
MAGNA EDUCATIO AD MELIORA
UNIVERSIDAD DE SANTANDER
Universidad de Santander
UDES
VIGILADA MINEDUCACIÓN | SNIES 2832

Estadística

Regresión lineal simple

Daniel Martínez Bello

Universidad de Santander
Maestria en Biotecnología

Diciembre de 2025

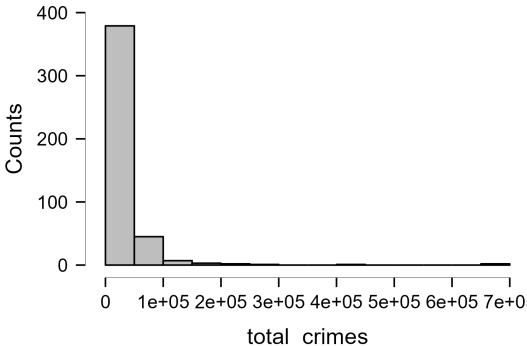
Cuando usted se vea involucrado en un estudio experimental o observacional y el interés se centre en hallar la relación entre un grupo de factores y una variable respuesta se puede utilizar una técnica llamada Regresión lineal.

Base crimenes

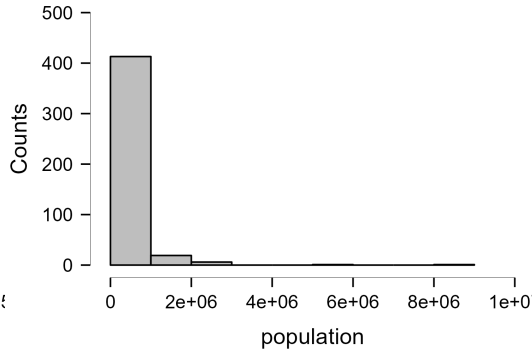
- total_crimes: conteo de crimenes por ciudad
- land_area: area de la ciudad
- population: población de la ciudad
- population_18_34: población de 18 a 34 años
- population_65_older: población mayor de 65 años
- active_physicians: número de médicos activos
- hospital_beds: número de camas hospitalarias
- percentage_high_graduates: porcentaje de graduados de high scholl
- percentage_bachelor_degrees: porcentaje de graduados universitarios de pregrado
- percentage_below_poverty: porcentaje de población por debajo de la línea de pobreza
- percentage_unemployment: porcentaje de desempleo
- percapita_income: ingresos per capita
- total_income: ingresos totales

	Median	Mean	Std. Deviation	Minimum	Maximum
total_crimes	11820,5	27111,6	58237,5	563,0	688936,0
land_area	656,5	1041,4	1549,9	15,0	20062,0
population	217280,5	393010,9	601987,0	100043,0	$8{,}9 \times 10^{+6}$
population_18_34	28,1	28,6	4,2	16,4	49,7
population_65_older	11,8	12,2	4,0	3,0	33,8
active_physicians	401,0	988,0	1789,7	39,0	23677,0
hospital_beds	755,0	1458,6	2289,1	92,0	27700,0
percentage_high_graduates	77,7	77,6	7,0	46,6	92,9
percentage_bachelor_degrees	19,7	21,1	7,7	8,1	52,3
percentage_below_poverty	7,9	8,7	4,7	1,4	36,3
percentage_unemployment	6,2	6,6	2,3	2,2	21,3
percapita_income	17759,0	18561,5	4059,2	8899,0	37541,0
total_income	3857,0	7869,3	12884,3	1141,0	184230,0

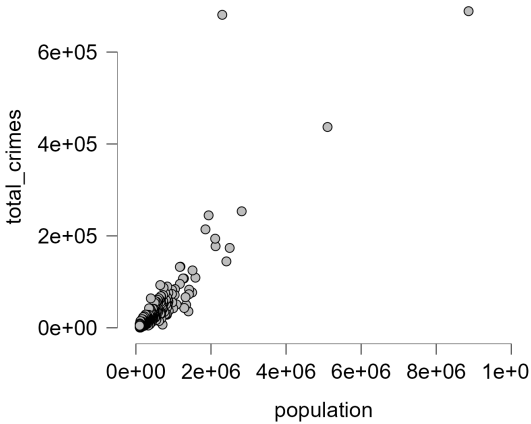
(a) Conteo de crimenes



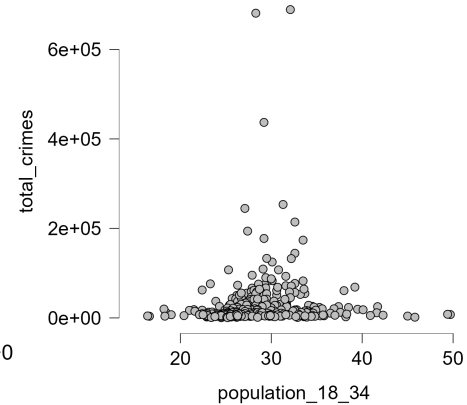
(b) Población total



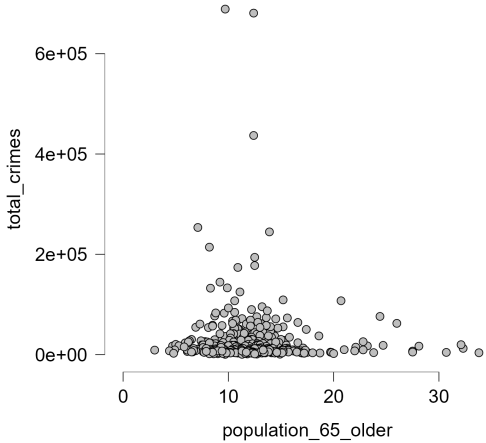
(a) Conteo de crimenes y población



(b) Crimenes y población entre 18 y 34 años



(a) Crimenes y población mayor de 65 años



(b) Crimenes y médicos

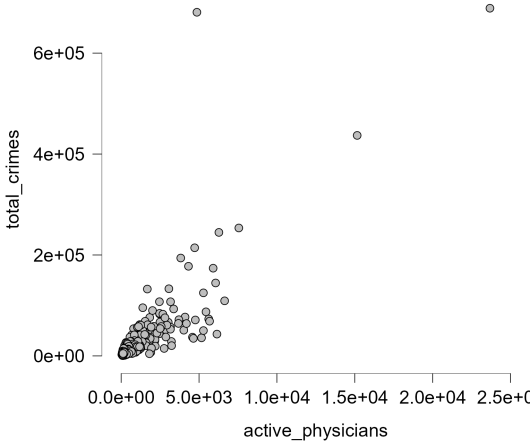


Figura: Gráfico de dispersión conteo de crimenes y población total

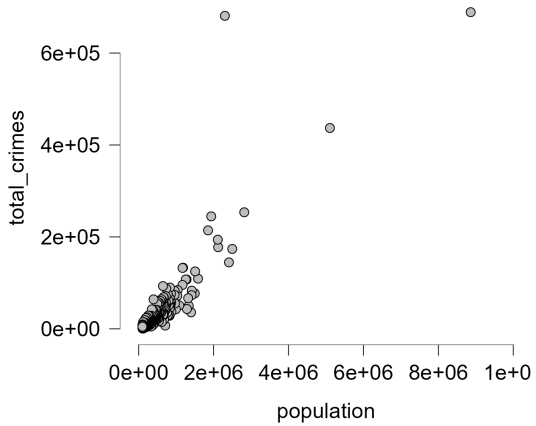
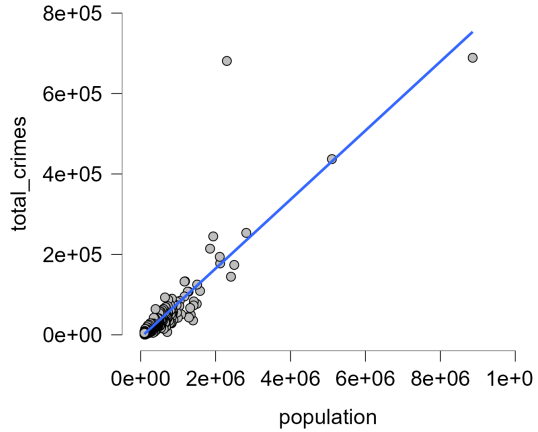


Figura: Recta de regresión conteo de crimenes y población total



Como generar la recta de regresión?

- La línea de regresión se genera de los datos, en este caso de la variable x que corresponde a población y la variable y que corresponde a crimenes.
- Se utiliza una técnica de la matemática denominada “estimación por mínimos cuadrados”.
- El método de estimación genera una ecuación que corresponde a una ecuación de la línea recta mas un componente denominado el “error”, que muestra como las observaciones se comportan con respecto a la línea recta que resume los datos.

Como generar la recta de regresión

- Aplicando los “mínimos cuadrados a los datos” obtenemos la ecuación de regresión y esa ecuación nos configura un MODELO del experimento.

- El modelo de regresión lineal simple es

$$y_i = \alpha + \beta * x_i + \epsilon_i$$

- Donde α es el intercepto (o sea el valor de y donde la recta cruza), β es el valor de la pendiente de la recta, y ϵ_i corresponde a la desviación de cada observación con respecto a la recta.

Para que sirve tener un modelo de los datos

- Ayuda a definir la relación o asociación entre dos variables, definidas de ahora en adelante como la variable predictora y la variable respuesta, o conocidas en otros contextos como la variable independiente y la variable dependiente.
- Utilizando un modelo de regresión podemos predecir una variable a partir de otra variable.

También podemos determinar el efecto de la variable independiente sobre la variable dependiente

- El parámetro de la variable independiente(predictora) denotada como β se denomina también como coeficiente de regresión. Este nos informa el grado de relación entre la variable predictora y la variable respuesta. Esto se hace utilizando una prueba de hipótesis de el coeficiente de regresión.

$$H_0 : \beta = 0;$$
$$H_A : \beta \neq 0$$

- La hipótesis de el coeficiente de regresión igual a 0 se prueba estimando un estadístico que tiene una distribución F, y se realiza a traves de la descomposición de la variabilidad de la ecuación de regresión estimada de los datos en tres partes, y esta descomposición se conoce como análisis de regresión.

Tabla de análisis de regresión

Fuente de Variacion	G.L.	Sumas de Cuadrados (SC)	Cuadrados Medios (CM)	F
Regresión	1	Suma de Cuadrados de la regresión	Cuadrados medios de la regresión	$\frac{\text{CM regresión}}{\text{CM del error}}$
Error	n-2	Suma de Cuadrados del error	Cuadrados medios del error	
Total	n-1	Suma de cuadrados Total		

Tabla de análisis de regresión

Fuente de Variación	G.L.	Sumas de Cuadrados (SC)	Cuadrados Medios (CM)	F
Regresión	1	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{1}$	$\frac{\text{CM regresión}}{\text{CM del error}}$
Error	n-2	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$	
Total	n-1	$\sum_{i=1}^{n}(y_i - \bar{y})^2$		

Donde, $\hat{y}_i$ corresponde a los valores ajustados, o predichos, $\bar{y}$ es la media de la variable dependiente y y_i corresponde al valor de cada variable dependiente.

Cálculo de las sumas de cuadrados

$$
\begin{aligned}
SC_{XX} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} \\
SC_{YY} &= \sum y_i^2 - \frac{(\sum y_i)^2}{n} \\
SC_{XY} &= \sum x_i y_i - \frac{\sum y_i \sum x_i}{n} \\
\bar{X} & \\
\bar{Y} &
\end{aligned}
$$

Como se estima la ecuación de regresión en una forma manual

- La ecuaciones para calcular los parametros son

$$\hat{\alpha} = \bar{y} - (\frac{SC_{XY}}{SC_{XX}}\bar{x})$$
$$\hat{\beta} = \frac{SC_{XY}}{SC_{XX}}$$

Tabla de análisis de regresión

Fuente de Variación	G.L.	Sumas de Cuadrados (SC)	Cuadrados Medios (CM)	F
Regresión	1	$\frac{(SC_{XY})^2}{SC_{XX}}$	$\frac{\frac{(SC_{XY})^2}{SC_{XX}}}{1}$	$\frac{\text{CM regresión}}{\text{CM del error}}$
Error	n-2	SC total - SC regresión	$\frac{SCerror}{n-2}$	
Total	n-1	SC_{YY}		

Donde, $\hat{y}_i$ corresponde a los valores ajustados, o predichos, $\bar{y}$ es la media de la variable dependiente y y_i corresponde al valor de cada variable dependiente.

Cuadro: Tabla de análisis de regresión, población total asociado a crimenes

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	$1{,}2 \times 10^{+12}$	1	$1{,}2 \times 10^{+12}$	1604,8	$< .001$
	Residual	$3{,}2 \times 10^{+11}$	438	$7{,}3 \times 10^{+8}$		
	Total	$1{,}5 \times 10^{+12}$	439			

Cuadro: Coeficientes de regresión lineal, población total asociado a crimenes

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	27111,6	2776,4		9,8	$< .001$
H_1	(Intercept)	−6587,4	1537,6		−4,3	$< .001$
	population	$8{,}6 \times 10^{-2}$	$2{,}1 \times 10^{-3}$	0,9	40,1	$< .001$

Estimación de las sumas de cuadrados

$$
\begin{aligned}
SC_{XX} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} \\
SC_{YY} &= \sum y_i^2 - \frac{(\sum y_i)^2}{n} \\
SC_{XY} &= \sum x_i y_i - \frac{\sum y_i \sum x_i}{n} \\
\bar{X} & \\
\bar{Y} &
\end{aligned}
$$

Como se estima la ecuación de regresión?

Es necesario hacer unos cálculos intermedios

$$\sum x_i^2 = 227049830330903$$
$$(\sum x_i)^2 = 29902988184288000$$
$$\sum y_i^2 = 1812333068618$$
$$(\sum y_i)^2 = 142303713108544$$
$$\sum x_i = 172924805$$
$$\sum y_i = 11929112$$
$$\sum x_i y_i = 18329414537404$$
$$\bar{x} = 393010.92$$
$$\bar{y} = 27111.62$$
$$SC_{XX} = 159088493548430.00$$
$$SC_{YY} = -62611821434690700.00$$
$$SC_{XY} = 13641143250078.60$$

Estimación de las sumas de cuadrados

$$
\begin{aligned}
SC_{XX} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 159088493548430 \\
SC_{YY} &= \sum y_i^2 - \frac{(\sum y_i)^2}{n} = -62611821434690700 \\
SC_{XY} &= \sum x_i y_i - \frac{\sum y_i \sum x_i}{n} = 13641143250078{,}60 \\
\bar{X} &= 393010{,}92 \\
\bar{Y} &= 27111{,}62
\end{aligned}
$$

Como se estima la ecuación de regresión?

- A manera de ilustración se presenta el cálculo de la ecuación de regresión
- A partir de las sumas de cuadrados (SC_{XX}), (SC_{YY}), y (SC_{XY}).
- La ecuaciones para calcular los parámetros son

$$\hat{\alpha} = \bar{y} - (\frac{SC_{XY}}{SC_{XX}}\bar{x}) = 27111{,}62 - (\frac{13641143250078{,}6}{159088493548430}393010{,}92) = -6587{,}35$$

$$\hat{\beta} = \frac{SC_{XY}}{SC_{XX}} = \frac{13641143250078{,}6}{159088493548430} = 0{,}09$$

¿Para que sirve tener un modelo de los datos?

- El modelo de regresión simple lineal para la asociación de los crimenes y la población total:

$$y_i = -6587{,}43 + 0{,}086 * x_i$$

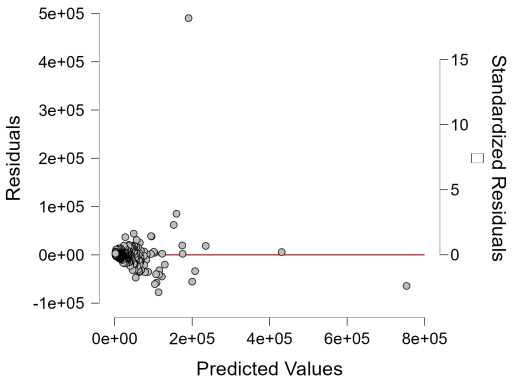
- Con este modelo podemos predecir cual es el número de crimenes a partir de un valor particular de población.
- $-6587{,}43 + 0{,}086 * 150000 = 6312$ crimenes.

De la tabla se extrae un elemento importante que corresponde a la suma de los cuadrados del error, denotada como SC_{error}. Este elemento se utiliza para hallar el cuadrado medio del error CM_{error}, que equivale a la SC_{error} dividido en los grados de libertad. Este elemento configura la varianza de los errores , o sea las desviaciones de los valores predichos con respecto a los valores observados, los errores también se conocen como residuales y se hallan utilizando la siguiente fórmula.

$$r_i = y_i - \hat{y}_i$$

donde los y_i son los valores observados de la respuesta y $\hat{y}_i$ son los valores predichos de la ecuacion de regresión $\hat{y}_i = \alpha + \beta * x_i$.

(a) residuales versus predichos



(b)

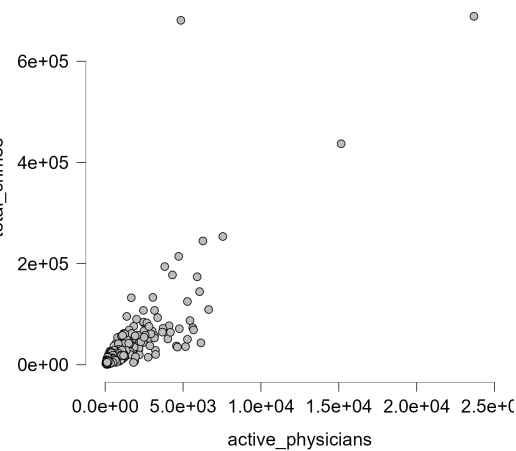


Figura: Crimenes y población

La varianza de los errores

Los residuales tienen una media de 0, o sea que la suma de todos los residuales es igual a 0, y una varianza que se estima de acuerdo al cuadrado medio del error (CM_{error})

$$Varianza(\epsilon) = Varianza_{error} = \frac{SC_{error}}{n-2}$$

La varianza del error se utiliza para encontrar el error standard de los coeficientes de regresión β y α

Que representan los residuales

- Los residuales agrupan todas las posibles fuentes de variabilidad en la respuesta entre los individuos de un experimento o estudio observacional. Las posibles fuentes de variabilidad entre otras son el error de medida (el error en que incurre el instrumento de medida que registra la variable respuesta), factores intrínsecos al individuo como su acervo genético, el cambio en tiempo en que se registra la variable respuesta para cada individuo.

Que representan los residuales o errores

- Los errores son importantes en regresión linear simple porque son la base de la validez de sus conclusiones, ya que un modelo de regresión debe cumplir con unos supuestos en relación con la distribución de los errores.
- Primero se asume que los errores tienen una distribución normal con media 0 y varianza σ_ϵ^2

$$\epsilon \sim N(0, \sigma_\epsilon^2)$$

- Además se asume que los errores tienen una varianza constante y que cada residual es independiente e identicamente distribuido.

La varianza de los errores y su relación con el error estándar del coeficiente de regresión

- La estimación de la varianza de β y su error standard se obtiene de la siguiente manera.

$$\text{Varianza}(\beta) = \frac{\text{Varianza}(\epsilon)}{(n-1)s_x^2}$$

- Donde Varianza(β) es la varianza del coeficiente de regresión, Varianza(ϵ) es la varianza de los errores y s_x^2 es la varianza de la variable predictora.
- La raíz cuadrada de la varianza del coeficiente de regresión β es igual a y corresponde al error estándar del coeficiente de regresión. Con este error estándar se pueden crear intervalos de confianza de β.