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

Minería de datos (*Data Mining*)
Regresión lineal (2ª Parte)

1

Profesora: *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.

Un ejemplo de lo anterior se muestra en la siguiente figura:

2

Año ($n = 6$)	Gastos de ID (X)	Ganancias anuales (Y)
1995	5	31
1994	11	40
1993	4	30
1992	5	34
1991	3	25
1990	2	20
	$\Sigma X = 30$	$\Sigma Y = 180$

Tabla 12-10

Cálculo de la suma de los errores individuales de la tabla 12-9

Y		$\hat{Y}$ (es decir, $20 + 2X$)		Error individual
31	−	[20 + (2)(5)]	=	1
40	−	[20 + (2)(11)]	=	−2
30	−	[20 + (2)(4)]	=	2
34	−	[20 + (2)(5)]	=	4
25	−	[20 + (2)(3)]	=	−1
20	−	[20 + (2)(2)]	=	−4
				0 ← Error total

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

Del ejemplo de la pizzería "Polito". La suma de errores sería: $\hat{y} = 60 + 5x$

3

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)

Row ID	NoEstud	Ventas	calculo	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

$\hat{y}_1$ = 60 + 5(2) = 70. Diferencia = 58 - 70 = -12

$\hat{y}_2$ = 60 + 5(6) = 90. Diferencia = 105 – 90 = 15

Anderson, Sweeney & Williams (2008) establece que a 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*.

4

SUMA DE CUADRADOS DEBIDA AL ERROR

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

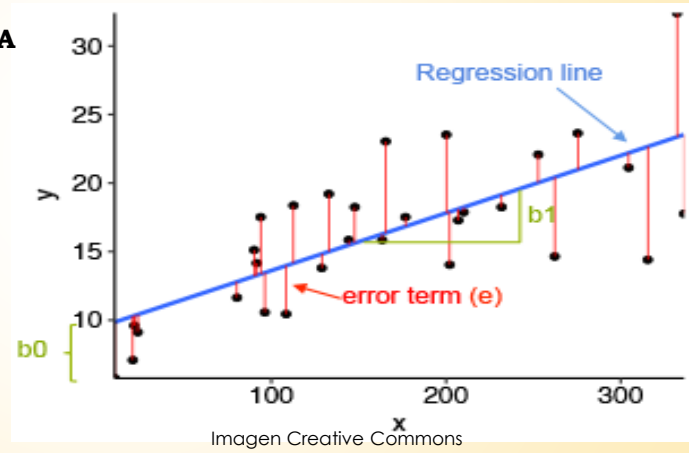


Imagen Creative Commons
En: http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/

Tabla 14.3 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

5

En el caso de la Pizzería "Polito", por ejemplo para $x_1 = 2$ y $y_1 = 58$ (valor real), el valor estimado para la pizzería número 1 es 70, el error al usar $\hat{y}$ del restaurant número 1 es -12

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

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

La estimación de la media de la pizzería "Polito" es

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

6

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

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

SUMA TOTAL DE CUADRADOS $\quad STC = \sum (y_i - \bar{y})^2$

- Los puntos se encuentran más agrupados en torno a la recta de regresión estimada.
- Considerando el punto y_{10} el error es más grande cuando se utiliza la media ($\bar{y}$ = 130) que cuando se utiliza le ecuación de estimación ($\hat{y}_{10}$ = 60 + 5(26) = 190).

7

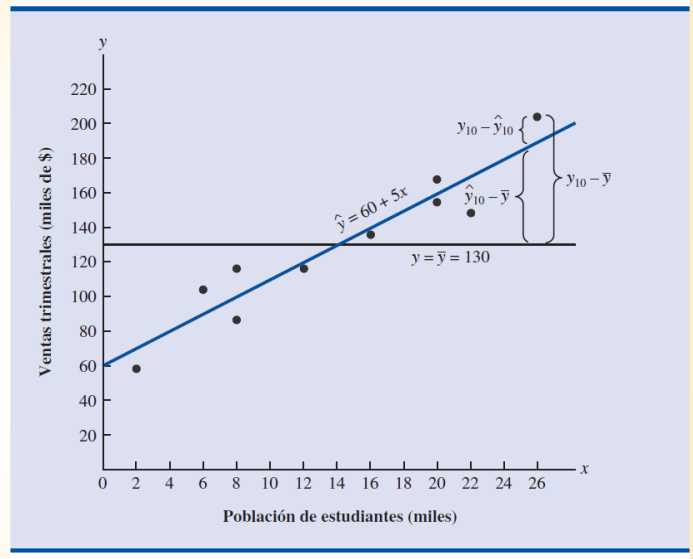


Figura 14.5. Desviación respecto a la línea de regresión estimada y la línea $y = \bar{y}$
Imagen tomada de Anderson, Sweeney & Williams (2008)

STC (SUMA TOTAL DE CUADRADOS) como una medida de qué tanto se agrupan las observaciones en torno a la recta $\bar{y}$

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

SCE (SUMA DE CUADRADOS DEBIDA AL ERROR) como una medida de qué tanto se agrupan las observaciones en torno de la recta $\hat{y}$

8

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

Para medir qué tanto se desvían de $\bar{y}$ los valores $\hat{y}$, de la recta de regresión, se calcula otra suma de cuadrados. A esta suma se le llama *suma de cuadrados debida a la regresión* y se denota SCR.

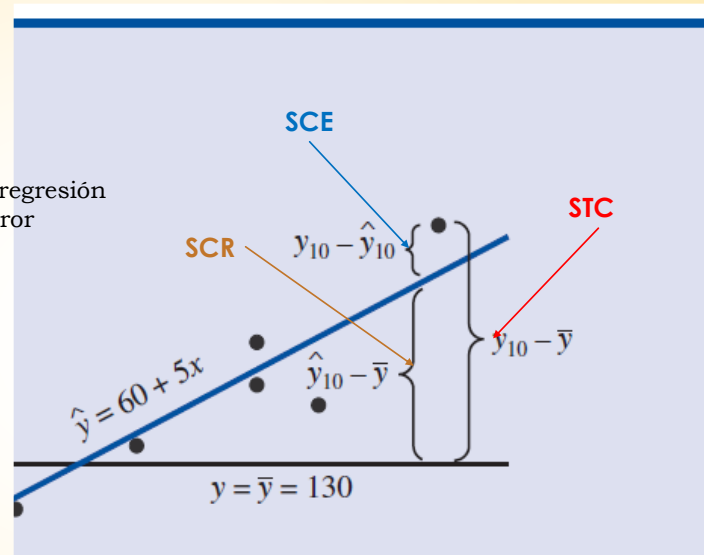
SUMA DE CUADRADOS DEBIDA A LA REGRESIÓN

$$SCR = \sum (\hat{y}_i - \bar{y})^2$$

9

STC = SCR + SCE

STC = Suma total de cuadrados
SCR = Suma de cuadrados debida a la regresión
SCE = Suma de cuadrados debida al error

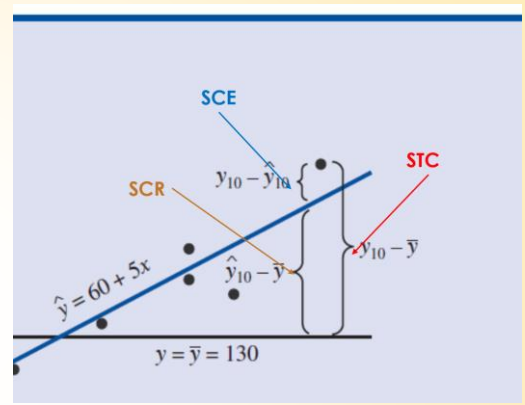


Porción de Imagen 14.5 tomada y modificada de Anderson, Sweeney & Williams (2008)

10

Consideraciones

1) La ecuación de regresión estimada se ajustaría perfectamente a los datos si cada uno de los valores de la variable independiente y_i se encontraran sobre la recta de regresión. En este caso para todas las observaciones se tendría que $yi - \hat{y}_i$ sería igual a cero, con lo que SCE = 0.



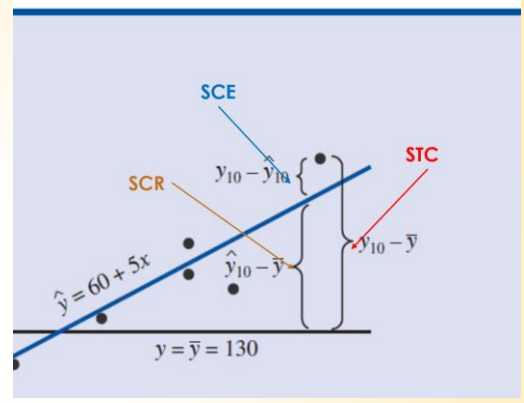
Porción de Imagen 14.5 tomada y modificada de Anderson, Sweeney & Williams (2008)

11

2) Como STC = SCR + SCE es necesario que para que haya un ajuste perfecto SCR debe ser igual a STC, y el cociente (SCR/STC) debe ser igual a uno.

Consideraciones (Continuación):

3) El cociente SCR/STC, que toma valores entre cero y uno, se usa para evaluar la bondad de ajuste de la ecuación de regresión estimada. A este cociente se le llama ***coeficiente de determinación*** **y se denota** r^2.



Porción de Imagen 14.5 tomada y modificada de Anderson, Sweeney & Williams (2008)

12

4) Cuando los ajustes son malos, se tendrán valores altos para SCE (residuales grandes).

El **COEFICIENTE DE DETERMINACIÓN** se denota con la letra r^2.

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

El coeficiente de determinación de la pizzería es:

$$r^2 = \frac{SCR}{STC} = \frac{14200}{15730} = 0.9027$$

13

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.

Coeficiente de correlación (r)

Anderson, Sweeney & Williams (2008) establece lo siguiente, Se tiene el coeficiente de correlación como una medida descriptiva de la intensidad de la relación entre dos variables x y y. Cuyos valores van desde -1 hasta +1.

El coeficiente de correlación muestral se calcula mediante la fórmula:

COEFICIENTE DE CORRELACIÓN MUESTRAL

$$r_{xy} = (signo\ de\ b_1)\sqrt{Coeficiente\ de\ determinación}$$
$$= (signo\ de\ b_1)\sqrt{r^2}$$

14

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

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

$$60 + 5x \text{ es } 0.9027$$

Como la pendiente de la ecuación de regresión estimada es positiva, la ecuación anterior indica que el coeficiente de correlación muestral es $+\sqrt{0.9027} = +0.9501$.

15

Con este coeficiente de correlación muestral, $r_{xy} = +0.9501$, se concluye que existe una relación lineal fuerte entre x y y.

16

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.