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

Profesora: *Dra. Fabiola Ocampo Botello*

1

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.

Imagen Creative Commons
En: https://ana-lacocinikadeana.blogspot.com/2012/10/dominos-pizza.html

Dra. Fabiola Ocampo Botello

2

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

No	NoEstud x	Ventas y
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

Imagen Creative Commons
En: https://pizzeria5tapas.blogspot.com/

3

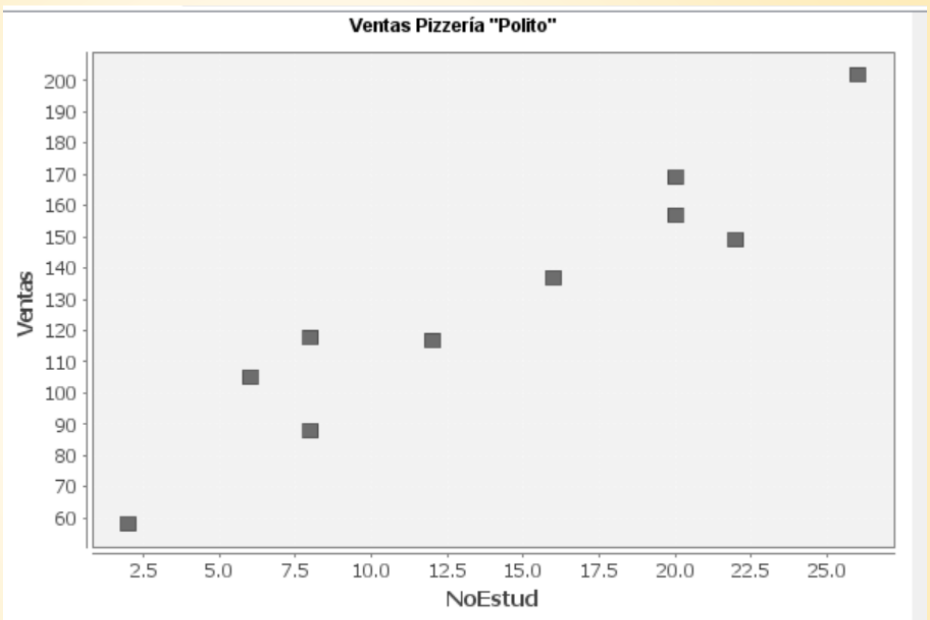
La pizzería número 1: $x_1 = 2$ y $y_1 = 58$ (2, 58) significa que está cerca de un campus con 2,000 estudiantes y reporta ventas de 58,000 pesos.

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.

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

Dra. Fabiola Ocampo Botello

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

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

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

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

Imagen tomada de Anderson, Sweeney & Williams (2008)

5

Dra. Fabiola Ocampo Botello

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

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

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
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

6

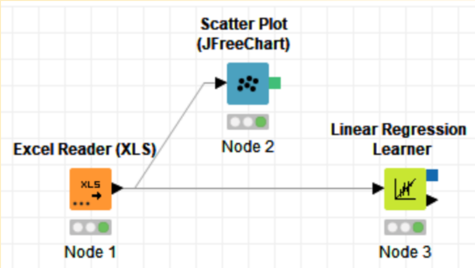
$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

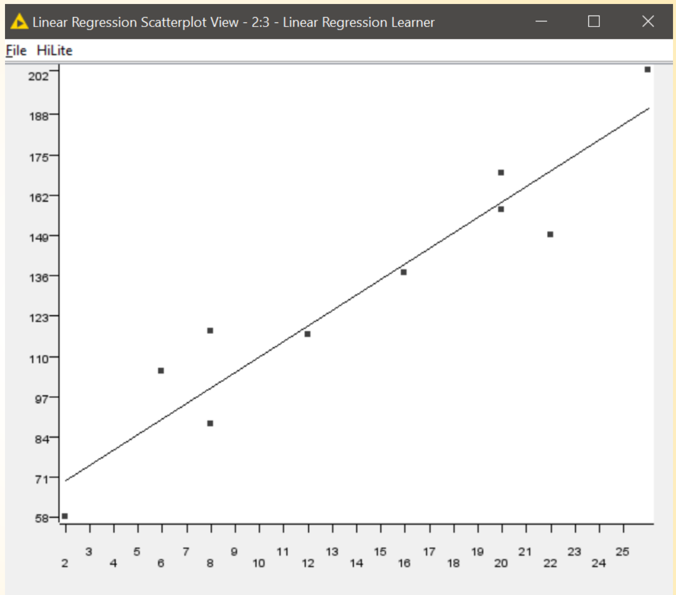
$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

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

Ecuación de regresión

Dra. Fabiola Ocampo Botello





7

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

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

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.

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
	Σx_i	Σy_i

Imagen tomada de Anderson, Sweeney & Williams (2008)

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

9

Dra. Fabiola Ocampo Botello

El error estándar de la estimación

¿Cómo evaluar la confiabilidad ecuación de estimación de regresión encontrada?

Levin et al (2004) establecen que el **error estándar de la estimación** mide la variabilidad o dispersión de los valores observados alrededor de la recta de regresión.

Anderson, Sweeney & Williams (2008) establecen que la diferencia que existe, en la observación *i*, entre el valor observado de la variable dependiente *yi*, y el valor estimado de la variable dependiente $\hat{y}_i$, se llama **residual *i***. El residual *i* representa el error que existe al usar $\hat{y}_i$ para estimar *yi*.

SUMA DE CUADRADOS DEBIDA AL ERROR

$$\text{SCE} = \Sigma(y_i - \hat{y}_i)^2$$

Imagen 14.8 tomada de Anderson, Sweeney & Williams (2008)

10

Dra. Fabiola Ocampo Botello

SUMA DE CUADRADOS DEBIDA AL ERROR (SCE)

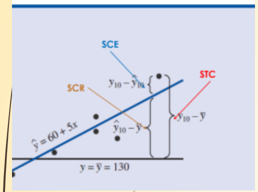


Tabla 14.3 e imagen tomada de Anderson, Sweeney & Williams (2008)

Restaurante i	x_i = población de estudiantes (miles)	y_i = ventas trimestrales (miles de $)	Ventas pronosticadas $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Error al cuadrado $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SCE = 1530

11

En el caso de la Pizzería "Polito", por ejemplo para $x_1 = 2$ y $y_1 = 58$, el valor estimado para la pizzería número 1 es 70, el error al usar $\hat{y}$ del restaurant número 1 es:

$$\hat{y}_1 = 60 + 5(2) = 70$$

Dra. Fabiola Ocampo Botello

SUMA TOTAL DE CUADRADOS (STC)

Se desea tener una estimación de las ventas trimestrales sin saber cuál es el tamaño de la población de estudiantes.

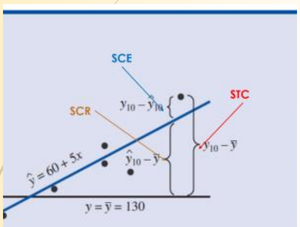


TABLA 14.4 CÁLCULO DE LA SUMA TOTAL DE CUADRADOS EN EL EJEMPLO DE ARMAND'S PIZZA PARLORS

Restaurante i	x_i = población de estudiantes (miles)	y_i = ventas trimestrales (miles de $)	Desviación $y_i - \bar{y}$	Desviación al cuadrado $(y_i - \bar{y})^2$
1	2	58	−72	5 184
2	6	105	−25	625
3	8	88	−42	1 764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1 521
9	22	149	19	361
10	26	202	72	5 184
				STC = 15 730

12

SUMA TOTAL DE CUADRADOS

$$\text{STC} = \Sigma(y_i - \bar{y})^2$$

Tabla 14.3 tomada de Anderson, Sweeney & Williams (2008)

Figura 14.9 tomada de Anderson, Sweeney & Williams (2008)

Dra. Fabiola Ocampo Botello

El coeficiente de determinación se denota con la letra r^2.

COEFICIENTE DE DETERMINACIÓN

$$r^2 = \frac{SCR}{STC}$$

Imagen 14.12 tomada de Anderson, Sweeney & Williams (2008)

El coeficiente de determinación de la pizzería "Polito" es:

r^2 = SCT/STC = 14200/15730 = 0.9027

r^2 se puede interpretar como el porcentaje de la suma total de cuadrados que se explica mediante el uso de la ecuación de regresión estimada.
En el ejemplo de la pizzería se concluye que 90.27% de la variabilidad en las ventas se explica por la relación lineal que existe entre el tamaño de la población de estudiantes y las ventas.

Dra. Fabiola Ocampo Botello

13

El coeficiente de correlación muestral se calcula mediante la fórmula:

COEFICIENTE DE CORRELACIÓN MUESTRAL

$$\begin{aligned} r_{xy} &= (\text{signo de } b_1)\sqrt{\text{Coeficiente de determinación}} \\ &= (\text{signo de } b_1)\sqrt{r^2} \end{aligned}$$

donde

b_1 = pendiente de la ecuación de regresión estimada $\hat{y} = b_0 + b_1x$

Figura 14.13 de Anderson, Sweeney & Williams (2008)

En el ejemplo de la Pizzería Polito, el valor del coeficiente de determinación correspondiente a la ecuación de regresión estimada:

60 + 5x es 0.9027

Como la pendiente de la ecuación de regresión estimada es positiva, la ecuación (14.13) indica que el coeficiente de correlación muestral es $+\sqrt{0.9027} = +0.9501$.
Con este coeficiente de correlación muestral, r_{xy} = +0.9501, se concluye que existe una relación lineal fuerte entre x y y.

Dra. Fabiola Ocampo Botello

14

Levin et al (2004) establecen que una forma de calcular el error Є es mediante el error estándar de la estimación, *mide la variabilidad o dispersión de los valores observados alrededor de la recta de regresión*.
El cual tiene la siguiente fórmula:

Error estándar de la estimación

$$s_e = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}}$$

Ecuación 12-6.

donde,

- Y = valores de la variable dependiente
- $\hat{Y}$ = valores estimados con la ecuación de estimación que corresponden a cada valor de Y
- n = número de puntos utilizados para ajustar la línea de regresión

Imágenes tomadas de Levin, et. al (2004)

15

Dra. Fabiola Ocampo Botello

ERROR CUADRADO MEDIO (ESTIMACIÓN DE σ^2)

$$s^2 = \text{ECM} = \frac{\text{SCE}}{n - 2}$$

Figura 14.15 de Anderson, Sweeney & Williams (2008)

ERROR ESTÁNDAR DE ESTIMACIÓN

$$s = \sqrt{\text{ECM}} = \sqrt{\frac{\text{SCE}}{n - 2}}$$

Figura 14.16 de Anderson, Sweeney & Williams (2008)

Para el caso de la Pizzería "Polito", se tiene:

$$s^2 = \text{ECM} = \frac{1530}{8} = 191.25$$

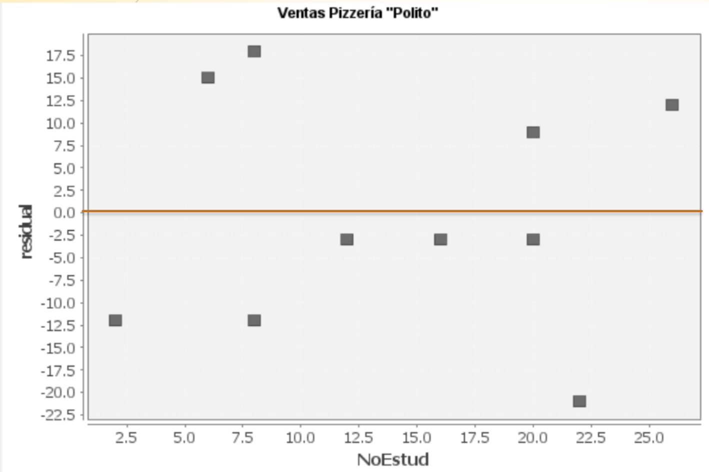
$$s = \sqrt{191.25} = 13.82$$

16

Dra. Fabiola Ocampo Botello

El residual de la observación i es la diferencia entre el valor observado de la variable dependiente (yi) y el valor estimado de la variable dependiente (ŷi).

Los residuales proporcionan información acerca del error.

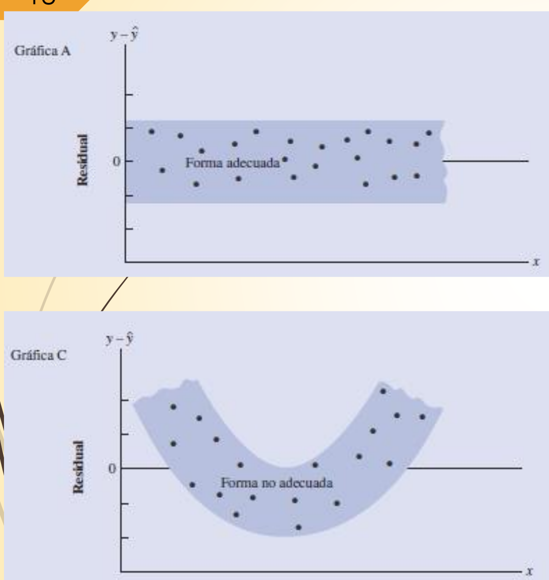


En la **gráfica de residuales**, para cada residual se grafica un punto La primera coordenada de cada punto está dada por el valor xi y la segunda coordenada está dada por el correspondiente valor del residual yi - ŷi.

Dra. Fabiola Ocampo Botello

GRÁFICAS DE LOS RESIDUALES CORRESPONDIENTES A TRES ESTUDIOS DE REGRESIÓN

18



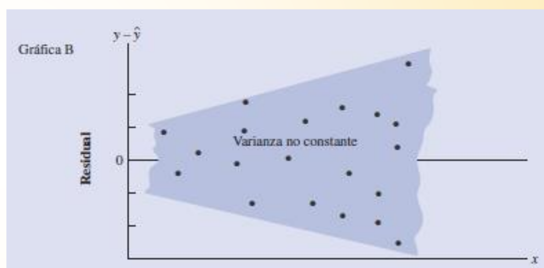


Figura 14.12 de Anderson, Sweeney & Williams (2008)

Dra. Fabiola Ocampo Botello

19

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.

Levín, 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