# Lección 3. Análisis conjunto de dos variables

**Estadística Descriptiva**

Parcialmente financiado a través del PIE13-024 (UMA)

**JULIA DE HARO GARCÍA**

## TEMA 3. ANÁLISIS CONJUNTO DE DOS VARIABLES

**3.1 COVARIANZA Y COEFICIENTE DE CORRELACIÓN LINEAL SIMPLE**

**3.2 REGRESIÓN LINEAL**

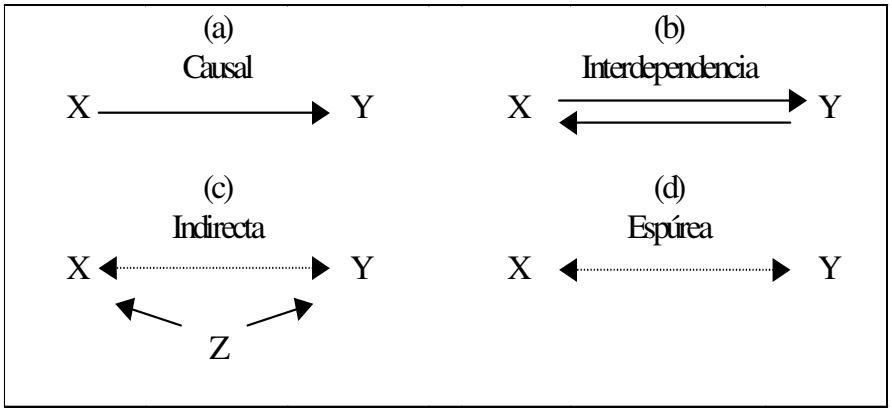
## Objetivos

- Saber representar e interpretar una distribución bidimensional mediante una nube de puntos.
- Interpretar la correlación lineal como una medida de la relación lineal existente entre dos variables.
- Determinar e interpretar la recta de regresión.
- Analizar la bondad del ajuste realizado. Coeficiente de determinación.
- Realizar predicciones, de forma crítica, a partir de una recta de regresión.



## Tipos de covariación

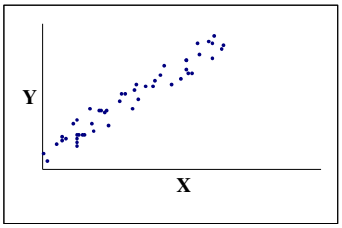
*Relaciones entre dos variables*



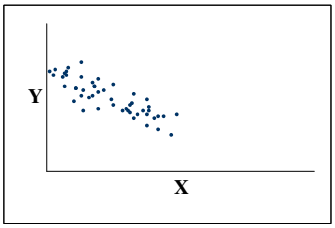
## Covariación. Análisis gráfico

### Nubes de puntos

a) *Asociación lineal creciente*



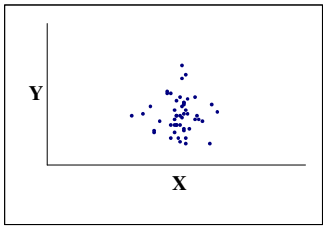

b) *Asociación lineal decreciente*



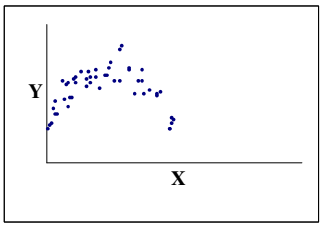

JULIA DE HARO 5

## Covariación. Análisis gráfico

c) *Ausencia de asociación*




d) *Asociación no lineal*




JULIA DE HARO 6

## 3.1 Covarianza

La **covarianza,** $s_{XY}$, es una medida absoluta que cuantifica la covariación lineal entre dos variables X e Y.

$$s_{XY} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N}$$

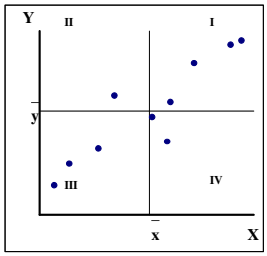
$$s_{XY} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \bar{x} \cdot \bar{y}$$

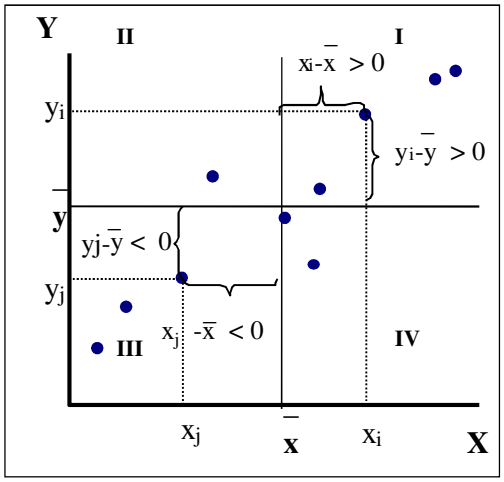


## 3.1 Covarianza

- El **signo** de la covarianza indica el tipo de asociación lineal entre X e Y.
  - Una **covarianza positiva** indica una relación lineal positiva o directa (Los valores bajos de ambas variables se presentan a la vez al igual que los elevados, pero no sabemos si hay una relación causal entre ellas) .
  - Una **covarianza negativa** es un reflejo de una relación lineal inversa o decreciente o negativa.
  - Una **covarianza nula** indica que no existe covariación o relación lineal entre las variables, pero puede o no existir otro tipo de covariación como exponencial, parabólica etc.

## 3.1 Covarianza






## 3.1 Coeficiente de correlación lineal simple

Una **medida relativa** que expresa el grado de asociación lineal entre las variables es el **coeficiente de correlación lineal simple**, $r_{XY}$ :

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \qquad -1 \le r_{XY} \le 1$$

Si $r_{XY} = 1$, la relación lineal es perfecta y directa

Si $r_{XY} = -1$, la relación lineal es perfecta e inversa

Si $r_{XY} = 0$, no existe relación lineal entre las variables ($s_{XY}=0$)

(Puede que sean independientes o que exista otro tipo de relación entre ellas)

## 3.2 Regresión lineal

Entre los diversos tipos de relaciones entre dos variables, la más estudiada es **la relación de causalidad**, es decir, una de ellas es la causa de la otra.

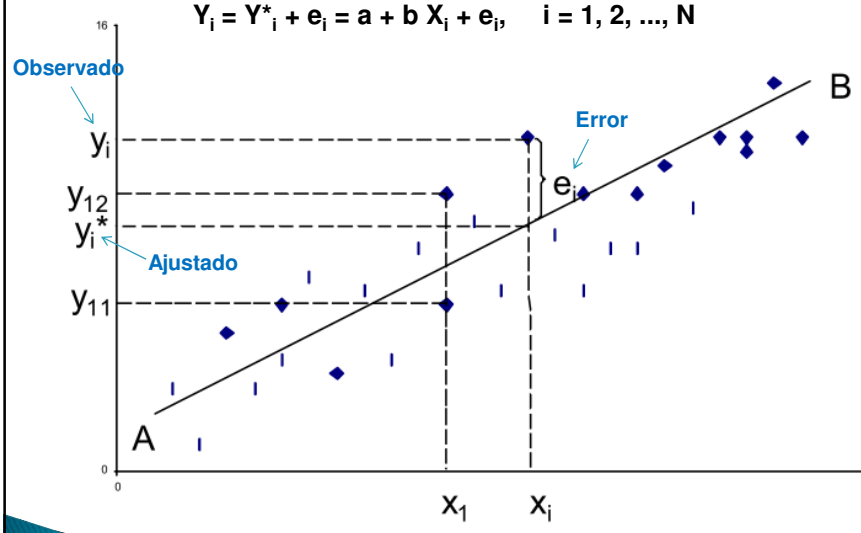
Aquella variable que depende de otra se denomina variable dependiente, endógena o explicada (Y), mientras que la variable que es el origen de los cambios se denomina variable independiente, exógena o explicativa (X).

El análisis de la regresión consiste en obtener la línea ideal, línea de regresión, hacia la que tienden los puntos del diagrama de dispersión; es decir, trata de determinar la dependencia exacta que se haya contenida en la dependencia estadística, mediante la eliminación de factores aleatorios.



## 3.2 Regresión lineal

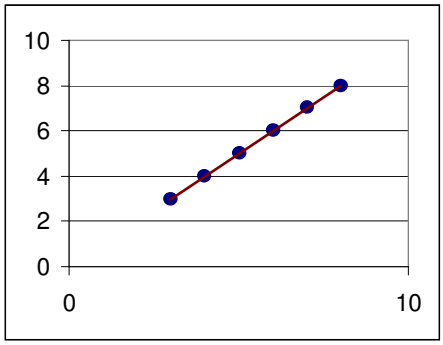
$$Y_i = Y^*_i + e_i = a + b\,X_i + e_i, \quad i = 1, 2, \ldots, N$$



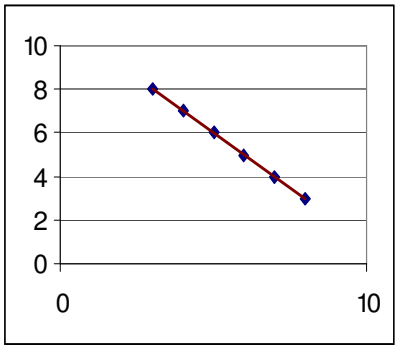

12

## 3.2 Regresión lineal

Sólo en el caso particular de que la relación fuese "perfecta", la nube de puntos estaría sobre una recta y todos los errores $e_i$ serían nulos.




$r_{XY} = 1$




$r_{XY} = -1$



## 3.2 Regresión lineal

De entre las infinitas rectas que sería posible ajustar a una nube de puntos, nos quedamos con la resultante de aplicar el criterio de mínimos cuadrados ordinarios, de tal forma que la recta obtenida es aquella que hace mínima la suma de los errores, $e_i$, al cuadrado.

$$\text{Min } \Sigma\, e_i^2 = \Sigma\, (Y_i - Y^*_i)^2$$

Tras la resolución de un sistema de ecuaciones normales, en el que se ha impuesto este criterio, los parámetros a y b de la recta de regresión son:

$$a = \overline{y} - b\,\overline{x}$$

$$b = \frac{s_{XY}}{s_X^2}$$

## 3.2 Regresión lineal. Interpretación de coeficientes

**a:** Es el **término independiente,** la ordenada en el origen.

Indica el valor de Y cuando X toma el valor cero.

**b:** Es la pendiente de la recta de regresión. Coeficiente de regresión.

Se puede interpretar diciendo que por término medio, cuando X se **incrementa** en una unidad, la variable Y aumenta (o disminuye, según el signo) en b unidades, según el modelo estimado.



## 3.2 Regresión lineal

- si b<0, **la dependencia lineal es inversa o negativa** (recta decreciente)
- si b=0, **no existe dependencia lineal** ($s_{XY}=0$) (recta horizontal, se cumple, $y_i^* = a$ )
- si b>0, **la dependencia lineal es directa o positiva** (recta creciente)

## 3.2 Regresión lineal

Entre las propiedades de mayor interés de la recta de regresión tenemos:

- pasa por el centro de gravedad de la nube de puntos, por el punto $(\overline{x}, \overline{y})$.
- la suma de los residuos resultantes es cero ($\Sigma e_i = 0$).



## 3.2 Regresión lineal. Elasticidad

Un concepto muy importante en economía es el concepto de **elasticidad**, que es la variación porcentual que experimenta la variable Y cuando X aumenta un 1%.

$$\text{Elasticidad} = \frac{\partial Y / Y}{\partial X / X} = \frac{\partial Y}{\partial X}\frac{X}{Y} = b\frac{X}{Y}$$

Elasticidad media:

$$E = b\frac{\overline{X}}{\overline{Y}}$$

## 3.2 Bondad del ajuste. Coeficiente de determinación

Para el **ajuste lineal** por el método de los mínimos cuadrados, se demuestra que el **coeficiente de determinación**, $R^2$, es igual al **coeficiente de correlación lineal simple al cuadrado, $r^2_{xy}$.**

$$R^2 = \frac{s^2_{Y*}}{s^2_Y} = 1 - \frac{s^2_e}{s^2_Y} = r^2_{XY} \qquad 0 \leq R^2 \leq 1$$

**$R^2$** indica la proporción de las variaciones de Y que vienen explicadas por X, según el modelo lineal.

Por lo tanto, **1-$R^2$** es la proporción de las variaciones de Y que no explica X según el modelo lineal ajustado. Lo no explicado por el modelo.



## Predicción

El análisis de la regresión, además de describir la dependencia causal entre las variables sirve para estimar o predecir el valor de la variable dependiente (Y) dado un valor de la variable independiente (X). Para ello basta con sustituir en el modelo estimado un valor de la variable exógena $X_p$ y el resultado es la predicción que el modelo ofrece para la variable endógena $Y_p$ :

$$Y_p = a + b\, X_p$$

Si el valor de X elegido para predecir el correspondiente valor de Y, está dentro del recorrido de X, a la predicción se le denomina **interpolación** y si está fuera del recorrido se le llama **extrapolación**.

## Predicción

**Fiabilidad de la predicción**

- Sólo deben realizarse predicciones si el modelo ofrece un **grado de ajuste** satisfactorio.
- La **interpolación** (predicción para valores dentro de la nube de puntos) tendrá un grado de fiabilidad mayor que la extrapolación, y en el caso de **extrapolaciones**, (predicción para valores fuera de la nube de puntos) cuanto más lejos estemos de los valores observados, menor precisión se obtiene en el valor de predicción.



## Ejercicio

Los ingresos y los gastos anuales de 10 familias, seleccionadas de forma aleatoria en un pueblo, en miles de euros, son:

| Ingresos | Gastos |
|---|---|
| 12 | 11 |
| 15 | 13 |
| 24 | 20 |
| 25 | 23 |
| 30 | 25 |
| 26 | 26 |
| 12 | 14 |
| 36 | 32 |
| 20 | 18 |
| 24 | 23 |

## Ejercicio

Haciendo uso de la anterior información, se pide:

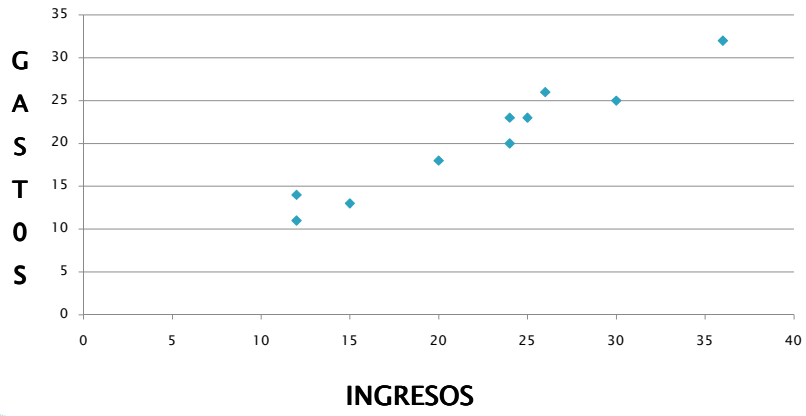
a) Represente gráficamente la nube de puntos correspondiente.
b) Calcule la covarianza entre ambas variables. Interprétela.
c) Calcule el coeficiente de correlación lineal. Interprételo.
d) Ajuste por el método de los mínimos cuadrados una recta en la que el gasto sea función de los ingresos.
e) ¿Cuáles son los significados estadístico y económico de los coeficientes de la línea ajustada?
f) Proporcione una medida de la bondad del ajuste. Interprétela.
g) Estime el gasto de una familia con 12500 € de ingresos. Comente la fiabilidad de dicha predicción.
h) Determine la variación porcentual que experimenta el gasto de una familia de ingresos 12500 € ante un incremento porcentual unitario de estos últimos.



## Ejercicio

Solución:

a)



## Ejercicio

b)

| | INGRESOS (X) | GASTOS (Y) | XY | X^2 | Y^2 |
|---|---|---|---|---|---|
| | 12 | 11 | 132 | 144 | 121 |
| | 15 | 13 | 195 | 225 | 169 |
| | 24 | 20 | 480 | 576 | 400 |
| | 25 | 23 | 575 | 625 | 529 |
| | 30 | 25 | 750 | 900 | 625 |
| | 26 | 26 | 676 | 676 | 676 |
| | 12 | 14 | 168 | 144 | 196 |
| | 36 | 32 | 1152 | 1296 | 1024 |
| | 20 | 18 | 360 | 400 | 324 |
| | 24 | 23 | 552 | 576 | 529 |
| TOTALES | 224 | 205 | 5040 | 5562 | 4593 |



## Ejercicio

b) La expresión de la covarianza es:

$$s_{XY} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \overline{x} \cdot \overline{y}$$

$$S_{XY} = \frac{5040}{10} - 22{,}4 * 20{,}5 = 44{,}8 * 10^6 €$$

Al obtener un valor positivo nos indica una relación lineal positiva o directa entre el ingreso y el gasto. Aspecto que ya se podía apreciar en la nube de puntos, apartado a.

## Ejercicio

c) La expresión del coeficiente de correlación es:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \qquad -1 \leq r_{XY} \leq 1$$

$$r_{XY} = \frac{44{,}8}{\sqrt{54{,}44}\sqrt{39{,}05}} = 0{,}97$$

Al obtener un valor positivo nos indica una relación lineal positiva o directa entre el ingreso y el gasto. Al estar su valor próximo a 1, esta relación es bastante intensa.



## Ejercicio

d) La recta de regresión a estimar es

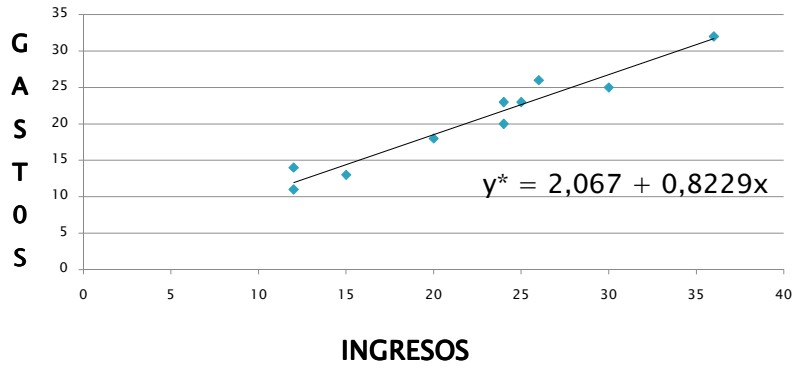
$$y_i^* = a + bx_i \quad \text{donde} \quad a = \bar{y} - b\bar{x}, \quad b = \frac{s_{XY}}{s_X^2}$$

$$a = 20{,}5 - 0{,}8229 * 22{,}4 = 2{,}067$$

$$b = \frac{44{,}8}{54{,}44} = 0{,}8229$$

## Ejercicio

d) La recta ajustada gráficamente:




JULIA DE HARO 29

## Ejercicio

e) Significado estadístico:

**a:** Ordenada en el origen. Término independiente.

**b:** Pendiente de la recta de regresión. Coeficiente de regresión.

Significado económico:

**a:** Para un ingreso nulo el gasto estimado es de 2067€. Podría ser una situación de endeudamiento.

**b:** Si aumentamos los ingresos en 1000€ el gasto aumenta en 822,9€, por término medio, según el modelo estimado.

JULIA DE HARO 30

## Ejercicio

f) Bondad del ajuste

$$R^2 = r_{XY}^2 \qquad 0 \leq R^2 \leq 1$$

$$R^2 = 0{,}97^2 = 0{,}94$$

Interpretación:

El ingreso explica el 94% de las variaciones del gasto según el modelo estimado. Es un buen ajuste (0,94 > 0,75). Solo deja sin explicar el 6%, (1- $R^2$).



## Ejercicio

g) Predicción

$Y_p^* = 2{,}067 + 0{,}8229\ (12{,}5) = 12{,}3532^* 10^3$ €

Fiabilidad de la predicción:

Como hemos visto el modelo tiene un grado de ajuste satisfactorio ($R^2 = 0{,}94 > 0{,}75$) y como 12,5 es un valor de los ingresos dentro del recorrido de dicha variable (interpolación), podemos concluir que la predicción realizada es fiable.

## Ejercicio

h) Elasticidad $$E = b\frac{X}{Y}$$

Este es el caso de elasticidad para un valor concreto de los ingresos, 12,5 ( no es elasticidad media), por ello el valor de Y (gasto) viene dado por:

$Y_p^* = 2{,}067 + 0{,}8229\ (12{,}5) = 12{,}3532^* 10^3$ €

$$E = 0{,}8229\ \frac{12{,}5}{12{,}3532} = 0{,}83$$

Interpretación: Al aumentar el ingreso en un 1%, aumenta el gasto un 0,83%.



## Ejercicio

h) En el caso de elasticidad media se procede igual sólo que sustituyendo los valores de x e y por sus puntos medios ( no hay que realizar la predicción del gasto).

$$E = b\frac{\overline{X}}{\overline{Y}}$$