

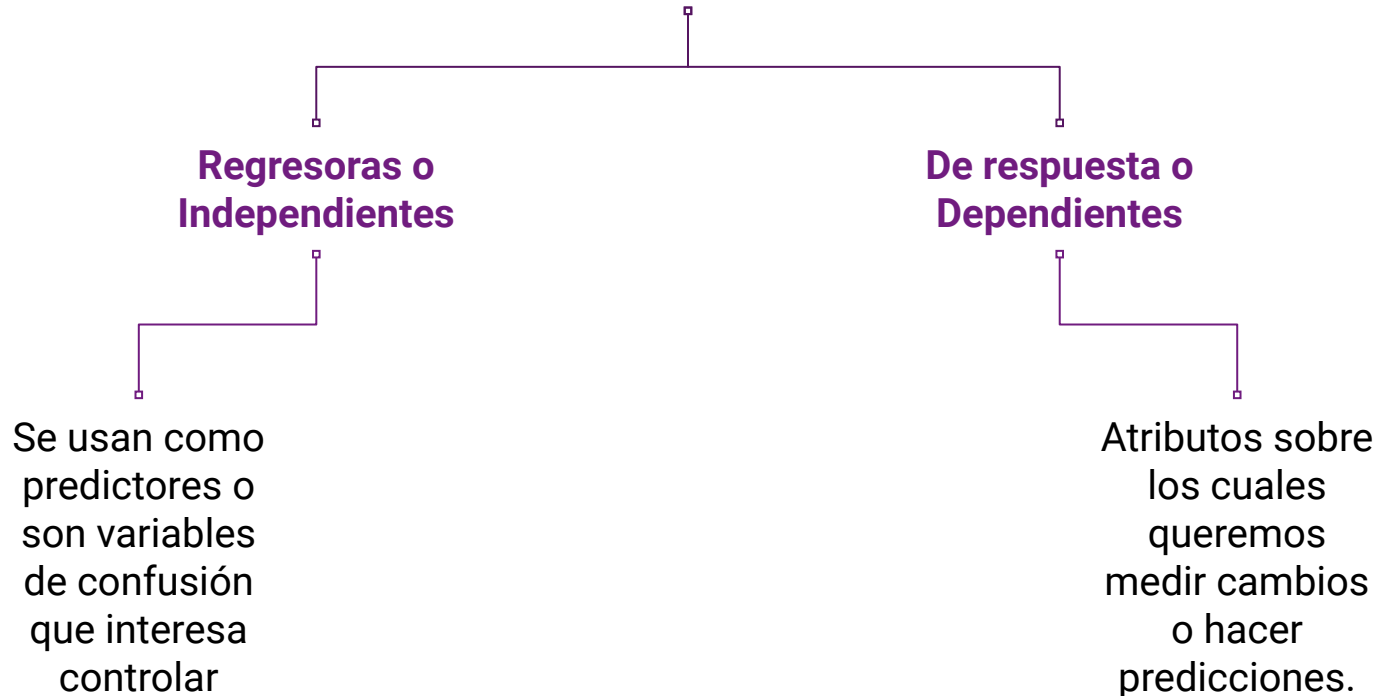
Regresión Lineal

EQUIPO 1:

Alfonso de Jesús García Gutiérrez	1853849
Angela Scarlett Cerecero Peña	1851124
José Antonio Aguilar Sánchez	1851167
José Manuel Romero Banda	1851528
Isabel Alejandra Rangel Vallejo	1848655

Tipos de variables en regresión lineal

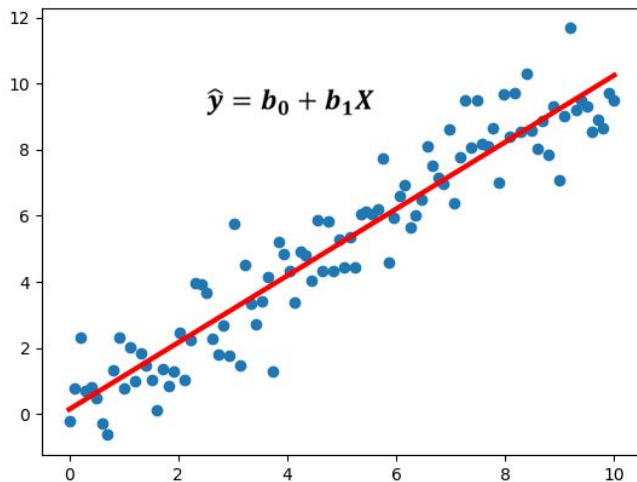
Variables



Regresión lineal simple

Un modelo de regresión lineal simple es un modelo de regresión donde interviene una variable regresora X , que tiene una relación con una respuesta Y , donde la relación es una línea recta. Este modelo es:

$$y = \beta_0 + \beta_1 x + e$$



Regresión lineal simple

Donde la ordenada al origen β_0 y la pendiente β_1 son constantes desconocidas, y \mathbf{e} es un componente aleatorio de error.

A los parámetros β_0 y β_1 se les suele llamar coeficientes de regresión. Estos se pueden estimar por el método de mínimos cuadrados con las siguientes fórmulas:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}$$

Datos para regresión lineal simple

Supongamos que se dispone de n observaciones, con y_i la i -ésima respuesta observada, y x_i la i -ésima observación.

OBSERVACIÓN	RESPUESTA	REGRESOR
i	Y	X
1	y_1	x_1
2	y_2	x_2
3	y_3	x_3
\vdots	\vdots	\vdots
n	y_n	x_n

Modelos Linealizables

Modelo	Función Linealizable	Transformación	Modelo Lineal Asociado
Potencia	$y = \beta_0 x^{\beta_1}$	$y^* = \ln(y) \quad x^* = \ln(x)$	$y^* = \ln(\beta_0) + \beta_1 x^*$
Exponencial	$y = \beta_0 e^{\beta_1 x}$	$y^* = \ln(y)$	$y^* = \ln(\beta_0) + \beta_1 x$
Logaritmo	$y = \beta_0 + \beta_1 \ln(x)$	$x^* = \ln(x)$	$y = \beta_0 + \beta_1 x^*$
Recíproco	$y = \frac{x}{\beta_0 x - \beta_1}$	$y^* = \frac{1}{y} \quad x^* = \frac{1}{x}$	$y^* = \beta_0 - \beta_1 x^*$

Prueba de significancia del modelo

Prueba de hipótesis

H0: $\beta_1 = 0$ (la regresión no es significativa, las variables involucradas no muestran relación)

H1: $\beta_1 \neq 0$ (la regresión es significativa, las variables involucradas muestran relación)

Rechazamos H0 si p valor $< \alpha$

Donde α = Nivel de significancia de la prueba.

Coefficiente de determinación (R^2)

El coeficiente de determinación se define como la proporción de variabilidad de la variable dependiente que es explicada por la regresión :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

Los valores de R^2 cercanos a 1 implican que la mayor parte de la variabilidad de y está explicada por el modelo de regresión.

Nivel de desempeño según R^2

Nivel de desempeño	Coefficiente de determinación
MUY BUENO	$0.9 < R^2 < 1$
BUENO	$0.7 < R^2 < 0.9$
MODERADO	$0.4 < R^2 < 0.7$
BAJO	$0.2 < R^2 < 0.4$
NULO	$0 < R^2 < 0.2$

Regresión lineal múltiple

Un modelo de regresión donde interviene más de una variable regresora, supongamos **k**, se llama modelo de regresión múltiple un modelo de regresión múltiple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos $\beta_0, \beta_1, \dots, \beta_k$.

En general, se puede relacionar la respuesta y con los **k** regresores, o variables predictivas bajo el modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e$$

Datos para regresión lineal múltiple

Supongamos que se dispone de $n > k$ observaciones con y_i la i -ésima respuesta observada, y x_{ij} la i -ésima observación o nivel del regresor x_j con $j=0, 1, 2, \dots, k$.

OBSERVACIÓN	RESPUESTA	REGRESORES			
i	y	X_1	X_2	...	X_k
1	y_1	x_{11}	x_{12}	...	x_{1k}
2	y_2	x_{21}	x_{22}	...	x_{2k}
3	y_3	x_{31}	x_{32}	...	x_{3k}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
n	y_n	x_{n1}	x_{n2}	...	x_{nk}

Multicolinealidad

Un problema serio que puede influir mucho sobre la utilidad de un modelo de regresión es la multicolinealidad, o dependencia casi lineal entre las variables de regresión.

Se mide con los factores de inflación de varianza (VIF, de Variance Inflation Factors)

$$VIF_j = \frac{1}{1 - R_j^2}$$

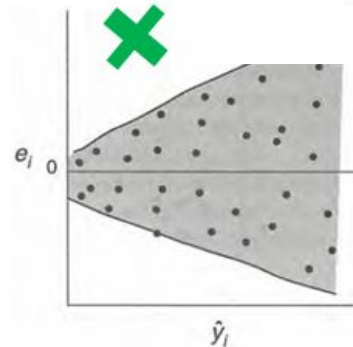
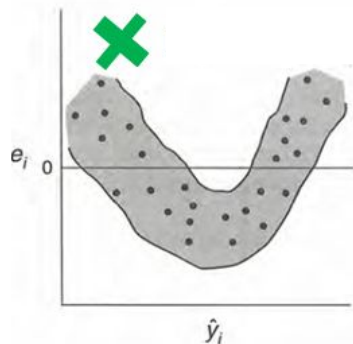
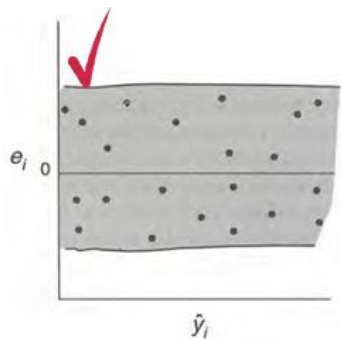
R_j^2 = Coeficiente de determinación múltiple obtenido haciendo la regresión x_j sobre las demás variables regresoras.

Si $VIF > 10$ hay problemas graves de multicolinealidad.

Cumplimiento de los supuestos

Después de elegir el mejor modelo significativo ya sea en base a regresión simple o múltiple hay que verificar el cumplimiento de los supuestos que son los siguientes:

1. El término del error tiene **media cero**.
2. El término del error tiene **varianza constante**.

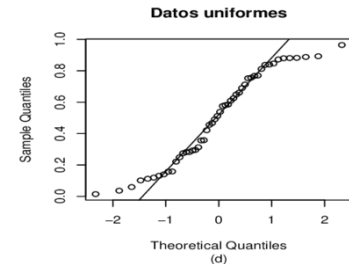
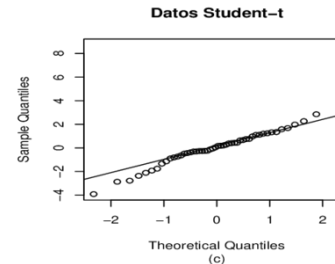
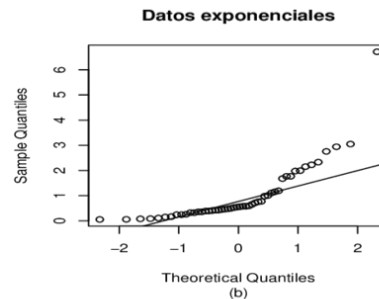
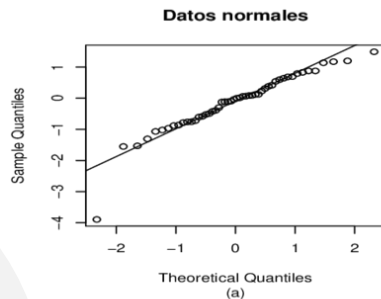


3. Los errores **no están correlacionados**.

Hipótesis	$H_0: \rho = 0$ equivalente a la incorrelación de los datos analizados		$H_1: \rho > 0$
Estadístico de prueba	$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$		
Criterio de decisión	Rechazo H_0 si $d < d_L$	No Rechazo H_0 si $d > d_U$	Prueba no concluyente si $d_L < d < d_U$

		$k = \text{Number of Regressors (Excluding the Intercept)}$									
		1		2		3		4		5	
Sample Size	Probability in Lower Tail (Significance Level = α)	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	.01	.81	1.07	.70	1.25	.59	1.46	.49	1.70	.39	1.96
	.025	.95	1.23	.83	1.40	.71	1.61	.59	1.84	.48	2.09
	.05	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21

4. Los errores tienen **distribución normal**.



Bibliografía

- J.A.R. (2016, 11 julio). *RPubs - Regresión Lineal Múltiple en R*. RPubS by RStudio.

https://rpubs.com/Joaquin_AR/226291?fbclid=IwAR0HLj4emZiRUGTHXkgiFRru-kUwBMD3QN5xBuWCn_vkkFpRZfn_Mwkw9bM

- Orellana, L. (2008). Regresión Lineal Simple. UBA. <http://web.dm.uba.ar/>
- Cerda, A..(2021). Análisis de Regresión Lineal Múltiple.
- Cerda, A .(2021). Verificación de cumplimiento de los supuestos del modelo.