Modelación estadística: La regresión lineal simple

Gabriel Cavada Ch.[1]
[1]*División de Bioestadística, Escuela de Salud Pública, Universidad de Chile.*

Statistical modeling: Simple linear regression

Cuando se observa un efecto, es inherente al pensamiento científico, buscar la o las causas que lo produjeron; esto es debido al estilo de pensamiento que poseemos de que "un conjunto de causas genera un efecto". Al simplificar esta estructura cognoscitiva podemos pensar que una respuesta es generada por una causa; lo que podemos representar, cuando causa y efecto son medibles numéricamente, por una relación funcional:

$$Y = f(X)$$

Efecto o respuesta — Causa

Particularmente, nos interesa modelar la respuesta cuando la relación funcional entre la respuesta y la causa es lineal, es decir, de la forma:

$$Y = \beta_0 + \beta_1 \cdot X$$

Obviamente, antes de ajustar un modelo como el propuesto es necesario saber si la variable respuesta se asocia linealmente con la variable independiente, cuando ambas se miden en n unidades de análisis, esto es, cuando se tiene una muestra de la forma:

Observación	X	Y
1	x1	y1
2	x2	y2
3	x3	y3
4	x4	y4
…	…	…
…	…	…
…	…	…
n	xn	yn

Para ello, definimos el Coeficiente de Correlación de Pearson entre X e Y como:

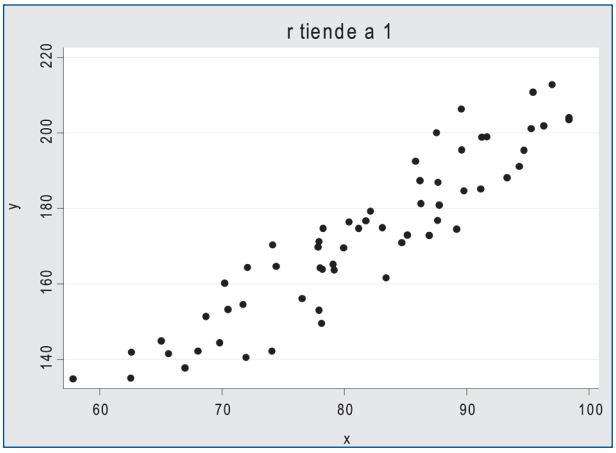
$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2}\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$$

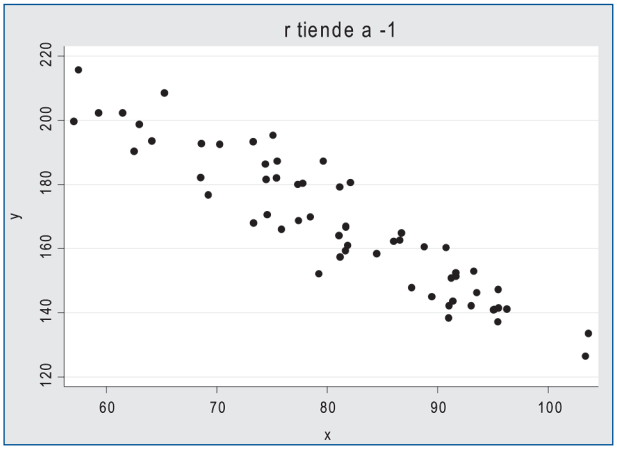
que mide el grado de asociación lineal entre X e Y, pudiendo demostrarse que:

$$-1 \le r_{xy} \le 1$$

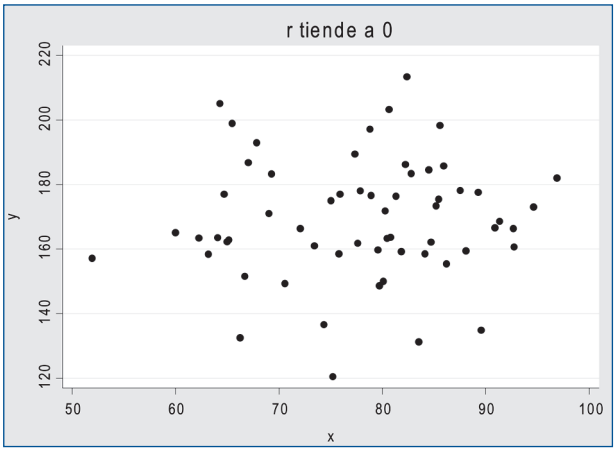
y que si:

- r_{xy} tiende a 1 la asociación es directa
- r_{xy} tiende a -1 la asociación es inversa
- r_{xy} tiende a 0 no existe asociación lineal





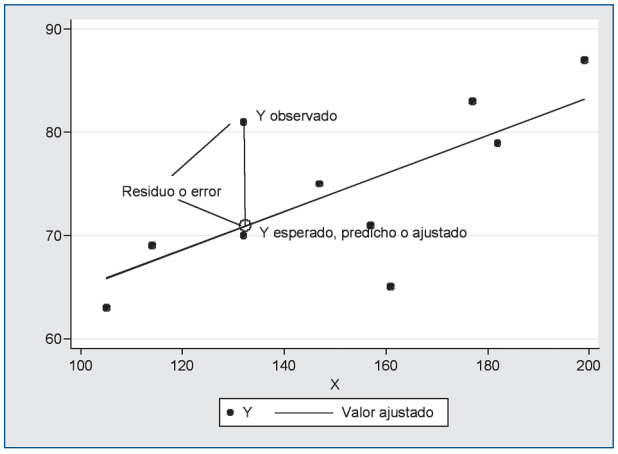
Rincón de la Bioestadística



Estimación de la recta de regresión

I. El método de los mínimos cuadrados: Se basa en la minimización de la suma de los errores (distancia entre el valor observado de Y y el respectivo valor estimado), es decir suponemos que:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$



Se trata de minimizar la función:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i)^2$$

Los valores que minimizan la función Q, se obtienen resolviendo el sistema de ecuaciones:

$$\frac{\delta Q}{\delta \beta_0} = 0$$

$$\frac{\delta Q}{\delta \beta_1} = 0$$

$$\begin{cases} n\beta_0 + \beta_1 \sum X_i = \sum Y_i \\ \beta_0 \sum X_i + \beta_1 \sum X_i^2 = \sum X_i Y_i \end{cases}$$

Estas ecuaciones reciben el nombre de ecuaciones normales, cuyas soluciones son:

$$\hat{\beta}_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \cdot \overline{X}$$

Que corresponden a los estimadores mínimo cuadráticos. Es fácil probar que:

$$\hat{\beta}_1 = \frac{S_y}{S_x} r_{xy}$$

II. Estimación por máxima verosimilitud: Se basa en suponer que la respuesta Y condicionada a un valor de X sigue una distribución normal, cuya esperanza está sobre la recta de regresión y cuya varianza, σ^2, es constante (homocedasticidad), en símbolos:

$$Y_i \sim N(\beta_0 + \beta_1 \cdot X_i, \sigma^2)$$

El supuesto anterior es equivalente a decir que los errores o residuos de la estimación siguen una distribución normal, cuya esperanza es 0 y su varianza es σ^2, es decir:

$$\varepsilon_i \sim N(0, \sigma^2)$$

Los estimadores para el intercepto y la pendiente de la recta, bajo estos supuestos, son idénticos a los estimadores conseguidos por el método de los mínimos cuadrados; la ventaja del supuesto distribucional es que podemos plantear inferencias y test de hipótesis sobre los parámetros estimados.

Una vez ajustado el modelo, la tabla de datos se extiende así:

Observación	X	Y Valor observado de Y	$\hat{Y} = a + b \bullet X$ Valor estimado de Y	$\varepsilon = Y - \hat{Y}$ Residuo o Error
1	x_1	y_1	$\hat{Y}_1$	ε_1
2	x_2	y_2	$\hat{Y}_2$	ε_2
3	x_3	y_3	$\hat{Y}_3$	ε_3
4	x_4	y_4	$\hat{Y}_4$	ε_4
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
n	x_n	y_n	$\hat{Y}_n$	ε_n

Una vez ajustado un modelo de regresión, es necesario conocer la calidad del mismo, para ello la variabilidad total de Y, que no depende del modelo ajustado, puede descomponerse en las llamadas "SUMAS DE CUADRADO" (SC), correspondientes a la varianza total, residual y debida al modelo o regersión:

$$\sum(Y-\overline{Y})^2=\sum(Y-\hat{Y})^2+\sum(\hat{Y}-\overline{Y})^2$$

SCTotal = SCResidual + SCRegresión

Los componentes de esta descomposición también reciben el nombre de Varianza total, no explicada y explicada, de modo que:

Varianza Total = Varianza no explicada + Varianza explicada

La varianza total es una cantidad fija, pues es sólo la varianza de la respuesta, no siendo así en la no explicada y explicada, pues dependen del modelo ajustado. Si el modelo ajustado fuera perfecto la varianza no explicada sería 0 y por consiguiente la varianza explicada sería igual a la varianza total. Este hecho nos lleva a definir como medida de la calidad del modelo el coeficiente de determinación como:

$$R^2=\frac{\sum(\hat{Y}-\overline{Y})^2}{\sum(Y-\overline{Y})^2}=\frac{\text{Varianza explicada}}{\text{Varianza no explicada}}$$

que en el caso de la regresión lineal simple coincide con r^2_{xy}.

La descomposición de la variabilidad es posible resumirla en la conocida Tabla de Análisis de la Varianza (Tabla ANOVA)

Fuente de Variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F
Regresión	1	$\sum(\hat{Y}-\overline{Y})^2$	$CMreg=\frac{\sum(\hat{Y}-\overline{Y})^2}{1}$	$F=\frac{CMreg}{CMres}$
Residuo	n-2	$\sum(Y-\hat{Y})^2$	$CMres=\frac{\sum(Y-\hat{Y})^2}{n-2}$	
Total	n-1	$\sum(Y-\overline{Y})^2$		

Asociada a la descomposición de la variabilidad y por ende a la calidad del modelo, se tiene la siguiente dócima:

$$H_0:\Re^2=0$$

$$H_1:\Re^2>0$$

Cuya estadística de prueba es:

$$F=\frac{CMreg}{CMres}\sim F(1,n\text{-}2)$$ (distribución F-Snedecor)

La estimación de la varianza del error es:

$$S^2=CMres=\frac{\sum(Y-\hat{Y})^2}{n-2}$$

La interpretación de la pendiente de la recta de regresión

Si el modelo que se estima es de la forma:

$$y=\alpha+\beta\bullet X$$

Considerando:

a) $y\,(X+1)=\alpha+\beta\,(X+1)=\alpha+\beta X+\beta$

b) $y\,(X)=\alpha+\beta\bullet Xy(X)=\alpha+\beta\bullet X$

Restando a) y b)
Se obtiene:

$$y\,(X+1)-y\,(x)=\beta$$

Es decir la pendiente de la recta, es el cambio de "Y" por unidad de "X".

Ejemplo: Para los siguientes pares de tallas en centímetros y pesos en kilos, ¿Cómo y cuánto explica la talla al peso? (Ver tabla)

id	talla	peso	id	talla	peso	id	talla	peso	id	talla	peso	id	talla	peso
1	147	75	11	128	70	21	190	91	31	139	75	41	201	97
2	157	71	12	211	96	22	166	86	32	112	72	42	162	69
3	177	83	13	175	72	23	139	76	33	147	81	43	208	86
4	132	81	14	142	86	24	122	75	34	186	88	44	164	77
5	182	79	15	143	82	25	192	81	35	188	93	45	161	79
6	105	63	16	199	98	26	173	73	36	176	72	46	209	86
7	161	65	17	133	70	27	166	85	37	209	98	47	193	87
8	132	70	18	149	68	28	151	73	38	145	76	48	192	76
9	114	69	19	190	89	29	193	84	39	156	82	49	173	88
10	199	87	20	190	84	30	115	60	40	206	75	50	204	85

Rincón de la Bioestadística

Para estos datos, el coeficiente de correlación es r = 0,7112 y la salida de STATA versión 10.1 para la estimación del modelo es:

Tabla ANOVA (Source: fuente de variación; SS: Suma de cuadrados; df: grados de libertad; MS: cuadrado medio; F: valor de la estadística F; Prob>F: p-value asociado a la dócima de existencia del modelo; R-squared: coeficiente de determinación y Root MSE: estimación de la desviación estándar residual).

Source	SS	df	MS	Number of obs = 50
-------	-------	-------	-------	F (1, 48) = 49.13
Model	2104.73925	1	2104.73925	Prob > F = 0,0000
Re-sidual	2056.14075	48	42.8362656	R-squared = 0,5058
-------	-------	-------	-------	Adj R-squared = 0,4955
Total	4160.88	49	84.9159184	Root MSE = 6,5449

En negritas se muestra el p-value asociado a la dócima de existencia del modelo, es decir el coeficiente de determinación es significativamente mayor que 0.

Estimación de la pendiente e intercepto del modelo (Coef.: pendiente y cons es el intercepto, luego los errores estándar, la estadística t-student asociada a la dócima cuya hipótesis nula es pendiente igual a cero e intercepto igual a cero, los respectivos p-values y sus intervalos de confianza).

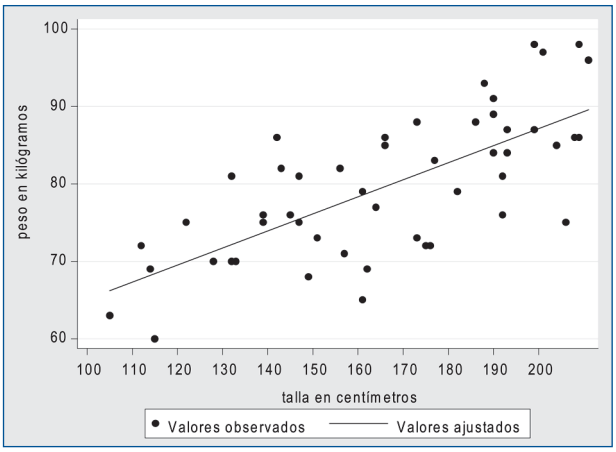
Peso	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
talla	,2206623	,03148	7,01	0,000	,1573675	,2839572
cons	43,0324	5,309501	8,10	0,000	32,35693	53,70787

Estos valores señalan que la pendiente de la recta se estimó en 0,22, es decir que por cada centímetro de talla se espera un incremento en el peso de 0,22 kilógramos. Los p-values en negritas señalan que tanto la pendiente como el intercepto son significativamente distintos de 0.

La recta de regresión estimada es:

$$peso = 43{,}03 + 0{,}22 \bullet talla$$

El siguiente gráfico describe la situación:

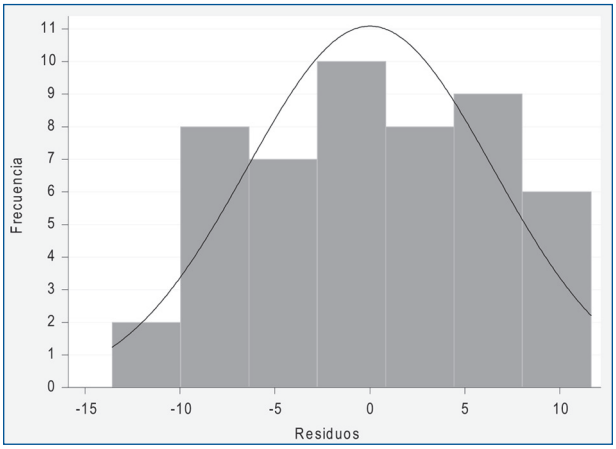


Diagnóstico en regresión lineal simple

Cuando se ajusta un modelo de regresión lineal simple, inicialmente sólo son creíbles las estimaciones de los parámetros, no así las inferencias hechas sobre ellos (test de hipótesis e intervalos de confianza). Para tener algún grado de certeza sobre la calidad de las inferencias, es necesario revisar el cumplimiento de los supuestos del modelo:

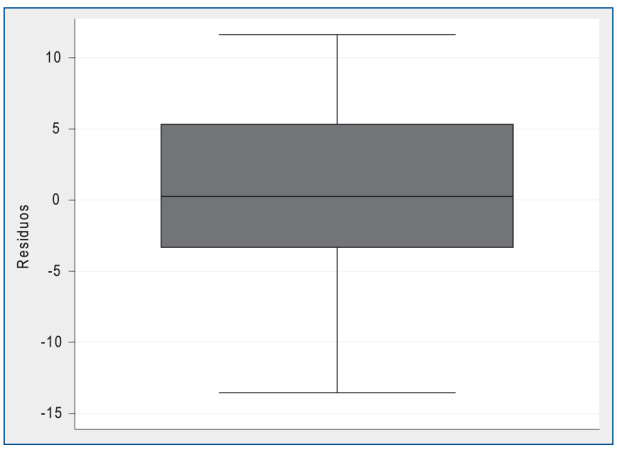
a) La normalidad de los residuos o errores: Luego de estimar el modelo se calculan los errores o residuos, que en este ejemplo llamaremos "resid".

Un histograma de estos residuos se muestra en la siguiente figura:

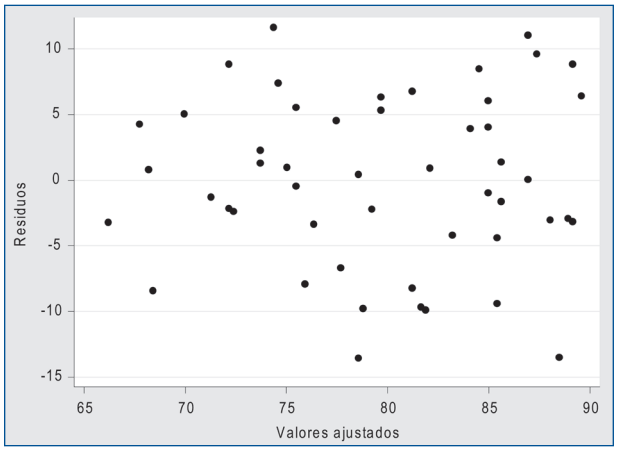


El histograma no deja claro que los residuos sean normales, sin embargo el test de Shapiro Wilk indica que no hay evidencia para decir que los residuos no son normales (p = 0,3315).

Sin embargo, la regresión lineal simple es "robusta" (esto es que se puede relajar el supuesto de normalidad de los errores) si es que estos tienen una distribución razonablemente simétrica, es decir si al rechazar la hipótesis nula en el test de Shapiro Wilk, aún se obtiene una distribución simétrica de los residuos, es decir una gráfico de cajas similar al siguiente:



b) La homocedasticidad de los residuos o errores: Para probar este supuesto se comienza observando el gráfico entre los residuos y los valores predichos, en dicho gráfico no debería observarse ningún patrón de comportamiento:



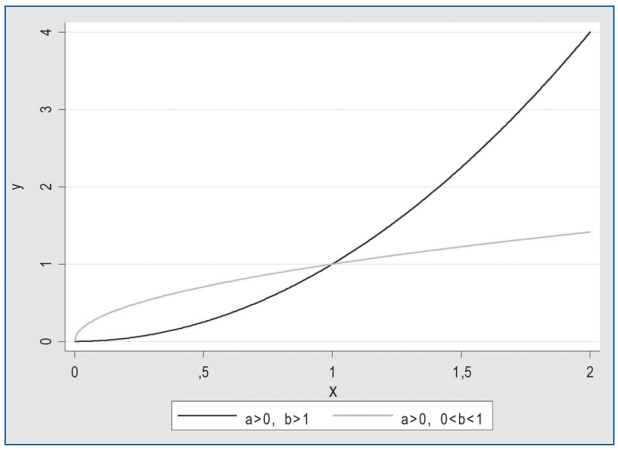
También podemos usar el test de herocedasticidad de Breusch-Pagan o Cook-Weisberg, cuya hipótesis nula es que la varianza de los residuos es constante, que en este caso entrega un p-value = 0,3269.

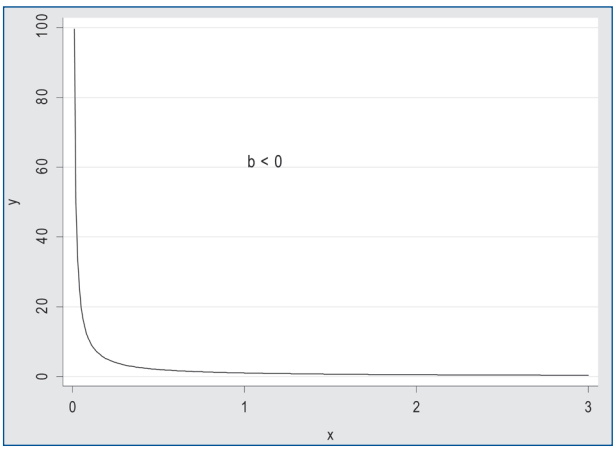
Ahora, podemos confiar en las inferencias del modelo.

Modelos linealizables

En muchas oportunidades una respuesta "Y" no depende linealmente de la causa "X", sin embargo, la relación funcional, mediante una sencilla transformación matemática la convierte en una relación lineal a la cuál le podemos aplicar la metodología de la regresión lineal simple. En aplicaciones a la medicina aparecen los siguientes modelos:

1. $Y = \alpha \bullet X^b$, modelo de potencias, cuya forma se muestra a continuación:

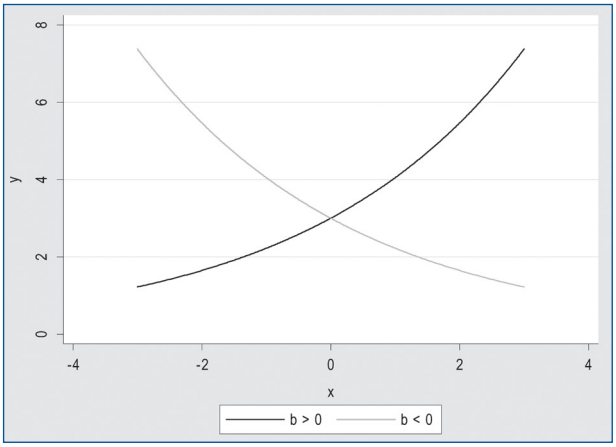




Tomando logaritmo al modelo:

1) $\ln(Y) = \ln(\alpha) + b \bullet \ln(X)$, los parámetros involucrados se estiman aplicando regresión lineal sobre el logaritmo de la respuesta y sobre el logaritmo de la causa.

2) $Y = \alpha \bullet e^{b \bullet X}$, modelo exponencial, cuya forma se muestra a continuación:



Tomando logaritmo al modelo:

$\ln(Y) = \ln(\alpha) + b \bullet X$, los parámetros involucrados se estiman aplicando regresión lineal sobre el logaritmo de la respuesta y sobre la causa.

Como se observa, la flexibilidad del modelo lineal simple es muy amplia, lo que lo convierte en un modelo muy recurrido y noble.