos y online para quienes no tienen experiencia de lectura de diarios en el hogar de origen, la

relación también desaparece. Dentro del grupo de jóvenes que no tienen el hábito de lectura, el 95% no lee diarios impresos actualmente. Esos valores son similares independientemente de la lectura de diarios por Internet.
En este caso podemos ver que la introducción de la variable de control nos indica que la relación original se diluye a la luz de la experiencia de lectura en el hogar, lo que demuestra la espuriedad de la relación.

En estas situaciones es interesante observar qué sucede en la relación entre la variable de control con cada una de las variables que componen la relación original (variable dependiente e independiente).

Cuadro 8: Lectura habitual de diarios impresos según experiencia de lectura en hogar de origen. Jóvenes de 20 a 28 años que residen en AMBA - año 2013.

		Experiencia de lectura de diarios en hogar de origen		Total
		Con experiencia	Sin experiencia	
Lectura de diarios impresos	Lee diarios impresos	80%	25%	57%
	No lee diarios impresos	20%	75%	43%
%		100%	100%	100%
n		220	300	520

Fuente: Datos ficticios

En el cuadro precedente volvemos a destacar que casi 6 de cada 10 entrevistados (57%) lee diarios impresos. Sin embargo, cuando analizamos esta variable en función de la experiencia de lectura en el hogar, ese porcentaje aumenta al 80%, disminuyendo al 25% entre quienes no tienen experiencia. La diferencia porcentual es del 55% entre una y otra categoría.
Este análisis demuestra la relación existente entre la experiencia de lectura en el hogar de origen y la lectura actual de diarios impresos, lo que nos permite afirmar que el hábito de lectura en el hogar familiar estimuló la lectura actual de diarios en los jóvenes.

Cuadro 9: Lectura habitual de diarios en Internet según experiencia de lectura en hogar de origen. Jóvenes de 20 a 28 años que residen en AMBA - año 2013.

		Experiencia de lectura de diarios en hogar de origen		Total
		Con experiencia	Sin experiencia	
Lectura de diarios en internet	Lee diarios en internet	28%	78%	52%
	No lee diarios en internet	72%	22%	48%
%		100%	100%	100%
n		220	300	520

Fuente: Datos ficticios

En el cuadro precedente observamos que más de la mitad de los entrevistados (52%) lee diarios en Internet. Cuando analizamos dicha variable en función de la experiencia de lectura observamos que entre los jóvenes que tienen experiencia de lectura en el hogar de origen el porcentaje de lectura de diarios por Internet es tan solo del 28%, aumentando dicho porcentaje al 78% entre los que no presentan dicha experiencia. La diferencia porcentual en este caso es del 50% entre una y otra categoría.
Este análisis demuestra que la experiencia de lectura en el hogar también está vinculada con la lectura actual de diarios por Internet, de forma tal que quienes traen consigo la experiencia de lectura de diarios impresos son menos proclives a la lectura de diarios en formato online.

Sintetizando todo el proceso anteriormente realizado, partimos de un análisis que ponía a prueba la hipótesis que los jóvenes que leen diarios por Internet no leen diarios impresos. Del análisis inicial de esta relación pudimos observar que existe covariación entre ambas variables de forma tal que entre aquellos que leen actualmente diarios en Internet no suelen repetir dicho hábito de lectura en formato impreso.
Sin embargo, la existencia de covariación no nos permite ratificar que leer diarios en formato digital explique realmente el abandono de lectura de diarios impresos en los jóvenes de entre 20 a 28 años de AMBA. Es por eso que decidimos introducir distintas variables de control que considerábamos teóricamente relevantes en nuestro análisis: sexo, nivel educativo y experiencia de lectura de diarios impresos en el hogar familiar de origen.

El control de cada una de estas variables en el análisis de la relación entre hábito de lectura de diarios impresos y online nos permite reconstruir tres posibilidades que pueden surgir del análisis:

- Que la variable sea **ajena** a la relación, como sucedió cuando controlamos la relación por la variable sexo.
- Que la relación **se especifique o se muestre fortalecida**, para alguna o algunas de las categorías de la variable de control, como sucede cuando se introduce la variable Nivel Educativo
- Que la relación original que observábamos sea **espuria**, como sucede cuando introducimos la variable de control Experiencia de lectura de diarios impresos en el hogar de origen. Es decir, que en aquella situación en donde observábamos una relación entre dos variables, éstas en realidad sólo covarían por la existencia de otra variable que es la que en realidad explica el comportamiento de nuestra variable dependiente: hábito de lectura de diarios impresos. Al intervenir la relación comprendemos que la relación original es una relación aparente y que estas variables sólo están vinculadas entre sí al estar ellas mismas asociadas con la variable de control.

Como cierre, cabe destacar que las tres alternativas anteriormente detalladas se presentan, con fines didácticos, como posibilidades referenciales. La mayoría de los análisis multivariados que realizamos en nuestra tarea de investigación suelen arribar a resultados que están, en mayor o menor medida, emparentados con las alternativas anteriormente detalladas.

3. Análisis multivariado para variables cuantitativas

Hasta aquí hemos visto el proceso que se efectúa en el análisis multivariado a partir de la utilización de variables cualitativas. Si nos viéramos en la necesidad de trabajar exclusivamente con variables cuantitativas o intervalares (las cuales son limitadas en sociología, a diferencia de lo que acontece con variables cualitativas) y establecer la relación entre ellas, esto nos llevará a trabajar con otros dos procedimientos: la correlación y la regresión lineal.

El análisis de correlación y regresión lineal, como cualquier medida estadística permite dimensionar el comportamiento empírico del fenómeno de estudio pero bajo las condiciones teóricas que el investigador determina. Entendemos por tales condiciones, en primer lugar a las variables, en tanto entidades construidas desde la teoría y en segundo lugar a las hipótesis en tanto expresión de la relación teórica entre aquéllas. Por ende los supuestos teóricos orientan a través de las hipótesis y las variables el análisis de la información obtenida y direccionan el uso e interpretación de las medidas estadísticas

El análisis de regresión lineal es una técnica estadística adecuada que se utiliza para estudiar la relación entre variables. En investigación social dicho análisis de regresión se aplica para predecir una amplia gama de fenómenos.

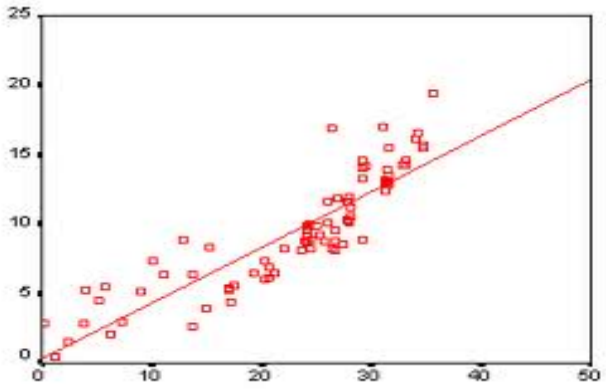
Según sea la cantidad de variables independientes con las que nos manejamos estaremos haciendo una distinción entre el análisis de regresión lineal simple y el de regresión lineal múltiple. El primero, tiene como función estimar el comportamiento de una variable dependiente a partir de la variable independiente. En el segundo se trabaja con un conjunto de variables independientes en donde se determina cuáles son las que más influyen en la variable dependiente.

Plantearse este tipo de conocimiento implica tener la necesidad de acceder a información que no se posee, pero exige partir de una condición muy importante: suponer que la información que se carece se puede registrar en una variable y que ésta se encuentra relacionada (covaría) con otra, de manera tal que a partir de ésta última se podrá llegar a la información requerida en forma confiable y válida.
Estos cálculos se realizan toda vez que necesitamos un dato que no hemos podido construir mediante el proceso teórico-empírico, sea porque corresponde a un momento futuro o porque, si bien la ocurrencia es presente, no se puede acceder a la información correspondiente en el momento de la recolección o porque la fuente de datos secundarios que consultamos no contiene el registro de interés.

Esta **estimación** solo es posible bajo ciertas condiciones:

- La relación de dependencia entre las variables debe ser suficientemente fuerte, de manera tal de minimizar la probabilidad de error en la **estimación**. La fuerza de la relación se expresa en la magnitud del resultado obtenido con el coeficiente “r” de Pearson (correlación). Cualquier relación entre variables, cuánto más fuerte, permite **predecir** más confiablemente el comportamiento de la variable dependiente, ya que a medida que aumenta dicha fuerza disminuye la libertad (independencia) de comportamiento de esta última variable.
- La elección de la variable independiente es de carácter teórico. Se supone (se asume como hipótesis) que esta variable es la que mejor da cuenta de las variaciones de la variable dependiente: es la estimadora, teóricamente, más apropiada. Por lo tanto su elección debe estar debidamente fundamentada.

Entendamos el **Análisis de regresión lineal** a partir de un diagrama de dispersión.



La nube de puntos adquiere su forma según la intensidad de la covariación. Así, cuánto más intensa o más fuerte sea la covariación, más densa y cercana a una recta será la “nube de puntos”, contrariamente cuanto menor sea la covariación más dispersa será la nube de puntos.

Por lo tanto cuando observamos diagramas muy dispersos sabremos que no podremos realizar **estimaciones precisas**, porque el comportamiento de la variable tiende a la independencia entre sí. Si por el contrario, el diagrama es más concentrado podemos confiar en esperar **mejores estimaciones**

El modelo de regresión lineal ha de cumplir una serie de supuestos que garanticen su correcta aplicación, a saber: a) linealidad; b) normalidad; c) homocedasticidad. Todos estos supuestos pueden ser estudiados mediante el recurso de las puntuaciones residuales que indican la diferencia entre las puntuaciones observadas y predichas por el modelo

Ya se ha visto la relación que pueden tener dos variables intervalares en la que una es considerada como una variable independiente en tanto la otra adopta el comportamiento de variable dependiente a los efectos de establecer la covariación existente. Es decir que esas variaciones pueden ser enteramente independientes o bien que dichas varianzas tengan alguna variación conjunta (o covarianza) o finalmente que toda la variación de las dos variables sea en efecto una variación conjunta.

Pero si nos atenemos a un escenario propio de las ciencias sociales en el que la realidad con la que se trabaja presenta no sólo dos variables sino una multiplicidad de relaciones entre los fenómenos; nos llevará forzosamente a considerar dos modelos de correlación ateniéndonos a la mayor cantidad de variables independientes que explicarían la variable dependiente. Dichos modelos son los denominados correlación parcial y correlación múltiple.

3.1. La correlación parcial

La correlación parcial mide el grado de relación existente entre dos variables pero en función del control que se ejerce sobre una o más variables. En otras palabras, es posible pensar que otras variables, por fuera del modelo bivariado presentado anteriormente, se encuentren influyendo en distinta medida en la relación original. Estas otras variables podrían incidir en la relación original siendo causantes de las variaciones presentadas en la correlación lineal.

El coeficiente de correlación parcial, en definitiva, debe ser considerado como la correlación que queda entre la variable independiente y la variable dependiente (ambas intervalares) una vez suprimidos los efectos de la variable de control. En definitiva, con la correlación parcial se procura explicar el comportamiento de la variable dependiente a partir de la variable independiente con una variable de control.

A partir de la fórmula que representa el modelo de correlación parcial se podrá entender de qué manera se trabaja con el control de una tercera variable.

$$r13.2 = \frac{r13 - (r12)(r23)}{\sqrt{(1 - r^2\,12)\,(1 - r^2 23)}}$$

El “r” expresado en la fórmula representa el coeficiente de correlación de Pearson, ya visto en el modelo de correlación lineal. La diferencia que se contempla en este caso es que en vez de medir la covariación entre dos variables, el coeficiente expresa la presencia de 3 variables.

Reemplazando los números, 1 y 3 representan la correlación simple (estableceremos que “1” es la variable independiente y “3” es la variable dependiente; en tanto “2” es la variable de control. En el numerador de la fórmula quedan expresadas tres correlaciones simples: r13 representa la correlación entre las dos variables originales (la independiente y la dependiente); r12 refleja la correlación entre la variable independiente y la variable de control en tanto que r23 establece la correlación entre la variable de control y la variable dependiente. Entretanto en el denominador se expresa el coeficiente de indeterminación (ya visto en el modelo de correlación simple); en este caso dicho coeficiente (que expresa en qué medida la variación de la variable dependiente no es explicada por la variación de la variable independiente) está presentado tanto para la indeterminación entre la variable independiente y la variable de control por un lado como por la indeterminación entre la variable de control y la variable dependiente por el otro.

El coeficiente de correlación parcial varía entre 1 y –1 y su resultado se interpreta de forma similar al de la correlación simple. Aquí lo que debe

considerarse es que si bien se explican las variaciones de la variable dependiente a partir de la independiente, dicha relación refleja la fortaleza de la variable de control con cada una de las dos variables de la relación original. Es necesario hacer notar que en el denominador de la fórmula se manifiestan los coeficientes de independencia “k” que miden la independencia estadística entre la variable independiente y la variable de control, por un lado y de la variable de control con la variable dependiente por el otro. Cuánto más fuerte sea la independencia entre las variables, mayor magnitud reflejará el denominador y por ende menor será la correlación final entre las variables. Entretanto en el numerador se está midiendo la correlación simple existente en la relación original y el producto de la correlación de ambas con respecto a la variable de control. Por ende, a medida que aumenta el denominador, (o la independencia estadística) disminuye la correlación simple entre las variables, disminuyendo el resultado final del modelo presentado.

A continuación presentaremos un ejemplo ilustrativo con fines didácticos en el que contemplaremos diversas posibilidades que podrían presentarse para graficar el procedimiento de correlación parcial. A tal efecto retomaremos el escenario planteado con antelación bajo el supuesto que nos encontramos realizando un estudio entre jóvenes de 20 a 28 años que residen en AMBA en el año 2013, con el objetivo de analizar el tipo de vínculo que poseen actualmente con la lectura de diarios en versión impresa, en un contexto de fuerte incremento de otros medios como fuentes de información, entretenimiento y comunicación.

Si quisiéramos conocer la relación que existe entre “cantidad de horas que lee diarios impresos en un mes” (variable dependiente) con “edad” (variable independiente) controlando “ingresos” (variable de control), utilizaremos los siguientes datos hipotéticos:

Correlación simple entre “edad” y “cantidad de horas que lee diarios impresos en un mes”: **0,70**
Correlación simple entre “cantidad de horas que lee diarios impresos en un mes” e “ingresos” **0,63**
Correlación simple entre “edad” y “ingresos” **0,49**
Al aplicar la fórmula el resultado del coeficiente de correlación parcial es 0,58; ¿qué puede sostenerse con estos valores? Pues que estamos frente a una **explicación de tipo parcial.** En este caso, la correlación parcial reduce solamente una parte comparada con la correlación bivariada original, es decir, que la variable de control afecta parcialmente la relación original. Siguiendo la misma lógica y a título de ejemplo; supongamos que si el coeficiente de correlación parcial hubiese arrojado un valor bastante más cercano a 0 entonces podría decirse que la variable de control **afecta** totalmente la relación. Por último, un resultado del coeficiente de correlación parcial muy cercano (o

igual) a 0,70 nos permitiría afirmar que la variable de control es **ajena** a la relación original planteada

3.2. La correlación múltiple

En reiteradas oportunidades hemos podido comprobar la relación existente entre una variable independiente y otra variable dependiente (ambas cuantitativas) lo que nos posibilita, en algunos casos, predecir los valores de una variable a partir de los valores observados en la otra. Pero ciertamente resulta difícil poder atribuir a una sola variable los resultados en otra; la realidad nos lleva a reconocer que para predecir con mayor precisión un determinado valor, es necesario observar e integrar en la predicción otras variables que también puedan estar relacionadas. Esto lleva a la necesidad de trabajar con más de dos variables (una independiente y otra dependiente) de manera simultánea y en el caso de una observación correlacional, requerirá de la aplicación de un modelo que permita pesar el grado de impacto que cada una de las variables observadas puede tener sobre los resultados de la predicción.

El procedimiento analítico que nos permite establecer cuánto de la variación en la variable observada (o dependiente) está asociado con la variación del conjunto de variables independiente que pretenden explicarla y predecirla se denomina: correlación múltiple.

El objetivo de la aplicación de dicho modelo apunta a la construcción de la mejor combinación del peso que cada variable independiente aporta en la medición de la variable dependiente que procura explicarse. Y esta mejor combinación sin duda tendrá una mayor correlación con la variable dependiente que la correlación que pueda tener cualquiera de las variables independientes con respecto a la dependiente, tomadas de manera individual.

Esta técnica supone que existen más de dos variables correlacionadas y que es posible determinar la forma como se comportan las diversas correlaciones a nivel bivariable a fin de conformar una correlación combinada o total.

El coeficiente R^2, símbolo de este tipo de correlación, se presenta en el modelo de la siguiente forma:

$$\mathbf{R^2123 = r^212 + r^213.2 (1- r^212)}$$

1 = variable dependiente
2 y 3 = variables independientes

$r^2$12: Proporción explicada de la variable dependiente por la variable independiente principal

$r^2$13.2: esto es r^2 de la correlación parcial, es la proporción explicada de la variable 3 controlada por la variable 2

(1- $r^2$12): Esto es el k^2, es la proporción no explicada de la variable dependiente por la variable independiente principal

La presentación de la fórmula (que considera un modelo de dos variables independientes y una variable dependiente) nos lleva a pensar en un análisis de los componentes que en cierta medida trasciende el mero resultado del coeficiente. Esta es una correlación en la que se permite descomponer la relación que hay entre las variables con un objetivo que difiere en cierto modo de lo presentado en la correlación parcial; en aquélla se apunta a la explicación de la variable dependiente a partir de una variable independiente con la presencia de una variable de control. En tanto en la correlación múltiple Lo que se procura es la captación de cómo explican a la variable dependiente un conjunto de variables independientes.

Para la aplicación del modelo de correlación múltiple es imprescindible saber desde qué hipótesis se está partiendo.
En el modelo más elemental (1 variable dependiente y 2 variables independientes) la primera hipótesis tiene que ver con elegir determinadas variables independientes para explicar a la variable dependiente. Pero el segundo nivel es el más importante, ya que debe suponerse y plantear de manera concreta cuál de las variables independientes se espera que tenga un mayor poder explicativo. Esto conlleva necesariamente a una jerarquización de las variables independientes.

- El primer elemento de la fórmula **($r^2$12)** incluye la variable dependiente y una de las variables independientes; esto nos va a indicar cuál es la proporción explicada de la variable dependiente por la variable independiente Nro. 2. (o variable principal). En este modelo que presento supongo que la variable independiente Nro. 2 tenga el mayor poder explicativo y va a cumplir esa función en toda la fórmula. Es decir, cuánto de la variable dependiente o sus variaciones se explican por la variable independiente (o variable Nro. 2).
- El segundo elemento de la fórmula **($r^2$13.2)** es el r^2 que mide la determinación que hay entre la variable dependiente (o variable Nro.1) y la otra variable independiente (o variable Nro. 3) controlando la variable independiente principal Esto indica cuál es la proporción de la variable dependiente por la otra variable independiente controlando la variable Nro. 2.
- El tercer elemento de la fórmula **(1- $r^2$12)** es el K cuadrado; es el coeficiente de indeterminación. Es la proporción no explicada de la variable dependiente por la variable independiente principal.

El coeficiente de correlación múltiple (R^2) varía entre 0 y 1; no hay valores negativos ya que todos los componentes son elevados al cuadrado. Pero más allá del resultado que arroja el coeficiente, la importancia del modelo radica en su desarrollo; esto es: cuánto explica la principal variable independiente; cuánto explica la otra variable independiente controlando la variable independiente principal y cuánto queda sin explicar.

En la correlación múltiple el R^2 indica qué proporción de las variaciones de la variable dependiente es explicada por el conjunto de variables independientes propuestas por la hipótesis. En segundo lugar, permite conocer el desempeño de las variables independientes al interior de la proporción explicada. En otras palabras, el conjunto de las variables independientes es tratado en forma agregada y desagregada. En tercer lugar, permite conocer la proporción de las variaciones no explicadas. La combinación de estas tres contribuciones del modelo a la comprensión del comportamiento de la variable dependiente, le ayuda al investigador no solamente a probar sus hipótesis sino también a especificar el alcance de las mismas. Cuenta con información suficiente como para aceptarla o rechazarla en su totalidad o parcialmente. Puede discriminar el desempeño de cada variable independiente en su intento por explicar el comportamiento de la dependiente.

Para graficar el procedimiento de correlación múltiple utilizaremos el mismo ejemplo anterior con la salvedad que trabajaremos con la variable “Ingresos” considerándola como variable independiente principal y “edad” como otra variable independiente. Entonces, si en un determinado grupo poblacional quisiéramos explicar el comportamiento de la variable “cantidad de horas que lee diarios impresos en un mes”” (variable dependiente o variable Nro. 1) y lo hacemos a partir de “ingresos” (variable independiente principal o variable Nro. 2) y de “edad” (segunda variable independiente o variable Nro. 3); debemos entonces calcular la correlación simple entre las dos primeras ($r^2 12 = 0,63$) y la correlación parcial entre 1 y 3 controlando 2 ($r^2 13.2) = 0,58$. Una vez obtenidos estos resultados procedemos a calcular los respectivos r2 y el coeficiente de independencia (indeterminación) k2 referidos a la correlación entre 1 y 2 ($1-r^2 12$)

A los coeficientes se les aplica sus respectivos cuadrados y el resultado será el siguiente:

$R^2 123 = 0,40 + 0,34 . (1 – 0,40) =$

$0,40 + 0,34 . 0,60 =$

$0,40 + 0,20 =$ **0,60**

De este ejemplo se puede concluir que “ingresos” explica en mucha mayor medida que la “edad”, las variaciones de la “cantidad de horas que lee diarios impresos en un mes”, siendo la proporción no explicada por “Ingresos” (k^2= 0,40)

4. A modo de síntesis

Como hemos visto en las páginas precedentes, el estudio de los fenómenos sociales requiere un abordaje que contemple su complejidad multidimensional. Es por esto que, para avanzar en su explicación, recurrimos a la realización de distintas técnicas de análisis multivariados.

Hemos visto también que es muy dificultoso en ciencias sociales aislar estos fenómenos reproduciendo situaciones típicas de laboratorio y detectar el grado de influencia que determinadas variables puedan tener en la relación original. Es por ello que la experimentación se re-elabora ex post facto, mediante el control de variables al momento del análisis.

Por último, cabe destacar que las distintas técnicas de análisis multivariado que hemos presentado según el nivel de medición de las variables utilizadas no son más que una primera aproximación a la temática.

Con distintos grados de complejidad y respondiendo a las hipótesis propuestas, existen también otras técnicas que nos permiten elaborar modelos multivariados de mayor complejidad (Modelo log-lineal, correspondencias múltiples, análisis discriminante, factorial, MANOVA, entre otros).

BIBLIOGRAFIA

- Archenti, Nélida: Cap. 16 “El proceso de análisis de tres variables categoriales”, en *Metodología de las Ciencias Sociales,* Marradi, A.; Archenti, N. y Piovani J. (comp.). Ed. Emecé. Buenos Aires. 2007
- Cohen, N. y Gómez Rojas, G. Cap VII. “Los objetivos, el marco conceptual y la Estrategia teórico-metodológica, triangulando en torno al problema de investigación”. *En torno de las metodologías: abordajes cualitativos y cuantitativos*, Lago Martínez, Gómez Rojas y Mauro (coord.), Proa XXI, Buenos Aires, 2003
- Cohen N., Di Virgilio M. y Martínez Mendoza R*. Correlación y regresión desde una perspectiva sociológica*. EUDEBA, Carrera de Sociología UBA. ISBN 950-23-0710.0 Buenos Aires, 1998.
- García Ferrando M. Cap. 9 "Medidas de asociación para variables de intervalo; Regresión y Correlación"; Cap. 12 "Estadística descriptiva III; tres o más variables" y Cap. 14 "Regresión y Correlación Múltiples. *El análisis de camino (parth análisis). Introducción a la Estadística en Sociología*. Editorial Alianza. España, 1985
- Hernández Sampieri, R. et. al. “Diseños experimentales en investigación”. *Metodología de la Investigación*. Ed. MacGraw-Hill. México, 1991
- Hyman, H.: “El modelo del experimento y el control de las variables” en M. Mora y Araujo et al: *El análisis de datos en la investigación social*, Nueva Visión, Buenos Aires, 1968.
- Lazarsfeld. Paul “La interpretación de las propiedades estadísticas como propiedad de investigación”. En Boudon, R. & Lazarsfeld, P. *Metodología de las ciencias sociales*. Tomo II: análisis empírico de la causalidad. Editorial Laia. (1966)
- Barcelona. Ponencia presentada en 1946 al Congreso de la Sociedad Americana de Sociología (ASA) en Cleveland.Mora y Araujo M. “El análisis de relaciones entre variables y la puesta a prueba de hipótesis sociológicas”, en Mora y Araujo (comp). *El Análisis de Datos en la Investigación Social*. Ed. Nueva Visión. Buenos Aires, 1984.
- Padua, J.: Cap. 2 “El proceso de investigación” (pp. 35a 45). *Técnicas de la investigación aplicada a las ciencias sociales*, Fondo de Cultura Económica, 1979.
- Sautú R.: Cap.1 “Formulación del objetivo de investigación” y Cap. 2 “El diseño de una investigación: teoría, objetivos y métodos*. Todo es Teoría. Objetivos y métodos de investigación*, Lumiere, 2005.
- Selvin H.C.: "El análisis multivariable *en El suicidio de Durkheim,.* Mora y Araujo M. Et.a Al., ob.cit. 1984.