

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

Minería de datos (*Data Mining*)  
*ejemplo Pizzería "Polito"*  
*Regresión lineal*

1

Profesora: *Dra. Fabiola Ocampo Botello*

Ejemplo adaptado de Anderson, Sweeney & Williams (2008).

*Se tienen los datos de 10 pizzerías (Pizzerías "Polito") ubicadas cerca de los campus universitarios. Tanto la cantidad de alumnos y las ganancias se expresan en miles, como se muestra en la siguiente tabla.*

2



Fuente de la imagen: La Cocinika de Ana  
<https://ana-lacocinikadeana.blogspot.com/2012/10/dominos-pizza.html>

Data Mining, ESCOM-IPN. *Dra. Fabiola Ocampo Botello*

Tabla No. 1. Ventas de la pizzería "Polito"

No	NoEstud	Ventas
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202



Fuente de la imagen: Pizzeria 5 tapas.  
<https://pizzeria5tapas.blogspot.com/>

3

La pizzería número 1:  $x_1 = 2$  y  $y_1 = 58$  (2, 58) significa que está cerca de un campus con 2,000 estudiantes y reporta ventas de 58,000 pesos.

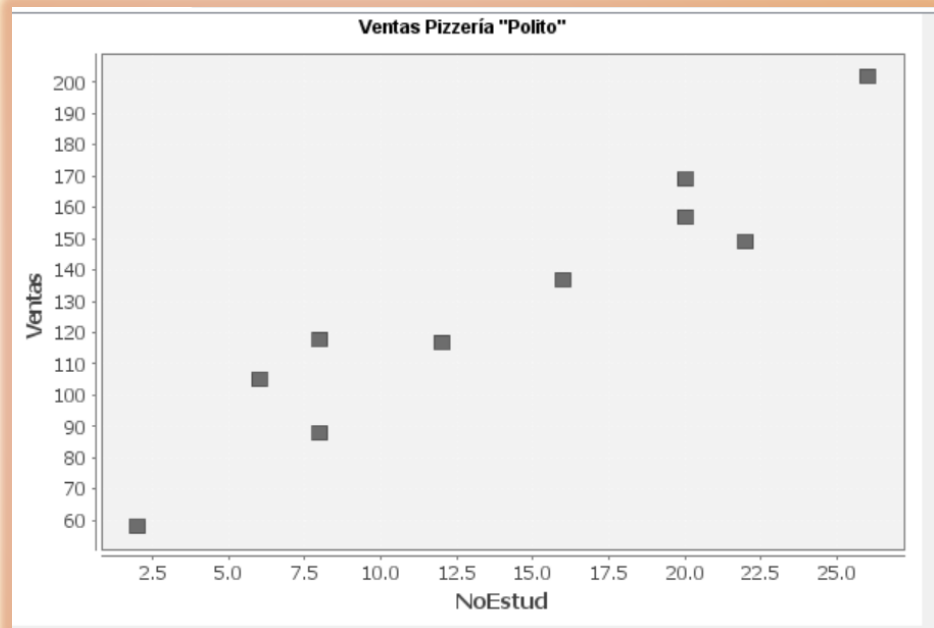
La pizzería número 2:  $x_2 = 6$  y  $y_2 = 105$  (6, 105) significa que está cerca de un campus con 6,000 estudiantes y reporta ventas de 105,000 pesos.

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

La variable independiente se coloca en el eje horizontal x (número de estudiantes).

La variable dependiente se coloca en el eje vertical y (ganancia).

4



Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

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

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

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

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

5

Imagen tomada de Anderson, Sweeney & Williams (2008)

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{1300}{10} = 130$$

Imágenes tomadas de Anderson, Sweeney & Williams (2008)

Restaurante $i$	$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totales	140	1300			2840	568
	$\sum x_i$	$\sum y_i$			$\sum(x_i - \bar{x})(y_i - \bar{y})$	$\sum(x_i - \bar{x})^2$

6

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

$$= \frac{2840}{568}$$

$$= 5$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$= 130 - 5(14)$$

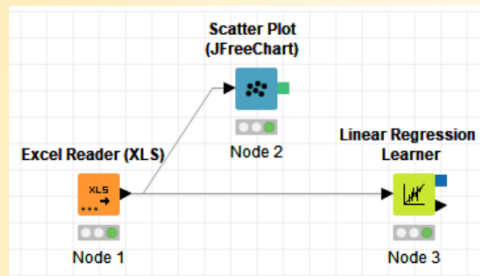
$$= 60$$

$$\hat{y} = 60 + 5x$$



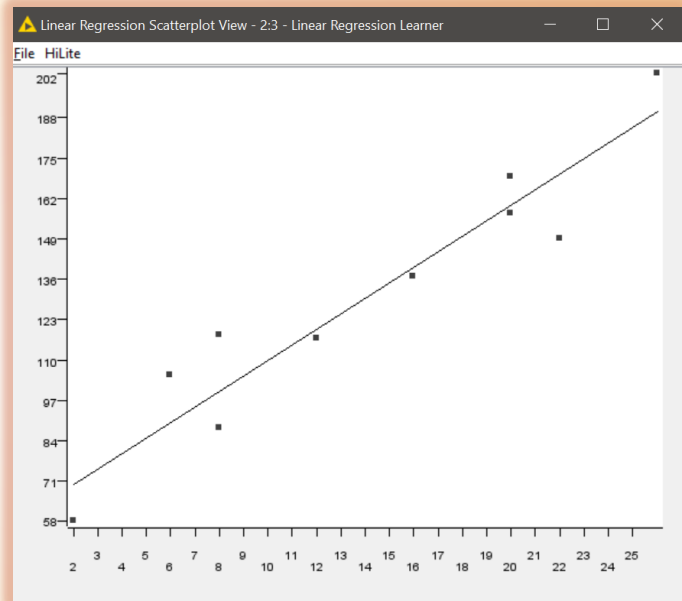
**Ecuación de regresión**

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello



7

Row ID	S Variable	D Coeff.
Row1	NoEstud	5
Row2	Intercept	60



Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

Suponga que se desean predecir las ventas de un restaurant que se encuentra cerca de un campus que tiene 16,000 estudiantes.

$$x = 16$$

$$\hat{y} = 60 + 5x$$

$$\hat{y} = 60 + 5(16) = 140$$

Se pronostica una venta de 140,000 pesos.

8

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

## Verificación de la ecuación de estimación

Levin, et. al (2004) establecen que un método para verificar la ecuación de estimación se fundamenta en una de las propiedades de la recta ajustada por el método de mínimos cuadrados, esto es, los errores individuales positivos y negativos deben sumar cero.

Imagen tomada de Anderson, Sweeney & Williams (2008)

9

Restaurante $i$	$x_i$	$y_i$
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202
Totales	140	1300
	$\Sigma x_i$	$\Sigma y_i$

Del ejemplo de la pizzería "Polito". La suma de errores sería:

$$\hat{y} = 60 + 5x$$

Row ID	I NoEstud	I Ventas	D calculo	D new column
1.0	2	58	-12	0
2.0	6	105	15	0
3.0	8	88	-12	0
4.0	8	118	18	0
5.0	12	117	-3	0
6.0	16	137	-3	0
7.0	20	157	-3	0
8.0	20	169	9	0
9.0	22	149	-21	0
10.0	26	202	12	0

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

10

### 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](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, Rubín, 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.

Dra. Fabiola Ocampo Botello