Investigación clínica XVIII

Del juicio clínico al modelo de regresión lineal

Lino Palacios-Cruz,[a,b] Marcela Pérez,[b] Rodolfo Rivas-Ruiz,[b] Juan O. Talavera[b]

Cada peculiaridad en un hombre es compartida por sus descendientes, pero en promedio, en un grado menor
Sir Francis Galton, 1886

From clinical judgment to linear regression model

When we think about mathematical models, such as linear regression model, we think that these terms are only used by those engaged in research, a notion that is far from the truth. Legendre described the first mathematical model in 1805, and Galton introduced the formal term in 1886. Linear regression is one of the most commonly used regression models in clinical practice. It is useful to predict or show the relationship between two or more variables as long as the dependent variable is quantitative and has normal distribution. Stated in another way, the regression is used to predict a measure based on the knowledge of at least one other variable. Linear regression has as it's first objective to determine the slope or inclination of the regression line: $Y = a + bx$, where "a" is the intercept or regression constant and it is equivalent to "Y" value when "X" equals 0 and "b" (also called slope) indicates the increase or decrease that occurs when the variable "x" increases or decreases in one unit. In the regression line, "b" is called regression coefficient. The coefficient of determination (R^2) indicates the importance of independent variables in the outcome.

Key words
linear models
models, statistical
statistics

El término *modelo matemático* nos remite a conceptos que solo atañen a quienes se dedican a investigar, noción que dista mucho de la realidad. En la práctica clínica y en la vida diaria realizamos asociaciones o predicciones que nos ayudan en nuestro desempeño cotidiano, por ejemplo, cuando evaluamos a un adolescente con un trastorno por uso de alcohol y deseamos inferir el efecto de los factores biológicos y medioambientales (y también indirectamente el pronóstico a mediano y largo plazo) mediante la asociación de la edad al inicio del consumo de alcohol y el número de familiares con el mismo hábito. En casa, no es infrecuente tratar de predecir la cantidad de un producto que debemos comprar para una semana si hay menos integrantes de la familia que los habituales.

Este método es valioso en escenarios clínicos o económicos, como cuando deseamos evaluar variables que, por su costo o dificultad para su obtención, requieren alguna aproximación clínica previa, como la estimación de la densidad mineral ósea a partir de la medición de peso, talla y la ultrasonografía ósea.

En este punto, por más que el lector juró "no volver a tratar con las matemáticas y alejarse lo más posible de ellas", tal vez comienza a sospechar que la relación es útil, ya que de manera natural aplica modelos matemáticos simples y complejos, en un nivel de menor conciencia pero con un nivel equiparable de utilidad.

Los métodos estadísticos, como la regresión lineal, permiten predecir o disminuir esa incertidumbre.[1] El análisis de regresión se define como "el estudio de la dependencia" o de cómo una respuesta o variable depende de uno o más predictores o variables independientes. Al considerar este modelo en un proyecto de investigación o análisis de la información es importante tomar en cuenta dos aspectos básicos:

- Que la dependencia de la respuesta sobre los predictores se lleva a cabo mediante el promedio, por lo tanto, se requiere que esta variable tenga una distribución normal.
- Que el promedio de la variable dependiente dadas las variables independientes es una función lineal, es decir, la variable dependiente se incrementa o disminuye conforme se incrementan o disminuyen los valores de las variables independientes o predictoras.[2] Dicho de otra manera: debe existir una relación en la que el incremento o disminución de una variable sea proporcional en cada punto.

Se considera *regresión lineal simple* si se relacionan solo dos variables, de las cuales la dependiente es cuantitativa. Cuando se utilizan dos o más variables para predecir una variable cuantitativa se considera *regresión lineal múltiple*. Las variables independientes pueden combinar variables cuantitativas y cualitativas.

Resumen

Pensamos que los modelos matemáticos, como la regresión lineal, son conceptos que solo atañen a quienes se dedican a investigar, noción que dista de la realidad. La primera descripción de un modelo matemático fue realizada por Legendré, en 1805, y la introducción formal del término fue hecha por Galton, en 1886. La regresión lineal es útil para predecir la relación entre dos o más variables, siempre y cuando la variable dependiente sea cuantitativa y cuente con una distribución normal. Su desarrollo tiene como primer objetivo determinar la pendiente o inclinación de la línea de regresión: $Y = a + bx$, donde "a" es la "constante de regresión" que equivale al valor de "Y" cuando "x" es igual a 0 y "b", también llamada pendiente de la recta, indica el incremento o decremento que se produce en "Y" cuando la variable "x" aumenta o disminuye una unidad. En la línea de regresión, "b" recibe también el nombre de coeficiente de regresión. El coeficiente de determinación (R^2) define la magnitud de la capacidad para predecir el efecto de las variables independientes sobre el resultado.

Palabras clave
modelos lineales
modelos estadísticos
estadística

Probablemente, la regresión lineal es, junto con la regresión logística, el modelo de regresión más aplicado, tanto en las investigaciones de las ciencias naturales y sociales como en la vida diaria.

Historia de la regresión lineal

Si bien la primera descripción documentada sobre un método de regresión lineal fue publicada por Legendré en 1805, en el método de mínimos cuadrados con el que abordaba una versión del teorema de Gauss-Márkov,[2-4] fue sir Francis Galton, médico y primo de Charles Darwin, quien introdujo el término *regresión*, en su artículo "Regression towards mediocrity in hereditary stature", publicado en 1886 en el *Journal of the Anthropological Institute*[5] y que menciona de nuevo en su libro *Natural Inheritance*, de 1889.[6] En ese trabajo clásico, Galton centró su descripción en los rasgos físicos de los descendientes (variable dependiente) a partir de los rasgos de sus padres (variable independiente). Analizó la altura de 205 padres y 930 hijos adultos a partir de sus registros familiares y llegó a la conclusión de que los padres muy altos tenían una tendencia a tener hijos que heredaban parte de esta altura, pero que se revelaba también una tendencia a regresar a la estatura media. A partir de estas observaciones, Galton señaló esta tendencia bajo la "ley de la regresión universal". Al final del mismo siglo XIX, Pearson y Yule aportaron muchas de las nociones modernas acerca de la correlación, que han contribuido al metalenguaje que nos permite el entendimiento del fenómeno de la dependencia entre variables.

Teoría y conceptos de la regresión lineal

Como se ha señalado, la regresión se utiliza para predecir una medida o variable dependiente (también llamada de desenlace o variable "y") basándonos en el conocimiento de al menos otra variable independiente (o variable relacionada con la maniobra o variable "x")[7] y un término aleatorio ε.[8]

El proceso de regresión lineal tiene como primer paso determinar la pendiente o inclinación de la línea de regresión, cuya representación algebraica para la regresión lineal simple es de la siguiente forma:[1]

$$E\,(Y/X) = a + bx$$

Donde:

El estimador de "Y" dado un valor de "X" es igual a $a + b$ que multiplica "x", asumiendo que la distribución de "Y" para una "x" determinada es normal y, además, que las varianzas de ambas variables son homogéneas, fenómeno conocido como *homocedasticidad*.[1,9,10]

La manera más popular de la representación matemática de la regresión lineal simple es como sigue:

$$y_t = B_0 + B_1 X_1 + \varepsilon$$

Donde:

y_t = variable explicada o dependiente. La "y", a diferencia de la "Y", que es un valor real dentro de la población, es el indicador de "Y" o el valor estimado a partir de una muestra que trata de predecir "Y".

B_0 = intercepto, ordenada al origen o constante de regresión, que es la altura a la que la recta corta al eje "Y", equivalente al valor de "y" cuando "x" es igual a cero.

B_1 = parámetro que mide la influencia que tienen las variables explicativas sobre la variable explicada o dependiente.

Ejemplo

En un grupo de 40 adultos jóvenes (67.5 % mujeres y edad promedio de 31.5 ± 4.71) que acuden a un servicio de salud mental, se desea predecir la edad de inicio del trastorno depresivo mayor (11 ± 3.63 años), a partir de la edad de inicio del trastorno por déficit de atención con hiperactividad (10.3 ± 1.63 años) (cuadro I). Al trazar para cada individuo el valor que obtienen en la edad de inicio para depresión (en el eje de las "Y") y para el trastorno por déficit de atención con hiperactividad (en el eje de las "X"), obtendríamos una gráfica similar a la de la figura 1.

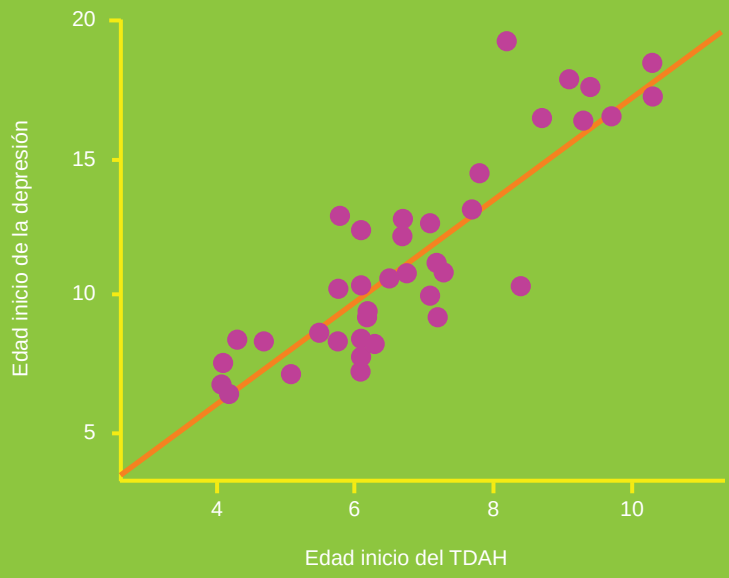


Figura 1 Regresión lineal para predecir la edad de inicio de depresión a partir de la edad de inicio del trastorno por déficit de atención con hiperactividad

Cuadro I Edad de inicio (en años) de los trastornos depresivo mayor y por déficit de atención con hiperactividad en 40 adultos jóvenes

TDM				TDAH			
Paciente	Edad	Paciente	Edad	Paciente	Edad	Paciente	Edad
1	19.2	21	10.3	1	8.2	21	6.1
2	6.7	22	9.9	2	4.1	22	7.1
3	8.3	23	9.2	3	4.7	23	7.2
4	6.4	24	16.3	4	4.2	24	9.3
5	7.1	25	10.2	5	5.1	25	5.8
6	11.2	26	16.4	6	7.2	26	8.7
7	14.4	27	17.5	7	7.8	27	9.4
8	12.7	28	9.3	8	6.7	28	6.2
9	9.1	29	8.2	9	8.1	29	6.3
10	13.1	30	11.1	10	7.7	30	7.2
11	10.6	31	12.6	11	6.5	31	7.1
12	8.3	32	10.8	12	4.3	32	6.8
13	7.5	33	9.1	13	4.1	33	6.2
14	17.8	34	7.7	14	9.1	34	6.1
15	10.3	35	8.3	15	8.4	35	5.8
16	17.2	36	8.6	16	10.3	36	5.5
17	16.4	37	8.4	17	9.7	37	6.1
18	12.3	38	10.8	18	6.1	38	7.3
19	12.8	39	7.3	19	5.8	39	6.1
20	12.1	40	18.4	20	6.7	40	10.3

TDM = trastorno depresivo mayor, TDAH = trastorno por déficit de atención con hiperactividad

Para la regresión lineal simple, B_1 corresponde a la pendiente de la línea y da la proporción de cambio en "y" por cada unidad de cambio en "x". Al término, a B_1 se le conoce como *coeficiente de regresión*. La positividad (relación directa) o negatividad (relación inversa) de la línea de regresión depende de B_1: si el valor de B_1 es positivo, "y" aumenta conforme "x" aumenta; si el valor de B_1 es negativo, "y" disminuye conforme "x" aumenta. Finalmente, si *b* es igual a cero, no hay cambios en "y", por lo tanto, la pendiente se mantiene en forma horizontal sin predecir "y" a partir de "x".[1]

La regresión lineal simple se puede ampliar para considerar simultáneamente los efectos de dos o más variables explicativas (es decir, regresión lineal múltiple). Existen otros modelos de regresión, como cuando la variable dependiente tiene un desenlace binario (regresión logística) o se trata de un desenlace binario en el transcurso del tiempo (riesgos proporcionales de Cox) o de recuento (regresión de Poisson).[9] En una regresión lineal múltiple podría expresarse como:

$$Y_t = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots + B_pX_p + \varepsilon$$

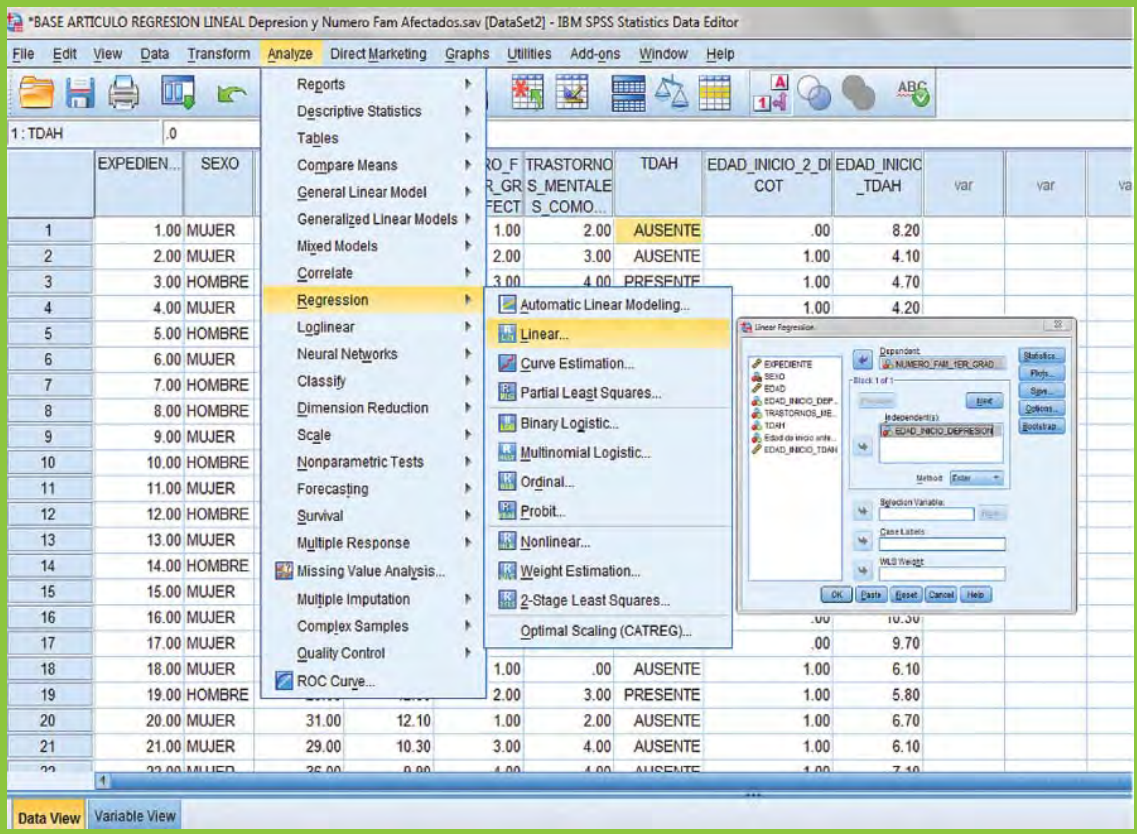
De manera similar, Y_t es la variable explicada o dependiente.

B_1, B_2, B_3... son los parámetros que miden la influencia que tienen la variables explicativas sobre la variable explicada o dependiente. Las B_i ($i > 0$) son los parámetros respectivos de cada variable independiente y *p* es el número de parámetros independientes por considerar en la regresión.

De regreso al ejemplo

Ahora bien, si se desea determinar si la edad de inicio del trastorno depresivo mayor (el grupo podría dividirse en dos subgrupos: sujetos que iniciaron con el trastorno antes de los 12 años y sujetos cuyo inicio del trastorno fue a partir de los 13 años, o bien, dejar la variable a nivel dimensional) predice el número de familiares en primer grado afectados con el mismo trastorno, en la figura 2 se observan los pasos para realizar el análisis en un programa estadístico.

Figura 2 Pasos para realizar una regresión lineal en el programa SPSS



En la predicción mediante la regresión lineal, un paso importante es la suma de los mínimos cuadrados, que corresponden a las cantidades que minimizan la suma de cuadrados de la varianza $(y - Y)^2$, ecuación que representa la recta con la menor distancia de "y" a "Y" (distancia entre un valor estimado y un valor real), pero elevada al cuadrado con el fin de no obtener un valor de 0, dado que "Y" se distribuye por igual forma a cada lado de la línea de regresión.[1,11]

En el estudio de la distancia del valor real y el valor estimado es necesario calcular un valor mínimo y un valor máximo del coeficiente de regresión, es decir, determinar el intervalo de confianza de 95 %, que proporciona las desviaciones explicadas por la pendiente de regresión.[1,7,12]

Al realizar la regresión lineal con los datos del ejemplo y al procesarlos con el programa ahora llamado IBM SPSS® (figura 3), podemos observar que la suma de cuadrados del modelo de regresión —es decir, el nivel de fluctuación de la variable Y_t que el modelo es capaz de explicar— tiene un valor de 21.24 y la suma de cuadrados de los modelos residuales —es decir, el indicador del nivel de error del modelo— es de 23.16 (o el porcentaje no explicado por el modelo). El valor de la constante, es decir el valor de "Y" cuando "X" es igual a 0, es de 4.12 y el coeficiente de regresión es de –0.204, con un intervalo de confianza de 95 % que va de –0.274 a –0.134. Es decir, mediante este modelo de regresión puede establecerse que a menor edad al iniciar el trastorno depresivo mayor, existe una mayor probabilidad de tener un mayor número de familiares en primer grado con el mismo trastorno.

Si bien mediante la línea de regresión se intenta predecir el valor de una variable a partir de otra, esta no proporciona en forma directa el porcentaje de la asociación de "y" (en el ejemplo, número de familia-

Model Summary

Model	R	R Square	Adjusted R Square	Std.Error of the Estimate
1	.692[a]	.478	.465	.78070

a. Predictors (Constant), EDAD_INICIO_DEPRESION

- Suma de cuadrados de la regresión (***sum of squares, regression***): **indica qué tanta variación de la variable dependiente explica nuestro modelo.**
- Suma de cuadrados de los residuales (***sum of squares, residual***): **indica qué tanta variación de la variable dependiente no explica nuestro modelo.**

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	21.239	1	21.239	34.848	.000[b]
	Residual	23.161	38	.609		
	Total	44.400	39			

a. Dependent Variable: NUMERO_FAM_1ER_GRADO_AFECT
b. Predictors: (Constant), EDAD_INICIO_DEPRESION

Coefficients[a]

Model		Unstandardized Coefficients		Stdandardized Coefficients		
		B	Std. Error	Beta	*t*	Sig.
1	(Constant)	4.119	.412		10.003	.000
	EDAD_INICIO_DEPRESION	–.204	.035	–.692	–5.903	.000

a. Dependent Variable: NUMERO FAM 1ER GRADO AFECT

Figura 3 Resultados de la regresión lineal

res en primer grado con trastorno depresivo mayor) a partir de "x" (en el ejemplo, edad de inicio del trastorno depresivo mayor). La intensidad de la asociación o el porcentaje de explicación del modelo se define con el coeficiente de determinación (R^2), que puede ser corregido o no y que equivale al cuadrado del coeficiente de correlación "R" (figura 3), que en nuestro modelo fue igual a 0.692.

La R^2 corregida para este modelo fue de 0.465, cuya traducción es que el número de familiares en primer grado que podrían tener el trastorno depresivo mayor se relaciona en 46.5 % con la edad de inicio de ese trastorno en el caso índice; en 53.5 % de los casos se debe a otros factores no incluidos en este modelo ($1 - R^2$).

Uno de los últimos pasos de este método de análisis es el establecimiento de la significación de la curva de regresión mediante la prueba de hipótesis, en la que se supone que el coeficiente de regresión no es igual a 0. Si bien el programa SPSS permite obtener automáticamente este resultado, anteriormente se debía analizar la varianza de la regresión, en la que si el coeficiente de regresión no es igual a 0, el valor de "F" observado es mayor que el valor crítico de "F", lo que se corresponde con un valor de $p < 0.05$. De esta forma, se rechaza la hipótesis nula, se acepta la hipótesis alterna y, de esa manera, se determina que la pendiente sí permite predecir "y" a partir de "x", es decir, que existe significación estadística.

Es importante recordar que la predicción de la variable dependiente a partir de una o más variables independientes no significa causalidad y que esta solo deberá considerarse si se cumplen las condiciones enunciadas en mayo de 1965 por Austin Bradford Hill: plausibilidad, especificidad, temporabilidad, etcétera.[13]

Conclusiones

El análisis de regresión es solo una herramienta más dentro de las opciones para que el clínico y el investigador se acerquen a la naturaleza de los resultados.[14,15] Los aspectos importantes que deben recordarse al momento de realizar y leer correctamente los resultados del análisis de regresión lineal son:

- La variable dependiente debe ser continua.
- La o las variables independientes pueden ser continuas o categóricas.
- El intercepto.
- El coeficiente de regresión.
- El coeficiente de determinación (R^2), importante para definir la magnitud de la relación de la o las variables predictoras sobre la variable resultante o predicha.
- El intervalo de confianza.
- El valor de "F" del análisis de la varianza de la regresión.

[a]Instituto Nacional de Psiquiatría "Ramón de la Fuente Muñiz", Secretaría de Salud, Distrito Federal, México
[b]Centro de Adiestramiento en Investigación Clínica, Coordinación de Investigación en Salud, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, Distrito Federal, México

Correspondencia: Lino Palacios-Cruz
Correo electrónico: palacioslino@gmail.com

Referencias

1. Talavera-Piña JO, Antonio-Ocampo A, Castellanos-Olivares A, Wacher-Rodarte NH. Regresión lineal simple. Rev Med IMSS. 1995;33(3):347-51.
2. Weisberg S. Linear hypothesis: regression (basics). En: Neil JS, Paul BB, editores. International encyclopedia of the social & behavioral sciences. Oxford: Pergamon; 2001. p. 8884-8.
3. Wikipedia. La Enclopedia Libre. Regresión lineal. [Consultado el 2 de septiembre de 2013]. Disponible en http://es.wikipedia.org/w/index.php?oldid=67582255.
4. Lam GK. On the unit of dose equivalent and the linear hypothesis. Health Phys. 1989;57(3):495-6.
5. Galton F. Regression towards mediocrity in hereditary stature. J Anthropol Inst. 1886;15:246-63.
6. Galton F. Natural inheritance. London-New York: MacMillan; 1889.
7. Pearce J. Regression, linear and nonlinear. En: Rob K, Nigel T, editores. International encyclopedia of human geography. Oxford: Elsevier; 2009. p. 302-8.
8. Canavos GC. Probabilidad y estadística. Aplicaciones y métodos. México: McGraw-Hill; 1988.
9. Devore JL. Probabilidad y estadística para ingeniería y ciencias. México: International Thomson Editores; 2005.
10. Dawson P, Trapp R. Bioestadística médica. México: Manual Moderno; 1990. p. 239-365.
11. Seber GAF. The linear hypothesis: a general theory. London: Griffin; 1966.
12. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. J Am Stat Assoc. 1997;92(437):179-91.
13. Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965;58(5):295-300.
14. Talavera JO, Rivas-Ruiz R. Investigación clínica IV. Pertinencia de la prueba estadística. Rev Med Inst Mex Seguro Soc. 2011;49(4):401-5.
15. Talavera JO, Rivas-Ruiz R. Investigación clínica XIV. Del juicio clínico al modelo estadístico. Rev Med Inst Mex Seguro Soc. 2013;51(5):170-5.