Florencia Gonzalez

Regresión y correlación de variables

Correlación de variables

El análisis de correlación de variables nos permite analizar en forma conjunta dos o más variables, para luego inferir resultados sobre una de ellas a partir de la otra (y otras).

Situación: Acabamos de salir de un exámen y nos encontramos con la noticia de que nos ha ido mal en el mismo.

Análisis de las acciones llevadas a cabo.

- Horas de estudios.
- Notas anteriores.
- Fuentes consultadas

Relevar a nuestro alrededor a los pares e indagamos sobre los mismo parámetros.

DEBEMOS ENCONTRAR UNA CORRELACIÓN ENTRE LO QUE CONSIDERAMOS PARÁMETROS HABITUALES Y PERTINENTES CON RESPECTO A LA CONSIGNA (rendir correctamente un exámen) Y SU CUANTIFICACIÓN (horas de estudios, cantidad de ejercicios, etc).

A partir de allí uno saca algunas conclusiones:

- Este exámen hay que prepararlo con más tiempo y/o hay que hacer casi todos los ejercicios de la práctica

REGRESIÓN LINEAL

Comenzaremos haciendo un análisis de correlación entre variables.

EJEMPLO INTRODUCTORIO

Una cadena de restaurantes de comida italiana tiene algunas sucursales cerca de campus de universidades.

Los gerentes creen que las ventas trimestrales de estas sucursales (que se denotan por *y*) están directamente relacionadas con el tamaño de la población estudiantil (que se denota *x*), o sea que en las sucursales cerca de campus que tienen una población grande, se genera una venta más grande.

¿Cuál será la relación entre las variables *x* e *y*?

REGRESIÓN LINEAL

IDEA PRINCIPAL

Cuando es posible tener datos, puede emplearse un procedimiento estadístico llamado ***análisis de regresión***, para obtener una ecuación que dictamine cuál es la relación entre ambas variables.

- Variable a predecir: variable dependiente.
- Variables que se utilizan para predecir: variables independientes.

Estudiaremos el tipo más sencillo, cuando las variables se relacionan en una línea recta, aproximadamente. El mismo se conoce como <u>regresión lineal simple</u>.

<u>Modelo de regresión lineal simple</u>: ecuación con que se describe cómo se relacionan x e y, más un error.

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

β_0 y β_1 son los parámetros del modelo, y ε es una variable aleatoria que se conoce como término del error.

REGRESIÓN LINEAL

Ecuación de regresión lineal simple: $E(y) = \beta_0 + \beta_1 x$

- Su gráfica es una recta.
- β_0 es la intersección de la recta con el eje y.
- β_1 es la pendiente.

POSIBLES RECTAS DE REGRESIÓN.

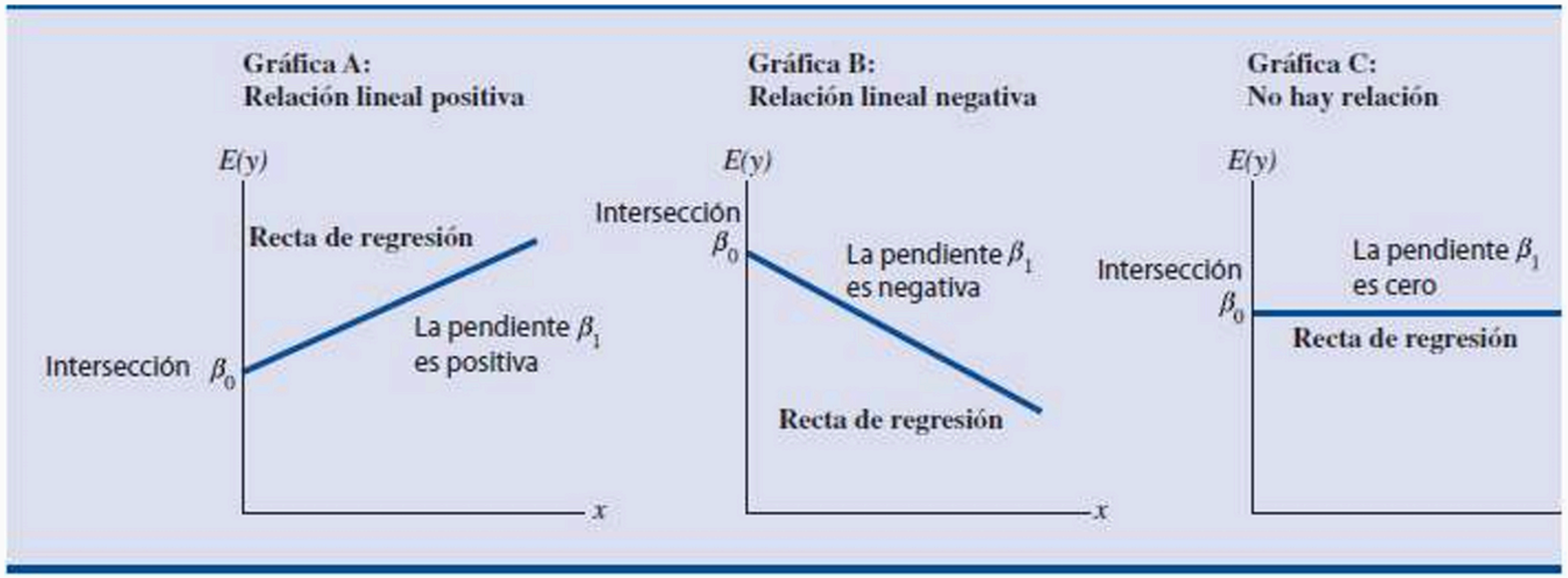


Figura 14.1 – Anderson – Pág. 546

ECUACIÓN DE REGRESIÓN ESTIMADA.

- Si se conocen β_0 y β_1, los reemplazamos en la ecuación dada y obtenemos lo requerido.
- Si no se conocen β_0 y β_1

Se los estima con datos muestrales

Se calculan estadísticos muestrales, b_0 y b_1, como estimaciones de los parámetros poblacionales β_0 y β_1

Se obtiene la ecuación de regresión lineal simple estimada

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

REGRESIÓN LINEAL

El métodos de mínimos cuadrados

- Método en el que se usan los datos muestrales para obtener la ecuación de regresión estimada.
- Siguiendo con el ejemplo…

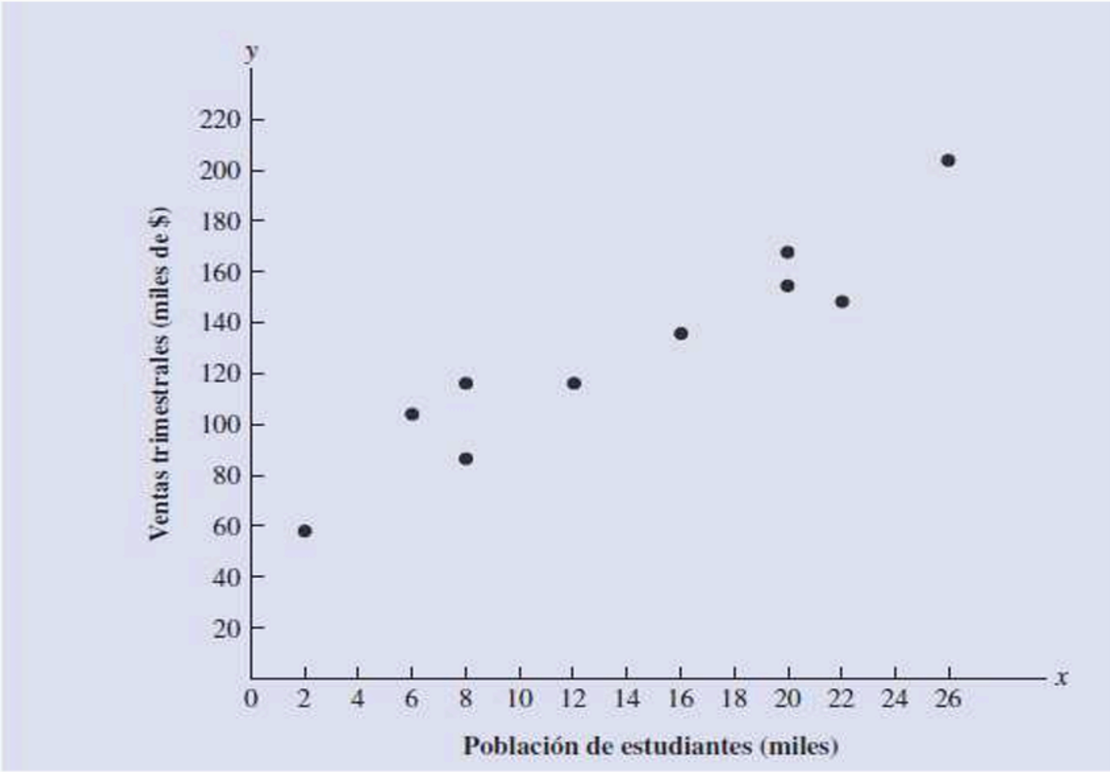
  Se recolectan datos de una muestra de 10 restaurantes ubicados cerca de campus universitarios. Los mismos están en la tabla:

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

REGRESIÓN LINEAL

Diagrama de dispersión

Pasamos estos datos a un diagrama de dispersión.



La relación entre x e y parece poder aproximarse mediante una línea recta. Se elige el modelo lineal simple, debemos buscar b_0 y b_1.

REGRESIÓN LINEAL

El métodos de mínimos cuadrados

Para el restaurante i, tendremos: $\widehat{y_i} = b_0 + b_1 \cdot x_i$

- y_i denota las ventas reales observadas.
- $\widehat{y_i}$ denota las ventas estimadas.

La diferencia entre los valores observados y los estimados debe ser pequeña, para cada restaurante.

Se quieren encontrar los valores de b_0 y b_1 que minimicen la suma de los cuadrados de las desviaciones.

$$Criterio\ de\ mínimos\ cuadrados\colon \min \sum (y_i - \hat{y}_i)^2$$

Se puede demostrar que, los valores que minimizan esa expresión son:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

REGRESIÓN LINEAL

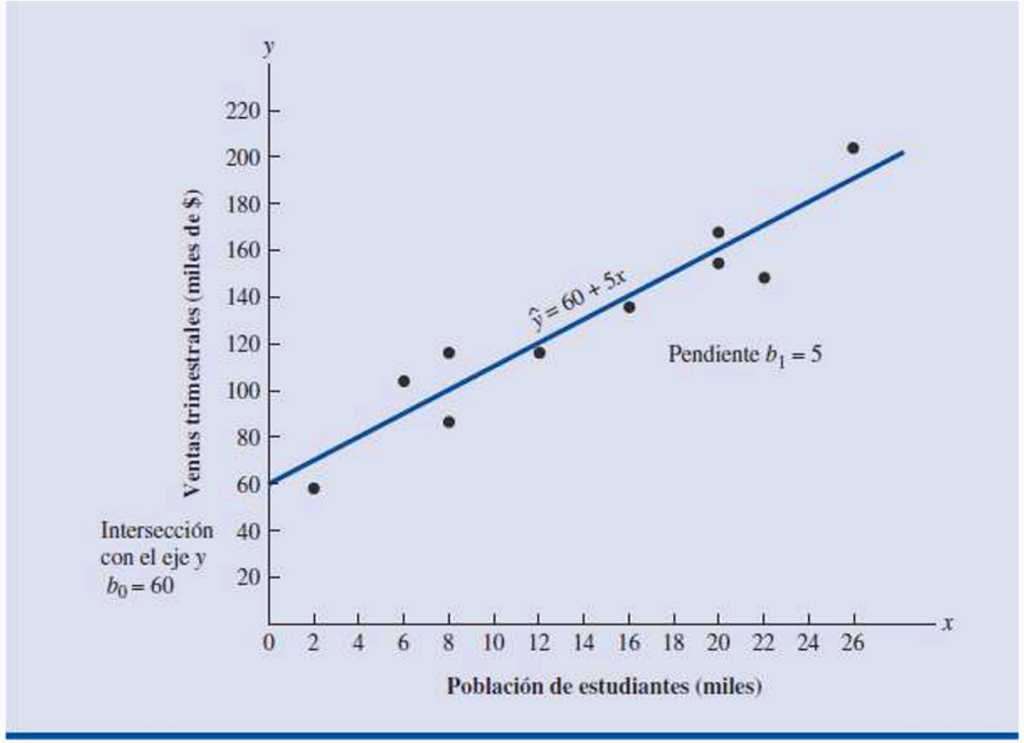
El métodos de mínimos cuadrados

En nuestro ejemplo…

$n = 10, \bar{x} = 14, \bar{y} = 130, b_1 = 5, b_0 = 60$

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

¿Qué se puede interpretar?



REGRESIÓN LINEAL

Error de la Estimación

Para continuar nuestro análisis de la regresión, calcularemos a continuación qué tan confiable es la ecuación hallada, lo cual haremos mediante el error estándar de la estimación, que puede calcularse con la siguiente fórmula:

$$S_e = \sqrt{\frac{\sum_1^n y_i{}^2 - a\sum_1^n y_i - b\sum_1^n x_i y_i}{n-2}}$$

Restaurante (i) |Población de estudiantes (xi)|ventas trimestrales (miles) (yi)|x.y| x^2|y^2

REGRESIÓN LINEAL

Error de la Estimación