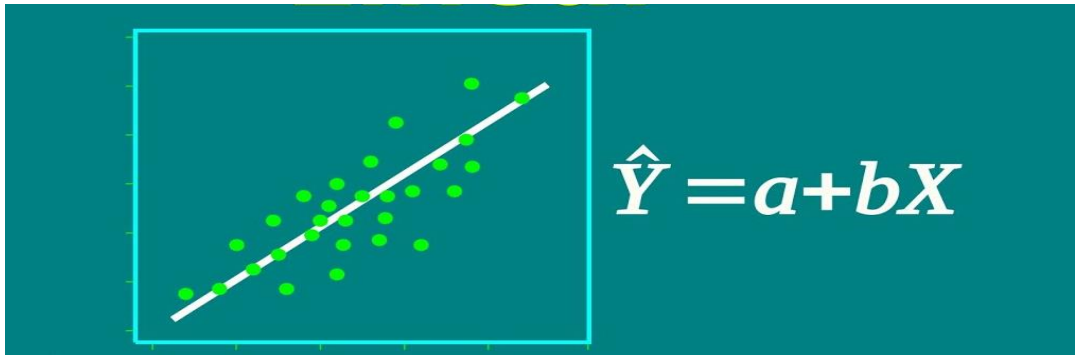
Regresión lineal Simple



Diplomado en ciencia de los datos

Contenido

- Relación entre variables
- Análisis de regresión y terminología
- Ecuación lineal simple
- Método de mínimos cuadrados
  - Caso Restaurante Pizzas
  - Caso Publicidad y Ventas
  - Caso correlación y regresión lineal de millas costo de aerolinea
- Coeficiente de determinación,
- Coeficiente de determinación ajustado
- Coeficiente de correlación
- Pruebas de significancia y contraste de hipótesis
- Prubas ANOVA
- Intérvalos de confianza e intérvalos de predicción

Relación entre variables

- Las decisiones suelen basarse en la relación entre dos o más variables
- La relación entre el gasto en publicidad y las ventas puede permitir a un gerente de mercadotecnia tratar de predecir las ventas.

Relación entre variables

- Las decisiones suelen basarse en la relación entre dos o más variables
- La relación entre la temperatura diaria y la demanda de electricidad para predecir la demanda de electricidad considerando las temperaturas diarias

Análisis de regresión

- *analisis de regresión* para obtener una ecuación que indique cuál es la relación entre las variables.

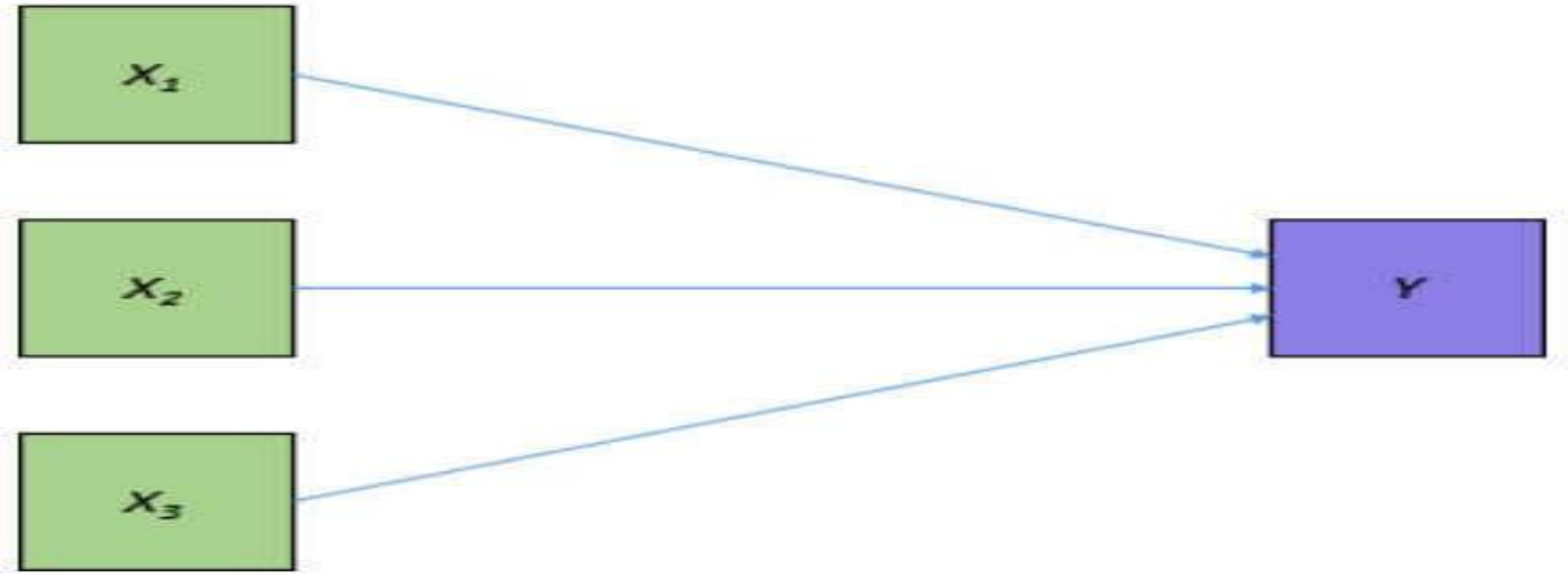
¿O lo dejamos a la intuición ?

Terminología

- A la variable que se va a predecir se le llama **variable dependiente**.
- A la variable o variables que se usan para predecir el valor de la variable dependiente se les llama **variables independientes**.

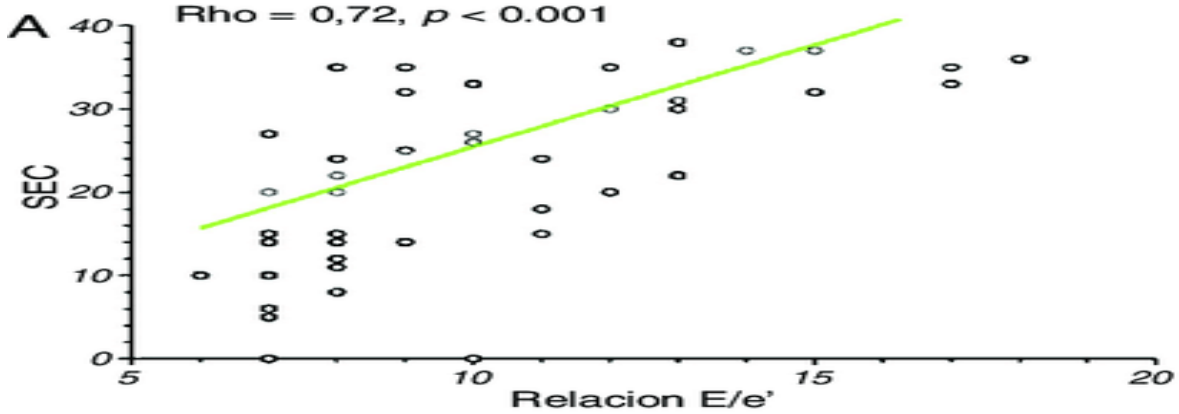
Terminología

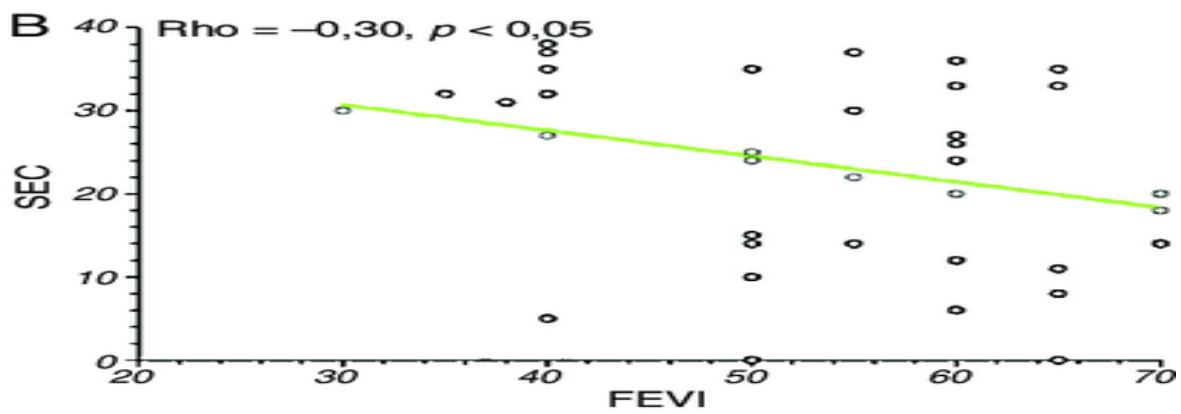
- A la variable que se va a predecir se le llama **variable dependiente**.
- A la variable o variables que se usan para predecir el valor de la variable dependiente se les llama **variables independientes**.

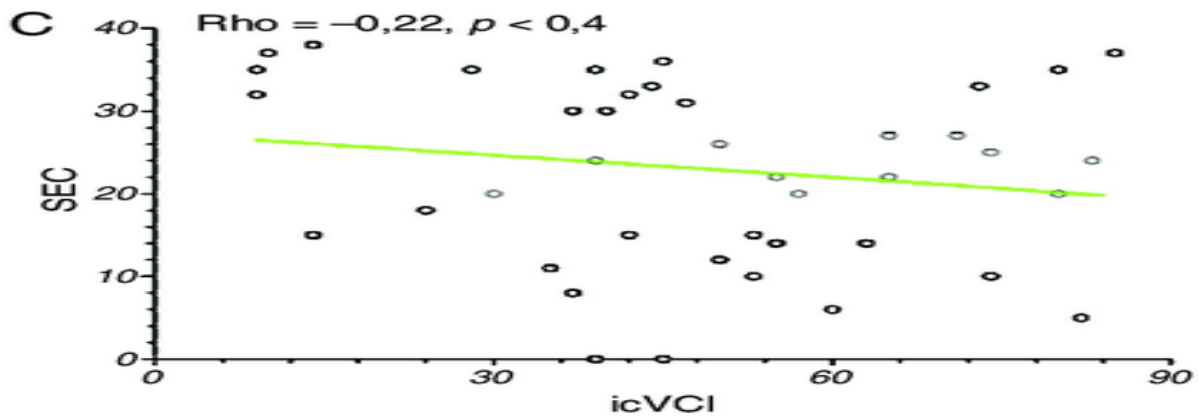


Terminología

- análisis de regresión en el que interviene una variable independiente y una variable dependiente y en el que la relación entre estas variables es aproximada mediante una línea recta. A este tipo de análisis de regresión se le conoce como **regresión lineal simple**.

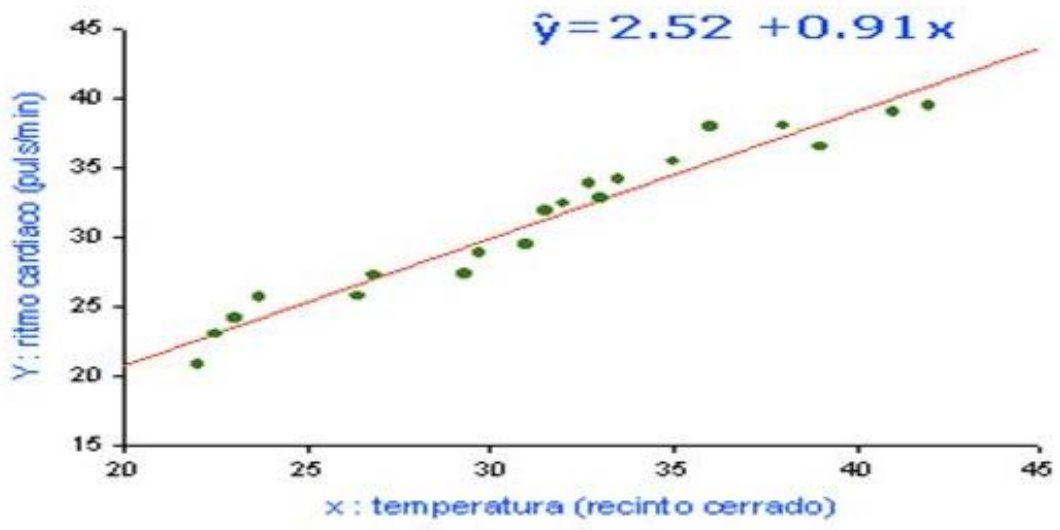






Terminología

- *En la regresion lineal simple, cada observacion consta de dos valores: uno de la variable independiente y otro de la variable dependiente.*



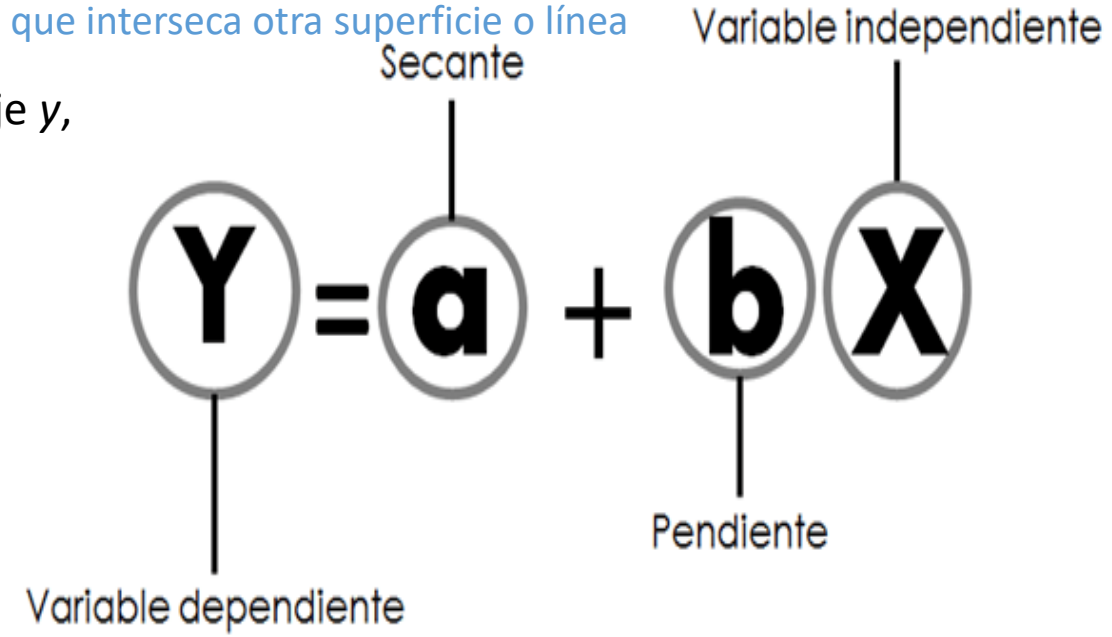
Ecuación

MODELO DE REGRESIÓN LINEAL SIMPLE

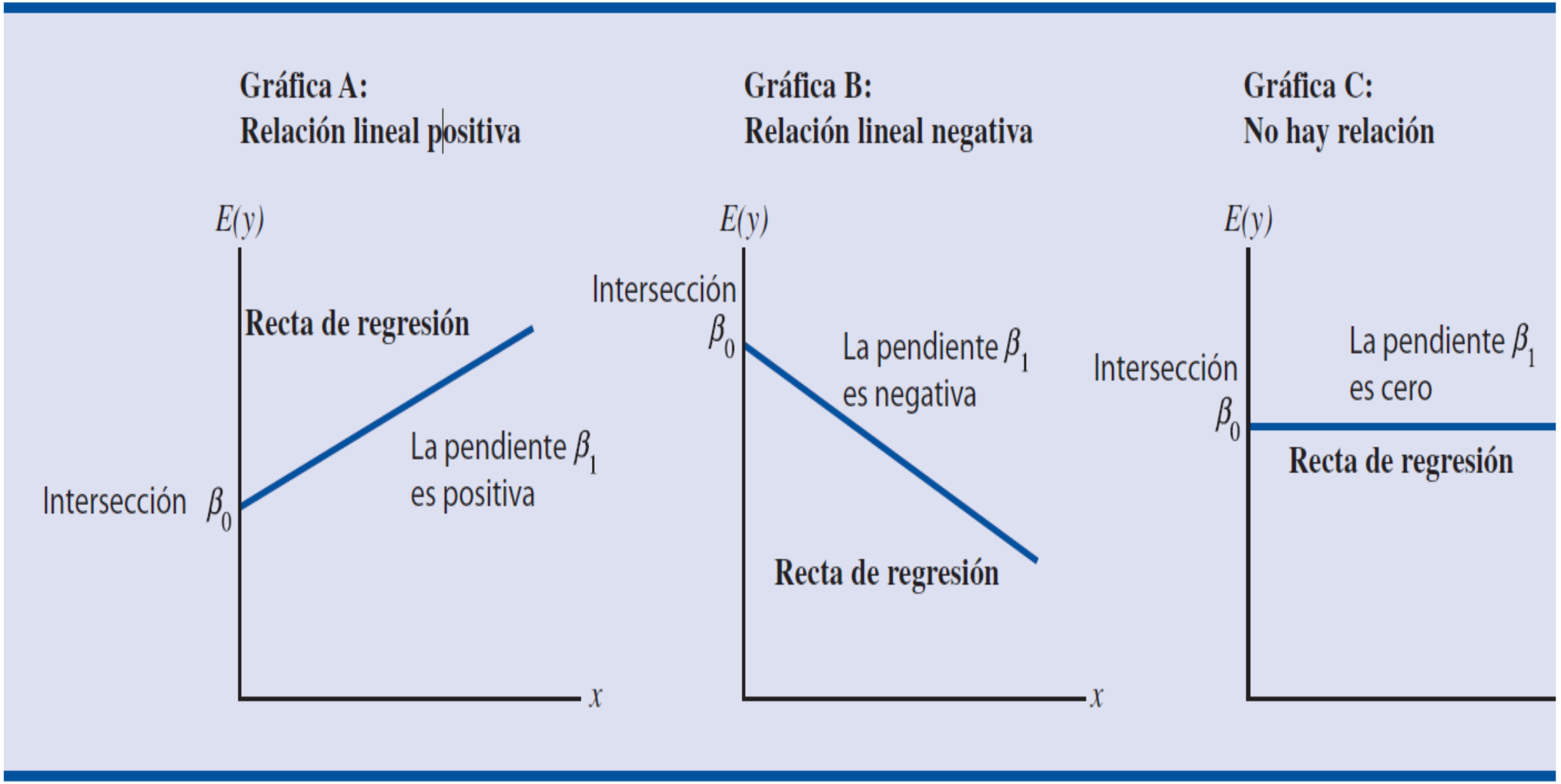
$$y = \beta_0 + \beta_1 x + \epsilon$$

Secante, es un concepto que, en la geometría, refiere a la superficie o la línea que interseca otra superficie o línea

Bo es la intersección de la recta de regresión con el eje *y*,
B1 es la pendiente
X Valor de variable independiente
Y es la media o valor esperado
E Error estadístico



Ecuación


Gráfica A:
Relación lineal positiva
E(y)
Recta de regresión
La pendiente β_1
es positiva
Intersección β_0
x
Gráfica B:
Relación lineal negativa
E(y)
Intersección
β_0
La pendiente β_1
es negativa
Recta de regresión
x
Gráfica C:
No hay relación
E(y)
Intersección
β_0
La pendiente β_1
es cero
Recta de regresión

Comentarios

El análisis de regresión no puede entenderse como un procedimiento para establecer una relación de causa y efecto entre las variables.
Este procedimiento sólo indica cómo o en qué medida las variables están relacionadas una con otra

Comentarios

Acerca de una relación causa
y efecto deben basarse en los
conocimientos de
los especialistas en la aplicación de
que se trate

Método de mínimos cuadrados

- El **método de mínimos cuadrados** es un método en el que se usan los datos muestrales para hallar la ecuación de regresión estimada.

- Ejemplo: se recolectan datos de una muestra de 10 restaurantes de Pizza ubicados todos cerca de diversas escuelas de educación superior.
- Las ventas dependen del número de estudiantes de cada escuela

Método de mínimos cuadrados

- El **método de mínimos cuadrados** es un método en el que se usan los datos muestrales para hallar la ecuación de regresión estimada.

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

¿Como determinar b0 y b1?

donde

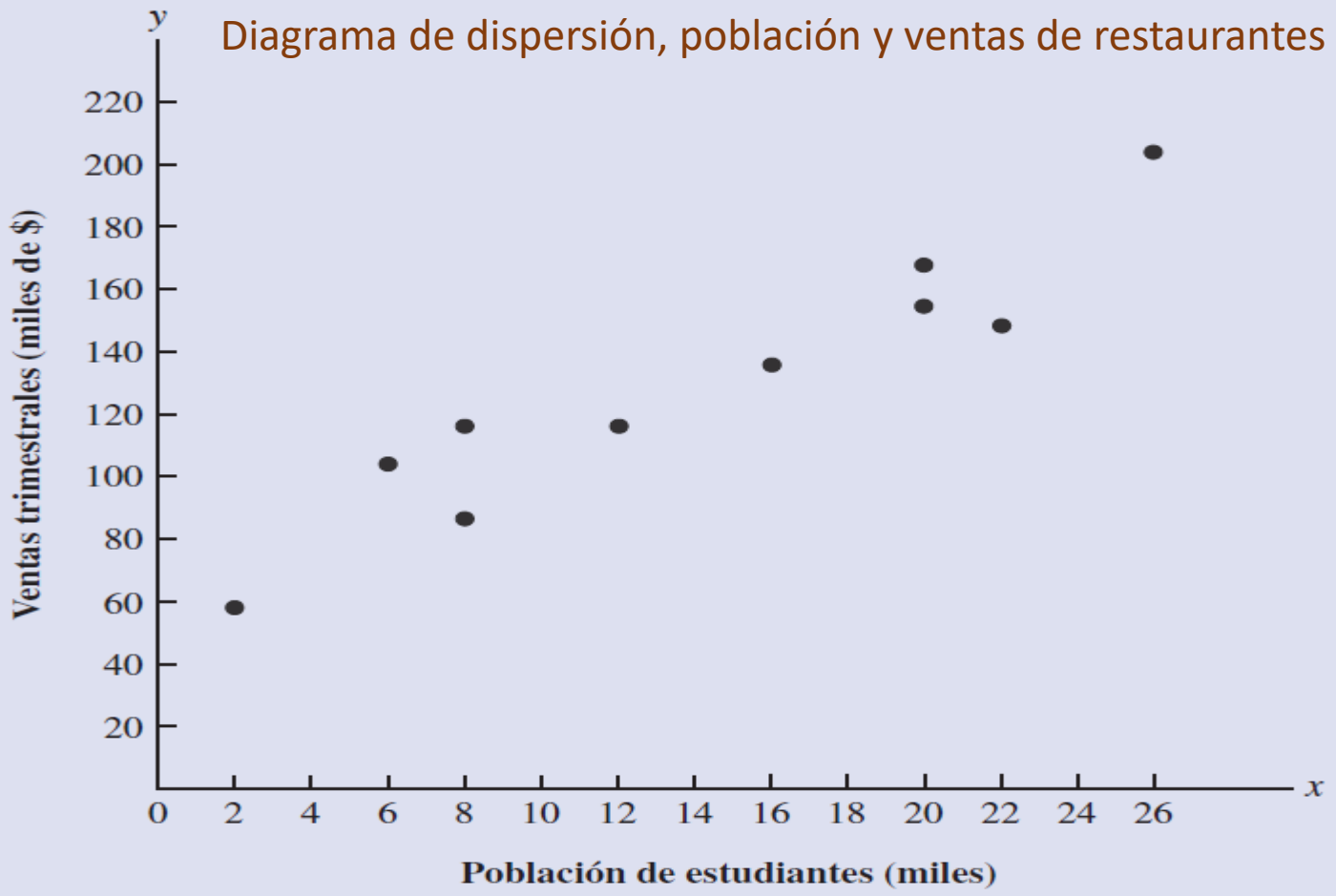
$\hat{y}_i$ = valor estimado de las ventas trimestrales (en miles de dólares) del restaurante i
b_0 = intersección de la recta de regresión con el eje y
b_1 = pendiente de la recta de regresión
x_i = tamaño de la población de estudiantes (en miles) del restaurante i

Método de mínimos cuadrados

Restaurante i	Población de estudiantes (miles) x_i	Ventas trimestrales (miles de $) 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

Método de mínimos cuadrados

Diagrama de dispersión, población y ventas de restaurantes Pizzas



Método de mínimos cuadrados

- En el método de mínimos cuadrados se usan los datos muestrales para obtener los valores de *b*0 y *b*1 que minimicen la *suma de los cuadrados de las desviaciones (diferencias)* entre los valores observados de la variable dependiente y_i y los valores estimados de la variable dependiente.

CRITERIO DE MÍNIMOS CUADRADOS

$$\min \Sigma(y_i - \hat{y}_i)^2$$

donde

y_i = valor observado de la variable dependiente en la observación i
$\hat{y}_i$ = valor estimado de la variable independiente en la observación i

Método de mínimos cuadrados

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

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

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

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

Método de mínimos cuadrados

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

$\bar{x}$ = 14 $\bar{y}$ = 130

Método de mínimos cuadrados

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

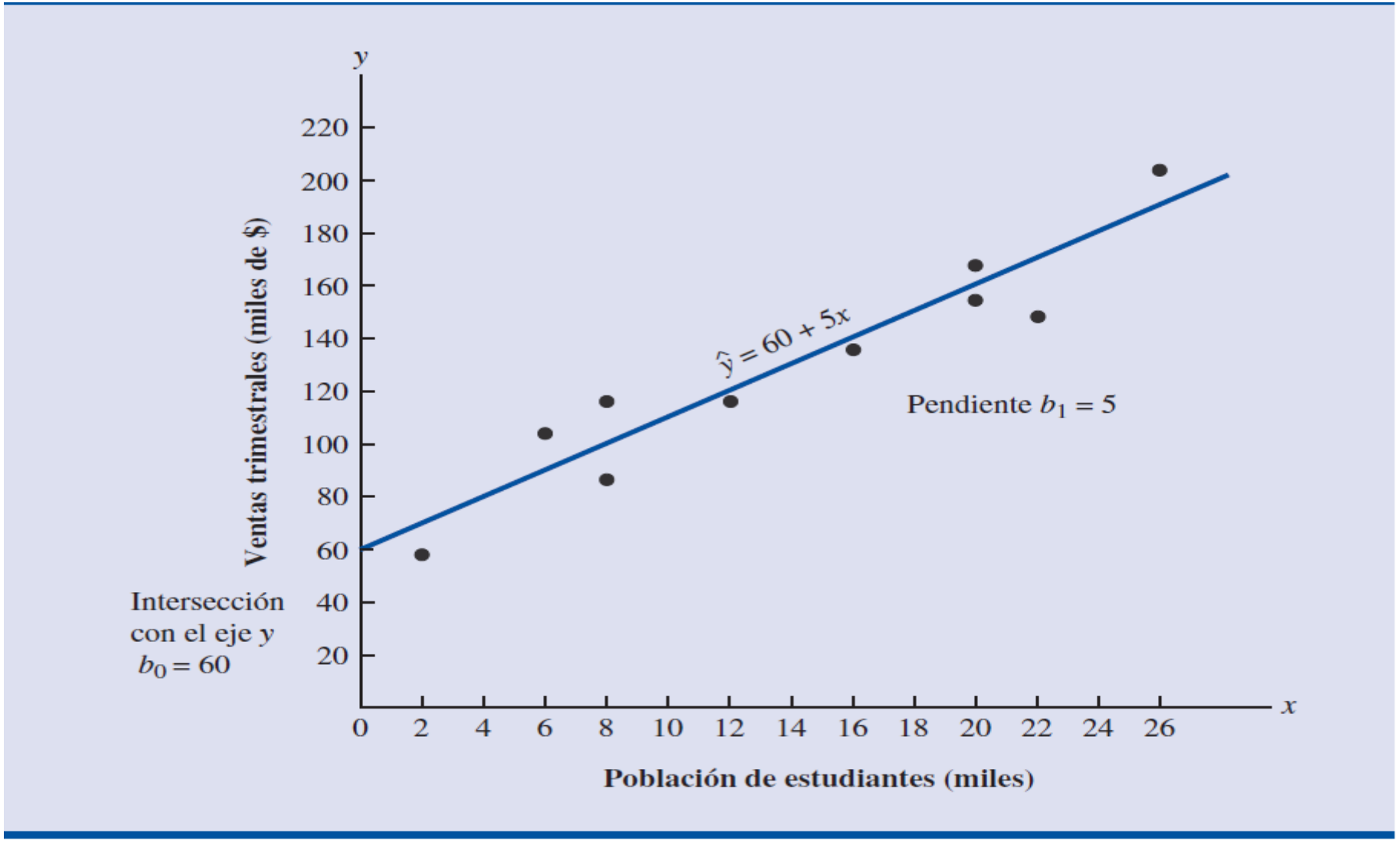
La intersección con el eje y (b_0) se calcula como sigue.

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

Por lo tanto, la ecuación de regresión estimada es

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

Método de mínimos cuadrados



Método de mínimos cuadrados

Por ejemplo, si se quisieran predecir las ventas trimestrales de un restaurante ubicado cerca de un campus de 16 000 estudiantes, se calcularía

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

Las ventas trimestrales pronosticadas para este restaurante serían de $140 000.

https://rpubs.com/rpizarro/575072 Estudiantes y Ventas

Ejemplos en R

https://rpubs.com/rpizarro/575072 Caso: Estudiantes y Ventas

https://rpubs.com/rpizarro/575057 Caso: Publicidad y Ventas

https://rpubs.com/rpizarro/576144 Caso: Millas de vuelo de aerolíneas y costos

Coeficiente de determinación. SCE

- ¿qué tan bien se ajusta a los datos la ecuación de regresión estimada?
- **coeficiente de determinación es** una medida de la bondad de ajuste de la ecuación de regresión estimada (lo bien que se ajusta la ecuación a los datos.
- A la diferencia que existe, en la observación *i*, entre el valor observado de la variable dependiente y_i, y el valor estimado de la variable dependiente $\hat{y}$ (que son las ventas pronosticadas), se le llama **residual *i***.

SUMA DE CUADRADOS DEBIDA AL ERROR

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

- El valor de **SCE** es una medida del error al utilizar la ecuación de regresión estimada para estimar los valores de la variable dependiente en los elementos de la muestra.

Coeficiente de determinación. SCE. residuales

residuales

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

menos

Por lo tanto, SCE 1530, mide el error que existe al utilizar la ecuación de regresión estimada

```
modelo$residuals
```

Coeficiente de determinación. STC

- La suma de las desviaciones al cuadrado que se obtiene cuando se usa la media muestral (130 para el ejemplo de Restaurantes Pizas) para estimar el valor de las ventas trimestrales de cada uno de los restaurantes de la muestra.
- Para el i-ésimo restaurante de la muestra, la diferencia $y_i - \bar{y}$ proporciona una medida del error que hay al usar para estimar las ventas.
- La correspondiente suma de cuadrados, llamada *suma total de cuadrados*, se denota **STC**.

SUMA TOTAL DE CUADRADOS

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

Coeficiente de determinación. STC

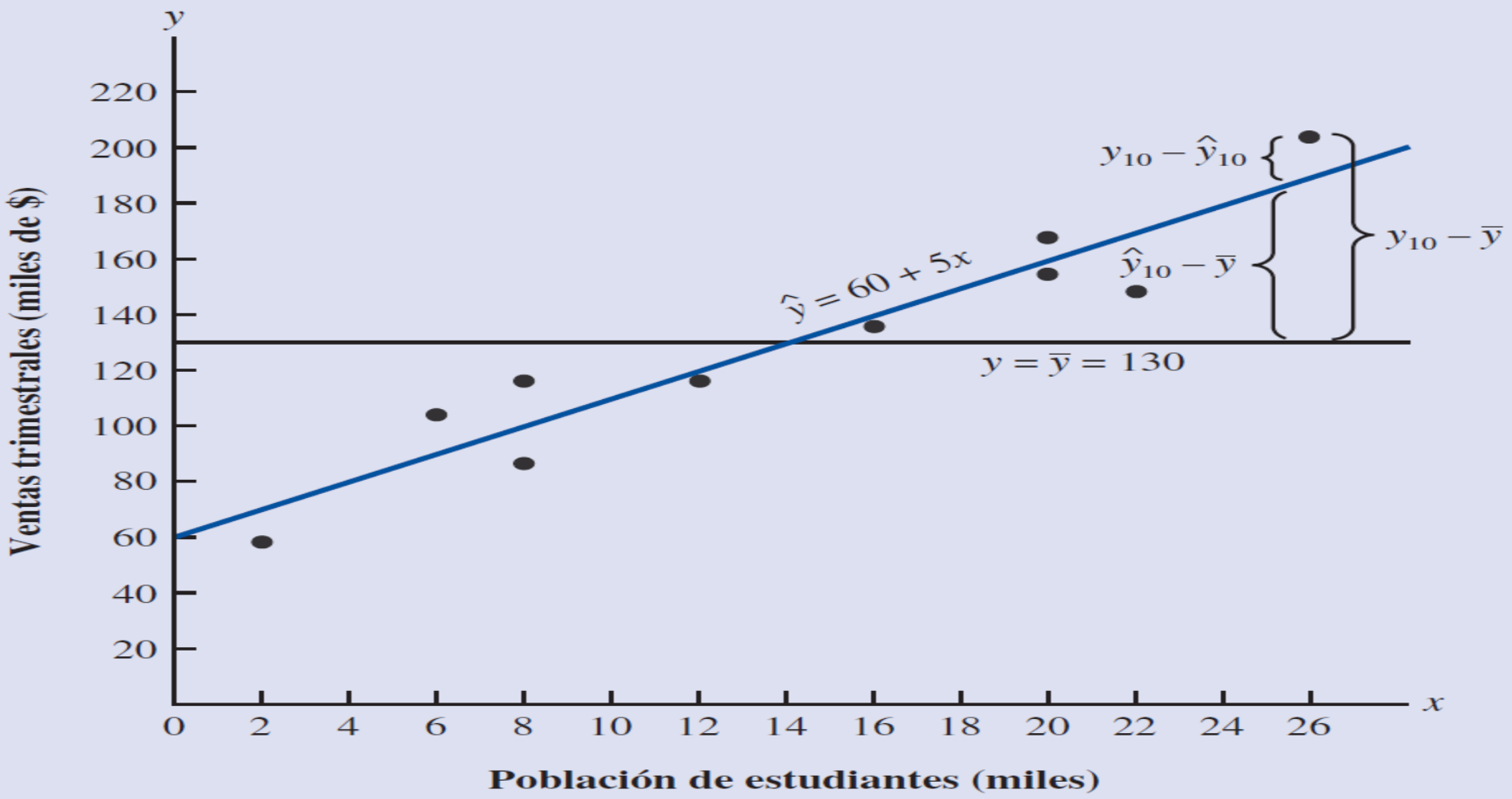
$\bar{y} = \Sigma y_i / n = 1300/10 = 130$

menos

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

Coeficiente de determinación

Como STC = 15 730 y SCE = 1530, la linea de regresión estimada se ajusta mucho mejor a los datos que la linea $y = \bar{y}$.



Coeficiente de determinación

- Se puede entender **STC** como una medida de qué tanto se agrupan las observaciones en torno a la recta $\bar{y}$, por otra parte,

- **SCE** como una medida de qué tanto se agrupan las observaciones en torno de la recta $\hat{y}$.

- Para medir qué tanto se desvían de los valores , 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

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

Coeficiente de determinación

RELACIÓN ENTRE STC, SCR Y SCE

$$STC = SCR + SCE$$

donde

STC = suma total de cuadrados

SCR = suma de cuadrados debida a la regresión

SCE = suma de cuadrados debida al error

Coeficiente de determinación CD

Se usan estas tres sumas de cuadrados, STC, SCR y SCE, para obtener una medida de la bondad de ajuste de la ecuación de regresión estimada

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.

COEFICIENTE DE DETERMINACIÓN

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

Es el porcentaje de varianza justificado por las variables independientes
Porcentaje de variabilidad de la variable dependiente con respecto a las independientes

Coeficiente de determinación

Si se expresa el **coeficiente de determinación** en forma de porcentaje, *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.

$$r^2 = \frac{\text{SCR}}{\text{STC}} = \frac{14\ 200}{15\ 730} = 0.9027$$

En el ejemplo de Armand's Pizza Parlors, 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 determinación ajustado (para regresión lineal múltiple)

- Muchos analistas prefieren ajustar *R*2 al número de variables independientes para evitar sobreestimar el efecto que tiene agregar una variable independiente sobre la cantidad de la variabilidad explicada por la ecuación de regresión estimada.
- Siendo *n* el número de observaciones y
- *p (ó k, en algunas fórmulas)* el número de variables independientes,
- el **coeficiente de determinación ajustado** se calcula como sigue.

COEFICIENTE DE DETERMINACIÓN MÚLTIPLE AJUSTADO

$$R_a^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

Coeficiente de determinación

https://rpubs.com/rpizarro/575072

```
summary(modelo)
```

```
##
## Call:
## lm(formula = ventas ~ estudiantes, data = datos)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.00  -9.75  -3.00  11.25  18.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.0000     9.2260   6.503 0.000187 ***
## estudiantes   5.0000     0.5803   8.617 2.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.83 on 8 degrees of freedom
## Multiple R-squared:  0.9027, Adjusted R-squared:  0.8906
## F-statistic: 74.25 on 1 and 8 DF,  p-value: 2.549e-05
```

coeficiente de determinación

coeficiente de determinación ajustado

Coeficiente de Correlación entre variables x , y, cor()

- El **coeficiente de correlación** es una medida descriptiva de la intensidad de la relación lineal entre dos variables *x* y *y*.
- Los valores del coeficiente de correlación son valores que van desde 1 hasta -1.
- El valor 1 indica que las dos variables *x* y *y* están perfectamente relacionadas en una relación lineal positiva; es decir, los puntos de todos los datos se encuentran en una línea recta que tiene pendiente positiva.
- El valor -1 1 indica que *x* y *y* están perfectamente relacionadas, en una relación lineal negativa, todos los datos se encuentran en una línea recta que tiene pendiente negativa.
- Los valores del coeficiente de correlación cercanos a cero (0), indican que *x* y *y* **no** están relacionadas linealmente.

COEFICIENTE DE CORRELACIÓN MUESTRAL

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

donde

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

Coeficiente de Correlación entre variables x , y

- En el caso de una relación lineal entre dos variables, tanto el **coeficiente de determinación** como el **coeficiente de correlación** muestral proporcionan medidas de la intensidad de la relación.

- El coeficiente de determinación proporciona una medida cuyo valor va desde **cero hasta uno**, mientras que el coeficiente de correlación muestral proporciona una medida cuyo valor va desde **–1 hasta +1**.

- El **coeficiente de correlación** lineal está restringido a la relación lineal entre dos Variables

- El **coeficiente de determinación** puede emplearse para relaciones no lineales y para relaciones en las que hay dos o más variables independientes.

- Por tanto, el **coeficiente de determinación** tiene un rango más amplio de aplicaciones.

Coeficiente de Relación entre variables x , y

- −0.90 = Correlación negativa muy fuerte.
- −0.75 = Correlación negativa considerable.
- −0.50 = Correlación negativa media.
- −0.25 = Correlación negativa débil.
- −0.10 = Correlación negativa muy débil.

- * 0.00 = No existe correlación alguna entre las variables.

- +0.10 = Correlación positiva muy débil.
- +0.25 = Correlación positiva débil.
- +0.50 = Correlación positiva media.
- +0.75 = Correlación positiva considerable.
- +0.90 = Correlación positiva muy fuerte.
- +1.00 = Correlación positiva perfecta (“A mayor X, mayor Y” o “a menor X, menor Y”, de manera
- proporcional. Cada vez que X aumenta, Y aumenta siempre una cantidad constante).

Metodología de la Investigación. Roberto Hernández Sampieri

Prueba de Significancia. Error Cuadrado Medio ECM

El **error cuadrado medio** (ECM) proporciona una estimación de $\sigma 2$; esta estimación es SCE dividida entre sus grados de libertad.

S2 = Q2

ERROR CUADRADO MEDIO (ESTIMACIÓN DE σ^2)

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

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

Prueba de Significancia. Error Estándar de Estimación ECM

- Para estimar σ se saca la raíz cuadrada de $s2$. Al valor que se obtiene, s, se le conoce como el **error estándar de estimación**.
- Raiz de Error Cuadrado Medio (**ECM**)

ERROR ESTÁNDAR DE ESTIMACIÓN

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

- **SCE** es una medida de la variabilidad de las observaciones reales respecto a la línea de regresión estimada
- El error estándar de estimación (**s**) se emplea acerca de las pruebas de significancia de la relación entre x y y.
- Prueba t, y prueba F

Prueba de Significancia. Error Estándar de Estimación ECM

En el ejemplo de Armand's Pizza Parlors, $s = \sqrt{ECM} = \sqrt{191.25} = 13.829$

Coeficiente de determinación

```
summary(modelo)
```

https://rpubs.com/rpizarro/575072

```
##
## Call:
## lm(formula = ventas ~ estudiantes, data = datos)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.00  -9.75  -3.00  11.25  18.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.0000     9.2260   6.503 0.000187 ***
## estudiantes   5.0000     0.5803   8.617 2.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.83 on 8 degrees of freedom
## Multiple R-squared:  0.9027, Adjusted R-squared:  0.8906
## F-statistic: 74.25 on 1 and 8 DF,  p-value: 2.549e-05
```

Prueba de Significancia. Prueba t

$$y = \beta_0 + \beta_1 x + \epsilon$$

- Si *x* y *y* están relacionadas linealmente, entonces β_1 es diferente de 0.
- El objetivo de la prueba *t* es determinar si se puede concluir que β_1 es diferente de cero y descartar la hipótesis nula

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

- Si se rechaza *H*0, se concluirá que β_1 es diferente de 0 y que entre las dos variables existe una relación estadísticamente significante

Prueba de Significancia. Prueba t

- Dado que el valor-*p* es menor a $\alpha = 0.01$ se rechaza *H*0 y se concluye que β_1 no es igual a cero.
- Esto es suficiente evidencia para concluir que existe una relación significativa entre la población de estudiantes y las ventas trimestrales.

PRUEBA t DE SIGNIFICANCIA PARA LA REGRESIÓN LINEAL SIMPLE

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

ESTADÍSTICO DE PRUEBA

$$t = \frac{b_1}{s_{b_1}}$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si valor-$p \leq \alpha$

Método del valor crítico: Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

donde $t_{\alpha/2}$ se toma de la distribución t con $n - 2$ grados de libertad.

Prueba de Significancia. Prueba F

- Una prueba ***F***, basada en la distribución de probabilidad *F* puede emplearse también para probar la significancia en la regresión.

- Cuando sólo se tiene una variable independiente, la prueba *F* lleva a la misma conclusión que la prueba *t*; es decir, si la prueba *t* indica que β_1 es diferente de 0, por tanto,

- La prueba ***F*** también indicará que existe una relación significante.

- Pero cuando hay más de una variable independiente, sólo la prueba ***F*** puede usarse para probar que existe una relación significante general.

Prueba de Significancia. Prueba F

PRUEBA F DE SIGNIFICANCIA EN EL CASO DE LA REGRESIÓN LINEAL SIMPLE

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

ESTADÍSTICO DE PRUEBA

$$F = \frac{\text{CMR}}{\text{ECM}}$$

REGLA DE RECHAZO

Método del valor-p:	Rechaza H_0 si valor-$p \leq \alpha$
Método del valor crítico:	Rechaza H_0 si $F \geq F_\alpha$

donde F_α es un valor de la distribución F con 1 grado de libertad en el numerador y $n - 2$ grados de libertad en el denominador.

Prueba de Significancia. Prueba F

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	Valor-*p*
Regresión	14 200	1	$\frac{14\,200}{1} = 14\,200$	$\frac{14\,200}{191.25} = 74.25$	0.000
Error	1 530	8	$\frac{1530}{8} = 191.25$		
Total	15 730	9			

anova(modelo) # Usando R

Comentarios al margen

El análisis de regresión, se usa para identificar la relación entre las variables, apoyados de análisis de correlación.

no puede emplearse como evidencia de una relación de causa y efecto.

Es una herramienta de apoyo matemático estadístico para estimaciones y predicciones.

Lo que se puede decir es que las variables están relacionadas y que la relación lineal explica parcialmente y de manera significante la variabilidad de y sobre el rango de los valores de x observados en los datos

Prácticas en R

1. Cargar librerías
2. Cargar datos
3. Explorar datos
4. Diagrama de dispersión
5. Aplicar modelo
6. Analizar modelo
7. Interpretar modelo
8. Resultados