Rincón iconoclasta

¿Cómo y cuándo realizar un análisis de regresión lineal simple? Aplicación e interpretación

Arturo Reding Bernal,* Mireya Zamora Macorra,*** Juan Carlos López Alvarenga**

Al realizar un proyecto de investigación los estudiantes de pregrado y posgrado y los investigadores se enfrentan con diversos retos a medida que van realizando el proyecto. Una de las etapas difíciles es cuando deben hacer el análisis estadístico de los datos, ya que el lenguaje técnico y rebuscado de los expertos los intimida en ocasiones, de ahí que es posible que los resultados obtenidos los analicen con desánimo. Por eso, dicho análisis normalmente se lo asignan al experto matemático o estadístico, a quien le dejan la tarea de analizar los resultados.

El objetivo de este artículo es puntualizar de manera práctica una de las técnicas estadísticas comúnmente utilizadas en ciencias de la salud: la regresión lineal simple, la cual es más conveniente que otros métodos. Iniciaremos con la descripción de conceptos fundamentales para entender y utilizar esta técnica.

En las diversas áreas de la salud con frecuencia buscamos identificar el efecto de condiciones adversas o potencialmente benéficas para el estado general de salud. Por medio del análisis estadístico y epidemiológico deseamos conocer si alguna variable de exposición se asocia con algún efecto en la salud. Este último también es conocido como "variable de respuesta predicha" o "variable de respuesta dependiente", mientras que la exposición es conocida como "variable explicativa predictora" o "variable explicativa independiente".

Las variables pueden adquirir distintos valores, según la escala de medición:

- Nominales: la variable adquiere valores categóricos sin ningún orden jerárquico, como en las dicotomías en las que se indica si un evento está presente o ausente, por ejemplo, si una adolescente está o no embarazada. Otra variable nominal que no tiene orden jerárquico es la procedencia de los pacientes, que corresponde a múltiples posibilidades. Dicha procedencia puede ordenarse de muchas formas, por ejemplo, en orden alfabético, cuya forma es arbitraria.
- Ordinales: aquí la variable adquiere valores categóricos que tienen un orden jerárquico, por ejemplo, las variables que registran el estadio tumoral o los grados de retinopatía hipertensiva o de nefropatía diabética (clasificación de Mogensen). En cada uno de estos ejemplos puede identificarse el orden de los estadios, aunque éstos se hayan clasificado por grupos de características cualitativas, que a pesar de su carácter subjetivo tienen alta reproducibilidad.
- Continuas: la variable puede adquirir cualquier valor numérico, ya sean números positivos o negativos de presión arterial, de concentraciones de colesterol, de días de hospitalización, etcétera.

De acuerdo con las necesidades de nuestro estudio, la pregunta de investigación y la escala de medición de la variable de respuesta determinarán el tipo de análisis

* Investigador en Ciencias Médicas A. Dirección de Investigación.
** Director de Investigación.
Hospital General de México, México, DF.
*** Escuela Nacional de Salud Pública, Epidemiología, Instituto Nacional de Salud Pública.

Correspondencia: Dr. Arturo Reding B. reding_79@yahoo.com
Recibido: julio, 2011. Aceptado: septiembre, 2011.

Este artículo debe citarse como: Reding-Bernal A, Zamora-Macorra M, López-Alvarenga JC. ¿Cómo y cuándo realizar un análisis de regresión lineal simple? Aplicación e interpretación. Dermatol Rev Mex 2011;55(6):395-402.

estadístico por realizar. En algunas ocasiones puede ser que lo único que nos interese sea conocer la asociación que existe entre dos variables, y con una correlación estadística –como la de Pearson o la de Spearman, dependiendo de si las variables se calculan en una escala de medición continua u ordinal– puede ser suficiente para conocer dicha asociación. Ahora que si el objetivo es conocer una diferencia de promedio entre grupos y si la variable de respuesta se calcula en una escala de medición continua, pueden utilizarse otras pruebas que no son el objetivo del presente artículo, como la prueba de la t de Student, el análisis de variancia de una vía y el análisis de covariancia, e incluso, también podría utilizarse el análisis de regresión lineal. Esto podría causar sorpresa a algunos investigadores, ya que clásicamente se piensa en correlación de dos variables de tipo continuo; sin embargo, con una regresión puede obtenerse la misma información que con las pruebas de contraste clásicas, ya que los modelos de regresión lineal son una generalización de estas técnicas. En cualquiera de estos casos existe un supuesto de normalidad subyacente, el cual se abordará en forma detallada más adelante.

MODELO DE REGRESIÓN LINEAL SIMPLE

Aspectos generales de cuándo utilizar un modelo de regresión lineal simple

La regresión lineal simple es útil para encontrar la fuerza o magnitud de cómo se relacionan dos variables: una independiente, que se representa con una *X*, y otra dependiente, que se identifica con una *Y*; sin embargo, la regresión lineal simple se distingue de otras pruebas, pues con ella puede estimarse o predecirse el valor de la variable de respuesta a partir de un valor dado a la variable explicativa. Para asociar estas dos variables se propone una línea recta –que describe la tendencia de los datos–, de ahí el nombre de regresión lineal. Dicha recta se expone en un plano y su grado de inclinación representa la pendiente, y una inclinación muy destacada indica grandes cambios en la variable dependiente.

En los modelos de regresión lineal pueden analizarse varias variables explicativas, pero en el modelo de regresión lineal simple sólo se considera una variable independiente para predecir el resultado. En este artículo sólo analizaremos el modelo de regresión lineal simple.

Una cuestión idónea es que cuando un investigador plantee una hipótesis de trabajo la haga después de analizar profundamente la bibliografía; en el análisis podría preguntarse lo siguiente: ¿existe alguna relación biológica factible que explique lo que pretendo encontrar?, ¿existen situaciones de esta naturaleza en la realidad?, pues el hecho de que él localice valores significativos en la regresión no implica que exista una asociación genuina, porque en muchas ocasiones se originan asociaciones espurias, es decir, originadas por el azar.

Hay que diferenciar entre correlación y regresión. La *correlación* es un coeficiente que nos permite evaluar la fuerza de asociación entre dos variables. Para ello debe considerarse el valor del coeficiente, cuyos límites son -1 a 1 y el valor de *p* obtenido. Una correlación que se acerque a 1 indica que el valor de *Y* aumenta a medida que aumenta el valor de *X* (Figura 1A). En caso de que el valor se acerque a 1 pero con signo negativo (-1) indica que la correlación es negativa o inversa; esto es, el valor de *Y* disminuye a medida que aumenta el valor de *X* (Figura 1B). Finalmente, si el valor es cercano a cero, esto indica que la relación entre *X* y *Y* produce una línea horizontal, y por tanto, no hay correlación; es decir, la variable dependiente *Y* permanece constante a medida que cambia *X* (Figura 1C). Respecto a la *regresión*, ésta aporta información adicional, ya que permite estimar el cambio promedio de unidades de la variable de respuesta *Y* por el cambio de unidades ocurrido en la variable explicativa *X*; además, permite hacer una predicción del comportamiento de las variables estudiadas en un determinado punto o momento.

Aspectos técnicos del modelo de regresión lineal simple

En esta sección se abordarán los aspectos técnicos del modelo de regresión logística; asimismo, se analizarán e interpretarán los resultados de un ejemplo dado. Para lo anterior se seleccionó una muestra de un estudio transversal que, entre 2010 y 2011, se realizó en el Hospital General de México. En dicho estudio a los participantes se les tomó una muestra de sangre, la presión arterial y diversas medidas antropométricas con el objetivo de analizar algunos factores asociados con los componentes del síndrome metabólico. Con la información obtenida generamos un modelo de regresión lineal simple para ilustrar el análisis y la interpretación del modelo.

Antes de realizar dicho modelo, un primer paso útil para conocer la relación entre dos variables es explorar los datos mediante un diagrama de dispersión, en el que

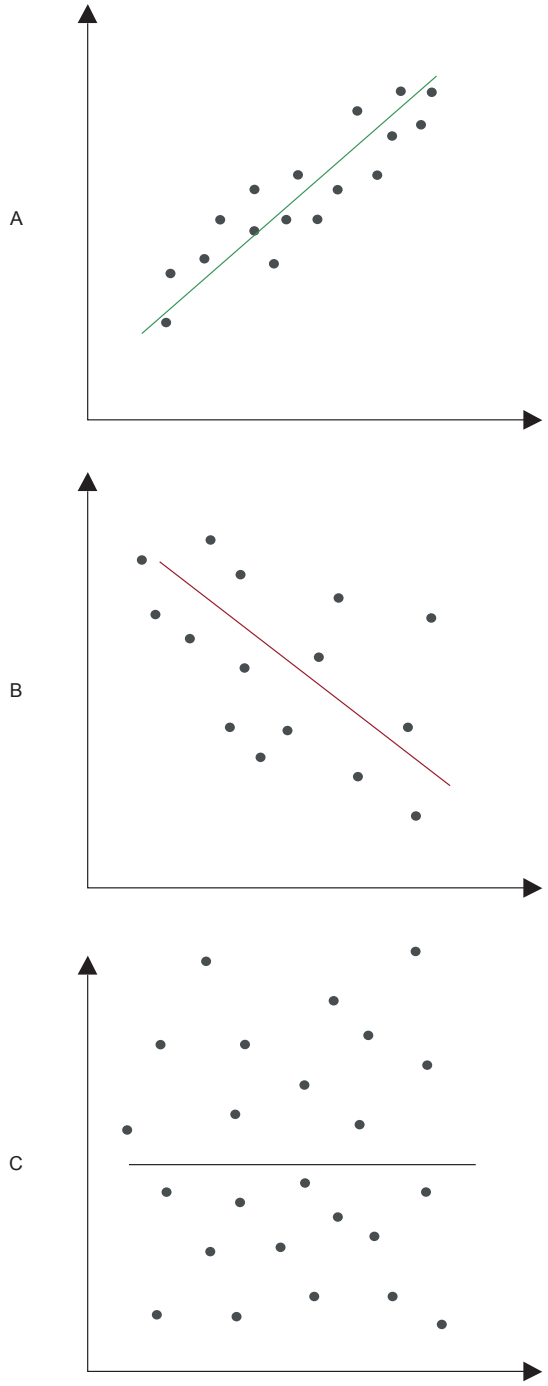

Figura 1. Regresión y correlación lineal simple. En A se observa que la distribución de la regresión es positiva o directa porque el coeficiente de correlación es cercano a 1. En B se observa que la distribución de la regresión es negativa o inversa porque el coeficiente de correlación es cercano a -1. Mientras que en C se muestra que no hay correlación entre las variables porque el coeficiente de correlación es cercano a 0 y porque *Y* permanece constante y muy dispersa a medida que aumenta *X*.

los valores de la variable independiente *X* son asignados al eje horizontal y los valores de la variable dependiente *Y* son asignados al eje vertical. El patrón que se obtiene a partir de este diagrama sugiere, en general, la distribución básica y la fuerza de la asociación entre las dos variables.[1] Si en términos gráficos se observa una relación aproximadamente lineal, entonces es adecuado proponer un modelo de regresión lineal simple. Una asociación aproximadamente lineal significa que por cada unidad que aumenta la variable explicativa se espera que suceda un efecto igual en la variable de respuesta, independientemente del valor que tenga la variable independiente. Sin embargo, cabe aclarar que esta aseveración sólo puede aplicarse a los límites de valores de la variable explicativa en estudio. Por ejemplo, si se desea determinar y cuantificar la asociación lineal entre la presión arterial sistólica (variable de respuesta o dependiente) y la edad de las personas (variable explicativa o independiente), primero se realizará una exploración gráfica de los datos mediante un diagrama de dispersión. En la Figura 2 puede observarse que la presión arterial sistólica se incrementa cuando la edad de las personas aumenta y que esta relación es aproximadamente lineal.

El método que generalmente se usa para obtener la recta deseada se conoce como "mínimos cuadrados"; asimismo, la recta resultante se conoce como "recta de mínimos cuadrados". De acuerdo con conceptos básicos de álgebra,

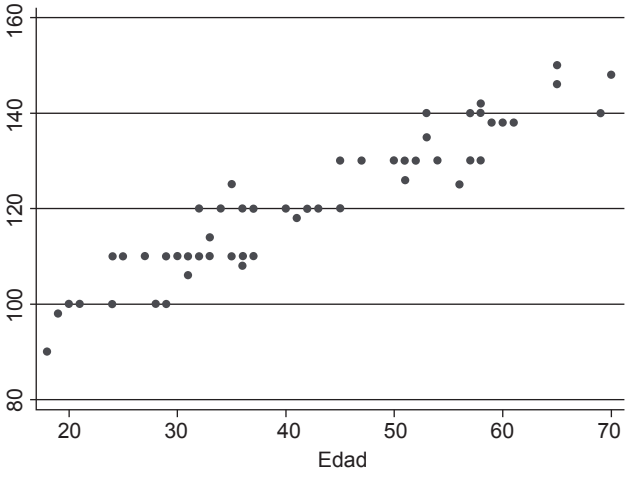

Figura 2. Diagrama de dispersión en el que la relación entre la edad y la presión arterial sistólica es aproximadamente lineal.

la ecuación general de una recta está dada en la siguiente expresión: $Y = \beta_0 + \beta_1 X$, donde Y son los valores correspondientes al eje vertical, β_0 es la ordenada al origen, β_1 es la pendiente, y X son los valores correspondientes al eje horizontal. El objetivo de un modelo matemático es que, en su mayor medida, se ajuste a los datos; sin embargo, esto no siempre es posible. En el ejemplo de la presión sistólica *vs* la edad se observa que los datos no forman exactamente una línea recta; esta diferencia entre el valor observado de Y y el de la línea recta ($\beta_0 + \beta_1 X$) se conoce como "error" (ε) [Figura 3]. Éste es un error estadístico; es decir, es una variable aleatoria que explica por qué el modelo no se ajusta exactamente a los datos.[2] La aleatoriedad del término *error* hace que la variable dependiente Y también sea una variable aleatoria. Por tanto, el modelo teórico utilizado para establecer la asociación lineal entre la variable de respuesta Y y la variable explicativa X es: $Y = \beta_0 + \beta_1 X + \varepsilon$ (ecuación 1).

A esta ecuación 1 se le llama "modelo de regresión lineal simple", puesto que sólo tiene una variable explicativa. En este modelo al suponer que el promedio y la variancia de ε son cero y σ^2, respectivamente, se tiene que el valor esperado de Y –dado un valor fijo de X– está dado en la siguiente expresión: $\mu_{Y/X} = E(Y/X) = \beta_0 + \beta_1 X$. En otras palabras, con estos supuestos para cada valor particular de la variable X se espera que el valor de la variable Y sea ($\beta_0 + \beta_1 X$), donde los parámetros a estimar son β_0 y β_1, que –como se mencionó antes– son la ordenada al origen y la pendiente de la recta, respectivamente. En la Figura 3 puede observarse una representación gráfica de estos dos parámetros y del error aleatorio ε sobre una recta de regresión lineal simple.

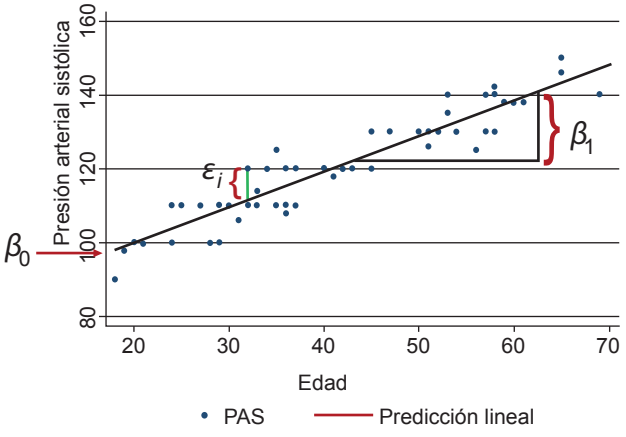

Figura 3. Representación de la recta, de los parámetros y del error de un modelo de regresión lineal simple. PAS: presión arterial sistólica.

En los modelos de regresión, como en la mayor parte de las pruebas de estadística aplicada, uno de los problemas que surge es el de la representatividad de los datos utilizados en el análisis, ya que por cuestiones de logística, de tiempo o de costos en la gran mayoría de los casos es imposible realizar un censo de la población diana. En estos casos se realiza el levantamiento de una muestra, la cual puede o no ser probabilística, aunque lo que se pretende con las técnicas de muestreo es que la muestra sea probabilística para poder extrapolar los resultados obtenidos al resto de la población de la cual se obtuvo la muestra.

Adicional al método de muestreo, es importante que los investigadores comprendan la naturaleza de las poblaciones de las que están interesados en realizar inferencia estadística. Este conocimiento respecto a la población diana debe ser suficiente para que sean capaces de elaborar un modelo matemático que la represente o para que determinen si su modelo se ajusta razonablemente a un modelo ya establecido. Además, no es de esperarse que el modelo sea una representación fiel de la realidad. Afortunadamente la práctica ha mostrado que el modelo tiene utilidad para ser aplicado en la clínica, aunque no se ajuste perfectamente a los datos. Los investigadores deben ser capaces de distinguir si el modelo se ajusta suficientemente a los datos para poder dar soporte o rechazar las hipótesis del estudio.[1]

En el caso particular del análisis de regresión lineal simple se debe estar seguro de que este modelo es el adecuado para proporcionar una representación al menos aproximada de la población. Según la hipótesis y el alcance de la investigación, en ocasiones un modelo de regresión lineal simple no será suficiente para explicar la relación entre las variables independiente y dependiente, por lo que será necesario realizar un análisis de regresión lineal múltiple en el que se incluya más de una variable explicativa. Pero como ya mencionamos en este artículo nos centraremos únicamente en el análisis de regresión lineal simple.

Supuestos del modelo de regresión lineal simple

Para hacer inferencias estadísticas a partir de la muestra, hay supuestos que fundamentan el modelo de regresión

lineal simple y que subyacen en el modelo de la ecuación 1, y son los siguientes:

- Normalidad: los errores tienen una distribución normal con media de cero y con variancia constante de σ^2. Esto quiere decir que los valores de *Y* siguen una distribución normal. Cuando este supuesto no se satisface, antes de realizar un modelo de regresión podría realizarse una transformación de la variable *Y*, en la que la nueva variable se distribuya aproximadamente en forma normal.
- Independencia: esto quiere decir que dos observaciones diferentes cualesquiera –los errores ε_i y ε_j– son estadísticamente independientes; en otras palabras, el valor de un error no depende del valor de cualquier otro error, y por consiguiente, los valores de *Y* de una muestra elegidos y los valores específicos de *X* dados también son independientes. Este supuesto puede ser violado cuando diferentes observaciones se realizan en el mismo individuo en diferentes momentos; por ejemplo, si se tomara el peso de un individuo en diferentes momentos, es de esperarse que los pesos estén relacionados en cada individuo. Cuando este supuesto no se cumple, pueden obtenerse conclusiones estadísticas no válidas.
- Homocedasticidad (homogeneidad de la variancia): este supuesto nos indica que la variabilidad del error es constante y es la misma para todos los errores ε_i, y como consecuencia la variancia de *Y* es la misma para diferentes valores fijos de *X*.
- Linealidad: indica que, una vez dados los valores fijos de *X*, las medias de *Y* forman una línea recta. Esta suposición se expresa simbólicamente así: $_{Y/X} = \beta_0 + \beta_1 X$, donde β_0 es la intercepción del valor promedio de la variable de respuesta *Y* cuando la variable explicativa *X* vale cero. Cuando los valores de la variable explicativa analizados no incluyen al cero, la interpretación de β_0 no tiene sentido. β_1 es la pendiente de la recta.

Estos supuestos son "requisitos" que tienen que cumplirse para que el modelo de regresión lineal simple se considere apropiado para el estudio correspondiente. Cuando alguno de ellos no se cumple, es necesario recurrir a otros métodos estadísticos que no lo requieran,[3] o para el caso de la no normalidad puede realizarse alguna transformación.

Estimación e interpretación de los parámetros del modelo de regresión lineal simple

Existen distintos métodos estadísticos para estimar los parámetros de un modelo de regresión lineal simple, como el de mínimos cuadrados o el de máxima verosimilitud. El método de mínimos cuadrados, que es un primer acercamiento para escoger la recta que mejor se ajuste a los datos, consiste en cuantificar las diferencias entre los valores observados y los valores predichos de todas las posibles rectas; la mejor será la que minimice dichas diferencias.[4]

Si los datos se obtienen a partir de una muestra ($\hat{\beta}_0$ y $\hat{\beta}_1$), se les denomina comúnmente "estimadores de los coeficientes de regresión", y a la diferencia entre el valor observado y el valor predicho ($Yi - \hat{Y}$, $i = 1,2,\ldots,n$) se le conoce como "residuo"; estas expresiones son de gran utilidad para evaluar si los supuestos del modelo se cumplen.

Para estimar los parámetros en el ejemplo de presión arterial sistólica y edad utilizamos el paquete estadístico STATA®. En este paquete el comando para realizar una regresión es *regress*, el cual va seguido de la variable de respuesta y posteriormente de las variables explicativas. Para el modelo de regresión lineal simple sólo se cuenta con una variable explicativa, por lo que la sintaxis queda de la siguiente manera: *regress var_respuesta var_explicativa*. En esta salida obtenemos tres secciones: una donde se aborda un análisis de variancia, otra donde se exponen varias estadísticas y la prueba global F, y otra donde se exponen los coeficientes de regresión estimados, los errores estándar, las pruebas de hipótesis útiles para determinar si los coeficientes de regresión son estadísticamente diferentes de cero, así como los intervalos de confianza (Cuadro 1).

Al modelo construido a partir de los estimadores de los parámetros y de las variables de interés se le denomina "modelo ajustado", y arriba de cada estimador se agrega un símbolo conocido como "signo de intercalación" o *hat*; también desaparece el término de *error*, de ahí que la expresión quede de la siguiente manera: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 Edad$, y una vez sustituidos los valores de la salida de STATA®, se tendría lo siguiente: $\hat{Y}_i = 80.80839 + 0.9547841 * Edad_i$.

La interpretación de este modelo sería que la presión arterial sistólica aumenta en 0.95 mmHg por cada año que se incrementa la edad; esta relación es estadísticamente significativa, ya que puede observarse que el valor de *p* = 0.000 (< 0.05), por lo que con una confianza de 95%

Cuadro 1. Modelo de regresión lineal simple. Presión arterial sistólica explicada por la edad

regress systolic edad

Source	SS	df	MS	Number of obs.	=	64
				F(1, 62)	=	549.46
Model	12155.386	1	12155.386	**Prob. > F**	**=**	**0.0000**
Residual	1371.59841	62	22.1225551	**R-squared**	**=**	**0.8986**
				Adj. R-squared	=	0.8970
Total	13526.9844	63	214.714038	Root MSE	=	4.7035

systolic	Coef.	Std. err.	*t*	*P* > |*t*|	[95% Conf.	interval]
edad	**0.9547841**	0.0407323	23.44	0.000	0.8733615	1.036207
_cons	80.80839	1.771692	45.61	0.000	77.26683	84.34995

se rechaza la hipótesis nula de que el estimador de β_1 sea cero. Cabe aclarar que para que esta interpretación tenga una validez estadística completa es necesario evaluar el modelo y el cumplimiento de los supuestos mediante los residuos. A continuación se explican estas características del modelo de regresión lineal simple.

Estadístico F

Es importante probar si el uso de una recta es adecuado para describir la relación entre las variables de estudio. Para determinar si existe dicha relación, se realiza una tabla ANOVA en el modelo de regresión, la cual prueba la diferencia que hay entre los datos observados y los datos esperados.

El estadístico con el que prueban es una *F* con 1 y $^{(n-2)}$ grados de libertad. Básicamente, la hipótesis que se prueba en el modelo de regresión lineal simple es la siguiente:

- H_0: la variable independiente del modelo no explica de manera significativa la variación de *Y*. Otra forma de interpretar esto es que el coeficiente de correlación es igual a cero o que el coeficiente β_1 es cero; ambos significan que no existe correlación entre las variables.
- H_a: la variable independiente del modelo explica significativamente la variabilidad de *Y*. Otra forma de interpretar esto es que el coeficiente de correlación y el valor β_1 son diferentes de cero.

Para evaluar la hipótesis de trabajo se verifica el valor *p* asociado con *F*, tal como se muestra en la salida anterior de STATA®; para nuestro ejemplo, con una confianza de 95% rechazamos la hipótesis nula, pues la $p = 0.000$ (< 0.05), con lo que concluimos que la variable independiente sí explica de manera significativa la variabilidad de *Y*.

Coeficiente de determinación R^2

Una vez que se traza la recta de regresión en el diagrama de dispersión, es importante disponer de una medida que indique si el modelo ha ajustado bien y que, además, permita decidir si el ajuste lineal es suficiente o si deben buscarse modelos alternativos; para ello se calcula el coeficiente de determinación.

La R^2 nos da una medida cuantitativa de qué tan bien el modelo ajustado predice a la variable independiente, y la medida usualmente se expresa en forma de porcentaje. Para el modelo anterior obtuvimos una R^2 de 0.8986; esto significa que la edad explica 89% de la variabilidad total de la presión arterial sistólica; entonces surge la pregunta de si esto… ¿es mucho o es poco? La respuesta es: depende. Por lo general, son muchos los factores que influyen en las enfermedades complejas, por lo que encontrar un coeficiente de determinación de esta magnitud es bastante decoroso; en la mayor parte de los casos el investigador establece si la variabilidad explicada es suficiente.

Prueba de hipótesis para el coeficiente de regresión β_1

Al realizar un modelo de regresión lineal simple es importante evaluar si la variable independiente realmente está explicando algo sobre el comportamiento de la variable de respuesta. Continuando con el ejemplo anterior, se busca probar si el coeficiente de regresión de la edad explica algo de la variabilidad de la presión arterial sistólica, o en otras palabras, si el coeficiente es distinto de cero. Para ello

se aplica una prueba de hipótesis en la que se supone lo siguiente: H_0: $\beta_1 = 0$ y H_a: $\beta_1 \neq 0$. Para probar lo anterior, se utiliza un estadístico de prueba *t*, y con un valor *p* se determina si se acepta o no la hipótesis.

Con la salida anterior de STATA®, mediante esta prueba de hipótesis, se observa que la edad sí aporta información para la predicción de la presión arterial sistólica, ya que con el valor de $p = 0.000$ (< 0.05), con un intervalo de confianza de 95%, se rechaza la hipótesis nula (H_0), y se concluye que el coeficiente de la edad es diferente de cero.

Prueba de hipótesis para el coeficiente de regresión β_0

La β_0 u ordenada al origen, que representa el valor de *Y* cuando $X = 0$, pocas veces obtiene una interpretación de interés o razonable, ya que los datos de la variable *X* no están cerca del origen. En nuestro ejemplo el valor del coeficiente β_0 –es decir, el valor de la presión arterial sistólica cuando la edad es cero– es aproximadamente 80.8, pero recordemos que es incorrecto hacer estimaciones que abarquen límites de edad que no se analizaron, por lo que en este caso la interpretación β_0 no tiene sentido. En STATA® aparece en el cuadro de resultados como: *_cons*. Los criterios para probar la hipótesis de que β_0 es distinta de cero son equivalentes a los utilizados en la prueba de hipótesis de β_1.

Intervalos de confianza

En muchos estimadores estadísticos –además de la estimación puntual–, es importante conocer el intervalo de confianza de éstos. Respecto a los estimadores de los coeficientes de regresión lineal simple, los intervalos de confianza pueden determinar –además de indicar el límite en el que se encuentran los coeficientes (β_0 y β_1)– si los coeficientes de regresión lineal simple son estadísticamente diferentes de cero o no. Si el intervalo incluye el cero, se dice que los coeficientes son estadísticamente distintos de cero, mientras que cuando el intervalo no incluye el cero, se dice que los coeficientes no son estadísticamente distintos de cero. Las fórmulas para calcular los intervalos de confianza de cada estimador son las siguientes:

$\beta_0 \pm t^{1-\alpha/2}_{n-2} *$ error estandar de β_0, y

$\beta_1 \pm t^{1-\alpha/2}_{n-2} *$ error estandar de β_1, donde $t^{1-\alpha/2}_{n-2}$ es el valor del estadístico *t* con $n-2$ grados de libertad y con una confianza usual de 95%. En nuestro ejemplo, en la salida de STATA®, pudimos observar que el intervalo de confianza de 95% para el coeficiente β_1 va de 0.87 a 1.04, lo que indica que el verdadero coeficiente está entre estos valores.

Validación de supuestos para el ejemplo de presión arterial sistólica y edad

Como se mencionó anteriormente, para que el modelo sea válido es importante probar algunos supuestos de regresión lineal; para ello se generan los residuos del modelo, los cuales son la diferencia entre el valor observado y el valor predicho. Con estos residuos se prueban todos los supuestos, a excepción del supuesto de independencia.

Supuesto de normalidad

Se considera que cada residuo tiene una distribución normal; para probar lo anterior pueden aplicarse varias pruebas, y una de ellas es la de Shapiro-Francia (Cuadro 2). En las pruebas de normalidad la hipótesis es la siguiente:

- H_0: los residuos se distribuyen como una función normal.
- H_a: los residuos no se distribuyen como una función normal.

En este caso nos interesa no rechazar la hipótesis nula (H_0), y cuando el valor de *p* sea mayor que 0.05, no rechazaremos la hipótesis nula (H_0). En la salida de STATA® de nuestro ejemplo pudimos observar que el valor de $p = 0.91$ (> 0.05), por lo que concluimos que los residuos se distribuyen de manera normal, y por consiguiente, nuestra variable de respuesta también se distribuye de manera normal.

Supuesto de homocedasticidad

Para probar este supuesto se aplica la prueba estadística de Breusch-Pagan, la cual se distribuye como una prueba de la ji al cuadrado. En esta prueba la hipótesis nula (H_0) indica que las variancias son iguales, mientras que la hipótesis alterna (H_a) indica que las variancias son distintas.

De igual manera que en el supuesto de normalidad nos interesa no rechazar la hipótesis nula (H_0), y cuando el valor de *p* sea mayor que 0.05, no rechazaremos la hipótesis nula (H_0). En nuestro ejemplo el valor de *p* =

Cuadro 2. Prueba de Shapiro-Francia para datos normales

Variable	Obs.	W'	V'	z	Prob. > z
residuos_sts.	64	0.99194	0.508	-1.369	**0.91454**

0.8293 (> 0.05), por lo que no rechazamos la hipótesis nula (H_0), y concluimos que las variancias de los errores son homocedásticas, y como consecuencia, la variancia de *Y* es la misma para los diferentes valores fijos de la edad (Cuadro 3).

Cuadro 3. Prueba de Breusch-Pagan/Cook-Weisberg para heterocedasticidad

H_0:	Variancia constante
Variables:	Valores ajustados a la presión sistólica
chi2(1)	= 0.05
Prob. > chi2	= 0.8293

Supuesto de linealidad

Para verificar este supuesto normalmente se busca mediante un diagrama que no forme patrones en la nube de puntos. En la distribución de los puntos de la Figura 4 no se observa ningún patrón específico, por lo que se concluye que el supuesto de linealidad sí se cumple.

EPÍLOGO

La regresión lineal simple es una técnica estadística que permite evaluar la fuerza de asociación entre dos variables. El coeficiente de determinación es la medida más importante para dar un sentido clínico a la importancia de la fuerza de asociación entre dos variables. Este coeficiente es una medida de la variancia que comparten las dos variables.

No basta con calcular los coeficientes de correlación y regresión, pues también deben evaluarse el ajuste de los

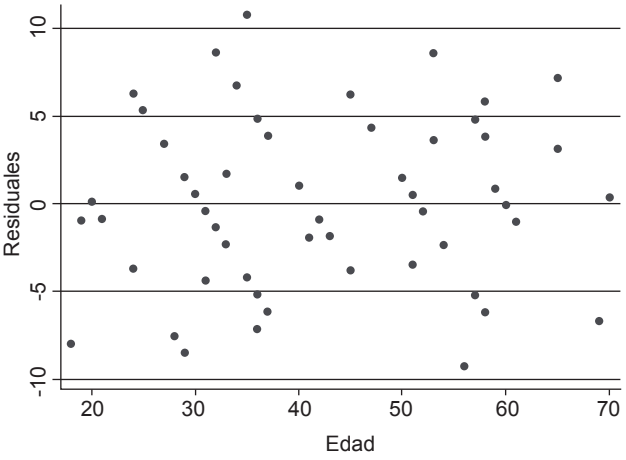

Figura 4. Diagrama de dispersión para probar el supuesto de linealidad.

datos y el cumplimiento de los supuestos de normalidad, homocedasticidad y linealidad que la regresión requiere.

La técnica de regresión lineal múltiple es una herramienta que puede utilizarse para disminuir el error de la regresión simple.

REFERENCIAS

1. Daniel WW. Bioestadística: base para el análisis de las ciencias de la salud. 4ª ed. Ciudad de México: Limusa Wiley, 2011.
2. Motgomery DC, Perk EA, Vining GG. Introducción al análisis de la regresión lineal. 3ª ed. Ciudad de México: CECSA, 2004.
3. Draper NR, Smith H. Applied regression analysis. Wiley Series in Probability and Statistics. 3rd ed. Washington: Wiley-Interscience, 1998.
4. Sotres D, Téllez Rojo MM. Regresión lineal simple. En: Hernández M, editor. Epidemiología: diseño y análisis de estudios. 1ª ed. Ciudad de México: Editorial Médica Panamericana, 2007.