Estadística

Universidad Europea de Madrid

Laureate International Universities

Estadística descriptiva bivariante

- Vamos a medir dos características (variables) sobre cada individuo.
- Las variables pueden ser cuantitativas o cualitativas combinadas de todas las formas posibles (cuantitativa- cuantitativa, cuantitativa-cualitativa, continua-discreta,etc.)
- Una muestra de la población consistirá en pares ordenados (x,y) de ambas características observadas sobre cada individuo.
- Llamaremos distribución conjunta de frecuencias de dos variables (X e Y) a la tabla que representa los valores observados $\longrightarrow$ frecuencias absolutas o relativas de cada par.

Universidad Europea de Madrid

Descripción de los datos

Tabla bivariante

Si para cada individuo tenemos dos datos, podemos construir una tabla de doble entrada.

EJEMPLO

Se han estudiado el peso (X) y la altura (Y) de 70 individuos obteniéndose los datos de la siguiente tabla (cuantitativa-cuantitativa)

Descripción de los datos

Pesos / Tallas	159 - 161	161 - 163	163 - 165	165 - 167	167 - 169	169 - 171
48	3	2	2	1	0	0
51	2	3	4	2	2	1
54	1	3	6	8	5	1
57	0	0	1	2	8	3
60	0	0	0	2	4	4

Descripción de los datos

Pesos / Tallas	159 – 161	161 – 163	163 – 165	165 – 167	167 – 169	169 – 171
48	3	2	2	1	0	0
51	2	3	4	2	2	1
54	1	3	6	8	5	1
57	0	0	1	2	8	3
60	0	0	0	2	4	4

En cada celda tenemos la frecuencia conjunta, F_{ij} , es decir, el número de individuos que presentan simultaneamente las características x_i e y_j

Universidad Europea de Madrid

Cuestiones

- Construir la tabla bivariante anterior con las frecuencias relativas
- En una tabla de doble entrada, la suma de todas las frecuencias relativas es ....
- En una tabla de doble entrada, la suma de todas las frecuencias absolutas es ....

Universidad Europea de Madrid

Descripción de los datos

Pesos / Tallas	159 – 161	161 – 163	163 – 165	165 – 167	167 – 169	169 – 171	Total Filas
48	0,04	0,03	0,03	0,01	0,00	0,00	0,11
51	0,03	0,04	0,06	0,03	0,03	0,01	0,20
54	0,01	0,04	0,09	0,11	0,07	0,01	0,34
57	0,00	0,00	0,01	0,03	0,11	0,04	0,20
60	0,00	0,00	0,00	0,03	0,06	0,06	0,14
Total columnas	0,09	0,11	0,19	0,21	0,27	0,13	1,00

Descripción de los datos

Distribuciones marginales

- Se obtienen al estudiar una de las variables de forma independiente de la otra.
- La distribución de la variable X se calcula sumando, para cada fila y sobre todas las columnas, las frecuencias conjuntas.
- La distribución de la variable Y se calcula sumando, para cada columna y sobre todas las filas, las frecuencias conjuntas.

Descripción de los datos

Pesos / Tallas	159 – 161	161 – 163	163 – 165	165 – 167	167 – 169	169 – 171	Total Filas
48	3	2	2	1	0	0	8
51	2	3	4	2	2	1	14
54	1	3	6	8	5	1	24
57	0	0	1	2	8	3	14
60	0	0	0	2	4	4	10
Total columnas	6	8	13	15	19	9	70

Frecuencias marginales F_Y F_X

Descripción de los datos

Pesos	
48	8
51	14
54	24
57	14
60	10
Total	70

Cuestiones

- Construir la tabla de frecuencias de la variable PESO
- En una tabla de doble entrada, la suma de las frecuencias marginales relativas de una de las variables es....
- La suma de las frecuencias marginales absolutas de una de las variables es...

Descripción de los datos

Distribuciones condicionadas

Se obtienen al estudiar un conjunto más pequeño de los datos formado por aquellos que tienen, por ejemplo, la característica x_i, $i = 1, \dots, r$, o la carecterística y_j, $j = 1, \dots, s$

Descripción de los datos

Distribuciones condicionadas

- Si consideramos los datos que tienen la característica y_j , la variable X definida sobre este conjunto se denomina **variable condicionada** y se suele denotar mediante $X/Y = y_j$

Universidad Europea de Madrid

Descripción de los datos

Pesos / Tallas	159 – 161	161 – 163	163 – 165	165 – 167	167 – 169	169 – 171	Total Filas
48	3	2	2	1	0	0	8
51	2	3	4	2	2	1	14
54	1	3	6	8	5	1	24
57	0	0	1	2	8	3	14
60	0	0	0	2	4	4	10
Total columnas	6	8	13	15	19	9	70

Frecuencias absolutas condicionadas
(al valor de la fila o la columna)

Universidad Europea de Madrid

Descripción de los datos

Frecuencias relativas de la TALLA condicionada al PESO=54

Y|X=54	$f(y_j \mid X = 54)$
159 – 161	1/24
161 – 163	3/24
163 – 165	6/24
165 – 167	8/24
167 – 169	5/24
169 – 171	1/24
Total	1

Descripción de los datos

Independencia entre variables

- El gran interés de analizar dos variables conjuntamente es conocer si existe o no relaciones entre ellas.
- Los dos casos extremos en la relación entre dos variables son:
  - *La ausencia de relación:* ***Independencia****.*

    El conocimiento de una variable no permite conocer nada sobre la otra variable.
  - *El caso de* ***dependencia funcional Y=f(X)****.*

    Y depende funcionalmente de X, si el conocimiento de X permite conocer, de forma exacta, los valores que toma Y.

Descripción de los datos

Independencia entre variables

- Entre estos dos casos extremos anteriores, existen un tipo de relaciones, que son las que nos van a interesar estudiar en muchos casos:
  - Sabemos que dos variables están relacionadas, pero no existe una dependencia funcional exacta.

- Este es el caso de ***dependencia estadística***, *en el que se puede* describir, aproximadamente, el comportamiento de una variable a partir de otra u otras variables $Y \approx f(x)$.

Descripción de los datos

Independencia entre variables

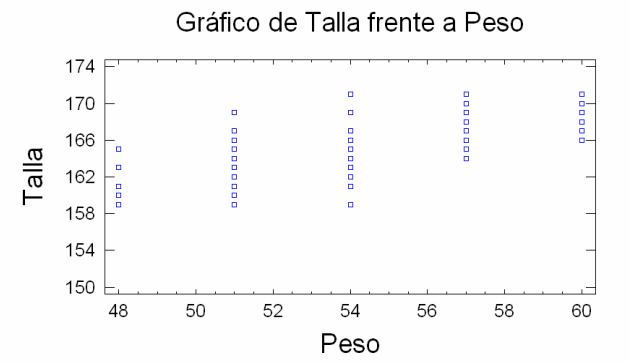
- Toda la información sobre la relación entre dos variables la provee la función de distribución conjunta. Además, basada en dicha distribución, existen una serie de medidas que nos van a servir para analizar esta relación.

Universidad Europea de Madrid

Descripción gráfica de los datos

La representación más útil para mostrar la relación entre dos variables continuas sin agrupar es el diagrama de dispersión.

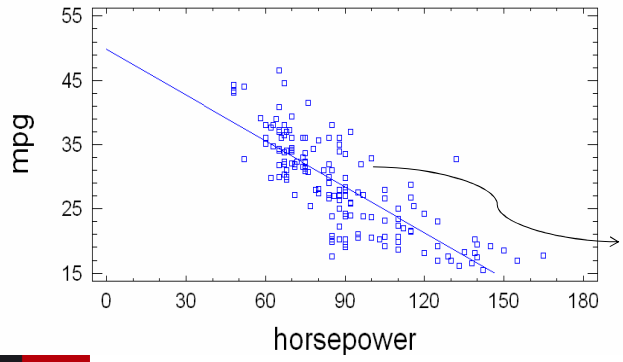
Cada par (x_i, y_j) se representa como un punto del plano cartesiano



Universidad Europea de Madrid

Descripción gráfica de los datos

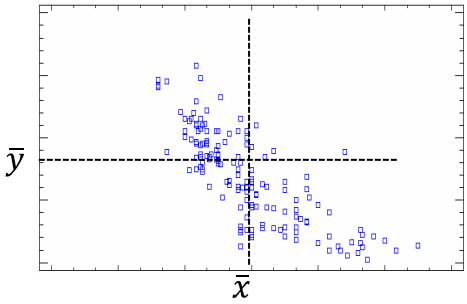
Del fichero cardata.sf, analizamos los caballos de fuerza (horsepower) de 155 coches frente al gasto de gasolina (mpg = milla por galón)



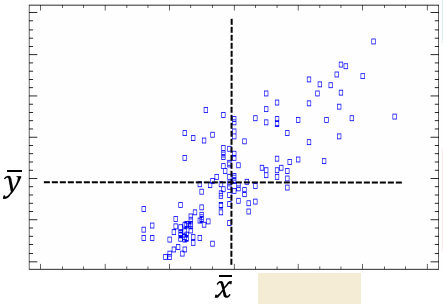
La millas recorridas parecen reducirse al aumentar el número de caballos

Universidad Europea de Madrid

Descripción gráfica de los datos



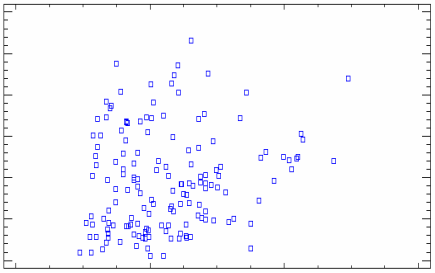
Relación lineal negativa



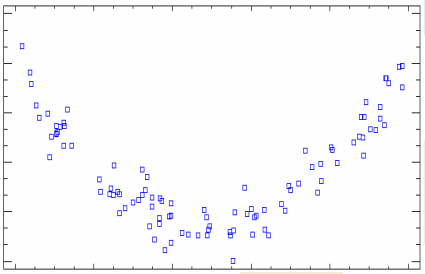
Relación lineal positiva

Universidad Europea de Madrid

Descripción gráfica de los datos



Ausencia de relación



Relación no lineal

Universidad Europea de Madrid

Medidas de dependencia lineal

Las dos medidas más utilizadas para cuantificar el grado y sentido de la dependencia **lineal** son:

- Covarianza
- Correlación

Medidas de dependencia lineal

Covarianza

- Nos indica si la relación entre las variables es positiva o negativa.
- Su magnitud depende de las unidades
- Si las variables son estadísticamente independientes entonces la covarianza es nula (el recíproco no es válido en general)

Medidas de dependencia lineal

Covarianza

- Para datos agrupados

$$S_{xy} = \sum_i \sum_j f_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \left(\sum_i \sum_j f_{ij} x_i y_j \right) - \bar{x}\bar{y}$$

- Para datos no agrupados

$$S_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \left(\sum_{i=1}^{n} x_i y_i \right) - \bar{x}\bar{y}$$

Universidad Europea de Madrid

Medidas de dependencia lineal

Correlación

- Mide la magnitud y la dirección de la dependencia lineal.
- Es adimensional.

Universidad Europea de Madrid

Medidas de dependencia lineal

Correlación

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

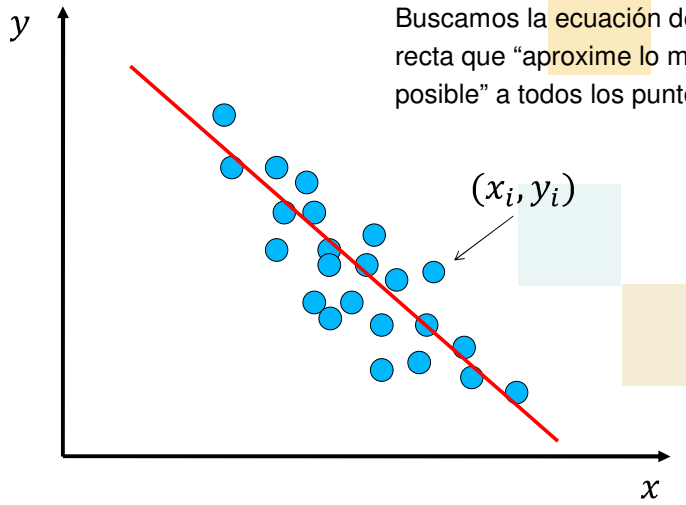
- Tiene el mismo signo que la covarianza.
- No mide las relaciones no-lineales.

Medidas de dependencia lineal

Correlación

- $-1 \leq r_{xy} \leq 1$.
- Decimos que las variables son incorreladas $\Leftrightarrow r = 0$
- Hay relación lineal perfecta $\Leftrightarrow r = 1 \; o \; r = -1$
- Cuanto más cerca esté de 1 o -1 mejor será el grado de relación lineal.

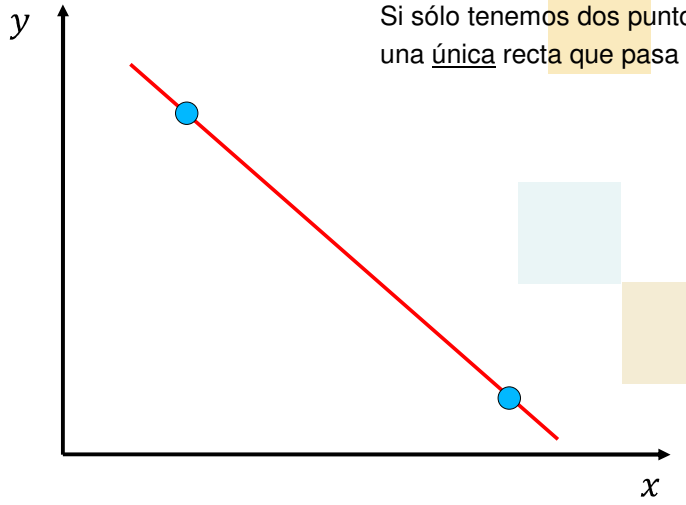
Recta de regresión



Buscamos la ecuación de una recta que "aproxime lo mejor posible" a todos los puntos

Universidad Europea de Madrid

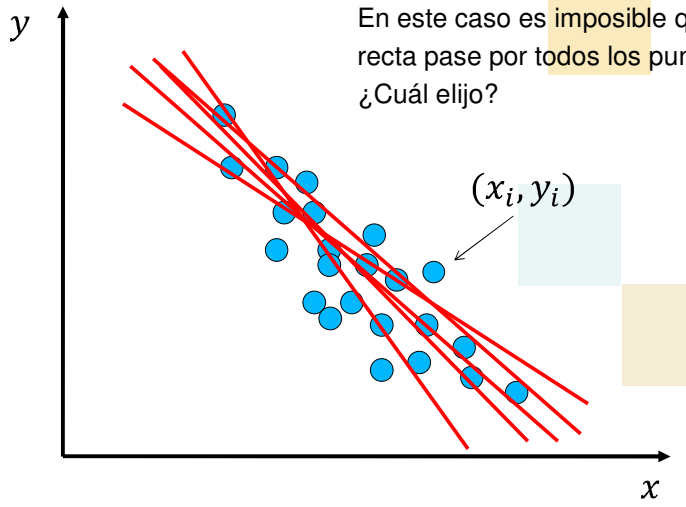
Recta de regresión



Si sólo tenemos dos puntos, hay una única recta que pasa por ellos

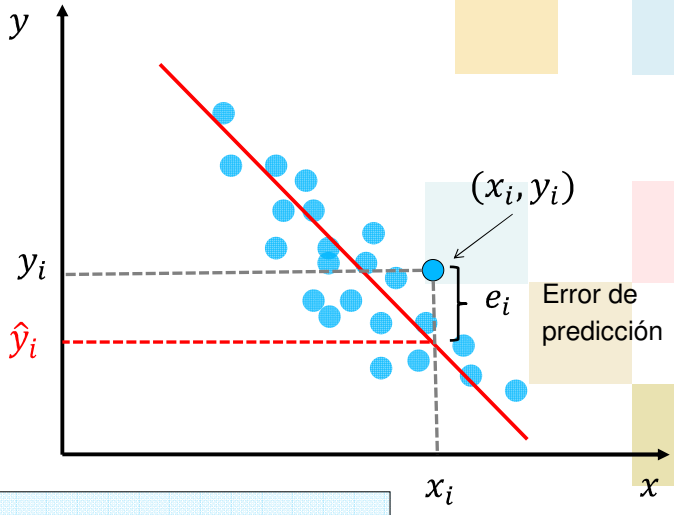
Universidad Europea de Madrid

Recta de regresión



En este caso es imposible que una recta pase por todos los puntos.
¿Cuál elijo?

Recta de regresión



Valor observado

Valor predicho por la recta

Error de predicción

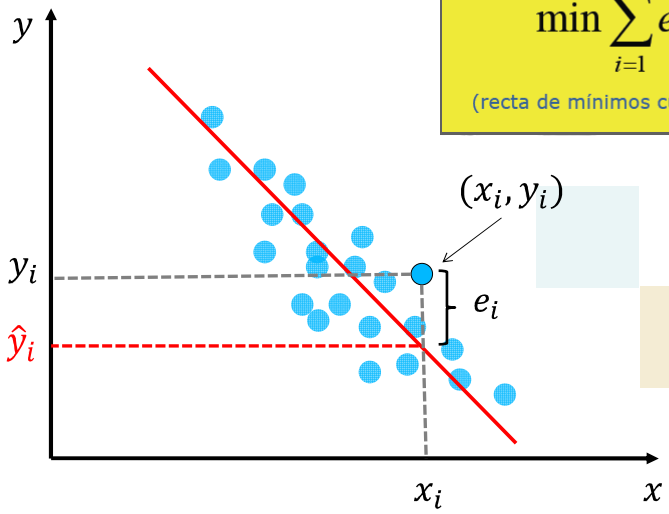
Error de predicción = $y_i - \hat{y}_i$

Recta de regresión

Buscamos la recta que minimiza los errores de predicción:

$$\min \sum_{i=1}^{N} e_i^2$$

(recta de mínimos cuadrados)

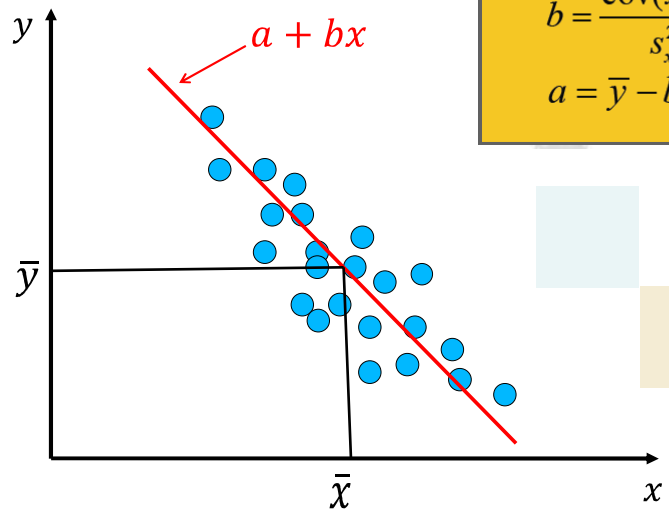


Recta de regresión

SOLUCIÓN

$$b = \frac{\text{cov}(x, y)}{s_x^2}$$

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



Observaciones

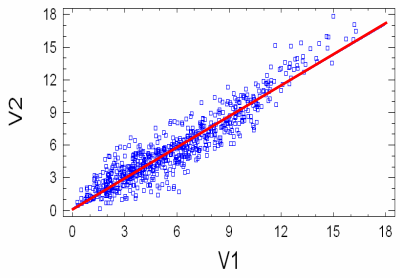
- La recta de regresión siempre pasa por el punto $(\bar{x}, \bar{y})$
- La pendiente de la recta es $\frac{\sigma_{xy}}{\sigma_x^2}$. Por lo que el signo de la covarianza marcará la pendiente de la recta.
- Si la pendiente es positiva (si la covarianza es positiva), valores grandes de la variable X, se relacionarán con valores grandes de la variable Y (dependencia positiva). Y si es negativa valores grandes de la variable X se relacionarán con valores pequeños de Y (dependencia negativa).

Universidad Europea de Madrid

Recta de regresión

Ejemplo

La variable V1 tiene la velocidad del viento registrada en la localización 1, mientras que la variable V2 tiene las velocidades registradas en esos mismos instantes en la localización 2. Se tiene un total de 115 pares de medidas.



Loc. 1
media: 2,51
varianza: 1,91

Loc. 2
media: 3,28
varianza: 2,36

cov(V1,V2) = 1,995

En la localización 1 se va a establecer un sistema informático para la telemedida de la velocidad del viento, pero no para la localización 2. Se quiere calcular la recta de regresión que permita predecir la velocidad de la Localización 2 sabiendo la de la Localización 1.

Universidad Europea de Madrid

Recta de regresión

Ejemplo

$$b = \frac{cov(X,Y)}{S_x^2} = \frac{1{,}995}{1{,}91} = 1{,}045$$

$$a = \bar{x} - b\bar{y} = 3{,}28 - 1{,}045 \times 2{,}51 = 0{,}657$$

$$\Rightarrow \quad \hat{V}_2 = 0{,}657 + 1{,}045 \times V_1$$

Si, por ejemplo, en la Localización 1 se mide una velocidad de viento de 5 m/s, la predicción en la Localización 2 es de un viento de **0.657+1.045x5=5.88 m/s**

Universidad Europea de Madrid

Ejemplo

- Una multinacional dedicada a la venta de gafas quiere averiguar si existe relación lineal entre el dinero que gasta en anuncios de televisión y sus ventas locales (en miles de euros). Para ello hizo un estudio durante siete meses, obteniendo los siguientes datos

MES	VENTAS	GASTOS
ENERO	50	0.5
FEBRERO	90	0.75
MARZO	30	0.4
ABRIL	90	0.7
MAYO	91	0.8
JUNIO	95	0.9
JULIO	95	0.95

Universidad Europea de Madrid

Ejemplo

- Averiguar si existe relación lineal entre las variables.
- Calcular la recta de regresión.
- ¿ Cuánto tendrá que gastarse en anuncios para obtener unas ventas de 97.000 euros?

Universidad Europea de Madrid