REGRESIÓN LINEAL

Jorge Dagnino S.[1]

- La regresión lineal permite predecir el comportamiento de una variable (dependiente o predicha) a partir de otra (independiente o predictora).
- Tiene presunciones como la linearidad de la relación, la normalidad, la aleatoridad de la muestra y homogeneidad de las varianzas.
- La regresión no prueba causalidad.
- Un artículo que usa regresión debe mencionar o mostrar que se analizó la "nube de puntos" y que se hizo un análisis de los residuales.
- La línea de regresión no debe extenderse más allá de los datos obtenidos.

En artículos anteriores se ha presentado las características de una población que sigue una distribución Normal, con parámetros tales como una media μ y una desviación estándar σ que la describen exactamente. Se vio que a través de una muestra de esa población se obtenían estimaciones puntuales de esos parámetros poblacionales y que era factible dimensionar también la precisión de esas estimaciones a través de los intervalos de confianza. Hasta aquí se ha tratado de la descripción de una variable única. Para explorar las relaciones entre dos o más variables, las alternativas dependerán si son nominales o numéricas. En el primer caso se usa el riesgo relativo, la razón de productos cruzados, la estadística kappa, entre otros. Si son numéricas, la exploración puede hacerse con una regresión o con una correlación. La regresión permite estimar cuánto crece o decrece una variable en relación a la otra, y la correlación, cuantificar la fuerza de la asociación. Cuál usar depende principalmente del objetivo de la comparación por un lado y de las características de las muestras que se han obtenido por el otro.

REGRESIÓN LINEAL

De un modo general se dice que existe regresión de los valores de una variable con respecto a los de la otra cuando hay alguna línea, denominada línea de regresión, que se ajusta más o menos claramente a los valores observados. La regresión se usa para la identificación de relaciones potencialmente causales o bien, cuando no existen dudas sobre su relación causal, para predecir una variable a partir de la otra. Cuando dos variables tienen una relación de tipo determinista, el valor de una define exactamente el valor de la otra; un ejemplo puede ser la relación entre la presión y el volumen de un gas a temperatura constante. En los fenómenos biológicos sin embargo, la relación entre las variables es de tipo aleatorio ya que existen múltiples factores, muchos de ellos desconocidos, que influyen sobre la relación que existe entre las dos variables. En el primer caso, al hacer las mediciones en una relación determinista, los resultados caerán muy cerca de la línea de regresión; la cercanía o coincidencia de los puntos observados con esa línea estará definida por el error de la medición. En el segundo caso, a esta variabilidad se debe agregar aquella introducida por esos múltiples factores adicionales.

El procedimiento a seguir puede dividirse en cuatro etapas:

1) La primera aproximación es a través de dibujar los puntos en un gráfico cartesiano que muestre la relación entre las dos variables.
2) Luego se determina la ecuación de la línea que mejor describa dichos puntos.
3) A continuación se calcula la variabilidad de la muestra en torno a la línea de regresión calculada.
4) Finalmente se pueden hacer inferencias.

Si se grafican los valores de las dos variables en un gráfico de coordenadas, se pueden obtener "nubes de puntos" como los de la Figura 1.

[1] Profesor Titular, División de Anestesiología, Pontificia Universidad Católica de Chile.

La designación de X como variable independiente y la de Y como dependiente es convencional, depende del objetivo de la regresión, y puede estar limitada por el tipo de muestreo usado como veremos luego. Una primera ojeada sobre la nube de puntos permite hacer algunas observaciones preliminares respecto del grado de asociación y la naturaleza de esta. Mientras más dispersos aparezcan los puntos, como en un círculo por ejemplo, podemos decir que es menos probable que estén relacionados. En el otro extremo, si todos los puntos están alineados podemos decir que existe una relación de tipo determinista. La segunda aproximación, cuando se pretende usar la regresión, es la de estimar si la relación es recta o curvilínea pues ello es crucial al momento de buscar cual es la mejor línea (relación) que representa las observaciones (Figura 2).

El cálculo de la recta que mejor represente los puntos observados se puede hacer de diversas maneras, buscando aquella que tenga la menor distancia desde los puntos dados. A ojo se pueden proponer alternativas que parecen igualmente válidas pero que son mutuamente excluyentes (Figura 3).

La manera más usada es la de medir la distancia vertical desde cada punto hasta la recta propuesta (Figura 4).

Para obviar el problema de los signos, de manera análoga a lo que es la varianza, se toman los cuadrados de estas distancias y por ello se conoce al método como el de los cuadrados mínimos. Las distancias también se conocen como residuales.

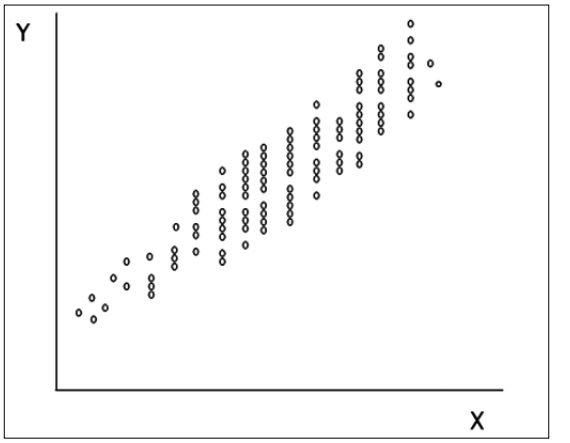

Figura 1. Nube de puntos. En las abscisas los valores de la variable independiente X y, en las ordenadas, los de la variable dependiente Y.

La ecuación obtenida (Figura 5) es del tipo:

$$\hat{Y} = \alpha + bX$$

La que puede leerse como:

Resultado predicho = intersección + (pendiente · valor del predictor).

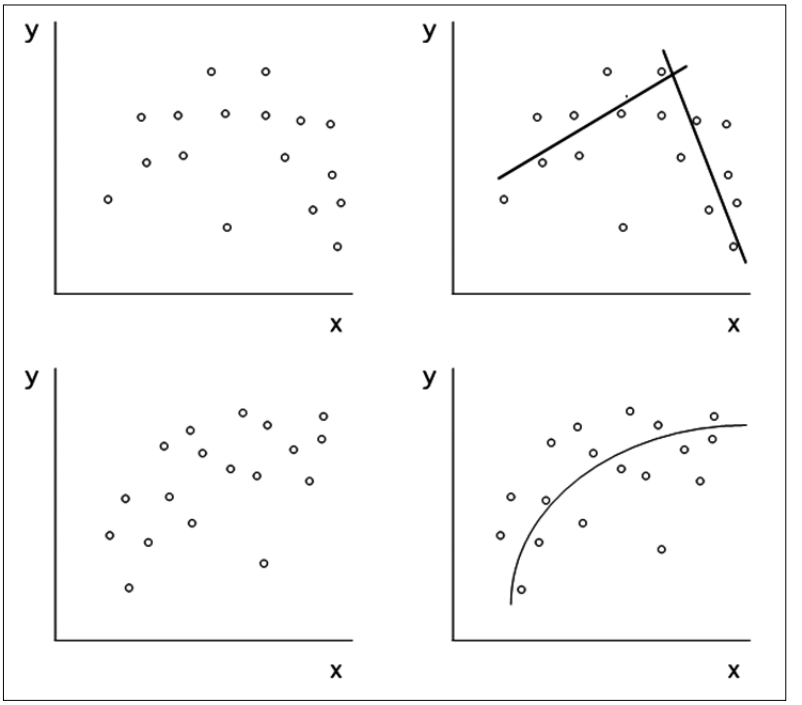

Figura 2. La inspección de la nube de puntos en los gráficos de la izquierda sugiere relaciones distintas de la de una simple recta, tal como se ha trazado en los gráficos de la derecha.

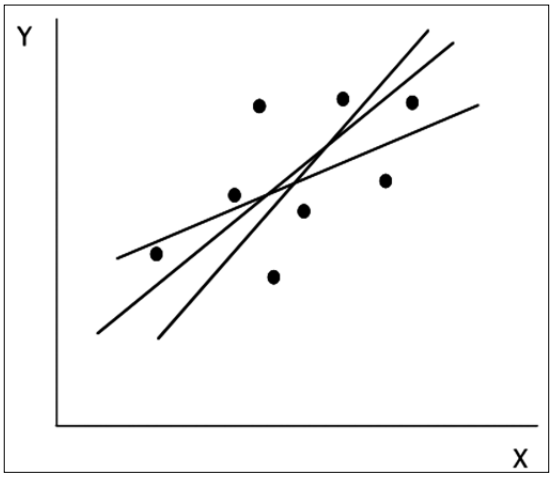

Figura 3. Tres posibles líneas trazadas a través de la nube de puntos.

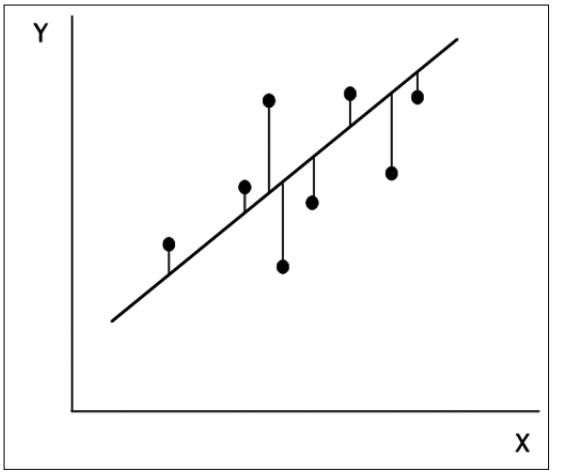

Figura 4. La distancia vertical medida desde cada observación hasta la línea estimada se denomina "residuo". Al igual que la varianza, sus cuadrados estiman la variabilidad de los puntos en relación a la línea.

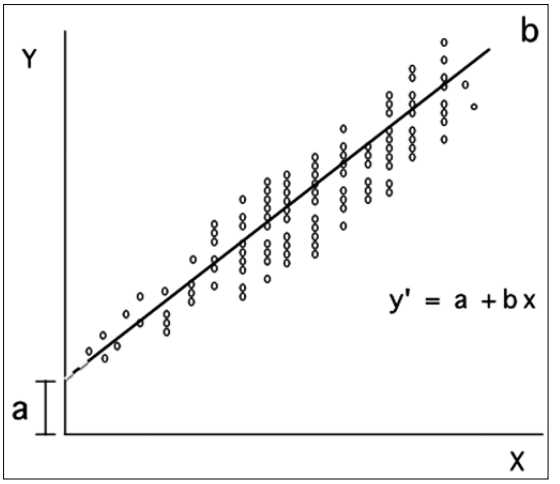

Figura 5. La línea calculada con la ecuación es la que mejor representa los datos, donde a es la intersección de la línea con el eje de las ordenadas y b es la pendiente de la línea.

Los valores de Y que se obtienen son predicciones y no valores reales por lo que se usa una notación diferente con un acento circunflejo sobre la Y (Ŷ). A los valores de a y b se les conoce como los coeficientes de la recta de regresión mínimo cuadrática (o de los cuadrados mínimos). En la población estos parámetros se denominan con sus letras griegas por lo que a y b son las mejores estimaciones de α y β, en el mismo sentido que la media es la mejor estimación de μ. Desgraciadamente, α y β también se usan para designar la magnitud del error Tipo I y II respectivamente, lo que no debe ocasionar confusión.

Los cálculos de a y de b son relativamente simples y hoy por hoy se hacen con un computador. Para fines prácticos, en medicina, el valor de a generalmente carece de importancia ya que rara vez, por no decir nunca, se miden los valores de Y cuando X es cero o cercana a cero. Un valor positivo de b indica que ambas variables aumentan conjuntamente; un valor negativo de b indica que un aumento de X ocasiona una disminución de Y.

La varianza mide el grado de variabilidad de los datos alrededor de la recta de regresión. Cuando la varianza es cero, todos los puntos Y coinciden con la recta, y X predice exactamente lo que debe valer Y. Alternativamente, mientras más grande es la varianza menos sirve X para predecir Y pues hay más incertidumbre. La varianza también se conoce como la varianza residual y para el cálculo de la varianza se usa el mismo concepto, usando la distancia de cada punto hasta la media y el tamaño de la muestra:

$$s^2 = \frac{1}{n-2}\sum d_i^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)$$

Se usa n-2, en vez de n-1 como se hace en el cálculo de la varianza de la media, pues se han debido estimar dos parámetros en lugar del único en aquella.

La fórmula puede simplificarse a:

$$s^2 = \frac{1}{n-2}\left[(YY) - b(XY)\right] = \frac{1}{n-2}\left[(YY) - \frac{(XY)^2}{XX}\right]$$

Donde la primera igualdad conviene a efectos conceptuales y la segunda al cálculo.

El término $(n-2)s^2$ se obtiene como la diferencia de (YY) o suma total de los cuadrados de la variable menos la cantidad b (XY) que puede entenderse como una disminución, debida a la regresión, en la suma total de los cuadrados. Con esto $(n-2)s^2$ será la parte de la suma de los cuadrados total que no ha sido explicada por la regresión, parte que depende de otras variables o influencias no consideradas en el modelo.

El punto puede clarificarse si se utiliza la siguiente partición:

$$(YY) = b(XY) + (n-2)s^2$$

La que puede leerse como:

s.c. total = s.c. regresión lineal + s.c. desviación de la regresión

(s.c. = suma de los cuadrados)

Conceptualmente entonces, está la variabilidad total de Y por un lado y la variabilidad de Y que X puede explicar.

El modelo de regresión parte de algunas presunciones acerca de las variables y de las muestras experimentales (Figura 6):

1) La media de la población de Y a un valor dado de X crece (o decrece) linealmente a medida que X aumenta.
2) Para cualquier valor de X, los posibles valores de Y se distribuyen normalmente.
3) La desviación estándar de la población de Y alrededor de su media a un valor dado de X es la misma que para todos los valores de X.

En otras palabras, el modelo supone aleatoridad de la muestra, Normalidad, linealidad y homogeneidad de las varianzas. En estas circunstancias, la línea de regresión puede ser pensada como aquella que une las medias de Y para cada valor de X. Es relativamente fácil verificar, una vez que se ha obtenido la recta, la efectividad de las presunciones usando para ello un gráfico de residuales enfrentando los valores de X en las abscisas con los de Y-Ŷ, los residuales, en las ordenadas en un diagrama cartesiano (Figura 7). La notación Ŷ denota el valor estimado de Y para cada valor de X usando la recta de regresión en oposición a los valores observados de Y; Ŷ frecuentemente se anota indistintamente como Y' también.

La nube obtenida debe ser paralela al eje horizontal (para verificar la linealidad) y debe tener un ancho homogéneo a lo largo (para verificar la homogeneidad de la varianza). Otros patrones sugieren que el modelo usado no produce el mejor ajuste.

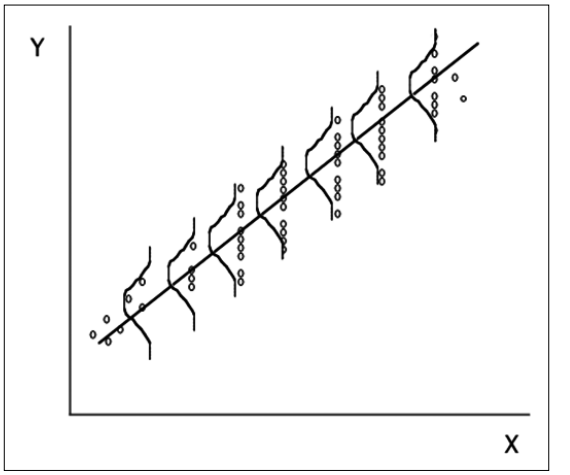

Figura 6. Los valores de Y para cada valor de X deben tener una distribución Normal y la variabilidad debe ser igual.

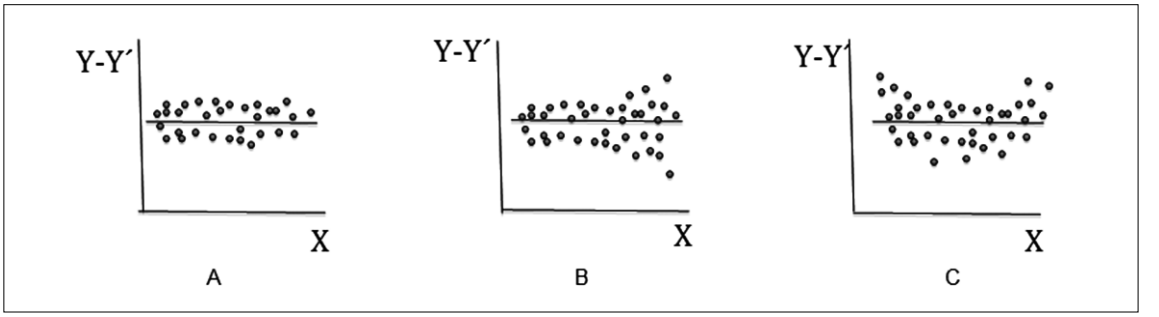

Figura 7. El análisis de los residuales permite verificar dos de las presunciones del modelo de regresión: la linearidad y la igualdad de las varianzas. En A, los residuos se distribuyen normalmente sobre y bajo la línea de identidad en todo el rango de X. En B, los residuos van aumentando en la medida que X aumenta. En C, los residuos no se distribuyen linealmente siendo mayores cuando X es menor o mayor.

La línea calculada (ajustada) explica una porción de la variabilidad de Y. Los residuos indican la cuantía de la variabilidad no explicada. Es por esto que el análisis de regresión habitualmente se presenta como una tabla de análisis de varianza. El valor F (estadístico del análisis de varianza) y la p obtenida se refieren a la hipótesis nula que plantea que la pendiente es igual a cero. El "*residual mean square*" es la varianza de los residuos, es decir, el cuadrado de la desviación estándar residual. Como ya dijimos, esta cuantifica la variación inexplicada por la línea de regresión y por lo tanto es un reflejo de cuan bueno es el ajuste de la ecuación obtenida. Una manera general de evaluar esta concordancia es considerar la proporción de la variabilidad total que el modelo es capaz de explicar. Esto usualmente se hace calculando la proporción de la suma de cuadrados que explica la recta sobre la suma de cuadrados total. A esta proporción se le conoce como R^2 que es el cuadrado del coeficiente de correlación como veremos en el artículo de correlación.

REGRESIÓN DE Y SOBRE X o DE X SOBRE Y

La definición de X como variable independiente y de Y como dependiente es convencional, pero es posible intercambiar sus posiciones si el objetivo lo admite, dependiendo de cual variable se quiere predecir en base a los cambios de la otra. Para poder hacer esto es indispensable que se haya tomado las muestras de ambas variables en forma aleatoria. Debe quedar claro, sin embargo, que las rectas que describen esas relaciones, Y sobre X o X sobre Y, son diferentes y por lo tanto, no son en ningún caso intercambiables.

Tal como se explicó, el modelo de regresión y la base del cálculo de la recta, nace de que el muestreo correcto de los datos es el de tomar valores de X fijados de antemano y en cada uno de ellos, tomar uno o varios valores de Y al azar (muestreo tipo II). Existe otra alternativa que es la de muestrear parejas de los valores X e Y en individuos elegidos al azar (muestreo tipo I). En este caso existe la exigencia de una distribución Normal no sólo de Y sino de X también (Figura 8).

Cada tipo de muestreo tiene sus ventajas e inconvenientes. El de tipo I permite realizar la regresión de Y sobre X o de X sobre Y indistintamente, en tanto que el de tipo II sólo admite la regresión de Y sobre X pues sólo las variables Y fueron tomadas al azar, (primera exigencia del modelo). La principal ventaja del tipo II es el control de los valores de X: un rango amplio de valores de X mejora la inferencia estadística, permite circunscribir los valores de X al rango que interesa estudiar para realizar las predicciones. Es importante destacar que sobre el rango no muestreado no hay información por lo que no se pueden extrapolar las predicciones fuera de ese rango, error muy común en la literatura.

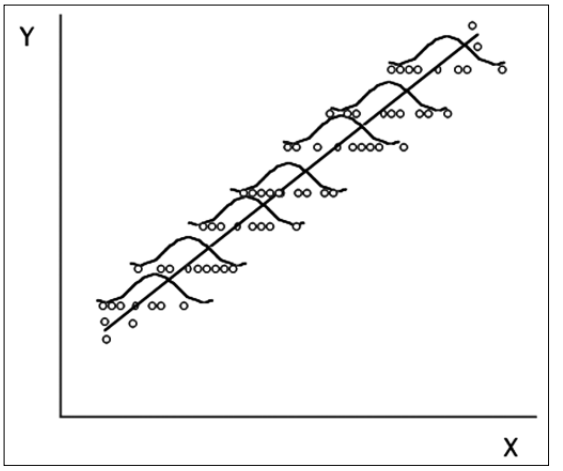

Figura 8. En el muestreo de tipo I también los valores de X deben tener una distribución Normal.

RESIDUOS E INTERVALOS DE CONFIANZA (Figura 9)

Las distribuciones de todos los valores posibles de a y b tienen una distribución Normal con medias a y b, y desviaciones estándar s_a y s_b que se denominan errores estándar de la intercepción y de la pendiente respectivamente. Estos errores estándar pueden ser usados tal como se usan los errores estándar de la media o de una proporción, para calcular intervalos de confianza y para las pruebas de hipótesis usando la distribución de t. Los intervalos de confianza de la línea de regresión se refieren a la línea de las medias y no a la población globalmente. Debe notarse que son angostos y se ensanchan hacia los extremos. No es infrecuente presentar estos como intervalos de confianza de toda la población, lo que no corresponde pues es análogo a usar el error estándar de la media como una medida de la variabilidad de la población en vez de la desviación estándar. Distinto es el caso de las bandas de confianza o intervalos de predicción para valores individuales, que son más anchas y que sí nos dan una idea de la variabilidad de la muestra y, por ende, del grado de imprecisión envuelto en la estimación. En el primer caso, nos dicen que podemos estar confiados en un 95% que el valor promedio de Y para un valor determinado de X está dentro de las líneas en cuestión. En el segundo caso, cuantifica la incerti-

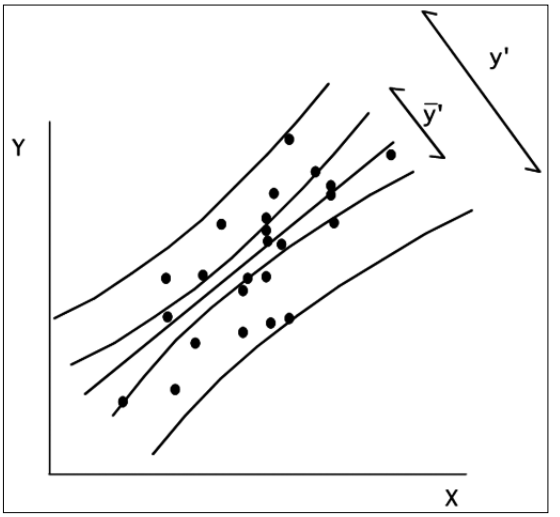

Figura 9. Los intervalos de confianza para la línea de regresión son más angostos que las bandas de confianza para la predicción individual. Ambas crecen en la medida que se alejan de la media de X lo que refleja que las predicciones son menos confiables en los extremos pues se basan en menos datos.

dumbre en la estimación de un valor individual de Y a partir de X que es lo que se quiere saber generalmente. El intervalo de confianza de las líneas de regresión puede hacerse más estrecho aumentando el tamaño de la muestra, pero no sucede lo mismo con el intervalo de predicción ya que este refleja fundamentalmente la variabilidad individual en torno a la recta calculada.

INFERENCIAS

Una vez trazada la recta, el siguiente paso es hacer inferencias en torno a ella. Debe quedar claro que la recta calculada en base a los datos obtenidos en una muestra es una estimación puntual de la recta poblacional ya que rectas calculadas en otra muestra de la misma población serán diferentes. Al igual que la media de cada muestra constituyen estimaciones de la media poblacional μ, la recta calculada a partir de una muestra es la mejor estimación de la verdadera recta poblacional.

La pregunta más importante y común es la de averiguar si la pendiente de una curva de regresión es o no diferente de cero, siendo la hipótesis nula que no es distinta de cero. Otra alternativa frecuente es la de comparar dos rectas de regresión para averiguar si corresponden a muestras de una misma población Aquí las preguntas posibles son tres y cada una da información diferente:

¿Son las varianzas iguales?
¿Son las pendientes iguales?
¿Son las alturas en el origen iguales?

En el caso de pendientes iguales se trata de comparar dos medias de una cierta variable y en dos poblaciones, pero quitando los posibles efectos de una segunda variable concomitante X que pudiera influir en el resultado. Este es el concepto del análisis de la covarianza.

USO PREDICTIVO Y DE CAUSALIDAD DE LA REGRESIÓN LINEAL

La evidencia de asociación entre dos variables en caso alguno implica que una de ellas sea la causa de la otra. Esta demostración de causalidad necesita de argumentación que no es matemática (ver artículo "Tipos de Estudio"). Así, cuando encontramos una asociación estadísticamente significativa entre dos variables, lo único que podemos afirmar es que existe una relación, pero no la naturaleza de la misma. En efecto, la relación encontrada puede deberse:

- A que el valor de X sea realmente causa del valor de Y. Por ejemplo, la dosis de un veneno puede ser la causa del porcentaje de muertes observado en un grupo de ratas.
- A que ambas variables se influyan mutuamente. Por ejemplo, las estaturas de pie y sentado.
- A que ambas variables dependan de una tercera variable común. Famoso es el caso que relacionaba inversamente la producción de acero en USA con la tasa de natalidad en Inglaterra (en la www existen sitios que contienen muchísimas de estas asociaciones espurias).

OTRAS REGRESIONES

Existen otras ecuaciones que pueden describir la relación entre dos variables en forma más adecuada: polinomiales, exponenciales logarítmicas, etc. Por otro lado, cuando hay más de una variable independiente y sólo una dependiente, se utiliza la regresión múltiple, para predecir con el conjunto de aquellas la variación de la segunda. Los conceptos son los mismos que los usados para la regresión única, pero los cálculos más largos. Esto explica por qué sólo con la creciente disponibilidad de computadores y programas para análisis estadístico vemos en la literatura médica cada vez más trabajos que utilizan las técnicas de regresión múltiple para analizar las relaciones entre múltiples variables independientes y una sola dependiente.

REFERENCIAS

1. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991.
2. Bland M. An Introduction to Medical Statistics. 3rd Ed., Oxford: OUP. 2006.
3. Dawson-Saunders B, Trapp RG. Bioestadística Médica. México D.F: Manual Moderno, 1993.
4. Farnsworth DL. Playing with residuals. Teaching Statistics 2009; 31: 81-84.
5. Glantz SA. Primer of Biostatistics. 3a edición, New York: McGraw-Hill, 1992.
6. Godfrey K. Simple Linear Regression in Medical Research. En Bailar III JC, Mosteller F. Medical uses of statistics. 2nd edition. Boston: NEJM Books, 1992.
7. Guyatt G, Walter S, Shannon H, Jaeschke R, Heddele N. Basic Statistics for Clinicans: 4. Correlation and Regression. Can Med Ass J 1995: 152: 487-504.
8. Portney LG, Watkins MP. Foundations of Clinical Research. Applications to practice. 2nd ed. Upper Saddle River: Prentice

Correspondencia a:
Dr. Jorge Dagnino S.
jdagnino@med.puc.cl