IDICSO
Instituto de Investigación en Ciencias Sociales
Facultad de Ciencias Sociales
Universidad del Salvador

ÁREA EMPLEO Y POBLACIÓN

El análisis de correlación y regresión lineal entre variables cuantitativas

por **Horacio Chitarroni***

Buenos Aires, DIC/2002

* **CHITARRONI, Horacio.** Lic. en Sociología, Universidad Nacional de Buenos Aires (UBA). Docente, Facultad de Ciencias Sociales, Universidad del Salvador (USAL). Docente de la Maestría en Ciencias Sociales del Trabajo, Facultad de Ciencias Sociales, UBA. Investigador Principal, Área Empleo y Población, IDICSO, USAL. Consultor del Consejo Nacional de Coordinación de Políticas Sociales, SIEMPRO (Sistema de Evaluación, Seguimiento y Monitoreo de Programas Sociales).

BREVE HISTORIA DEL IDICSO. Los orígenes del IDICSO se remontan a 1970, cuando se crea el "Proyecto de Estudio sobre la Ciencia Latinoamericana (ECLA)" que, por una Resolución Rectoral (21/MAY/1973), adquiere rango de Instituto en 1973. Desde ese entonces y hasta 1981, se desarrolla una ininterrumpida labor de investigación, capacitación y asistencia técnica en la que se destacan: estudios acerca de la relación entre el sistema científico-tecnológico y el sector productivo, estudios acerca de la productividad de las organizaciones científicas y evaluación de proyectos, estudios sobre política y planificación científico tecnológica y estudios sobre innovación y cambio tecnológico en empresas. Las actividades de investigación en esta etapa se reflejan en la nómina de publicaciones de la "Serie ECLA" (SECLA). Este instituto pasa a depender orgánica y funcionalmente de la Facultad de Ciencias Sociales a partir del 19 de Noviembre de 1981, cambiando su denominación por la de Instituto de Investigación en Ciencias Sociales (IDICSO) el 28 de Junio de 1982.

Los fundamentos de la creación del IDICSO se encuentran en la necesidad de:

- Desarrollar la investigación pura y aplicada en Ciencias Sociales.
- Contribuir a través de la investigación científica al conocimiento y solución de los problemas de la sociedad contemporánea.
- Favorecer la labor interdisciplinaria en el campo de las Ciencias Sociales.
- Vincular efectivamente la actividad docente con la de investigación en el ámbito de la facultad, promoviendo la formación como investigadores, tanto de docentes como de alumnos.
- Realizar actividades de investigación aplicada y de asistencia técnica que permitan establecer lazos con la comunidad.

A partir de 1983 y hasta 1987 se desarrollan actividades de investigación y extensión en relación con la temática de la integración latinoamericana como consecuencia de la incorporación al IDICSO del Instituto de Hispanoamérica perteneciente a la Universidad del Salvador. Asimismo, en este período el IDICSO desarrolló una intensa labor en la docencia de post-grado, particularmente en los Doctorados en Ciencia Política y en Relaciones Internacionales que se dictan en la Facultad de Ciencias Sociales. Desde 1989 y hasta el año 2001, se suman investigaciones en otras áreas de la Sociología y la Ciencia Política que se reflejan en las series "Papeles" (SPI) e "Investigaciones" (SII) del IDICSO. Asimismo, se llevan a cabo actividades de asesoramiento y consultoría con organismos públicos y privados. Sumándose a partir del año 2003 la "Serie Documentos de Trabajo" (SDTI).

La investigación constituye un componente indispensable de la actividad universitaria. En la presente etapa, el IDICSO se propone no sólo continuar con las líneas de investigación existentes sino también incorporar otras con el propósito de dar cuenta de la diversidad disciplinaria, teórica y metodológica de la Facultad de Ciencias Sociales. En este sentido, las áreas de investigación del IDICSO constituyen ámbitos de articulación de la docencia y la investigación así como de realización de tesis de grado y post-grado. En su carácter de Instituto de Investigación de la Facultad de Ciencias Sociales de la Universidad del Salvador, el IDICSO atiende asimismo demandas institucionales de organismos públicos, privados y del tercer sector en proyectos de investigación y asistencia técnica.

IDICSO

Departamento de Comunicación

Email: idicso@yahoo.com.ar

Web Site: http://www.salvador.edu.ar/csoc/idicso

El análisis de correlación y regresión lineal entre variables cuantitativas

Cuando se debe explorar la relación entre dos escalas cuantitativas (de intervalos iguales o de razones), es frecuente acudir al coeficiente de correlación lineal r de Pearson, también denominado coeficiente *producto-momento*. El concepto de correlación puede interpretarse como sinónimo de asociación entre las puntuaciones que asumen los casos en ambas variables, covariación o variación conjunta. Una medida que expresa la variación de una variable cuantitativa es su varianza: si decimos que dos variables covarían, esto significa que tienen una parte de sus varianzas en común.[1]

En la medida en que los puntajes de dos variables guarden alguna relación (como sería el caso de la estatura y el peso de las personas: en general, a medida que se es más alto, también se pesa más[2]), conociendo el valor que asume un caso cualquiera en una de las variables existirá la posibilidad de predecir su puntuación en la otra: esta predicción será tanto más precisa y segura cuanto más estrecha sea la relación entre ambas variables: si el peso no variara sino en función de la estatura (lo que implicaría que todas las personas de igual estatura pesarían lo mismo), entonces, bastaría con conocer la talla de alguien para predecir su peso sin temor a error alguno. El instrumento estadístico que permite hacer este tipo de estimaciones se denomina regresión lineal.

La correlación lineal

Tales casos –en que las varianzas de dos variables estén tan estrechamente asociadas de manera que solo haya un único valor de una de ellas para cada valor de la otra– no son empíricamente usuales. Sin embargo, en términos teóricos y si consideramos las variaciones de dos variables, expresadas en términos de sus varianzas o de sus desviaciones estándar, podemos imaginar tres situaciones:

- Que estas varianzas sean por entero independientes (esquema 1)
- Que estas varianzas tengan alguna parte en común: o sea que exista alguna variación conjunta o covarianza entre las variables (esquema 2)
- Que toda la variación de las dos variables sea variación conjunta (esquema 3)

ESQUEMAS DE COVARIANZAS

Esquema 1 Esquema 2 Esquema 3

[1] Se recordará que la varianza es la sumatoria de las desviaciones a la media elevadas al cuadrado y promediadas. La parte de arriba del cociente se denomina suma de cuadrados.
[2] Aunque, obviamente, no es la talla el único determinante de las diferencias en el peso…

De estas situaciones, la más frecuente y la que resulta más relevante desde el punto de vista del análisis, es la segunda: hay algo de varianza en común. Un ejemplo permitirá aclarar más esta noción: si computamos las tasas de actividad y las tasas de empleo de la Argentina durante los últimos diez años (o bien para todas las provincias argentinas en un año dado), se apreciará que a medida que aumenta la tasa de actividad, también tiende a aumentar el empleo.[3] Sin embargo, el empleo no se incrementa solamente porque más gente pretenda trabajar: puede ser que en condiciones de expansión de la economía –y sin que medien incrementos en la propensión a trabajar– la ocupación crezca porque se están creando puestos de trabajo que absorben desempleados preexistentes. De igual manera, podrían tener lugar incrementos en la fuerza de trabajo potencial sin que se altere la tasa de empleo, si esos incrementos no son absorbidos por un sostenido aumento de la demanda empresaria de mano de obra. Sería, pues, el caso en que las varianzas de ambas variables son parcialmente comunes: la zona sombreada del esquema 2 nos muestra esa covariación.

Una de las lecturas del coeficiente de correlación r de Person permite interpretarlo como una medida de covarianza o variación común de las variables en términos de su máxima variación total. Se trataría de un cociente cuyo numerador es la covarianza o variación común (la zona sombreada) y su denominador es la variación total de ambas variables (cuya expresión es el producto y no la suma de sus varianzas):

r = covarianza XY/(varianza X * varianza Y)

Si toda la varianza fuese común (esquema 1) el valor del coeficiente sería igual a la unidad. En cambio, si no existiera zona común, el coeficiente r valdría cero. Obviamente, entonces, el valor empírico de r ha de variar en un rango de cero a uno, según la proximidad a una u otra situación. Y la elevación del coeficiente r al cuadrado permite obtener el llamado coeficiente de determinación r^2: este coeficiente indica qué proporción de la varianza de una de las variables resulta explicada por la variación de la otra[4] (por ejemplo, qué proporción de la variación del peso depende de la variación de las estaturas). Así, un valor de r = 0,70 implicaría un de 0,49: alrededor de la mitad de la variación del peso dependería de la variación de las estaturas, en tanto que el 51% restante se explicará, seguramente, por otros factores: el ancho de los huesos, la masa muscular, el tejido adiposo, etc. Estos factores determinarán una variación adicional de los pesos (no explicada por las estaturas), que hará que no todos quienes miden lo mismo pesen igual. Esta varianza no explicada es medida por el coeficiente de no determinación $k^2 = 1 - r^2$.

Pero además, el coeficiente r de Pearson tiene otro atributo: puede asumir valores positivos o negativos (el signo es independiente del valor absoluto). Un

[3] Lo que resulta lógico, puesto que más gente está dispuesta a ingresar al mercado de trabajo y algunos de ellos consiguen hacerlo.
[4] Explicación, en términos estadísticos, no es equivalente a causalidad.

valor de r positivo significa que ambas escalas se correlacionan positivamente: a medida que aumenta la tasa de actividad tiende a incrementarse el empleo. En cambio, un valor de r negativo implica correlación inversa: cuando se incrementa la tasa de desempleo tienden a descender los salarios (habrá más gente dispuesta a trabajar por ingresos bajos).

La regresión lineal

Supongamos que la talla media de la población adulta fuera, en nuestro país, 1, 68 m, en tanto que el peso promedio alcanzara a 65 kg. Si alguien nos pidiera que estimáramos el peso de un individuo cualquiera sin más datos que esos, la mejor "predicción" que podríamos hacer sería, precisamente, estimarle un peso coincidente con la media de la población. Sería una gran casualidad que acertáramos y –seguramente– incurriríamos en un importante error de estimación.

Ahora bien, si alguien se hubiera tomado, previamente, el arduo trabajo de pesar y medir a cada uno de los habitantes adultos, habría comprobado que, aunque no todos los que tienen igual talla pesan lo mismo (hay una dispersión de pesos en torno a cada talla), las medias de peso para cada estatura diferirían entre sí. Y si conociéramos esas medias de pesos para cada talla y contáramos con alguna información adicional, por ejemplo si supiéramos que el sujeto mide 1,80, seguramente no estimaríamos el promedio general sino la media de peso para esa estatura, que resultaría en un peso bastante superior: por ejemplo, 80 kg. Tampoco allí acertaríamos, pero seguramente nuestra estimación será mucho más cercana que la anterior. Obviamente, esta posibilidad de "mejorar" la predicción existe porque ambas variables –peso y estatura– se hallan correlacionadas. Y, como se ha dicho al comienzo, si esa correlación fuera tan estrecha que todos los que tienen igual talla también tuvieran igual peso, las predicciones estarían exentas de error.

Sin embargo, pesar y medir a todos sería muy trabajoso, de manera que no resultaría factible conocer tales promedios de peso por cada estatura. Entonces, por una cuestión de mera practicidad, es posible imaginar que pesos y estaturas (o cualesquiera otras variables: por ejemplo las tasas de actividad y las tasas de desempleo, o bien los años de educación formal y los ingresos laborales) ajustan su relación a una función lineal. Esto quiere decir que si hubiéramos hecho tales mediciones y representáramos las medias de peso para cada estatura (o bien las medias de estatura para cada peso) sobre una ordenada y una abscisa, estas medias seguirían el curso de una recta. A esta recta, que se llamaría "recta de regresión de Y sobre X", porque se usa para predecir valores de Y basándose en valores de X) correspondería una ecuación lineal:

$$Y = a + b*X + e$$

ESQUEMA DE RECTA POSITIVA

En esta ecuación Y es el valor de un caso en la variable Y, que queremos predecir (por ejemplo, el peso de alguien), en tanto que X es el valor de ese caso en la variable X, que nos es conocido y tomamos como base para la predicción (por ejemplo, la estatura de alguien). ¿Qué cosa son los estadísticos a y b? El primero se denomina la "constante" o la "ordenada al origen" y puede interpretarse como el valor de Y cuando X vale cero: sería el punto en que la recta corta al eje vertical (aunque resulte absurdo, para seguir con el ejemplo, sería el peso de alguien cuya estatura es cero). En cuánto al coeficiente b, que se denomina la "pendiente" de la recta, sería el incremento (o, al revés, la disminución) que experimentan los puntajes de Y cada vez que X aumenta en una unidad. Así, si cada centímetro de estatura significara medio kilo adicional de peso, b valdría 0,50. Obviamente, cuando la correlación es negativa (cuando el coeficiente r tiene signo negativo), también b es negativo. Y la recta que mejor representa la relación entre las variables tiene inclinación inversa, como se aprecia en el esquema.

Finalmente, el término **e** representa el error en las predicciones: este error obedece a las variaciones de Y no producidas por X.

ESQUEMA DE RECTA NEGATIVA

Ahora bien, se ha dicho que hacemos el supuesto de que las medias de Y para cada X seguirían el curso de una recta, que usaremos para predecir. ¿Pero cuál recta hemos de usar?. Seguramente, no ha de ser cualquiera. Supongamos que tomamos una muestra de un centenar de personas adultas de la población de la Argentina y –a ellos sí– los pesamos y medimos. Y luego representamos las estaturas sobre el eje horizontal y los pesos sobre el eje vertical. Si trazamos los puntos correspondientes a las observaciones, tendremos lo que se da en llamar una "nube de puntos" o "diagrama de dispersión". La simple inspección visual de este diagrama nos permitirá intuir si la relación entre las variables se ajusta aceptablemente a una función recta. Ello dependerá de si podemos imaginar una recta que pase relativamente cerca de la mayoría de los puntos. La mejor recta de todas será la que cumpla la condición de minimizar la suma de las distancias medidas desde los puntos a la recta, elevadas al cuadrado. Por eso, se llama también "recta de cuadrados mínimos" o de "mejor ajuste".

Obviamente, en cualquier situación sería posible encontrar una recta de mejor ajuste, que minimice dichas distancias. Pero si aún esta recta deja muy lejos a gran parte de las observaciones, no nos serviría para hacer predicciones adecuadas: ello significa que la relación entre las variables no se ajusta bien a una función de esta clase. Pues bien: el coeficiente r de Pearson puede interpretarse asimismo como una medida de dicho ajuste. Si el coeficiente vale 1 (o bien –1) esto quiere decir que ese ajuste es máximo: todos los puntos se situarían sobre la recta.[5]

[5] Con lo que el error de estimación e, en la ecuación, sería cero.

Evidentemente, si r = 1 (independientemente su signo y de los valores de a y b) la totalidad de las observaciones caerían exactamente sobre la recta. Y pensaríamos que, salvo una enorme casualidad, ello se debe a que en la Argentina toda la gente de cierta talla pesa igual y a que esa relación es perfectamente lineal, razones por las cuales, al obtener una muestra, hemos hallado unos valores que se ajustan totalmente a una recta.

Seguramente no será así. Los puntos se dispersarán en torno a la recta, pero hallaremos de todas formas la mejor recta posible. Claro que esta recta la obtenemos a partir de las observaciones muestrales, de modo que los a y b correspondientes serán, en realidad, estimadores de los verdaderos coeficientes α y β poblacionales.

Las distancias entre los puntos y la recta (que esta recta de cuadrados mínimos permite achicar) pueden ser vistas como los errores de estimación de Y sobre X.[6] Y serían causadas por la parte de la varianza de Y que no resulta explicada por X (puesto que no habría estas dispersiones, si Y sólo variara en función de X).

Todos los razonamientos anteriores pueden invertirse cuando se trata de estimar X a partir de Y (la estatura conociendo el peso). En este caso, la ecuación será:

$$X = a + b*Y + e$$

Aquí, a sería la abscisa al origen (si estamos representando las estaturas sobre el eje horizontal), vale decir el punto donde la recta intersectará con dicho eje. Y b será el incremento de estatura para cada incremento de una unidad de peso. Estos coeficientes corresponderán a la "recta de regresión de X sobre Y" (que predice valores de X basándose en valores conocidos de Y). Cuando r = +-1 ambas rectas de mínimos cuadrados coinciden. De lo contrario, hay dos rectas, y en este caso, las predicciones no son simétricas. Esto quiere decir que si la predicción del peso para una persona de 1,80 de estatura resulta 80 kg., la predicción inversa: la estatura de alguien que pesa 80 kilos no resultará 1,80 m. (sino algo más o algo menos, pero la diferencia sería tanto más reducida cuanto mayor sea el valor de r).

Ahora bien, se ha explicado que a y b son estimadores de los verdaderos parámetros α y β. Y en rigor, el valor de r calculado en base a datos muestrales, también será un estimador del verdadero coeficiente. No puede dejar de considerarse la posibilidad de que solamente existiera cierta relación lineal en la muestra, por obra de la casualidad, sin que la hubiera en la población. ¿Cómo saberlo?. A partir de la correlación y la regresión pueden calcularse dos pruebas de significación: t de Student y F de Snedecor (F es, en realidad, el cuadrado de t), que permiten contrastar dos hipótesis de nulidad estrechamente vinculadas:

[6] De hecho, si las elevàramos al cuadrado, las promediáramos y obtuviéramos la raíz cuadrada de ese promedio, tendríamos el error estándar de las estimaciones de Y sobre X. Si no hubiera tales distancias, error estándar sería cero, de manera que podríamos obtener estimaciones totalmente certeras. Para ello, r debiera valer 1 o bien –1.

- la primera afirma que r = 0 en la población (no hay relación lineal alguna)
- la segunda afirma que b = 0 en la población (los incrementos de una variable no producen efecto alguno en la otra)

$$F = r^2 * (n - 1) / (1 - r^2)$$

Con un valor de F (o de t) lo suficientemente grande, es posible rechazar las hipótesis nulas. Obviamente, F (y t) serán tanto más elevados a medida que r^2 sea mayor. Pero conviene observar que, aún con moderados valores de r^2, podríamos obtener valores grandes de F (y la significación estadística necesaria para rechazar las hipótesis nulas) toda vez que la muestra sea lo suficientemente grande.[7] Esto no debiera sorprendernos: si hubiéramos observado cierta correlación entre peso y estatura sobre una muestra de 5 o 10 personas, aún cuando hubiera sido perfecta, podríamos dudar de que fuera así en la población total. Pero si nuestra muestra hubiera incluido un millón de personas, podríamos desdeñar la influencia del azar.[8]

Condiciones o supuestos del modelo

Por cierto que, como la mayor parte de los modelos estadísticos empleados, el uso paramétrico de este modelo de la regresión lineal conlleva exigentes supuestos. Además del ajuste a la linealidad y el uso de escalas de intervalos, debieran cumplirse en la población las condiciones propias de la distribución normal bivariada. Ello significa que, para cada valor de X las Y debieran distribuirse en forma normal y con similar varianza. Y, de igual modo, para cada valor de Y debieran distribuirse las X en forma normal y con varianzas semejantes.

Para que esto fuera cierto, al menos debiera verificarse que las distribuciones muestrales de las variables no se alejaran en exceso de la normalidad y que sus varianzas no fueran demasiado diferentes. Lo primero puede constatarse pos simple inspección del histograma o someterse a prueba empleando un test no paramétrico[9] que permite determinar si una distribución se aleja significativamente de la normalidad. En lo que respecta a las varianzas, existe el test de Levene, pero el SPSS no lo calcula. En este caso, la comparación de los coeficientes de variación de las dos variables (no influidos por las respectivas escalas) puede orientarnos.

Hay algunas cosas que pueden hacerse si las variables no se adecuan a estos requisitos. La falta de linealidad, así como el apartamiento de la normalidad y el exceso de varianza, por ejemplo, pueden mejorar apreciablemente si se

[7] Puesto que el tamaño muestral n figura en el numerador del cociente.
[8] Esto es así independientemente del tamaño de la población. Si sobre diez tiradas de moneda obtenemos siete caras, pensaríamos que es casual. En tanto que si obtenemos 700 caras sobre mil tiradas, ya no creeríamos en tal casualidad: la moneda está "cargada". La "población" de tiradas puede imaginarse, sin embargo, como infinita.
[9] Se trata de la prueba de Kolmorov-Smirnov.

sustituye una variable por su logaritmo. A veces, también la raíz cuadrada puede producir efectos semejantes.

Otras aplicaciones

Hemos empleado el ejemplo de pesos y estaturas porque, aunque burdo, se presta para una fácil comprensión. En ese ejemplo, las variables que se correlacionan están medidas de individuos. Pero también podrían haberse medido en el tiempo, a medida que una sola persona crece y aumenta de peso. En base a la correlación así, podríamos predecir cuánto pesará un niño cuando haya alcanzado la estatura de 1,60.

Pero vimos que podríamos hacer algo más: considerar al tiempo como una variable. Si anotamos los pesos (o las estaturas) que va alcanzado el niño año a año, podremos hallar una recta de regresión y un coeficiente b que estimará el incremento de peso (o de talla) por cada incremento de tiempo (por ejemplo, por cada año). De ese modo podríamos hacer una predicción para cuando tenga, por ejemplo, 13 años. Por cierto que estas predicciones serían defectuosas, porque justamente el crecimiento no es lineal: los niños dan "estirones" a ciertas edades. Pero tampoco sería tan disparatada.

Si ahora trasladamos estos razonamientos al mercado de trabajo, veremos que sería posible estimar el empleo conociendo, por ejemplo, la tasa de actividad[10]. Esto podríamos hacerlo computando ambas tasas para un conjunto de países o de provincias. Pero también, considerando la evolución de ambas tasas en la Argentina. De hecho, cuando se dice que si la economía creciera 6% anual en el próximo quinquenio el empleo ascendería cierta cantidad de puntos, se está calculando una regresión de ese tipo. Y si una variable –por ejemplo la tasa de actividad femenina– mostrara una tendencia relativamente lineal a lo largo de cierto período, podríamos hacer lo mismo que con el crecimiento del niño: calcular una regresión que estimara el incremento (o la disminución) de esa tasa por cada unidad de tiempo, a efectos de estimar su valor dentro de una década.

Un ejemplo

Dejaremos ahora de lado las cuestiones antropométricas, para proporcionar un ejemplo más vinculado a la temática que aquí interesa.

Las variables usadas en el trabajo práctico de operacionalización, medidas para las provincias argentinas, nos serán de utilidad.

Por ejemplo, podemos explorar la relación entre las tasas de actividad y empleo. ¿Cuándo de la variación del empleo dependerá de la variación de la propensión de la gente a insertarse en la actividad económica?

[10] Aunque sólo sería razonable hacerlo si el valor de *r* fuera alto.

Correlations

		TASACT	TASEMPL
TASACT	Pearson Correlation	1,000	,880*
	Sig. (2-tailed)	,	,000
	N	23	23
TASEMPL	Pearson Correlation	,880**	1,000
	Sig. (2-tailed)	,000	,
	N	23	23

**. Correlation is significant at the 0.01 level (2-tailed).

La primera tabla de resultados obtenidos con el SPSS nos muestra la matriz de correlaciones entre ambas variables. El coeficiente de correlación es alto y positivo (0,88). Y el nivel de significación (0,000) nos indica que podemos rechazar la hipótesis nula (que el coeficiente fuera cero en la población) con una probabilidad de error menor a 0,1%, vale decir una confianza superior a 99,9%. En rigor, aquí no estamos empleando una muestra, de modo que diríamos que ese coeficiente podría obtenerse azarosamente menos de una de cada mil veces. Los valores 1 en la matriz indican los coeficientes de correlación de cada variable con sí misma.

Al correr la regresión se obtiene la siguiente tabla:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,880[a]	,775	,765	1,7639

a. Predictors: (Constant), TASACT

Aparece nuevamente el valor del coeficiente r (0,88) y el coeficiente de determinación r cuadrado: 0,75 (el 77% de la varianza de las tasas de empleo resultaría explicado por la variación de la tasa de actividad).

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	225,340	1	225,340	72,428	,000
	Residual	65,336	21	3,111		
	Total	290,677	22			

a. Predictors: (Constant), TASACT

b. Dependent Variable: TASEMPL

Esta nueva tabla muestra las sumas de cuadrados explicadas por la regresión, no explicadas y total[11]. Si calculáramos aquí el cociente entre las varianzas explicada (de la regresión) y total, obtendríamos el valor de r cuadrado (0,77). También aparece el valor de F de Snedecor y lasignificación con que podemos rechazar la hipótesis nula.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,028	3,664		,553	,586
	TASACT	,795	,093	,880	8,510	,000

a. Dependent Variable: TASEMPL

Finalmente, tenemos otra tabla que muestra el coeficiente b = 0,795 (cada vez que la tasa de actividad aumenta un punto, el empleo aumenta, en promedio, 0,79 puntos). Y la constante del modelo es la **a** de la ecuación: la ordenada al origen. Adicionalmente, tenemos el valor de la t de Student correspondiente al coeficiente b (8,51) y su significación: podemos rechazar la hipótesis nula, que diría que beta es igual a cero en la población, con más de 99,9% de confianza (para este caso, en que no se trata de datos muestrales, diremos que este valor de **b** no se obtendría en forma casual sino menos de una de cada mil veces, lo que suena bastante convincente...)

Un segundo ejemplo será la posible relación entre informalidad e ingresos laborales. Se supone que al aumentar la informalidad, los ingresos disminuirán, puesto que las remuneraciones son menores en el sector informal.

Correlations

		INFORMAL	INGLAB
INFORMAL	Pearson Correlation	1,000	-,671*
	Sig. (2-tailed)	,	,000
	N	23	23
INGLAB	Pearson Correlation	-,671**	1,000
	Sig. (2-tailed)	,000	,
	N	23	23

**. Correlation is significant at the 0.01 level (2-tailed).

[11] Estos conceptos ya fueron vistos en relación con el análisis de varianza.

Aquí se observa que, efectivamente, la correlación es considerable y negativa. El coeficiente r de Pearson es –0,67 y resulta significativo al nivel de 0,01 (en rigor, la significación es tan elevada como en el caso anterior, por lo que nuevamente rechazaríamos la hipótesis nula: estos resultados no podrían ser casuales sino menos de una de cada mil veces).

Podría decirse que los ingresos laborales no son una variable que se distribuya en forma normal. Efectivamente, si calculamos la prueba no paramétrica de Kolmogoro-Smirnov para estas variables, veríamos que en el caso de la informalidad, podemos aceptar la hipótesis nula de que la distribución no se aparta de la normalidad (el error sería de 76% si la rechazáramos). Pero en el caso del ingreso laboral esto es dudoso: esta hipótesis nula podría rechazarse con menos de 10% de probabilidad de error.

One-Sample Kolmogorov-Smirnov Test

		INFORMAL	INGLAB
N		23	23
Normal Parameters[a,b]	Mean	47,4217	500,4783
	Std. Deviation	7,2123	169,3494
Most Extreme Differences	Absolute	,140	,259
	Positive	,106	,259
	Negative	-,140	-,174
Kolmogorov-Smirnov Z		,671	1,243
Asymp. Sig. (2-tailed)		,759	,091

a. Test distribution is Normal.

b. Calculated from data.

Como solución puede sustituirse la variable ingreso laboral por su logaritmo. La correlación no variaría sustancialmente:

Correlations

		INFORMAL	LOGING
INFORMAL	Pearson Correlation	1,000	-,678*
	Sig. (2-tailed)	,	,000
	N	23	23
LOGING	Pearson Correlation	-,678**	1,000
	Sig. (2-tailed)	,000	,
	N	23	23

**. Correlation is significant at the 0.01 level (2-tailed).

Y el test de Kolmogorov-Smirnov para esta última variable mostrará que ahora ya no se aparta de la normalidad. Asimismo, podemos ver que ha disminuido su

desviación estándar, que en la variable original era bastante mayor que la de la informalidad. Quiere decir que la transformación también contribuyó bastante a producir homocedasticidad.[12]

One-Sample Kolmogorov-Smirnov Test

		LOGING
N		23
Normal Parameters[a,b]	Mean	6,1722
	Std. Deviation	,2848
Most Extreme Differences	Absolute	,213
	Positive	,213
	Negative	-,124
Kolmogorov-Smirnov Z		1,021
Asymp. Sig. (2-tailed)		,249

a. Test distribution is Normal.

b. Calculated from data.

La correlación parcial

La existencia de correlación empírica entre dos variables, sin embargo, no nos muestra la relación "pura" entre ellas. Es usual que otras variables –ajenas al modelo bivariado– estén influyendo de distintas maneras en la correlación original. Por ejemplo, podría haber una tercera variable que causara las variaciones de las variables originales (las "hiciera mover" hacia arriba o hacia abajo), con lo que las "forzaría" a correlacionarse entre sí. Si esto sucediera, estaríamos viendo una correlación acaso inexistente (o superior a la real). Y en caso de que esa tercera variable cesara de ejercer sus efectos, dicha correlación tendería a desaparecer.

Algo similar ocurriría si la relación entre dos variables (X e Y) estuviera "intermediada" por una tercera variable Z. En este caso, si toda la acción de X sobre Y se produjera a través de Z, en caso de que Z dejara de variar también debiera esperarse la desaparición (o al menos la atenuación) de la relación original.

En cambio, supongamos que dos variables X e Y mostraran una correlación positiva muy baja: una tercera variable Z que "hiciera aumentar" a X e "hiciera bajar" a Y estaría impidiendo que ambas se correlacionaran positivamente: si pudiéramos suprimir su acción, entonces sería de esperar que la correlación entre ambas aumentara.

[12] Si calculáramos los coeficientes de variación respectivos (dividiendo los desvíos estándar por las medias) veríamos que el ingreso laboral tenía una variabilidad considerablemente más alta que la informalidad. En cambio, esto no sucede en el caso del logartimo del ingreso.

Si, por el contrario, X e Y hubieran mostrado una correlación negativa, podría ocurrir que dicha correlación inversa fuera producida –al menos en parte– por una variable Z que hiciera lo mismo que en el caso anterior: producir aumentos en los puntajes de Y y disminuciones en los de X: si así fuera, esta variable "ayudaría" a X e Y a mantener una correlación negativa. Suprimido su efecto, esta correlación negativa se reduciría o desaparecería.

Supongamos que una variable Z se correlacionara con Y pero no con X. Sería el caso de una variable ajena a X que explicaría parte de la varianza de Y. Si así fuera, esta variable Z formaría parte del término de error **e** en la ecuación lineal (Y = a+b*X+e). En este caso, si eliminamos esa fuente de varianza inexplicada, sería de esperar que las predicciones de Y desde X fueran más ajustadas (y que *r* aumentara).[13]

La correlación parcial posibilita este tipo de controles. Permite "descontar" de la correlación entre dos variables, X e Y, los posibles efectos de una o más variables de control (Z, T, W). Esto puede hacerse "paso a paso" (controlando las variables de a una) o bien se pueden controlar un conjunto de variables en forma simultánea. Lo primero resulta más aconsejable, porque permite comprender mejor el posible entramado de relaciones entre el conjunto de variables consideradas. El coeficiente de correlación parcial $r_{xy.z}$ debe ser interpretado como la correlación que queda entre X e Y una vez suprimidos los efectos de Z.[14]

Será de utilidad ver un ejemplo, basado en los mismos datos que ya hemos empleado. Como se ha visto, los aumentos de la tasa de actividad generan incrementos en la tasa de empleo. Esta correlación era alta y positiva. Sin embargo, el desempleo es una fuente de variación de la tasa de empleo: la hace bajar.[15] Y, además, el desempleo está afectado por la tasa de actividad: en este caso, la correlación es positiva.[16] ¿Qué sucederá, pues, con la correlación entre tasas de actividad y empleo si eliminamos los efectos del desempleo: *ceteris paribus*?[17] Es presumible que la relación original se mostraría más fuerte. Veámoslo:

[13] Supongamos que calculamos el coeficiente de correlación entre los años de educación formal de las personas y los ingresos laborales que obtienen. Este coeficiente sería moderadamente positivo, pero los ingresos varían también en función de otras variables: por ejemplo, hay quienes disponen de mejores redes sociales que les permiten obtener empleos más ventajosos, a igualdad de calificaciones educativas. Si fuera posible "neutralizar" los efectos de estas redes sociales, la correlación entre educación e ingresos aumentaría.

[14] Si lo eleváramos al cuadrado, obtendríamos un coeficiente de determinación que expresaría la varianza de Y explicada por X (o de X explicada por Y) una vez eliminada la influencia de Z.

[15] Efectivamente, si se destruyen puestos de trabajo bajará la tasa de empleo aunque no disminuya la tasa de actividad.

[16] Cuando aumenta la tasa de actividad, si no se crean suficientes puestos de trabajo, habrá más desocupados.

[17] En realidad, para apreciar la relación "pura" debiéramos ir controlando todas las demás variables que, conforme al modelo teórico adoptado, pudieran estar influyendo: el crecimiento económico, el nivel de los salarios, la intensidad en el uso de tecnología, etc.

```
P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N

Controlling for..    TASDESOC

               TASACT    TASEMPL

TASACT          1,0000      ,9993
               (    0)    (   20)
               P= ,       P= ,000

TASEMPL           ,9993     1,0000
               (   20)    (    0)
               P= ,000    P= ,
```

Efectivamente, descontando los efectos del desempleo, las variables originales intensifican su correlación, que ahora es casi perfecta (r = 0,999). Es lógico: de no haber desempleo, toda la variación de la tasa de empleo estaría determinada por la tasa de actividad.[18] Este coeficiente arroja, además, elevada significación (0,000).

La correlación y la regresión múltiples

Ya se ha visto que las ecuaciones de regresión bivariadas (del tipo Y = a + b*X + e) incluyen un término de error, determinado por las variaciones de Y que no están producidas por X sino por otras variables (Z, T, W, etc.). Pues bien: si se incorporan estas otras variables a una ecuación de regresión (vale decir, si usamos un conjunto de variables independientes en lugar de una sola), será de esperar que ese término de error se reduzca y las estimaciones de Y resulten más precisas. Una ecuación multivariada con n variables independientes sería de la forma:

$$Y = a + b_1*X + b_2*Z + b_3*T + e$$

En esta ecuación, X, Z, T son las distintas variables independientes. En tanto que b_1, b_2, etc. son los respectivos coeficientes que revelan las respectivas influencias o pesos de estas variables independientes sobre Y. Estos coeficientes pueden ser positivos o negativos: algunas de estas variables harán aumentar a Y, mientras que otras reducirán sus valores.

Supongamos que utilizamos una ecuación de este tipo para predecir los valores de Y. Estas predicciones serán, seguramente, mejores que si sólo empleáramos X, pero no serán exactas (quedarán otras variables no incorporadas al modelo: reduciremos el factor **e** sin llegar a eliminarlo). Si luego calculáramos un coeficiente de correlación entre los valores de Y resultantes de esta predicción y

[18] En efecto: aunque se expandiera la economía y creciera la demanda de mano de obra, en ausencia de desempleados sólo sería posible que aumentara el empleo si más gente se mostrara dispuesta a trabajar.

los valores realmente observados, obtendríamos el coeficiente de correlación múltiple entre la variable Y y las variables independientes X, Z y T.

Si eleváramos al cuadrado dicho coeficiente, tendríamos el coeficiente de determinación múltiple $R^2_{y.xzt}$, que nos dirá qué proporción de la varianza de Y explican simultáneamente todas las variables independientes incorporadas.

Debe tenerse en cuenta que este R^2 múltiple no es la simple sumatoria de cada uno de los r^2 individuales de las variables independientes con Y. Ello es así, porque las variables independientes también suelen estar correlacionadas entre sí y algunas de ellas explican la misma parte de la varianza de Y: vale decir, superponen sus efectos. La correlación múltiple elimina o descuenta estas duplicaciones.

De lo anterior se desprende que, cuando usamos variables independientes muy correlacionadas entre sí,[19] es presumible que no lograremos aumentar en demasía la varianza explicada de Y, puesto que superpondrán sus efectos y todas explicarán la misma parte de la varianza. En cambio, si las variables independientes no están correlacionadas, cada una de ellas será responsable de una parte diferente de la varianza de Y, de modo que lograrán adicionar nuevas proporciones de varianza explicada.

Exigencias del modelo

Por cierto, tanto la correlación parcial como la correlación múltiple plantean iguales requisitos que la correlación y regresión simples: normalidad y homocedasticidad. En la medida en que las variables se aparte mucho de ellos, no sería recomendable incorporarlas a la ecuación. Pero siempre existe la posibilidad de realizar algunas de las transformaciones a las que se ha aludido previamente (por ejemplo, logarítmicas).

Un ejemplo

A modo de ejemplo, podemos calcular la correlación múltiple entre la tasa de empleo (como variable dependiente) y las tasas de actividad y desempleo como variables predictoras o independientes.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,999[a]	,999	,999	,1374

a. Predictors: (Constant), TASDESOC, TASACT

Como era de esperar, aquí logramos explicar la totalidad de la varianza. Tanto r como su cuadrado son prácticamente iguales a uno.

[19] Este efecto se llama, en términos estadísticos, *multicolinealidad*.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	290,299	2	145,149	7689,483	,000
	Residual	,378	20	1,888E-02		
	Total	290,677	22			

a. Predictors: (Constant), TASDESOC, TASACT

b. Dependent Variable: TASEMPL

La segunda tabla permite ver las sumas cuadráticas: la correspondiente a la regresión es casi igual a la total. El valor de F es elevado y altamente significativo.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5,789	,293		19,789	,000
	TASACT	,851	,007	,943	115,988	,000
	TASDESOC	-,390	,007	-,477	-58,662	,000

a. Dependent Variable: TASEMPL

Y, finalmente, tenemos la tabla con los coeficientes b de ambas variables predictoras (además de la constante u ordenada al origen). Cuando la tasa de actividad crece un punto, la tasa de empleo lo haría en 0,85. En tanto que al crecer un punto el desempleo, la tasa de empleo bajaría 0,39 puntos. La significación de t de Student no permite abrigar dudas: estos coeficientes b no pueden ser iguales a cero en la población (en nuestro caso, la probabilidad de que aparecieran por casualidad sería menor a una de cada mil veces).

El caso de las variables *dummy*

A veces, para predecir los valores de una variable y explicar su varianza necesitamos incorporar al modelo de regresión múltiple variables independientes que no son numéricas sino categóricas. Supongamos, por ejemplo, que queremos predecir los ingresos[20] de las personas. Si usamos como predictoras los años de educación y la edad, tal vez nos resulte importante incorporar variables que no son cuantitativas: por ejemplo, la categoría

[20] En rigor, no es un ejemplo adecuado, porque los ingresos no se distribuyen normalmente. Sin embargo, podríamos acudir a una transformación logarítmica.

ocupacional (empleador, cuentapropista, asalariado, trabajador sin remuneración). En este caso, esta variable podría ser incorporada a la ecuación si se la transforma en una variable *dummy* (simulada). Ello consiste en generar n – 1 variables dicotómicas con valores cero y uno, siendo n el número de categorías de la variable original.

Para el caso de la variable categoría ocupacional, la transformación sería la siguiente:

Categoría ocupacional	Variables dummy		
	Empleador	Cuenta propia	Asalariado
Empleador	1	0	0
Cuenta propia	0	1	0
Asalariado	0	0	1
Trabajador sin remuneración	0	0	0

Crearíamos tres variables dicotómicas: la primera de ellas sería "Empleador". Quien lo sea tendrá valor 1 en esa variable y valor cero en las variables "Cuenta propia" y "Asalariado". Los cuentapropistas tendrán valor 1 en la segunda variable y cero en las otras, etc. No necesitamos crear, en cambio, una variable llamada "Trabajador sin remuneración": lo será quien tenga valores cero en las tres anteriores. Esta última es la categoría "base" de las *dummy*.[21]

Una vez realizada esta transformación, estas variables pueden ser incorporadas en una ecuación de regresión: sus valores sólo pueden variar entre cero y uno[22] y sus coeficientes b indicarán, en cada caso, cuanto aumentan o disminuyen los ingresos cuando una de estas variables pasa de cero a uno (por ejemplo, cuando alguien es un empleador).

Buenos Aires, DIC/2002

por **Horacio Chitarroni**
Investigador Principal, Área Empleo y Población, IDICSO, USAL.
Email: hchitarroni@siempro.gov.ar

[21] Obviamente, podríamos haber definido como base cualquiera de las cuatro categorías.
[22] Al haber un solo intervalo, no puede haber intervalos desiguales. Son, pues, "variables de intervalos iguales".