Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)
Regresión lineal (1ª Parte)

1

Profesora: *Dra. Fabiola Ocampo Botello*

Levín, Rubín, Balderas, Del Valle y Gómez (2004:510) establecen que el término de regresión fue utilizado por primera vez por Sir Francis Galton en el año de 1877 como un término estadístico. Sir Francis Galtón desarrolló un estudio que mostró que los niños nacidos de padres altos tienden a *regresar* a la estatura media de la población. Utilizó el término regresión para designar el *proceso general de predecir una variable* (en este caso la estatura de los niños) *a partir de otra variable* (la estatura de los padres).

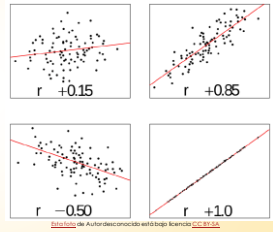
2

Las variables conocidas se llaman variables independientes y la variable desconocida a predecir se llama variable dependiente.

Levín y otros (2004) establecen que la correlación y la regresión muestran tanto la naturaleza como la fuerza de la relación entre dos variables.

Definición de correlación:

Una correlación existe entre dos variables cuando valores mayores de una variable van con valores consistentes de otra variable o cuando valores grandes de una variable corresponden de manera consistentes con valores menores de otra variable. (Bennet, Briggs & Triola, 2011:286).



3

Según Aguayo y Lora (2007), la correlación es una técnica matemática que evalúa el grado de asociación o relación entre dos variables cuantitativas, tanto en términos de direccionalidad como de fuerza o intensidad proporcionadas por un coeficiente.

El coeficiente de correlación puede tener valores que oscilan entre -1 y +1, considerando el cero.
Cuando el valor se acerca a +1, ambas variables (X y Y) se relacionan de manera muy estrecha. Existe una correlación positiva si cuando se incrementa el valor de X también se incrementa el de Y o cuando hay un decremento en el valor de X también hay un decremento en el valor de Y.

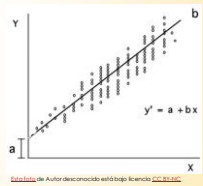
4

Del mismo modo, cuando el valor se acerca a -1 refleja que existe una relación de forma inversa, esto es, cuando aumenta el valor de X existe un decremento en el valor de Y y cuando X obtiene puntajes bajos Y alcanza puntajes altos.

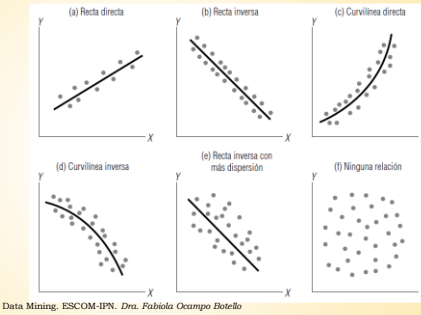
Regresión Lineal

La correlación sólo expresa la relación existente entre dos variables numéricas, no expresa causalidad. La regresión es una modelo de predicción.

Carollo (2012) establece que "*El objetivo de un modelo de regresión es tratar de explicar la relación que existe entre una variable dependiente (variable respuesta) Y y un conjunto de variables independientes (variables explicativas) X1,..., Xn. En un modelo de* ***regresión lineal simple*** *tratamos de explicar la relación que existe entre la variable respuesta Y y una única variable explicativa X.*"



Imágenes y ejemplo tomados de Levin, et. al (2004).

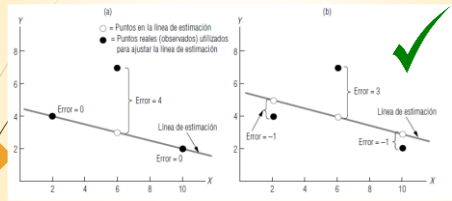


Se usa el método de mínimos cuadrados para encontrar los valores a y b,

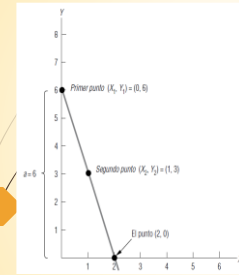
$$\hat{y}_i = b_0 + b_1 x_i$$

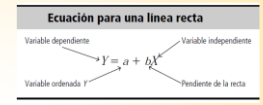
La línea de estimación

$$\hat{Y} = a + bX$$



Imágenes y ejemplo tomados de Levin, et. Al (2004)





La a se denomina la "ordenada de Y", por que cruza el eje Y.
La b es la "pendiente" de la recta.

Imágenes tomadas de Levin, et. Al (2004)

Debido al comportamiento que tienen los datos, se puede modelar mediante una ecuación de regresión simple. Lo cual es:

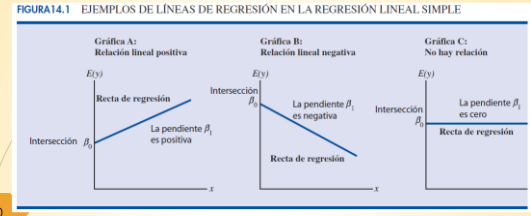
$$\hat{y} = b_0 + b_1 x_i$$

Donde:

$\hat{y}_i$	Valor estimado de las ventas trimestrales del restaurant i
b_0	Intersección de la recta de regresión con el eje y
b_1	Pendiente de la recta de regresión
x_i	Tamaño de la población de estudiantes del restaurante i

9

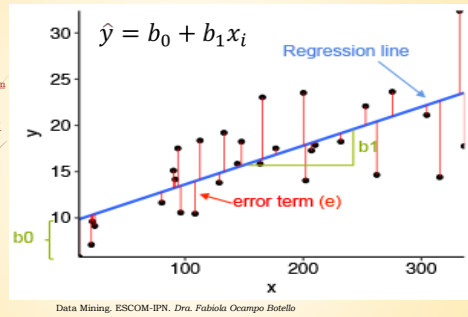
FIGURA 14.1 EJEMPLOS DE LÍNEAS DE REGRESIÓN EN LA REGRESIÓN LINEAL SIMPLE



10

Imagen tomada de Anderson, Sweeney & Williams (2008).

Imagen Creative Commons
En: http://www.stbda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/



11

PENDIENTE E INTERSECCIÓN CON EL EJE y DE LA ECUACIÓN DE REGRESIÓN ESTIMADA*

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \quad (14.6)$$

$$b_0 = \bar{y} - b_1\bar{x} \quad (14.7)$$

donde

x_i = valor de la variable independiente en la observación i
y_i = valor de la variable dependiente en la observación i
$\bar{x}$ = media de la variable independiente
$\bar{y}$ = media de la variable dependiente
n = número total de observaciones

12

Imagen tomada de Anderson, Sweeney & Williams (2008)

Ejemplo Número 1: (Adaptado de Levin, et. al (2004:522-523)

Suponga que la directora del Departamento de Salubridad de Chapel Hill está interesada en la relación que existe entre la antigüedad de un camión de basura y los gastos de reparación que hay que esperar.

13

Tabla 12-6 Gastos anuales de reparación de camiones

Número del camión	Antigüedad del camión en años (X)	Gastos de rep. durante el último año en cientos de dólares (Y)
101	5	7
102	3	7
103	3	6
104	1	4

Imagen y ejemplo tomados de Levin, et. al (2004)

Primer paso. Organizar los datos para calcular $\bar{x}$ y $\bar{Y}$.

Imágenes y ejemplo tomados de Levin, et. al (2004)

14

Pendiente de la recta de regresión de mejor ajuste

$$b = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

Ecuación 12.4

Ordenada $\bar{Y}$ de la recta de regresión de mejor ajuste

$$a = \bar{Y} - b\bar{X}$$

Ecuación 12.5

Camiones (n = 4) (1)	Antigüedad (X) (2)	Gastos de reparación (Y) (3)	XY (2) × (3)	X^2 $(2)^2$
101	5	7	35	25
102	3	7	21	9
103	3	6	18	9
104	1	4	4	1
	$\Sigma X = 12$	$\Sigma Y = 24$	$\Sigma XY = 78$	$\Sigma X^2 = 44$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{12}{4} = 3 \leftarrow \text{Media de los valores de la variable independiente}$$

$$\bar{Y} = \frac{\Sigma X}{n} = \frac{24}{6} = 6 \leftarrow \text{Media de los valores de la variable dependiente}$$

Tabla 12-7. Cálculo de los datos para las ecuaciones 12-4 y 12-5.

Segundo paso. Calcular $\bar{x}$ y $\bar{Y}$.

$$b = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2} = \frac{78 - (4)(3)(6)}{44 - (4)(3)^2} = \frac{78 - 72}{44 - 36} = \frac{6}{8} = 0.75 \leftarrow \text{Pendiente de la línea}$$

$$a = \bar{Y} - b\bar{X} = 6 - (0.75)(3) = 6 - 2.25 = 3.75 \leftarrow \text{Ordenada } \bar{Y}$$

$$\hat{Y} = a + bX = 3.75 + 0.75X$$

15

Con esta ecuación, la directora del Departamento de Salubridad puede estimar los gatos anuales de reparación. Si se tiene un camión de 4 años de antigüedad, se estima:

$$\hat{Y} = 3.75 + 0.75(4) = 3.75 + 3 = 6.75 \leftarrow \text{Gastos anuales de reparación esperados de \$675.00}$$

Imágenes y ejemplo tomados de Levin, et. al (2004)

Ejemplo Número 2: (tomado de Levin, et. al (2004:523-)

El vicepresidente de una compañía química y de fabricación de fibras cree que las ganancias anuales de la empresa dependen de la cantidad gastada en investigación y Desarrollo (ID), pero el nuevo presidente de la compañía no está de acuerdo, por lo que ha solicitado una ecuación para pronosticar los beneficios anuales derivados de la cantidad presupuestada para ID.

16

Tabla 12-8 Relación anual entre investigación, desarrollo y ganancias

Año	Millones de dólares gastados en investigación y desarrollo (X)	Ganancia anual (millones de dólares) (Y)
1995	5	31
1994	11	40
1993	4	30
1992	5	34
1991	3	25
1990	2	20

Imagen y ejemplo tomados de Levin, et. al (2004)

Primer paso. Organizar los datos como se muestran en la siguiente figura.

Tabla 12-9
Cálculo de los datos para las ecuaciones 12-4 y 12-5

Año (n = 6)	Gastos de ID (X)	Ganancias anuales (Y)	XY	X^2
1995	5	31	155	25
1994	11	40	440	121
1993	4	30	120	16
1992	5	34	170	25
1991	3	25	75	9
1990	2	20	40	4
	$\Sigma X = 30$	$\Sigma Y = 180$	$\Sigma XY = 1{,}000$	$\Sigma X^2 = 200$

$$\bar{X} = \frac{\Sigma X}{n} \quad [3\text{-}2]$$

$$= \frac{30}{6}$$

$= 5 \leftarrow$ Media de los valores de la variable independiente

$$\bar{Y} = \frac{\Sigma Y}{n} \quad [3\text{-}2]$$

$$= \frac{180}{6}$$

$= 30 \leftarrow$ Media de los valores de la variable dependiente

Imagen y ejemplo tomados de Levin, et. al (2004)

Segundo paso. Calcular $\bar{x}$ y $\bar{Y}$.

$$b = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

$$= \frac{1{,}000 - (6)(5)(30)}{200 - (6)(5)^2}$$

$$= \frac{1{,}000 - 900}{200 - 150}$$

$$= \frac{100}{50}$$

$= 2 \leftarrow$ Pendiente de la recta

$$a = \bar{Y} - b\bar{X}$$

$$= 30 - (2)(5)$$

$$= 30 - 10$$

$= 20 \leftarrow$ Ordenada Y

$$\hat{Y} = a + bX$$

$$= 20 + 2X$$

Si la compañía gastó 8 millones de dólares para ID en el año de 1996, entonces debió ganar aproximadamente 36 millones de dólares en ese año.

Imágenes y ejemplo tomados de Levin, et. al (2004).

Referencias bibliográficas

Anderson, Sweeney & Williams. (2008). Estadística para administración y economía, 10ª edición. Cengage Learning.

Bennet, Briggs & Triola (2011). Razonamiento estadístico. Pearson. México.

Carollo Limeres, M. Carmen. (2012). Regresión lineal simple. Apuntes del departamento de estadística e investigación operativa . Disponible en: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_%20simple_2011_12.pdf

Kerlinger, F. N. & Lee, H. B. (2002). Investigación del comportamiento. Métodos de investigación en ciencias sociales. 4ª ed. México: Mc. Graw Hill.

Levin, Rubin, Balderas, Del Valle y Gómez. (2004). Estadística para administración y economía. Séptima Edición. Prentice-Hall.

Mason, Lind & Marshal. (2000). Estadística para administración y economía. Alfaomega. 10ª edición.