ESTADÍSTICA

GRADO TURISMO

TEMA 3: ANÁLISIS DE DATOS TURÍSTICOS BIDIMENSIONALES

Prof. Rosario Martínez Verdú

TEMA 3: ANÁLISIS DE DATOS TURÍSTICOS BIDIMENSIONALES

1. Distribuciones bidimensionales de frecuencias y diagrama de dispersión.
2. Covariación y correlación.
3. Regresión lineal.
4. Análisis de la Bondad del Ajuste y predicción.

1.- Distribuciones bidimensionales de frecuencias y diagrama de dispersión

TIPOS DE DISTRIBUCIONES BIDIMENSIONALES CONJUNTAS:

- **Distribuciones con frecuencias conjuntas no unitarias**

Objetivo: Analizar dos variables simultáneamente o conjuntamente a partir de la ordenación de los datos en tablas de doble entrada o de contingencia.

Familia	X nº miembros	Y nº coches
1	1	0
2	3	1
3	1	1
4	5	2
5	5	2
6	3	2
7	1	0
8	3	0
9	5	1
10	1	1

a) Obtener la distribución conjunta de frecuencias de (X,Y).

b) Obtener las distribuciones marginales.

c) ¿Son X e Y independientes?

d) Obtener la distribución de frecuencias del nº de miembros de las familias sin automóvil.

e) Obtener la distribución de frecuencias del nº de coches de las familias de 3 miembros.

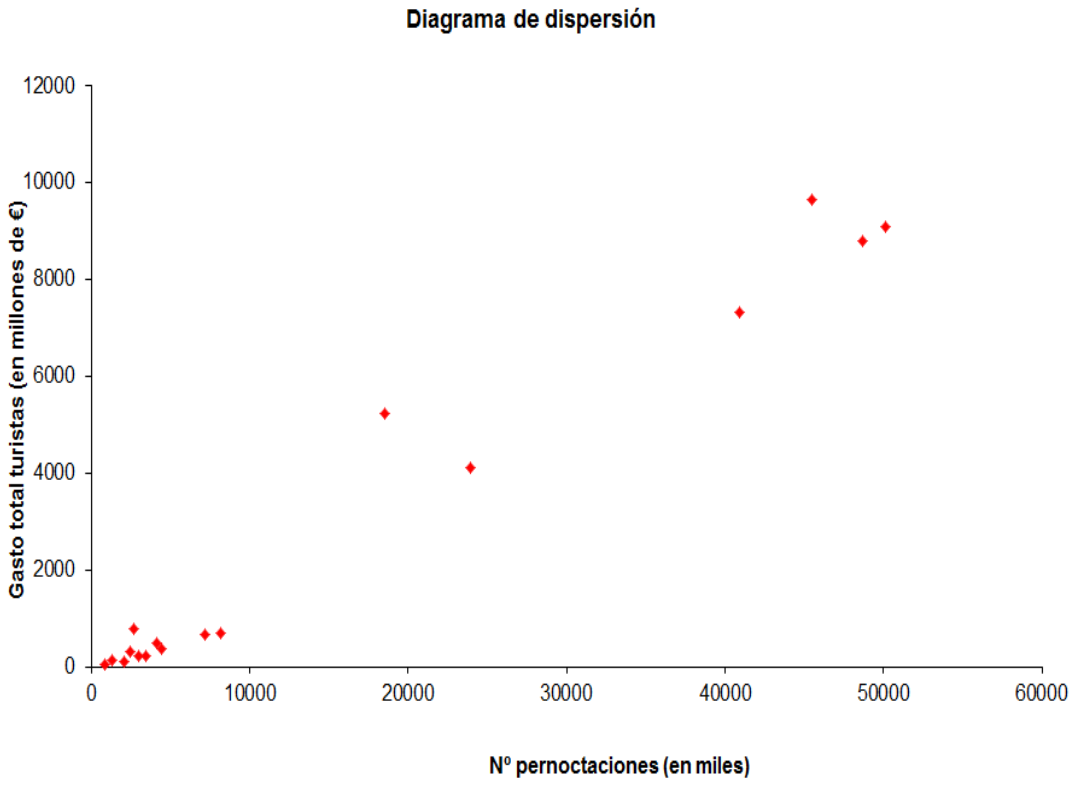
•Distribuciones con frecuencias conjuntas unitarias

Se dispone de información para 2009 sobre las N=17 Comunidades Autónomas sobre las siguientes variables:

•X: Nº de Pernoctaciones, en miles.

•Y: Gasto total de los turistas, en millones de euros.

CCAA	X Nº pernoctaciones	Y Gasto total turistas
Andalucía	40916	7337,7
Aragón	4417	365,9
Asturias	2996	212,6
Baleares	48676	8790,7
Canarias	50132	9082,3
Cantabria	2500	323,2
Castilla-La Mancha	3495	216,9
Castilla y León	7178	680,7
Cataluña	45484	9643,0
Comunidad Valenciana	23950	4101,8
Extremadura	2065	118,5
Galicia	8196	682,8
Madrid	18561	5226,9
Murcia	2715	775,1
Navarra	1363	143,0
País Vasco	4184	498,0
Rioja	899	43,3



Fuente: Encuesta de ocupación hotelera 2009, INE y Encuesta de Gasto Turístico (Egatur) 2009, IET.

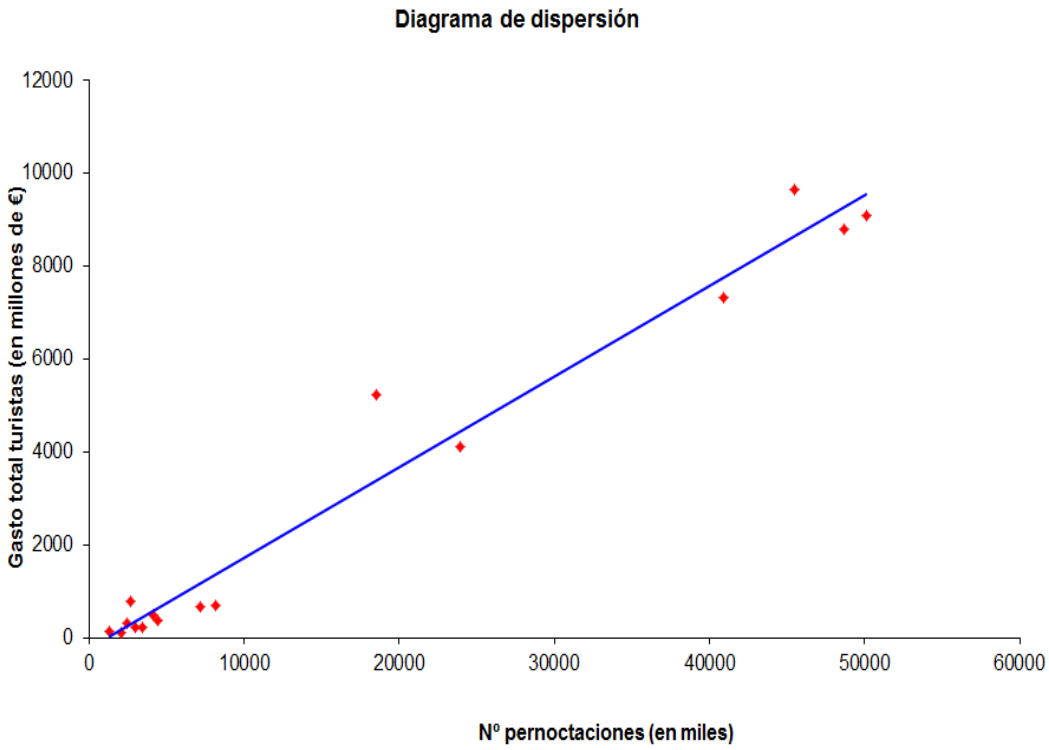
•Distribuciones con frecuencias conjuntas unitarias

Se dispone de información para 2009 de las N=17 Comunidades Autónomas sobre las siguientes variables:

•X: Nº de Pernoctaciones, en miles.

•Y: Gasto total de los turistas, en millones de euros.

CCAA	X Nº pernoctaciones	Y Gasto total turistas
Andalucía	40916	7337,7
Aragón	4417	365,9
Asturias	2996	212,6
Baleares	48676	8790,7
Canarias	50132	9082,3
Cantabria	2500	323,2
Castilla-La Mancha	3495	216,9
Castilla y León	7178	680,7
Cataluña	45484	9643,0
Comunidad Valenciana	23950	4101,8
Extremadura	2065	118,5
Galicia	8196	682,8
Madrid	18561	5226,9
Murcia	2715	775,1
Navarra	1363	143,0
País Vasco	4184	498,0
Rioja	899	43,3



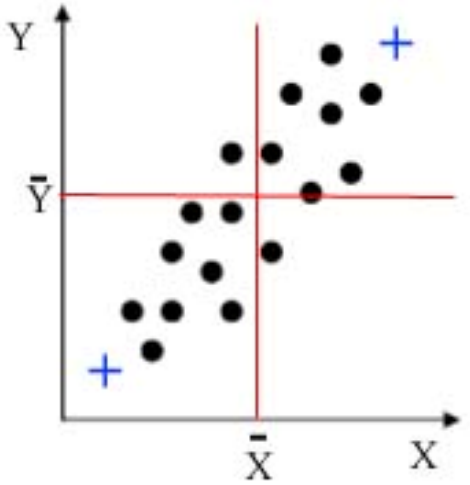
Fuente: Encuesta de ocupación hotelera 2009, INE y Encuesta de Gasto Turístico (Egatur) 2009, IET.

2.- COVARIACIÓN Y CORRELACIÓN

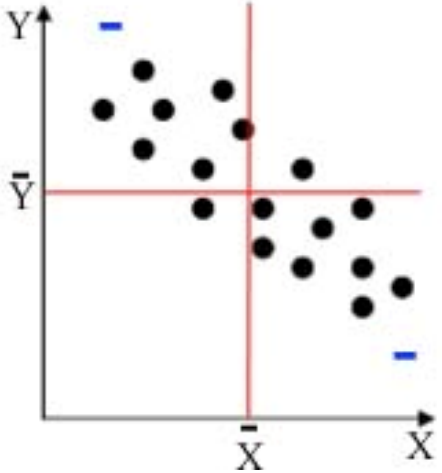
Objetivo: definir unas medidas estadísticas (**covarianza y coeficiente de correlación lineal**) que pongan de manifiesto la existencia o no de relación de tipo lineal entre dos variables. Para ello nos basamos en 2 características importantes de la distribución conjunta de (X,Y):

Vector de Medias: $\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix}$ Matriz de Varianzas-Covarianzas: $\begin{pmatrix} S_X^2 & S_{XY} \\ S_{XY} & S_Y^2 \end{pmatrix}$

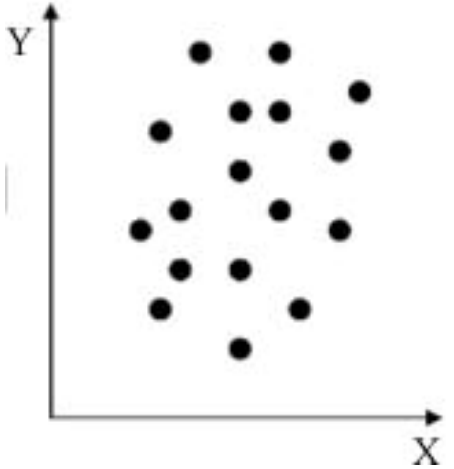
Covarianza: $S_{XY} = \frac{1}{N}\sum (X_i - \bar{X})(Y_i - \bar{Y})$ ⟹ S_{XY} sirve para medir la variación conjunta entre X e Y. Más que su valor, interesa analizar su signo.



$S_{XY}>0$ las variables varían en el mismo sentido



$S_{XY}<0$ las variables varían en sentido contrario



$S_{XY}=0$ no hay variación conjunta (incorrelación)

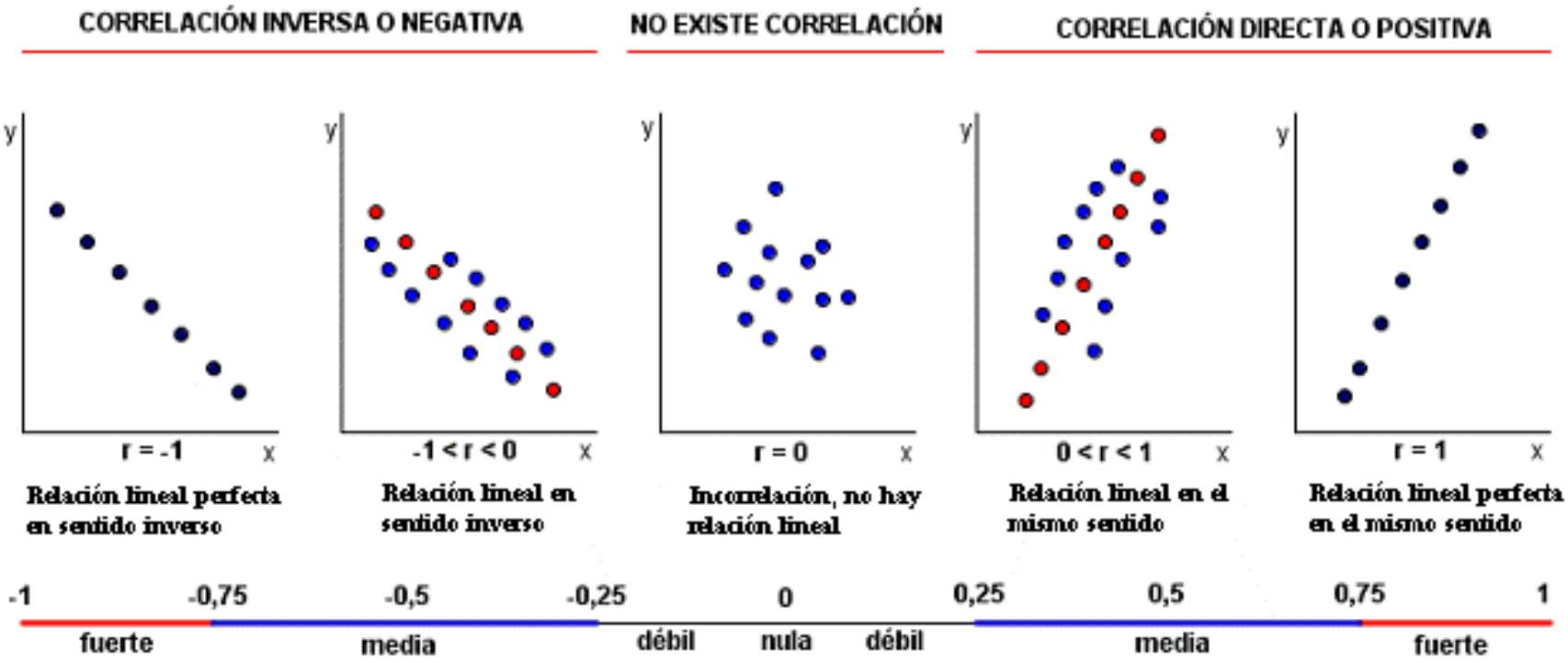
Coeficiente de correlación lineal r_{XY}

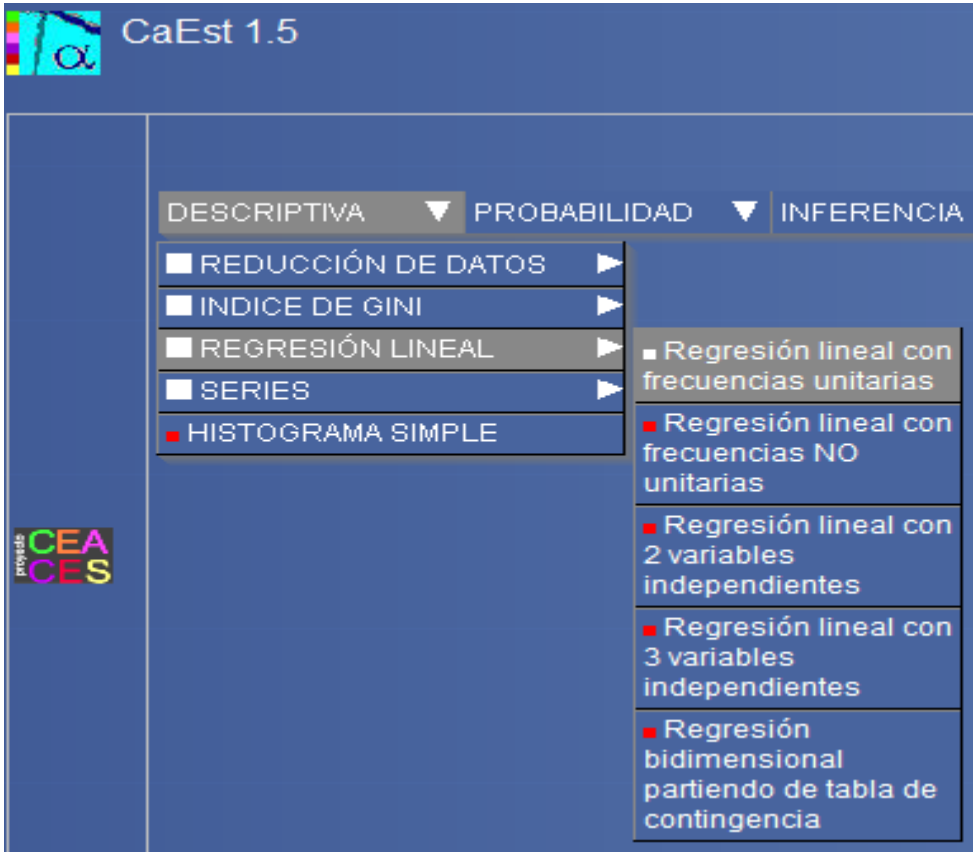
Está basado en la covarianza y mide el grado o intensidad de la relación lineal entre dos variables como también determina el sentido de dicha relación. Interesa interpretar tanto su valor como su signo. Se define como:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} \qquad -1 \leq r_{XY} \leq 1$$

Signo r_{XY} = signo de S_{XY}

Interpretación del valor y del signo de r_{XY}





Con la CaEst se pueden calcular todas estas medidas:

Ejemplo anterior:
X: Nº de Pernoctaciones
Y: Gasto total de los turistas

Vector de Medias: $\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix}$

Matriz de Varianzas-Covarianzas: $\begin{pmatrix} S_X^2 & S_{XY} \\ S_{XY} & S_Y^2 \end{pmatrix}$

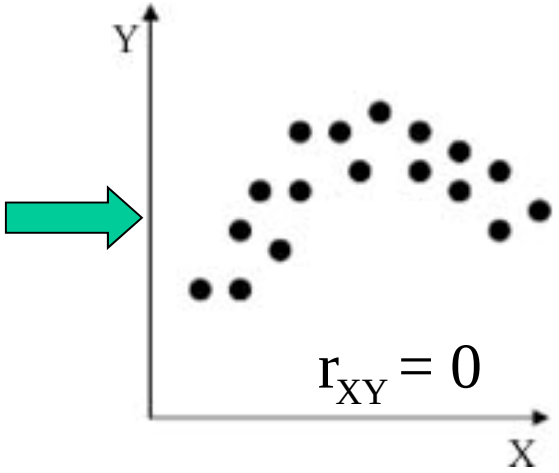
Resultados de cálculo de las Medidas con Caest

Indicadores	Y	X	
Media	2837.788	15748.647	
Varianzas y covarianza	12704865.039	324983247.522	63324218.433
Desv.Típica	3564.388	18027.292	
C.Correlación	0.985		

r_{XY}

S_{XY}

Si $r_{XY}=0$, **¿son las variables independientes?**
No necesariamente, lo único que se puede concluir es que no hay relación lineal entre las variables, pero las variables pueden tener otro tipo de relación.

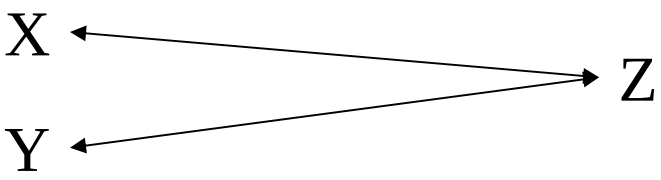


Correlaciones Espúreas o sin sentido

A veces es posible encontrar un coeficiente de correlación alto entre dos variables que no tienen relación justificada por ninguna teoría. Es lo que se llama correlación espúrea o sin sentido. Un ejemplo: Neyman en 1952 analizó la relación entre la tasa de nacimientos de niños y niñas y la población de cigüeñas en varias regiones, y encontró un alto coeficiente de correlación entre estas variables.

Correlación indirecta

A veces dos variables X e Y presentan un coeficiente de correlación lineal alto entre ellas, pero esta relación es aparente o indirecta ya que ambas variables están en realidad relacionadas con una tercera variable Z. Para medir la verdadera relación entre X e Y se puede calcular el **COEFICIENTE DE CORRELACIÓN PARCIAL**.



COEFICIENTE DE CORRELACIÓN PARCIAL

Es un coeficiente de correlación lineal entre X e Y en el que se elimina la influencia que ejerce una tercera variable Z sobre ambas variables.

EJEMPLO

CCAA	X nº de reclusos	Y nº de Bibliotecas	Z Población 2009
Andalucía	17495	869	8302923
Aragón	2644	374	1345473
Asturias	1547	159	1085289
Baleares	1937	184	1095426
Canarias	3198	208	2103992
Cantabria	724	71	589235
Castilla-La Mancha	7021	453	2081313
Castilla y León	2227	609	2563521
Cataluña	10531	830	7475420
Comunidad Valenciana	8240	624	5094675
Extremadura	1408	501	1102410
Galicia	4904	550	2796089
Madrid	10515	513	6386932
Murcia	967	129	1446520
Navarra	250	131	630578
País Vasco	1472	323	2172175
Rioja	405	51	321702

Fuente: INE y Ministerio del Interior.

$r_{XY} = 0{,}816$ $\quad r_{XZ} = 0{,}945$ $\quad r_{YZ} = 0{,}849$

¿Es real esta alta correlación positiva entre X e Y o hay una tercera variable Z (Población 2009) que es la responsable? Calculamos el coeficiente de correlación parcial entre X e Y:

$$r_{XY}^{p} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{\left(1-r_{XZ}^2\right)\left(1-r_{YZ}^2\right)}}$$

$$= \frac{0{,}816 - 0{,}945 \times 0{,}849}{\sqrt{\left(1-0{,}945^2\right)\left(1-0{,}849^2\right)}} = 0{,}079$$

Si se elimina la influencia de la variable población (Z), casi no hay relación lineal entre el nº de reclusos (X) y el nº de bibliotecas (Y).

3.- REGRESIÓN LINEAL

Vamos a suponer que entre las variables X e Y existe **una relación de causa-efecto.** Es decir, una variable (la X) es la causa y la otra (la Y) es el efecto. Variaciones en X (la causa) van a provocar variaciones en Y (el efecto).

Ejemplo: Para un conjunto de hogares, las variables Ingresos y Gasto en Turismo, ¿cuál sería X (la causa) y cuál sería Y (el efecto)?

Regresión Y/X (de Y respecto a X): Es una función matemática que nos va a explicar los valores de la Y a partir de los valores de la X: $Y = f(X)$

- X será la **variable independiente** o explicativa.
- Y será la **variable dependiente** o explicada.

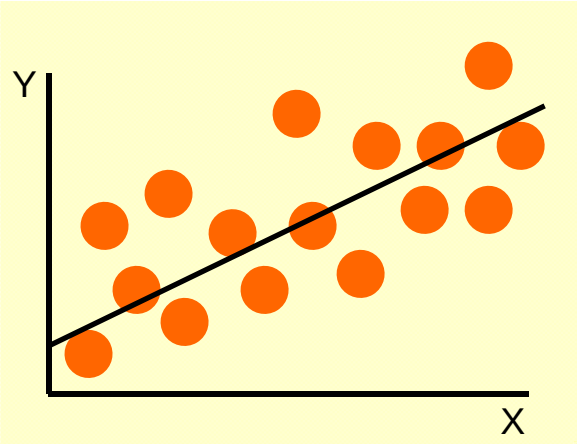
Utilidades de la regresión:

- Medir el efecto que una variación (aumento o disminución) de X provoca en Y.
- Hacer predicciones para la variable Y a partir de valores de X.

Modelo de Regresión Y/X (de Y respecto a X): función matemática que nos va a explicar los valores de la Y a partir de los valores de la X: Y= f(X)

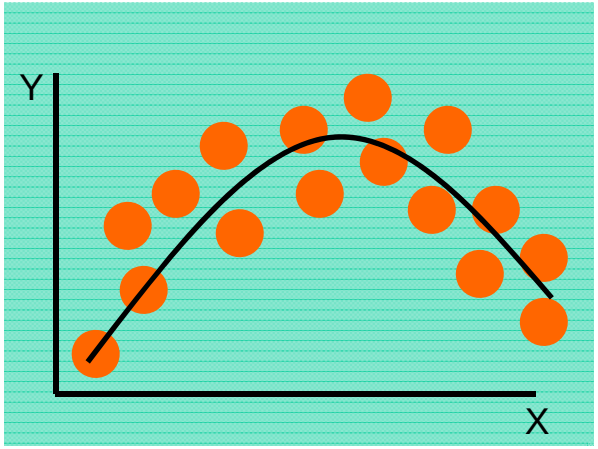
EJEMPLOS DE MODELOS DE REGRESIÓN

El diagrama de dispersión nos ayuda a determinar el tipo de relación existente entre 2 variables:



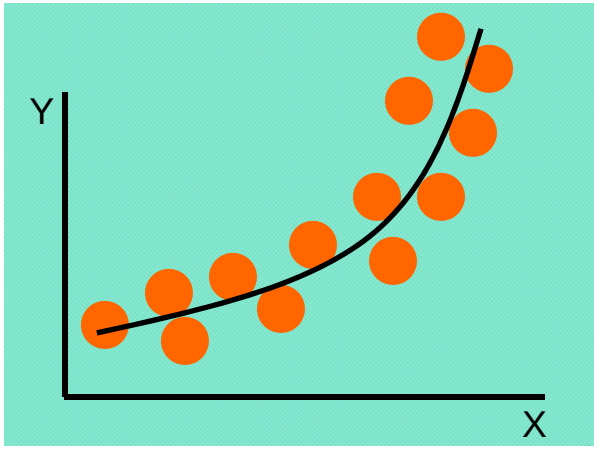
Regresión Lineal:

$Y^* = a + bX$



Regresión Parabólica:

$Y^* = a + bX + cX^2$



Regresión Exponencial:

$Y^* = a\, b^X$

Es la que veremos

MODELO DE REGRESIÓN LINEAL

PROBLEMAS DEL MODELO DE REGRESIÓN:

-1º Elegir una función matemática que relacione ambas variables.

Elegimos una función lineal (una recta) por → Sencillez
→ Más se usa

-2º ¿Cuál es la recta que mejor se ajusta a los puntos del diagrama de dispersión?

Ecuación de una recta: $Y^* = a + b\,X$

En definitiva, determinar los valores de los **coeficientes a y b** de la recta de regresión.

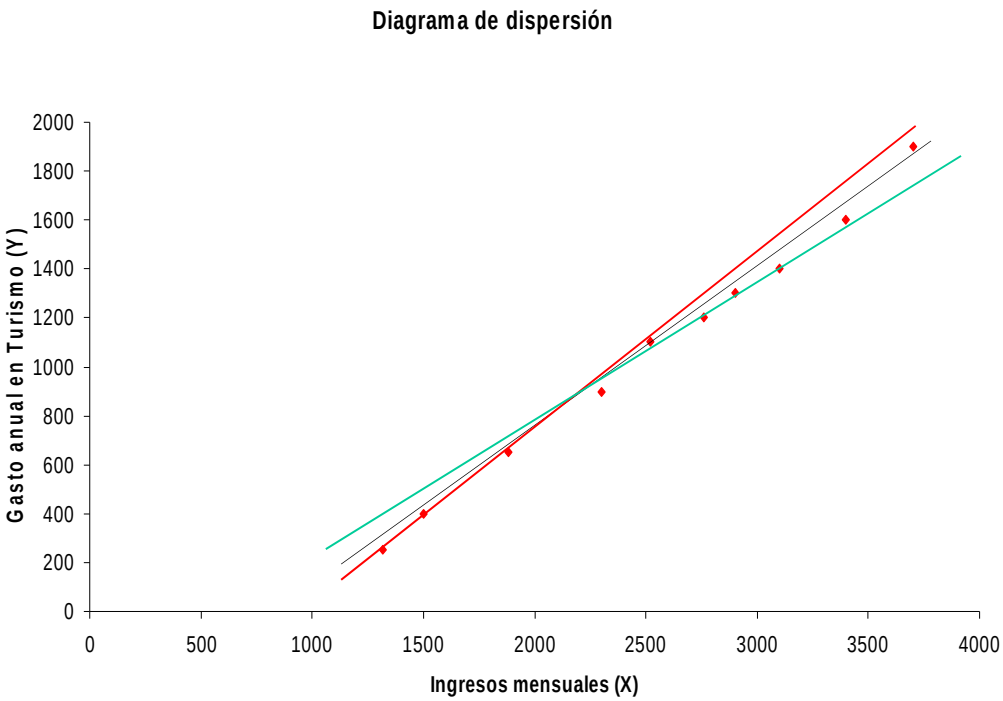
El **método minimocuadrático** permite determinar los valores de los coeficientes a y b de la recta de regresión:

$$b = \frac{S_{XY}}{S_X^2} \qquad a = \bar{Y} - b\,\bar{X}$$

¿Interpretación de los coeficientes a y b de la recta de regresión? lo veremos con un ejemplo.

•EJEMPLO:

Hogar	Ingresos mensuales en € (X)	Gasto anual en Turismo en € (Y)
1	1880	650
2	2300	900
3	3700	1900
4	2760	1200
5	3400	1600
6	2900	1300
7	1320	250
8	1500	400
9	2520	1100
10	3100	1400



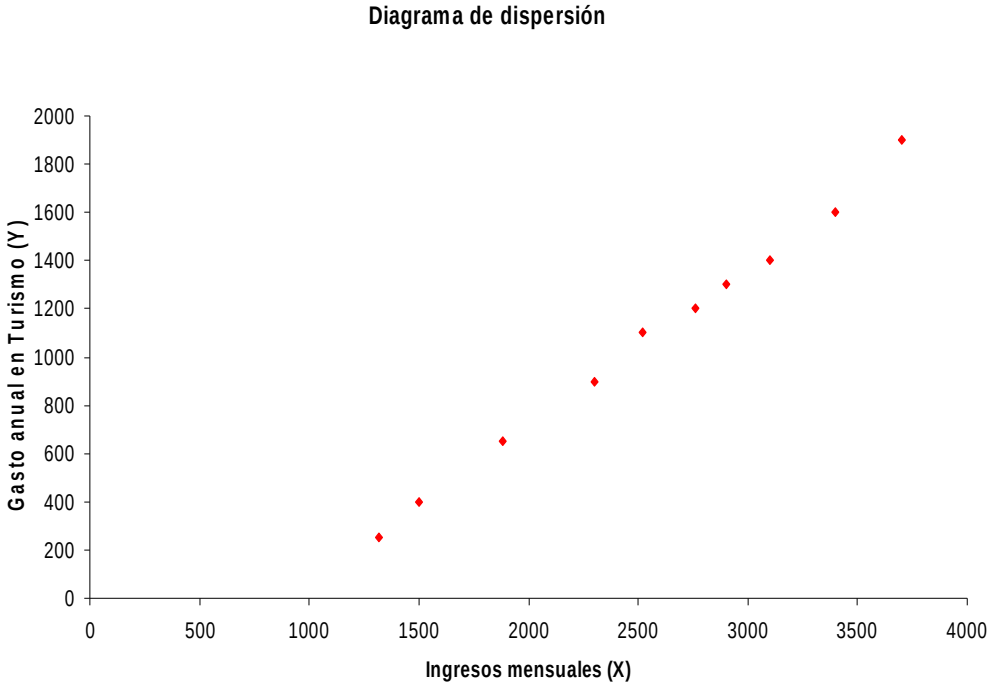
•EJEMPLO :

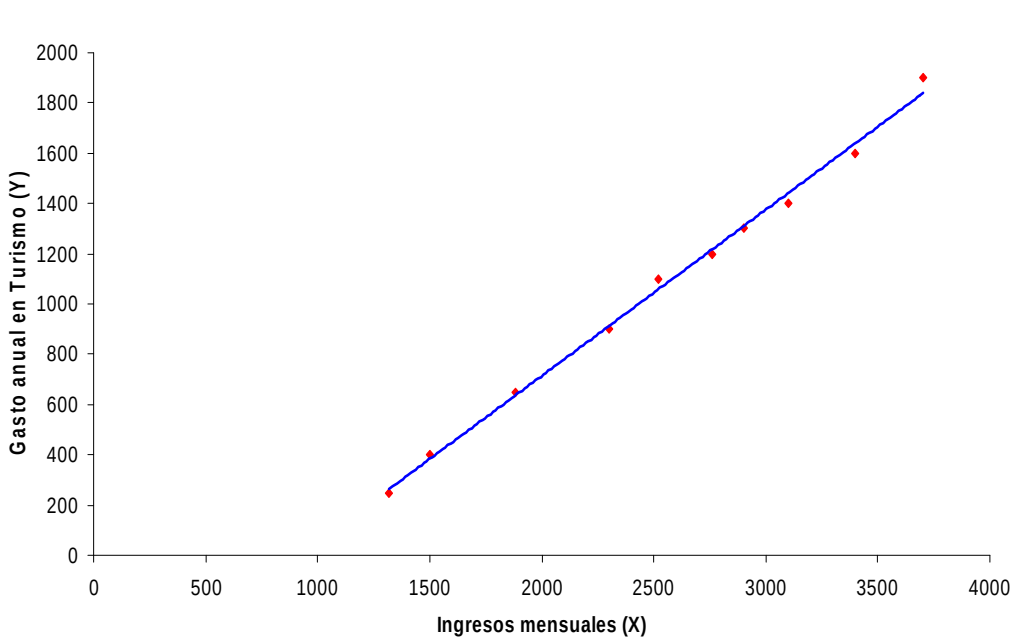
Hogar	Ingresos mensuales en € (X)	Gasto anual en Turismo en € (Y)
1	1880	650
2	2300	900
3	3700	1900
4	2760	1200
5	3400	1600
6	2900	1300
7	1320	250
8	1500	400
9	2520	1100
10	3100	1400

Vector de Medias: $\begin{pmatrix} \overline{X} = 2538 \\ \overline{Y} = 1070 \end{pmatrix}$

Matriz de Varianzas-Covarianzas:

$$\begin{pmatrix} S_X^2 = 564036 & S_{XY} = 372940 \\ S_{XY} = 372940 & S_Y^2 = 247600 \end{pmatrix}$$





Matriz de Varianzas-Covarianzas:

Vector de Medias: $\begin{pmatrix} \bar{X} = 2538 \\ \bar{Y} = 1070 \end{pmatrix}$

$$\begin{pmatrix} S_X^2 = 564036 & S_{XY} = 372940 \\ S_{XY} = 372940 & S_Y^2 = 247600 \end{pmatrix}$$

$$b = \frac{S_{XY}}{S_X^2} = \frac{372940}{564036} = 0{,}661$$

$$a = \bar{Y} - b\,\bar{X} = 1070 - 0{,}661 \times 2538 = -607{,}618$$

El Modelo de Regresión lineal de Y/X es:

$$Y^* = -607{,}618 + 0{,}661\,X$$

- ¿Qué es el coeficiente a?

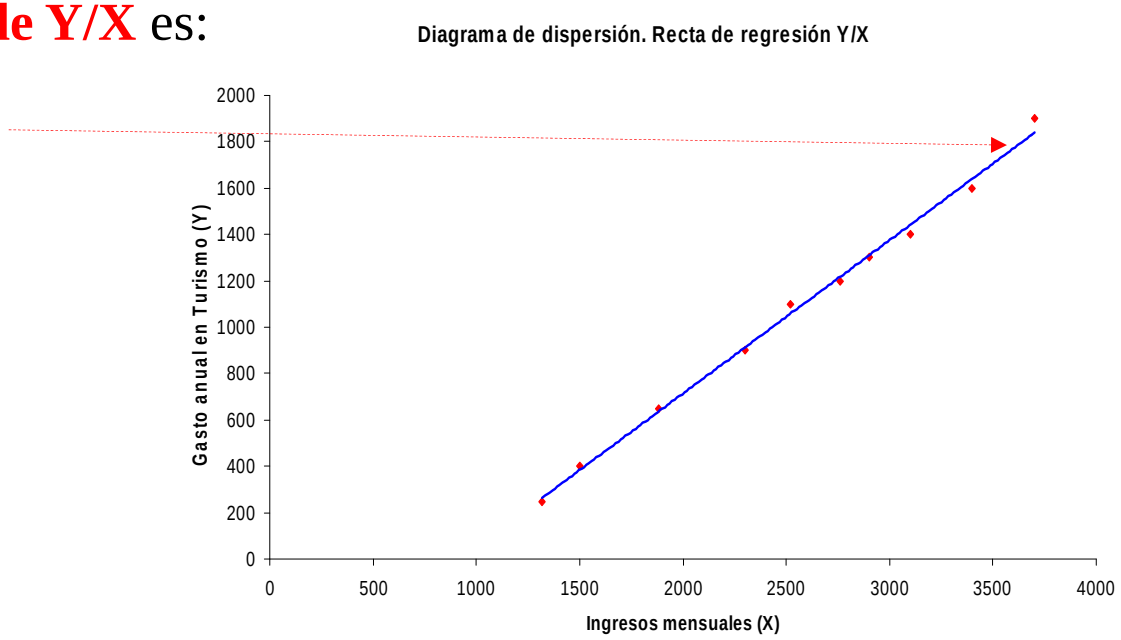
Si $X=0 \Rightarrow Y^* = -607{,}618$

¿Tiene sentido económico?

- ¿Qué es el coeficiente b?

La pendiente de la recta

¿Qué significa?



Aparte de X y de Y, se crean dos nuevas variables:

• **Y*** : La Y teórica o explicada. Son los valores estimados de Y que nos proporciona el modelo de regresión lineal. La parte de los valores de Y que podemos explicar a partir de los valores de X.

$$Y^* = a + b\,X = -607{,}618 + 0{,}661\,X$$

• **e** : La variable error o residuo. Son los errores que se cometen al ajustar la recta de regresión. Lo que no explica el modelo de regresión.

$$e = Y - Y^*$$

Hogar	Ingresos mensuales en € (X)	Gasto anual en Turismo en € (Y)	Y* teórica Y*=a+bX	error e=Y-Y*
1	1880	650	635.1	14.9
2	2300	900	912.7	-12.7
3	3700	1900	1838.1	61.9
4	2760	1200	1216.7	-16.7
5	3400	1600	1639.8	-39.8
6	2900	1300	1309.3	-9.3
7	1320	250	264.9	-14.9
8	1500	400	383.9	16.1
9	2520	1100	1058.1	41.9
10	3100	1400	1441.5	-41.5
	∑ suma	10700	10700	0

4.- ANÁLISIS DE LA BONDAD DEL AJUSTE Y PREDICCIÓN

	Y observada	Y* teórica	e error
Media	$\bar{Y}$	$\bar{Y}^* = \bar{Y}$	$\bar{e} = 0$
Varianza	S_Y^2	$S_{Y^*}^2$	S_e^2

Varianza explicada

Varianza residual

Relación entre las 3 varianzas: $S_Y^2 = S_{Y^*}^2 + S_e^2$

Coeficiente de determinación: $R^2 = \dfrac{S_{Y^*}^2}{S_Y^2}$ $\qquad 0 \le R^2 \le 1$

- R^2 es la parte de la varianza de Y que explica el modelo de regresión.
- 1- R^2 es la parte de la varianza de Y que no explica el modelo, que se debe a los errores que se cometen.

Propiedad de la regresión lineal: $R^2 = \dfrac{S_{XY}^2}{S_X^2 \, S_Y^2} = r_{XY}^2$

Resultados de la CaEst 1.5:

Indicadores	Y	X	
Media	1070	2538	
Varianzas y covarianza	**247600**	564036	372940
Desv.Típica	497.594	751.023	

REGRESIÓN

C.Correlación	0.998
C.Determinación	0.996
Varianza Explicada	**246609.6**
Varianza Residual	**990.4**
Coeficiente a	-607.618
Coeficiente b	0.661
RECTA	Y*= -607.618 + 0.661X

Regresión lineal del Ejemplo de la diapositiva 4

A partir de la información de 2009 para las N=17 Comunidades Autónomas sobre las siguientes variables:

•Nº de Pernoctaciones, en miles.

•Gasto total de los turistas, en millones de euros.

Se ha calculado, con ayuda de la CaEst (ver diapositiva 8), las siguientes características de la distribución bidimensional de ambas variables:

Vector de Medias: $\begin{pmatrix} 15749 \\ 2838 \end{pmatrix}$ Matriz de Varianzas-Covarianzas: $\begin{pmatrix} 324983248 & 63323918 \\ 63323918 & 12704865 \end{pmatrix}$

a) Si se desea realizar una regresión lineal de una variable en función de la otra, suponiendo una relación de causa-efecto, ¿qué variable tendría más sentido que fuera la dependiente (la Y) y qué variable la independiente (la X)? *Sol: Y: Gasto Turistas X: Pernoctaciones*

b) Obtén los coeficientes de la recta de regresión minimocuadrática de Y respecto a X e interpreta el valor de la pendiente. *Sol: $Y^* = -233{,}1 + 0{,}195\,X$*

c) Calcula e interpreta una medida de la bondad del ajuste efectuado. *Sol: $R^2 = 0{,}97$*

d) ¿Qué porcentaje de la variación de la variable Y no puede explicarse a partir del modelo de regresión ajustado? *Sol: 3%*

e) Si para el siguiente año una CCAA piensa que el nº de pernoctaciones, en miles, será de 25000, ¿cuál sería el correspondiente gasto total de los turistas? ¿Es fiable este resultado?
Sol: 4641,9 millones de €. Fiabilidad del 97% ceteris paribus

Más información sobre este tema en:

- PARRA, E; CALERO, F.J.: Estadística para Turismo. Ed. McGraw-Hill, Madrid, 2007. Capítulo 7.
- ESTEBAN, J.; y otros.: "Estadística Descriptiva y nociones de Probabilidad", Ed. Thomson, segunda impresión 2006. Capítulos 3 y 4.
- MONTIEL, A.M.; RIUS, F.; BARÓN F.J.: *Elementos básicos de Estadística Económica y Empresarial*. Ed. Prentice Hall, Madrid, 1997. Capítulos 5 y 6.
- RONQUILLO, A: Estadística Aplicada al Sector Turístico, Ed Ramón Areces, Madrid, 1997. Capítulo 6.
- http://www.uv.es/ceaces/descriptiva/simplem.htm
  http://www.uv.es/ceaces/base/regresion/simple.htm
- http://webpersonal.uma.es/de/J_SANCHEZ/Capitulo3.PDF