Tema 2: Estadística Descriptiva

Descripción conjunta de variables

Departamento Matemática Aplicada

Universidad de Málaga

Curso 2015-2016

Descripción conjunta de varias variables

Consideremos el estudio conjunto de dos caracteres de la población, aunque los métodos descritos resultan fácilmente generalizables a un mayor número de variables. Sea

- X variable con modalidades $x_1,\ x_2,\ \ldots$
- Y la variable con modalidades $y_1,\ y_2,\ \ldots.$

Una muestra de la variable bidimensional $(X,\ Y)$ está formada por los distintos valores $(x_i,\ y_j)$ que se pueden obtener al observar conjuntamente las dos variables.

- La frecuencia absoluta n_{ij} indica el número de veces que se repite el par de valores $(x_i,\ y_j)$.
- La frecuencia relativa f_{ij} indica la proporción de veces que se repite la pareja de valores $(x_i,\ y_j)$ sobre el total de datos de la muestra.

Representaciones

Si el número de observaciones es pequeño, podemos representar las variables en forma de **tabla simple**.

variable X	x_1	x_2	...	x_N
variable Y	y_1	y_2	...	y_N

Ejemplo

Se prueban cinco trozos experimentales de un material aislante bajo diferentes presiones. A continuación se presentan los valores (P) de presión (en Kg/cm^2) y las magnitudes (C) de compresión resultantes (en mm): (1,1), (2,1), (3,2), (4,2) y (5,4). Representar la distribución de frecuencias.

Se construye una tabla simple de valores con los pares de datos de la muestra.

P	1	2	3	4	5
C	1	1	2	2	4

Representación tabular simple

Si el número de observaciones es grande, pero tenemos pocas modalidades; podemos usar una **tabla simple con 3 filas o columnas** conteniendo las parejas de valores y sus frecuencias correspondientes.

variable X	variable Y	frecuencia absoluta	frecuencia relativa
x_1	y_1	n_1	f_1
x_2	y_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	y_i	n_i	f_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	y_k	n_k	f_k
		N	1

Ejemplo representación tabular simple

Ejemplo

Una empresa de software somete a sus programas a un proceso para depurar errores. El número de controles efectuados disminuye los posibles errores finales pero incrementa los costes de producción. Se observan conjuntamente el número de controles C efectuados y el número de errores graves E detectados al finalizar su desarrollo, obteniéndose la muestra: (0,0), (1,1), (1,1), (0,1), (1,1), (0,1), (1,1), (1,1), (1,0), (1,0), (1,1), (0,1), (1,1), (0,0), (1,0), (1,0), (2,0), (0,1), (1,1), (2,0). Crear una tabla estadística para representar la distribución de frecuencias.

C	E	n_i	f_i
0	0	2	0.1
0	1	4	0.2
1	0	4	0.2
1	1	8	0.4
2	0	2	0.1
		20	1

Tabla bidimensional

Si el número de observaciones y de modalidades es grande, utilizaremos una tabla de doble entrada, representando la frecuencia absoluta de una pareja $(x_i,\ y_j)$ en la casilla de cruce de cada fila y columna **(distribución conjunta)**

$x \backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_p	
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2p}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kp}	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot p}$	N

$$\begin{cases} n_{i\cdot} = \sum_j n_{ij} \\ n_{\cdot j} = \sum_i n_{ij} \\ N = \sum_i \sum_j n_{ij} \\ N = \sum_i n_{i\cdot} = \sum_j n_{\cdot j} \end{cases}$$

Distribuciones marginales: Son las frecuencias $(n_{i\cdot})$ de los valores de la variable X (sumando por filas) y las frecuencias $(n_{\cdot j})$ de los valores de la variable Y (sumando por columnas).

Ejemplo tabla bidimensional

Ejemplo

Representar en tablas de doble entrada las distribuciones de frecuencias absolutas y relativas para los datos del ejemplo anterior.

C	E	n_i	f_i
0	0	2	0.1
0	1	4	0.2
1	0	4	0.2
1	1	8	0.4
2	0	2	0.1
		20	1

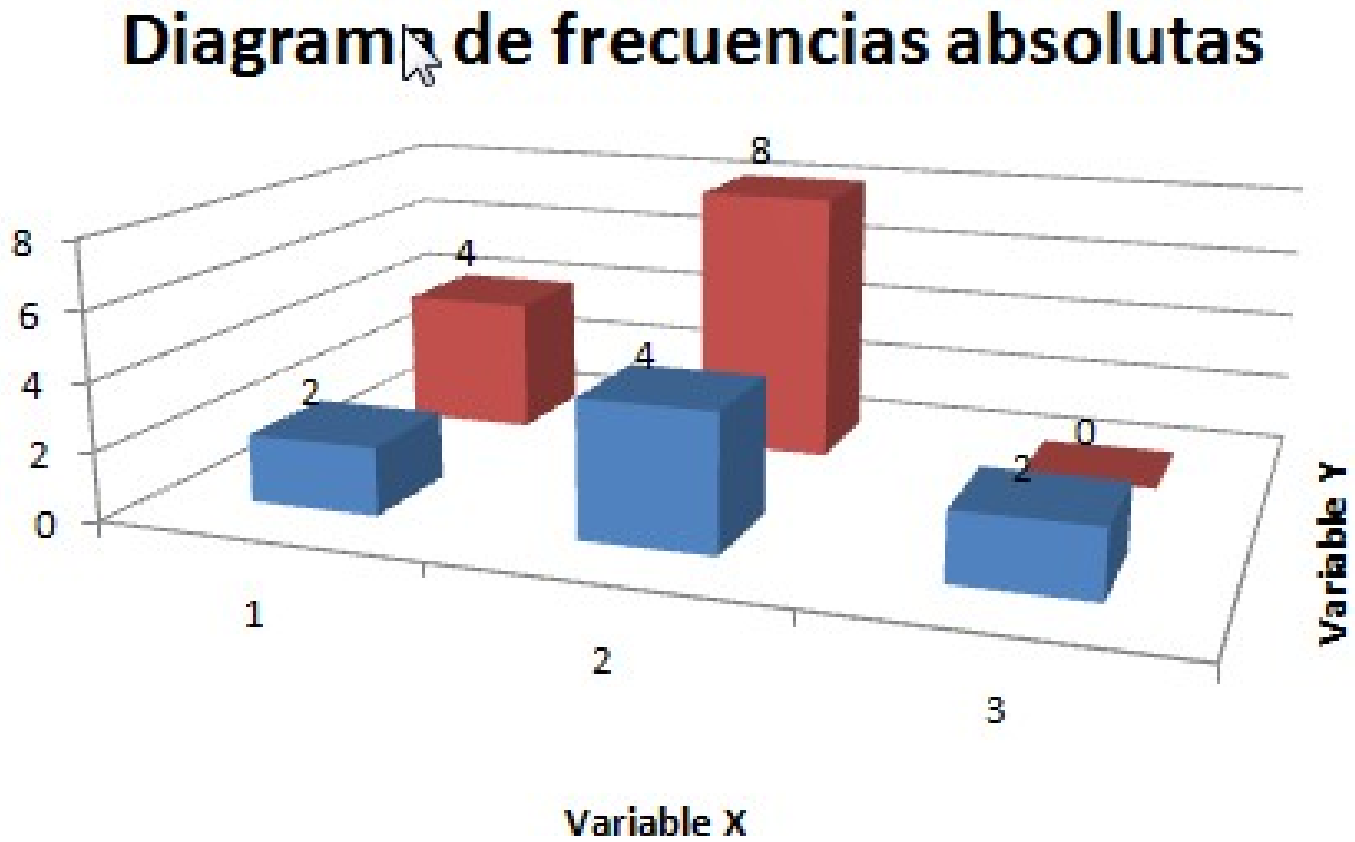
$\Longrightarrow$

n_{ij}	0	1	E
0	2	4	6
1	4	8	12
2	2	0	2
C	8	12	20

f_{ij}	0	1	E
0	0.1	0.2	0.3
1	0.2	0.4	0.6
2	0.1	0	0.1
C	0.4	0.6	1

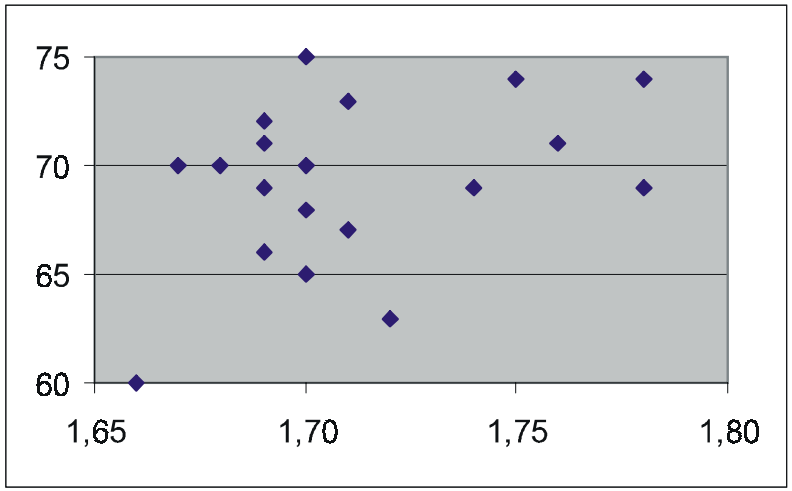
Representaciones gráficas

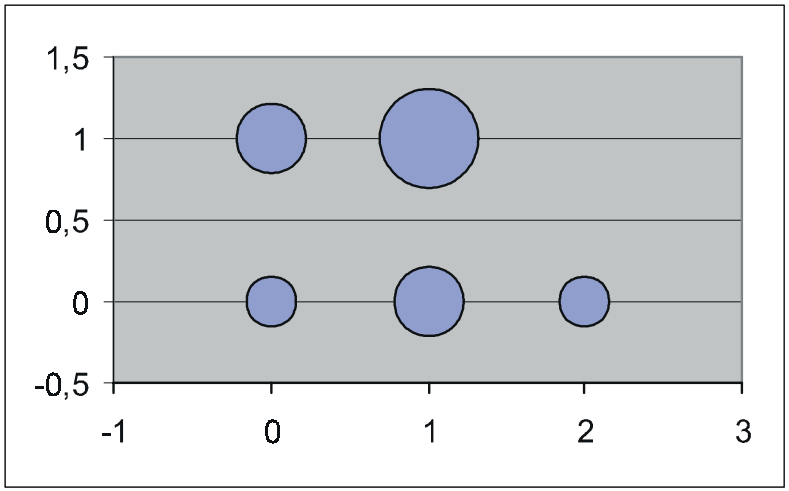
Diagrama de frecuencias. Caso discreto. Similar al diagrama de barras unidimensional. Es una representación tridimensional en la que el plano base representa los valores de las variables y la altura las frecuencias.



Representaciones gráficas-2

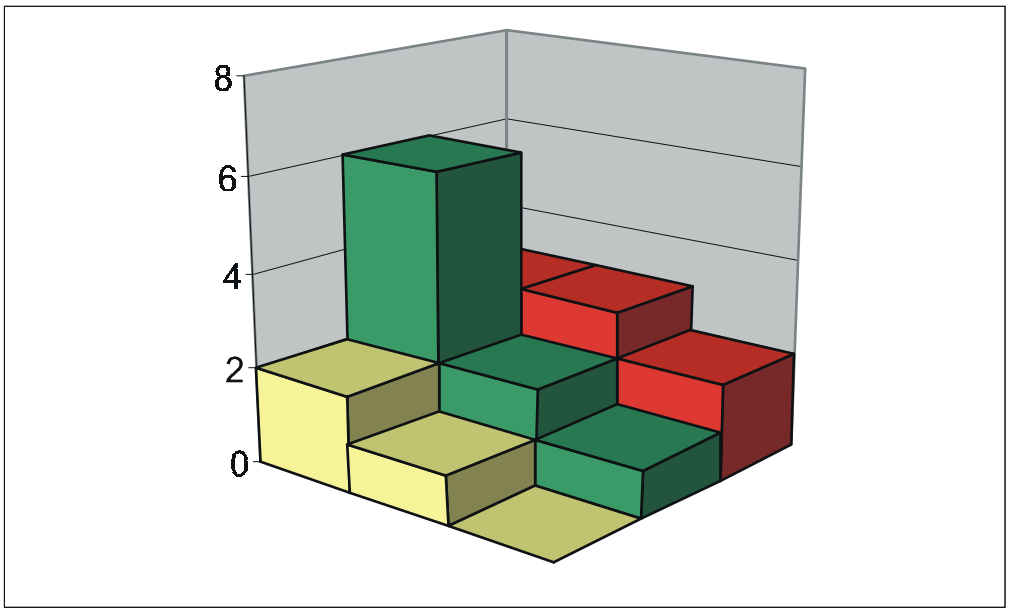
Diagrama de dispersión. Representamos los distintos pares de valores sobre unos ejes cartesianos, obteniéndose una nube de puntos. La frecuencia de cada par de puntos se puede representar usando distintos tamaños de puntos.





Representaciones gráficas-3

Estereograma. Cuando los datos de ambas variables se agrupan en intervalos. Se usa como base las regiones del plano correspondientes a los intervalos y la frecuencia queda representada por el volumen de un paralelepípedo, luego $h_{ij} = \frac{n_{ij}}{S_{ij}}$, donde S_{ij} es el área de la modalidad $(x_i,\ y_j)$.



Frecuencias Marginales

Se obtienen al estudiar una variable con independencia de la otra. Su nombre se debe a que la distribución se obtiene sumando en los márgenes de la tabla de la distribución conjunta.
Si se desea estudiar una de las variables de forma aislada, se tiene que separar la información relativa a dicha variable. Si X tiene modalidades $x_1,\ x_2,\ \ldots,\ x_k$ e Y modalidades $y_1,\ y_2,\ \ldots,\ y_p$ se obtienen las frecuencias marginales:

$$n_{i\cdot} = \sum_{j=1}^{p} n_{ij} \qquad f_{i\cdot} = \frac{n_{i\cdot}}{N}$$

$$n_{\cdot j} = \sum_{i=1}^{k} n_{ij} \qquad f_{\cdot j} = \frac{n_{\cdot j}}{N}$$

Distribuciones Marginales

Ejemplo

Calcular la distribución marginal de la variable C (número de controles efectuados a un software) del ejemplo anterior.

Eliminar la columna correspondiente a la variable E y agrupar las modalidades que sean iguales.

C	E	n_i	f_i
0	0	2	0.1
0	1	4	0.2
1	0	4	0.2
1	1	8	0.4
2	0	2	0.1
		20	1

$\Longrightarrow$

C	n_i	f_i
0	6	0.3
1	12	0.6
2	2	0.1
	20	1

Distribución unidimensional de la variable C

Ejemplo distribuciones Marginales

Dada la tabla bidimensional de frecuencias absolutas:

$X \backslash Y$	-2	-1	0	1	2	
0	2	7	12	10	4	35
1	5	14	23	15	7	64
2	12	31	23	8	3	77
3	20	18	8	2	1	49
	39	70	66	35	15	**225**

la marginal de X será:

X	$n_{i\cdot}$
0	35
1	64
2	77
3	49
	225

y la de Y:

Y	$n_{\cdot j}$
-2	39
-1	70
0	66
1	35
2	15
	225

Distribuciones Condicionadas

Surgen al considerar sólo aquellos valores de la muestra que presentan una determinada modalidad (o condición) en una de las variables.

Se llama distribución condicionada del carácter X respecto a la clase j del carácter Y, y se denota X/y_j, a la distribución unidimensional de la variable X cuando **sólo se consideran los individuos de la clase j de Y**.

$$n_i^j = n_{ij} \qquad \text{y} \qquad f_i^j = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}} \qquad i = 1, 2, ..., k$$

Análogamente se puede definir la distribución condicionada del carácter Y respecto a la modalidad i de X, denotada por Y/x_i.

$$n_j^i = n_{ij} \qquad \text{y} \qquad f_j^i = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}} \qquad j = 1, 2, ..., p$$

Ejemplo distribuciones Condicionadas

Dada la tabla bidimensional de frecuencias absolutas:

$X \backslash Y$	-2	-1	0	1	2	
0	2	7	12	10	4	35
1	5	14	23	15	7	64
2	12	31	23	8	3	77
3	20	18	8	2	1	49
	39	70	66	35	15	**225**

La distribución de Y condicionada a que X=2 será:

Y	n_j^3
-2	12
-1	31
0	23
1	8
2	3
	77

Y la distribución de X condicionada a que Y=-1 será:

X	n_i^2
0	7
1	14
2	31
3	18
	70

Momentos

Consideramos los datos agrupados en una tabla bidimensional.

Definición

Llamamos **momento de orden** $(\mathbf{r},\ \mathbf{s})$ **respecto al punto** $(\mathbf{a},\ \mathbf{b})$ *a:*

$$\mathbf{M_{rs}(a,\ b)} = \sum_{\mathbf{i=1}}^{\mathbf{k}} \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{(x_i - a)^r (y_j - b)^s f_{ij}}$$

Casos especiales:

- **Momentos ordinarios** (m_{rs}): Cuando $(a,\ b) = (0,\ 0)$.
- **Momentos centrales** (μ_{rs}): Cuando
$$(a, b) = (m_{10},\ m_{01}) = (\bar{x},\ \bar{y})$$

Momentos ordinarios y centrales

Definición

Llamamos **momento ordinario de orden (r, s):**

$$m_{rs} = \sum_{i=1}^{k} \sum_{j=1}^{p} (x_i)^r (y_j)^s f_{ij}$$

Definición

Llamamos **momento central de orden (r, s):**

$$\mu_{rs} = \sum_{i=1}^{k} \sum_{j=1}^{p} (x_i - \bar{x})^r (y_j - \bar{y})^s f_{ij}$$

Momentos importantes

Ordinarios:

$$m_{0,0} = 1 \qquad \mathbf{m_{1,0}} = \bar{\mathbf{x}} = \frac{1}{N}\sum_i n_{i\cdot} x_i$$
$$\mathbf{m_{0,1}} = \bar{\mathbf{y}} = \frac{1}{N}\sum_j n_{\cdot j} y_j \qquad m_{2,0} = \frac{1}{N}\sum_i n_{i\cdot} x_i^2$$
$$m_{0,2} = \frac{1}{N}\sum_j n_{\cdot j} y_j^2 \qquad m_{1,1} = \frac{1}{N}\sum_i \sum_j n_{ij} x_i y_j$$

Llamamos **centro de gravedad** de la distribución al punto:

$$\mathbf{G} = (\bar{\mathbf{x}},\ \bar{\mathbf{y}}) = (m_{1,0},\ m_{0,1})$$

Centrales:

$$\mu_{0,0} = 1 \qquad \mu_{1,0} = 0 \qquad \mu_{0,1} = 0$$
$$\mu_{\mathbf{2,0}} = \frac{1}{N}\sum_i n_{i\cdot}(x_i - \bar{x})^2 = \sigma_{\mathbf{x}}^{\mathbf{2}} = \mathbf{V(x)} = \mathbf{m_{2,0}} - \bar{\mathbf{x}}^{\mathbf{2}}$$
$$\mu_{\mathbf{0,2}} = \frac{1}{N}\sum_j n_{\cdot j}(y_j - \bar{y})^2 = \sigma_{\mathbf{y}}^{\mathbf{2}} = \mathbf{V(y)} = \mathbf{m_{0,2}} - \bar{\mathbf{y}}^{\mathbf{2}}$$
$$\mu_{\mathbf{1,1}} = \frac{1}{N}\sum_i \sum_j n_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \sigma_{\mathbf{xy}} = \mathbf{Cov(x,y)} = \mathbf{m_{1,1}} - \bar{\mathbf{x}}\bar{\mathbf{y}}$$

Ejemplo momentos

Ejemplo

La tabla representa el tiempo de establecimiento de la comunicación en s. (t), respecto a la distancia al servidor en Km. Hallar $\bar{t}$, $\bar{d}$, $V(t) = \sigma_t^2$, $V(d) = \sigma_d^2$, $Cov(t, d) = \sigma_{td}$ *y* $M_{1,2}(3, 2)$.

$t \backslash d$	$[0, 1)$	$[1, 2)$	$[2, 4)$	$[4, 8)$	*8 ó más Km.*	
$[0 - 2]$	3	12	15	10	22	62
$(2 - 5]$	9	8	5	9	12	43
$(5 - 10]$	7	5	3	8	8	31
10 ó más seg.	11	7	8	8	10	44
	30	32	31	35	52	**180**

Ejemplo momentos-2

Teniendo en cuenta el convenio sobre que los intervalos extremos de amplitud infinita tienen igual amplitud que su adyacente, calculamos las marcas de clase:

	I_d	$[0,1)$	$[1,2)$	$[2,4)$	$[4,8)$	$[8,\infty)$	
I_t	$t_i \backslash d_j$	0.5	1.5	3	6	10	
$[0,2]$	1	3	12	15	10	22	62
$(2,5]$	3.5	9	8	5	9	12	43
$(5,10]$	7.5	7	5	3	8	8	31
$(10,\infty)$	12.5	11	7	8	8	10	44
		30	32	31	35	52	**180**

$$\bar{t} = m_{1,0} = \frac{1(62)+3.5(43)+7.5(31)+12.5(44)}{180} = \frac{995}{180} \approx 5.52778$$

$$\bar{d} = m_{0,1} = \frac{0.5(30)+1.5(32)+3(31)+6(35)+10(52)}{180} = \frac{886}{180} \approx 4.92222$$

Ejemplo momentos-3

	l_d	$[0,1)$	$[1,2)$	$[2,4)$	$[4,8)$	$[8,\infty)$	
l_t	$t_i \backslash d_j$	0.5	1.5	3	6	10	
$[0,2]$	1	3	12	15	10	22	62
$(2,5]$	3.5	9	8	5	9	12	43
$(5,10]$	7.5	7	5	3	8	8	31
$(10,\infty)$	12.5	11	7	8	8	10	44
		30	32	31	35	52	**180**

$$m_{2,0} = \frac{1^2(62)+3.5^2(43)+7.5^2(31)+12.5^2(44)}{180} = \frac{9207.5}{180} \approx 51.152778 \Rightarrow$$

$$\mathbf{V(t) = m_{2,0} - \bar{t}^2 \approx 20.59645}$$

$$m_{0,2} = \frac{0.5^2(30)+1.5^2(32)+3^2(31)+6^2(35)+10(52)}{180} = \frac{6818.5}{180} \approx 37.88056 \Rightarrow$$

$$\mathbf{V(d) = m_{0,2} - \bar{d}^2 \approx 13.6523}$$

Ejemplo momentos-4

I_t	I_d / $t_i \backslash d_j$	$[0,1)$ 0.5	$[1,2)$ 1.5	$[2,4)$ 3	$[4,8)$ 6	$[8,\infty)$ 10	
$[0,2]$	1	3	12	15	10	22	62
$(2,5]$	3.5	9	8	5	9	12	43
$(5,10]$	7.5	7	5	3	8	8	31
$(10,\infty)$	12.5	11	7	8	8	10	44
		30	32	31	35	52	**180**

$\sum_i \sum_j n_{ij} t_i d_j = 1(0.5)(3) + 1(1.5)(12) + 1(3)(15) + 1(6)(10) + 1(10)(22) +$
$3.5(0.5)(9) + 3.5(1.5)(8) + 3.5(3)(5) + 3.5(6)(9) + 3.5(10)(12) + 7.5(0.5)(7) +$
$7.5(1.5)(5) + 7.5(3)(3) + 7.5(6)(8) + 7.5(10)(8) + 12.5(0.5)(11) + 12.5(1.5)(7) +$
$12.5(3)(8) + 12.5(6)(8) + 12.5(10)(10) = 4523.75$

$$m_{11} = \frac{4523.75}{180} \approx 25.131944 \Rightarrow$$

$$\mathbf{Cov(t,d)} = \mu_{11} = \mathbf{m_{1,1}} - \bar{\mathbf{t}}\bar{\mathbf{d}} \approx 25.131944 - (5.52778)(4.92222) \approx \mathbf{-2.077}$$

Ejemplo momentos-5

		l_d	$[0,1)$	$[1,2)$	$[2,4)$	$[4,8)$	$[8,\infty)$	
		d_j	0.5	1.5	3	6	10	
l_t	t_i	$t_i-3 \backslash d_j-2$	-1.5	-0.5	1	4	8	
$[0,2]$	1	-2	3	12	15	10	22	62
$(2,5]$	3.5	0.5	9	8	5	9	12	43
$(5,10]$	7.5	4.5	7	5	3	8	8	31
$(10,\infty)$	12.5	9.5	11	7	8	8	10	44
			30	32	31	35	52	**180**

$$\sum_i \sum_j n_{ij}(t_i-3)(d_j-2)^2 =$$
$$-2(-1.5)^2(3)-2(-0.5)^2(12)-2(1)^2(15)-2(4)^2(10)-2(8)^2(22)+0.5(-1.5)^2(9)+$$
$$0.5(-0.5)^2(8)+0.5(1)^2(5)+0.5(4)^2(9)+0.5(8)^2(12)+4.5(-1.5)^2(7)+$$
$$4.5(-0.5)^2(5)+4.5(1)^2(3)+4.5(4)^2(8)+4.5(8)^2(8)+9.5(-1.5)^2(11)+$$
$$9.5(-0.5)^2(7)+9.5(1)^2(8)+9.5(4)^2(8)+9.5(8)^2(10)=7877.875 \Rightarrow$$

$$\mathbf{M_{1,2}(3,2)=\frac{7877.875}{180}\approx 43.76597}$$

Relación entre variables

El objetivo de analizar conjuntamente dos variables diferentes de una población es establecer el tipo de relación existente entre ellas, diferenciando tres casos:

- **Independencia:** No hay relación alguna entre las variables, es decir, ninguna proporciona información alguna sobre la otra.
- **Dependencia funcional:** El valor de una variable queda determinado conociendo el valor de la otra variable para esa misma observación.
- **Dependencia estadística:** Una variable proporciona información sobre la otra, pero conociendo la modalidad de una de ellas no queda determinada la modalidad de la otra.

Independencia entre variables

Definición

- *Se dice que el carácter X es independiente de Y, si todas las frecuencias relativas condicionadas de X respecto a cualquier clase de Y coinciden con las de la marginal de X, es decir $f_i^j = f_{i\cdot}$ para todo j y para todo i.*
- *Análogamente se define la independencia de Y respecto a X si $f_j^i = f_{\cdot j}$ para todo i, j.*

	C_1	C_2	C_3	C_4	
A	4	6	10	2	22
B	2	3	5	1	11
	6	9	15	3	33

Independencia-2

Observación

Si X es independiente de Y entonces Y es independiente de X.

Si X es independiente de $Y \Rightarrow f_{i\cdot} = f_i^j,\ \forall i, j$.
Además siempre se verifica:

$$f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_{i\cdot}}\frac{n_{i\cdot}}{N} = f_j^i f_{i\cdot} \text{ y también } f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_{\cdot j}}\frac{n_{\cdot j}}{N} = f_i^j f_{\cdot j}$$

De esta última: $f_{ij} = f_i^j f_{\cdot j} = f_{i\cdot} f_{\cdot j} \Rightarrow f_j^i = f_{\cdot j}$ que es la condición para que Y sea independiente de X.

Observación

Las variables X e Y son independientes si y solo si $f_{ij} = f_{i\cdot} f_{\cdot j}\ \forall i, j$

Relación entre variables

Ejemplo

Comprobar si la siguiente tabla de frecuencias corresponde a dos variables independientes.

Consideremos la distribución de frecuencias relativas en forma de tabla de doble entrada

	y_1	y_2	y_3	y_4
x_1	1	3	2	4
x_2	3	9	6	12
x_3	2	6	4	8

	y_1	y_2	y_3	y_4	
x_1	$1/60$	$3/60$	$2/60$	$4/60$	$1/6$
x_2	$3/60$	$9/60$	$\mathbf{6/60}$	$12/60$	$\mathbf{3/6}$
x_3	$2/60$	$6/60$	$4/60$	$8/60$	$2/6$
	$1/10$	$3/10$	$\mathbf{2/10}$	$4/10$	1

Son independientes, pues observamos que el producto de las frecuencias de las distribuciones marginales **siempre** coincide con la frecuencia correspondiente de la distribución conjunta. Por ejemplo, $f_{2\cdot} \cdot f_{\cdot 3} = f_{23}$, es decir, $3/6 \cdot 2/10 = 6/60$.

Ejemplos de dependencia funcional

	C_1	C_2	C_3	C_4	
A_1	4	6	0	0	10
A_2	0	0	5	7	12
	4	6	5	7	22

La variable A depende funcionalmente de la C, pues conocida la clase de C se conoce la de A. Ej.: Si sabemos que es C_3, sabemos que es A_2. Al revés no es cierto, conocida la clase de A no se conoce la de C. Si sabemos que es A_2 podrá ser C_3 o C_4.
En el próximo ejemplo **la dependencia funcional es mutua**:

	C_1	C_2	C_3	C_4	
A_1	4	0	0	0	4
A_2	0	0	0	9	9
A_3	0	0	10	0	10
A_4	0	7	0	0	7
	4	7	10	9	30

Dependencia estadística

La dependencia funcional y la independencia son casos extremos de la relación posible entre dos variables. Generalmente, lo que se produce es una dependencia estadística, en la que el conocimiento de una variable da información válida sobre la otra (reduce incertidumbres).
Ejemplos:

- Estatura y peso. (Ambas cuantitativas continuas)
- Nacionalidad y Renta. (Cualitativa y cuantitativa continua).
- Familias por 'Número hijos' y 'Número de móviles'. (Ambas cuantitativas discretas).
- 'Marca router' y 'Compañía telefónica'. (Ambas cualitativas).

Regresión y correlación

Definición

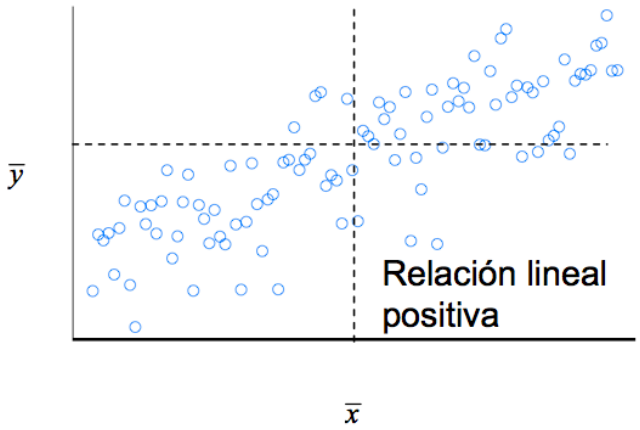
Correlación *es una medida del grado de dependencia entre las variables.*
La **regresión** *pretende encontrar un modelo aproximado de la dependencia entre las variables.*

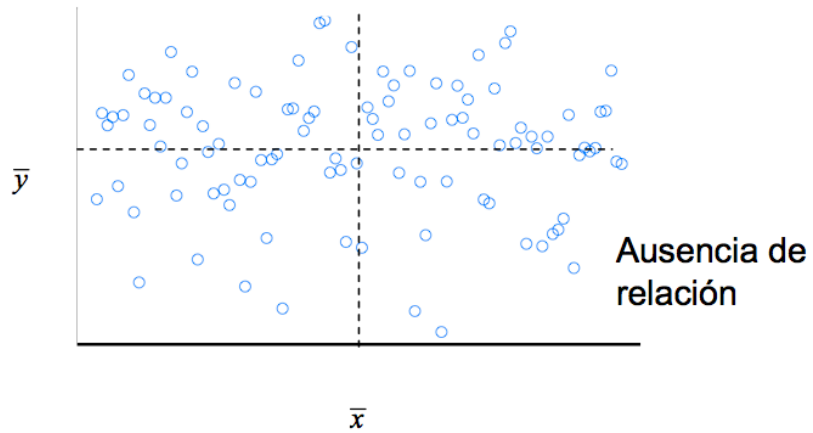
Representando los datos de la muestra de la variable bidimensional obtenemos una nube de puntos. Se llama *línea o curva de regresión* a la función que "mejor" se ajusta a esa nube de puntos.

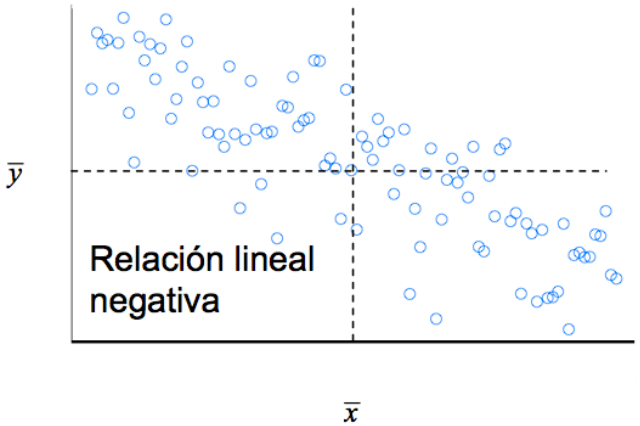
Si todos los valores de la variable satisfacen la ecuación calculada, se dice que las variables están perfectamente correladas. La ecuación de la curva de regresión para este caso nos permite predecir valores desconocidos.

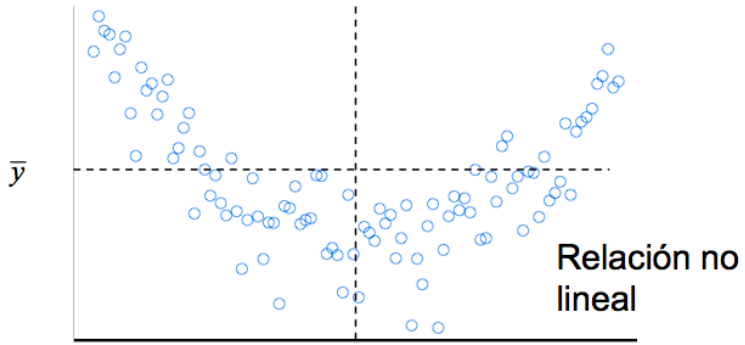
Regresión y correlación

El diagrama de dispersión muestra el tipo de relación existente:



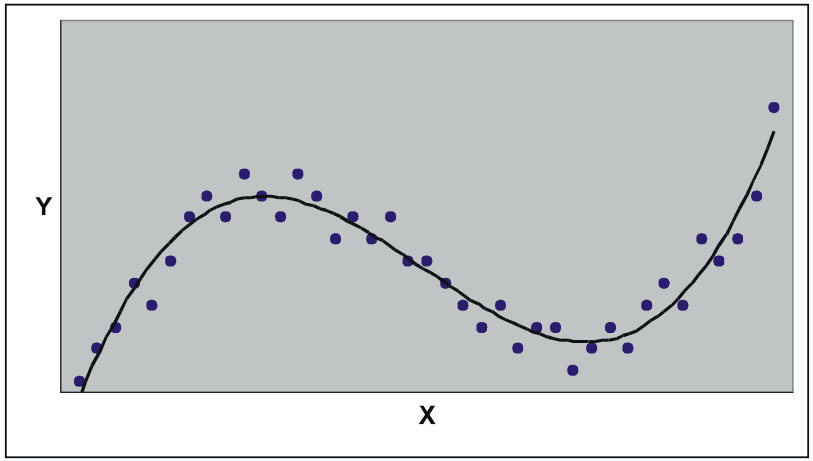






Curva de regresión

A la vista de la nube de puntos, podemos elegir el tipo de modelo a elegir: lineal, cuadrático, exponencial, etc.



La bondad del ajuste se puede observar representando conjuntamente la nube de puntos y la curva de regresión.

Regresión general

La regresión puede realizarse para todo tipo de variables, incluso cualitativas.

	C_1	C_2	Turista	Tripulación	
Español	3	10	25	2	40
Alemán	7	14	8	1	30
Francés	3	0	12	15	30

Regresión de Tipo respecto a Nacionalidad: Si es 'Español' lo más probable es que viaje como 'Turista'. Si es 'Alemán' en C_2 y si es 'Francés' sea de la 'Tripulación'.

Regresión de Nacionalidad respecto a Tipo: Si viaja en 'Clase-1´ o en 'Clase 2' lo más probable es que sea 'Alemán', si en 'Turista' que sea 'Español' y si es de la 'Tripulación' que sea 'Francés'.

Regresión general-2

Regresión de cualitativa con cuantitativa.

L.\ Hr.	[0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]	(0.8, 1]	
Málaga	0	0	3	26	2	31
Granada	1	4	8	12	6	31
Cádiz	0	0	7	11	13	31

Regresión de 'Humedad relativa' respecto a 'Localidad': Para Málaga y Granada lo más probable es (0.6, 0.8]. Para 'Cádiz' del (0.8, 1].

Regresión de 'Localidad' respecto a 'Humedad relativa': Si es inferior a 0.6 lo más probable es que se haya medido en 'Granada', si es de (0.6, 0.8] que sea de 'Málaga' y que sea de 'Cádiz' si es superior a 0.8.

Ajuste por el método de mínimos cuadrados

Sean los datos $\{(x_i,\ y_i)\}$ para dos variables estadísticas X e Y cuantitativas. El objetivo es determinar la función $y = f(x)$ de un subconjunto de las funciones reales (rectas, parábolas, hipérbolas, ...) que más se aproxime a los datos. Se trata, pues, de minimizar la **función objetivo mínimo-cuadrática**:

$$\mathbf{F} = \sum_{\mathbf{i}} \left(\mathbf{y_i} - \mathbf{y_i^{est}}\right)^2 = \sum_{\mathbf{i}} \left(\mathbf{y_i} - \mathbf{f(x_i)}\right)^2$$

$y_i^{est} = f(x_i)$ es el valor de y estimado por la regresión para x_i.
$e_i = y_i - y_i^{est}$ es el error cometido por el ajuste para el i-ésimo dato.
Minimizar la función objetivo significa minimizar el Error Cuadrático Medio $\left(ECM = \frac{\sum_i e_i^2}{N}\right)$ y la media cuadrática de los errores $\left(MC = \sqrt{\frac{\sum_i e_i^2}{N}}\right)$.

Tipos de ajuste

El tipo de ajuste de mínimos cuadrados está determinado por el tipo de función $y = f(x)$ elegido. Los más usados son:

- **Ajuste lineal:** $\mathbf{y = f(x) = a + bx}$ (parámetros a y b).
- **Ajuste parabólico:** $\mathbf{y = a + bx + cx^2}$ (parámetros a, b y c).
- **Ajuste hiperbólico:** $y = \frac{\mathbf{1}}{\mathbf{a+bx}}$ (parámetros a y b).
- **Ajuste exponencial:** $\mathbf{y = ae^{bx}}$ (parámetros a y b).

Un ajuste de mínimos cuadrados requiere del cálculo de los valores de los parámetros del modelo que minimizan la función objetivo $F(a, \cdots) = \sum_i (y_i - f(x_i))^2 = \sum_i e_i^2$.
Existen otros tipos de ajuste. En particular, se define la **curva general de regresión de** Y **sobre** X como la función que asigna a cada valor x_i de la variable X la media de la variable condicionada Y/x_i.

Ajuste de la recta Y/X

Dado un conjunto de puntos $\{(x_i,\ y_i)\}_{i\in\mathcal{I}}$ queremos calcular una recta de la forma $\mathbf{y = a + bx}$ que mejor se ajuste a esos datos en el sentido "de mínimos cuadrados", es decir que minimice la función:

$$\mathbf{F(a,\ b)} = \sum_{\mathbf{i}\in\mathcal{I}} (\mathbf{y_i - (a + bx_i)})^{\mathbf{2}}.$$

Los valores de los parámetros **a** y **b** que minimizan esa función se obtienen resolviendo el sistema de ecuaciones:

$$\nabla \mathbf{F} = \begin{bmatrix} \frac{\partial F}{\partial a} \\ \frac{\partial F}{\partial b} \end{bmatrix} = \mathbf{0} \Rightarrow \left\{ \begin{array}{l} \frac{\partial F}{\partial a} = -2\sum_i (y_i - a - bx_i) = 0 \\ \frac{\partial F}{\partial b} = -2\sum_i (y_i - a - bx_i)x_i = 0 \end{array} \right\} \Rightarrow$$

Recta de regresión Y/X
Sistema de
ecuaciones normales

$$\begin{array}{c} \sum_{\mathbf{i}} \mathbf{y_i} = \mathbf{Na} + \mathbf{b}\sum_{\mathbf{i}} \mathbf{x_i} \\ \sum_{\mathbf{i}} \mathbf{x_i y_i} = \mathbf{a}\sum_{\mathbf{i}} \mathbf{x_i} + \mathbf{b}\sum_{\mathbf{i}} \mathbf{x_i^2} \end{array}$$

Ajuste de la recta X/Y

Análogamente, si dado un conjunto de puntos $\{(x_i,\ y_i)\}_{i\in\mathcal{I}}$ queremos calcular una recta de la forma $\mathbf{x} = \mathbf{a}' + \mathbf{b}'\mathbf{y}$ que mejor se ajuste a esos datos en el sentido "de mínimos cuadrados", la función a minimizar es:

$$\mathbf{G}(\mathbf{a}',\ \mathbf{b}') = \sum_{\mathbf{i}\in\mathcal{I}} (\mathbf{x_i} - (\mathbf{a}' + \mathbf{b}'\mathbf{y_i}))^2.$$

Ahora los parámetros $\mathbf{a}'$ y $\mathbf{b}'$ deberán satisfacer las ecuaciones:

$$\nabla \mathbf{G} = \begin{bmatrix} \frac{\partial G}{\partial a'} \\ \frac{\partial G}{\partial b'} \end{bmatrix} = \mathbf{0} \Rightarrow \left\{ \begin{array}{l} \frac{\partial G}{\partial a'} = -2\sum_i (x_i - a' - b'y_i) = 0 \\ \frac{\partial G}{\partial b'} = -2\sum_i (x_i - a' - b'y_i)y_i = 0 \end{array} \right\} \Rightarrow$$

Recta de regresión X/Y
Sistema de ecuaciones normales

$$\begin{array}{c} \sum_{\mathbf{i}} \mathbf{x_i} = \mathbf{N}\mathbf{a}' + \mathbf{b}' \sum_{\mathbf{i}} \mathbf{y_i} \\ \sum_{\mathbf{i}} \mathbf{x_i}\mathbf{y_i} = \mathbf{a}' \sum_{\mathbf{i}} \mathbf{y_i} + \mathbf{b}' \sum_{\mathbf{i}} \mathbf{y_i^2} \end{array}$$

Ajuste lineal forma matricial

El sistema de ecuaciones normales en forma matricial para el caso de una regresión lineal es:

Recta de Y sobre X: (y=a+bx)

$$\begin{bmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}$$

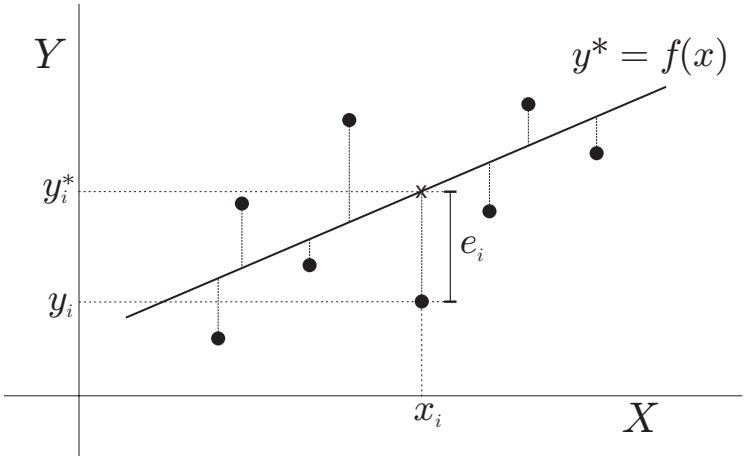
Recta de X sobre Y: (x=a'+b'y)

$$\begin{bmatrix} N & \sum_i y_i \\ \sum_i y_i & \sum_i y_i^2 \end{bmatrix} \begin{bmatrix} a' \\ b' \end{bmatrix} = \begin{bmatrix} \sum_i x_i \\ \sum_i x_i y_i \end{bmatrix}$$

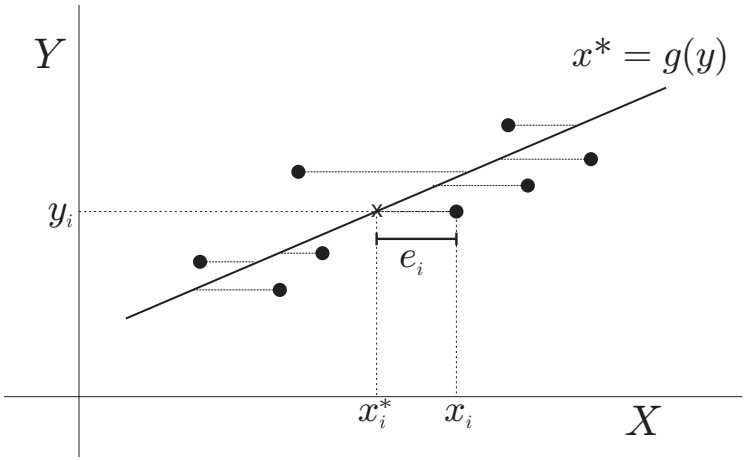
Nota: Las ecuaciones pueden ser fácilmente adaptadas para los casos en que se dispone de datos con frecuencias.

Significado de los ajustes Y/X y X/Y

Ajuste Y/X: Minimiza $F = \sum_i e_i^2 = \sum_i (y_i - y_i^*)^2$



Ajuste X/Y: Minimiza $G = \sum_i (e_i')^2 = \sum_i (x_i - x_i^*)^2$



Ejemplo ajuste lineal

Ejemplo

La tabla siguiente muestra la evolución de la población española de edad comprendida entre 80 y 89, entre los años 2002 y 2011.

y=Número	893218	926708	963513	1003857	1088204
x=Año	2002	2003	2004	2005	2006
y=Número	1126204	1126704	1166200	1202349	1239183
x=Año	2007	2008	2009	2010	2011

Ajustar las rectas de Y/X y de X/Y

Las ecuaciones normales para Y/X: $\left\{ \begin{array}{c} \sum_i y_i = Na + b\sum_i x_i \\ \sum_i x_i y_i = a\sum_i x_i + b\sum_i x_i^2 \end{array} \right.$

$$\Rightarrow \left\{ \begin{array}{c} 10736140 = 10a + 20065b \\ 21545296484 = 20065a + 40260505b \end{array} \right. \Rightarrow \left\{ \begin{array}{c} a = -77522182.7 \\ b = 39170.5939 \end{array} \right\}$$

La recta ajustada es: $\mathbf{y = -77522182.7 + 39170.5939x}$

Puede usarse para estimar el número previsto para 2012:

Número=-77522182.7+39170.5939(2012)=1289052.23

Ejemplo ajuste lineal - continuación

Las ecuaciones normales para X/Y:

$$\begin{cases} \sum_i x_i = Na' + b' \sum_i y_i \\ \sum_i x_i y_i = a' \sum_i y_i + b' \sum_i y_i^2 \end{cases}$$

$$\Rightarrow \left\{ \begin{array}{c} 20065 = 10a' + 1073614b' \\ 21545296484 = 1073614a' + 11655937654544b' \end{array} \right\} \Rightarrow$$

$\mathbf{a' = 1979.702,\ b' = 0.00002496051}$

La recta ajustada es: $\mathbf{x = 1979.702 + 0.00002496051y}$

Puede usarse para estimar el año en que se prevén 1300000 individuos de esa edad:

Año=1979.702+0.00002496051(1300000)=2012.1507

Así, si los datos están medidos a 1 de enero, se prevé esa cantidad para el 24/2/2012. (0.1507*366=55.16 días)

Ajuste lineal. Propiedades

Dividiendo por N las ecuaciones normales:

$$\left.\begin{array}{c} \frac{\sum_i y_i}{N} = a + b\frac{\sum_i x_i}{N} \\ \frac{\sum_i x_i y_i}{N} = a\frac{\sum_i x_i}{N} + b\frac{\sum_i x_i^2}{N} \end{array}\right\} \Rightarrow \left\{\begin{array}{c} \bar{y} = a + b\bar{x} \\ m_{11} = a\bar{x} + bm_{20} \end{array}\right\}$$

$$\left.\begin{array}{c} \frac{\sum_i x_i}{N} = a' + b'\frac{\sum_i y_i}{N} \\ \frac{\sum_i x_i y_i}{N} = a'\frac{\sum_i y_i}{N} + b'\frac{\sum_i y_i^2}{N} \end{array}\right\} \Rightarrow \left\{\begin{array}{c} \bar{x} = a' + b'\bar{y} \\ m_{11} = a'\bar{y} + b'm_{02} \end{array}\right\}$$

Deducimos que **el centro de gravedad $\mathbf{G} = (\bar{\mathbf{x}},\ \bar{\mathbf{y}})$ pertenece a ambas rectas**. Las rectas Y/X y X/Y se cortan en G.
Eliminando a en la de Y/X y a' en la de X/Y:

$$m_{11} - \bar{x}\bar{y} = b(m_{20} - \bar{x}^2) \Rightarrow b = \frac{Cov(x,\ y)}{\sigma_x^2} = \frac{\mu_{11}}{V(x)}$$

$$m_{11} - \bar{x}\bar{y} = b'(m_{02} - \bar{y}^2) \Rightarrow b' = \frac{Cov(x,\ y)}{\sigma_y^2} = \frac{\mu_{11}}{V(y)}$$

Coeficiente de correlación lineal de Pearson:

Definición

El **coeficiente de correlación lineal** *mide el grado de relación lineal (magnitud y dirección) entre las variables:*

$$\rho = \mathbf{r} = \frac{\mathbf{Cov(x,\ y)}}{\sigma_x \sigma_y} = \frac{\mu_{11}}{\sigma_x \sigma_y} \qquad (-1 \leq \mathbf{r} \leq 1)$$

Significado: Es la media geométrica de los coeficientes b y b', $\mathbf{r} = \sqrt{\mathbf{bb'}}$. (El signo de r será el de b: $r = \text{signo(b)}\sqrt{\text{bb'}}$.)

- **r>0** Correlación lineal directa.
- **r<0** Correlación lineal inversa.
- **r=0** Variables incorreladas.
- **r=1 ó r=-1** Correlación lineal perfecta (directa o inversa).

Expresiones del ajuste lineal

1. De $b = \frac{\mu_{11}}{V(x)} = \frac{\mu_{11}\sigma_y}{\sigma_y\sigma_x^2} = r\frac{\sigma_y}{\sigma_x} \Rightarrow b = r\frac{\sigma_y}{\sigma_x}$
2. De $b' = \frac{\mu_{11}}{V(y)} = \frac{\mu_{11}\sigma_x}{\sigma_x\sigma_y^2} = r\frac{\sigma_x}{\sigma_y} \Rightarrow b' = r\frac{\sigma_x}{\sigma_y}$
3. Como ambas rectas pasan por $G = (\bar{x}, \bar{y})$, las rectas quedan:

   Recta Y/X: $y - \bar{y} = b(x - \bar{x}) \Rightarrow y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$
   Recta X/Y: $y - \bar{y} = \frac{1}{b'}(x - \bar{x}) \Rightarrow y - \bar{y} = \frac{1}{r}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$
4. Llamando $\mathtt{Y} = \frac{y-\bar{y}}{\sigma_y}$, $\mathtt{X} = \frac{x-\bar{x}}{\sigma_x}$ queda:

   $$\text{Recta Y/X:} \qquad \mathtt{Y} = r\mathtt{X}$$

   $$\text{Recta X/Y:} \qquad \mathtt{Y} = \frac{1}{r}\mathtt{X}$$
5. Las pendientes de las rectas Y/X y X/Y son $m_{Y/X} = b$ y $m_{X/Y} = 1/b' \quad \Rightarrow |m_{X/Y}| \geq |m_{Y/X}|$.

Ejemplo

Ejemplo

Dada la tabla de doble entrada:

$C \backslash E$	0	1	
0	2	4	6
1	4	8	12
2	2	0	2
	8	12	20

1. *Ajustar las rectas de Y/X y de X/Y (sin usar las ecuaciones normales).*
2. *Calcular el coeficiente de correlación lineal de Pearson.*

Calculamos: $\bar{C} = \frac{6\cdot 0+12\cdot 1+2\cdot 2}{20} = 0.8, \quad \bar{E} = \frac{8\cdot 0+12\cdot 1}{20} = 0.6,$

$\sigma_C^2 = V(C) = \frac{6\cdot 0^2+12\cdot 1^2+2\cdot 2^2}{20} - 0.8^2 = \frac{20}{20} - 0.64 = 0.36,$

$\sigma_E^2 = V(E) = \frac{8\cdot 0^2+12\cdot 1^2}{20} - 0.6^2 = \frac{12}{20} - 0.36 = 0.24.$

Ejemplo-cont.

Calculamos las desviaciones típicas marginales:

$\sigma_C = \sqrt{0.36} = 0'6$ y $\sigma_E = \sqrt{0.24} = \frac{\sqrt{6}}{5}$.

Calculamos la covarianza:

$$\sigma_{CE} = \frac{2\cdot 0\cdot 0 + 4\cdot 0\cdot 1 + 4\cdot 1\cdot 0 + 8\cdot 1\cdot 1 + 2\cdot 2\cdot 0}{20} - 0'8\cdot 0'6 =$$

$$= -0'08\frac{8}{20} - 0'48 = 0'4 - 0'48 = -0'08$$ y aplicamos la fórmula para obtener el coeficiente de correlación lineal y las rectas:

$$r = \frac{-0.08}{0.6(\sqrt{6}/5)} = -\frac{2}{3\sqrt{6}}$$

Recta E/C: $$E - 0.6 = -\frac{2}{3\sqrt{6}}\frac{\sqrt{6}/5}{0.6}(C - 0.8) \Rightarrow E = \frac{7}{9} - \frac{2}{9}C$$

Recta C/E: $$E - 0.6 = -\frac{3\sqrt{6}}{2}\frac{\sqrt{6}/5}{0.6}(C - 0.8) \Rightarrow C = 1 - \frac{1}{3}E$$

Varianza residual. Coeficiente de determinación.

Dada una nube de puntos $\{(x_i,\ y_i)\}$, llamamos **vector residuo** $\vec{e} = (e_i)$ con $e_i = y_i - y_i^{est}$. Es decir, e_i es el error cometido por el ajuste para la i-ésima observación.

Definición

Llamamos **varianza residual** *a la varianza del vector residuo.*

$$\mathbf{V_r = \sum_i f_i(e_i - \bar{e})^2 = \sum_i f_i e_i^2 - \bar{e}^2}$$

Definición

Llamamos **coeficiente de determinación** *a:*

$$\mathbf{R^2 = 1 - \frac{V_r}{V(y)}}$$

Discusión

El coeficiente de determinación R^2 verifica $0 \leq R^2 \leq 1$.

Definición

Llamamos **varianza explicada** *por la regresión a* $\mathbf{V_e = R^2 V(y)}$

De $R^2 = 1 - \frac{V_r}{V(y)}$, obtenemos: $V_r = (1 - R^2)V(y)$, luego:

$$\mathbf{V(y)} = R^2 V(y) + (1 - R^2)V(y) = R^2 V(y) + V_r = \mathbf{V_e + V_r}$$

Así, $R^2 = \frac{V_e}{V(y)}$ representa la fracción de la varianza explicada por el ajuste.

- $R^2 = 1 \Rightarrow$ Ajuste perfecto.
- $R^2 = 0 \Rightarrow$ El ajuste no explica nada.

Coeficiente de determinación caso lineal

En el caso lineal los residuos verifican:

- $\sum_i \mathbf{e_i} = \mathbf{0} \quad \Leftrightarrow \quad \langle \vec{e},\ \vec{1} \rangle = 0 \quad \Leftrightarrow \quad \vec{e} \perp \vec{1}$

  $\sum_i e_i = \sum_i (y_i - y_i^{est}) = \sum_i (y_i - (a + bx_i)) = \sum_i y_i - Na - b\sum_i x_i = 0$
- $\sum_i \mathbf{e_i x_i} = \mathbf{0} \quad \Leftrightarrow \quad \langle \vec{e},\ \vec{x} \rangle = 0 \quad \Leftrightarrow \quad \vec{e} \perp \vec{x}$

  $\sum_i e_i x_i = \sum_i (y_i - y_i^{est}) x_i = \sum_i x_i (y_i - a - bx_i) = \sum_i x_i y_i - a\sum_i x_i - b\sum_i x_i^2 = 0$

DESCOMPOSICIÓN DE LA VARIANZA

En el caso lineal la varianza de y puede expresarse como suma de la varianza residual y la varianza de los puntos estimados y_i^{est}.

$$V(y) = \sum_i f_i (y_i - \bar{y})^2 = \sum_i f_i (y_i - y_i^{est})^2 + \sum_i f_i (y_i^{est} - \bar{y})^2 = V_r + V(y^{est})$$

Para comprobarlo los pasos son:

- $\bar{y_i^{est}} = \bar{y}$, pues ambos valen $a + b\bar{x}$
- $\sum_i (y_i - y_i^{est})(y_i^{est} - \bar{y}) = \sum_i e_i r \frac{\sigma_y}{\sigma_x} (x_i - \bar{x}) = 0$ ya que $\sum_i e_i = 0$ y $\sum_i e_i x_i = 0$

Simplificación varianza residual caso lineal

En el caso lineal $\mathbf{V_e} = V(\vec{y^{est}}) = \sum_i f_i(y_i^{est} - \bar{y})^2 = \sum_i f_i(a + bx_i - (a + b\bar{x}))^2 = b^2 \sum_i f_i(x_i - \bar{x})^2 = \mathbf{b^2 V(x)} \Rightarrow$

$$\mathbf{V_e} = b^2\sigma_x^2 = \left(r\frac{\sigma_y}{\sigma_x}\right)^2 \sigma_x^2 = \mathbf{r^2 V(y)}$$

Luego la varianza residual puede obtenerse desde el coeficiente de regresión lineal r:

- $\mathbf{R^2 = r^2}$
- $\mathbf{V_r = (1 - r^2)V(y)}$

Ajuste exponencial $y = ae^{bx}$

$\mathbf{y = ae^{bx}}$ (se introducen logaritmos para linealizar el modelo) $\Rightarrow$

$$ln(y) = ln(ae^{bx}) = ln(a) + bx$$

Llamando $\hat{y} = ln(y)$ y $A = ln(a)$ se tiene el modelo: $\hat{y} = A + bx$. Ajustando una recta a los datos $\{(x_i,\ \hat{y}_i = ln(y_i))\}$ se obtienen $A = ln(a)$ y b, por lo que (con $a = e^A$) se tienen los parámetros a y b del ajuste exponencial $y = ae^{bx}$ buscado.

Ejemplo

Ajustar una curva del tipo $y = ae^{bx}$ a los datos de la tabla:

x_i	0	1	2	3	6
y_i	7	5	4	3.5	3

Hallar la varianza residual y el coeficiente de determinación.

Ejemplo ajuste exponencial

x_i	0	1	2	3	6	12
y_i	7	5	4	3.5	3	22.5
$\hat{y}_i = ln(y_i)$	1.9459	1.6094	1.3863	1.2528	1.0986	7.293
x_i^2	0	1	4	9	36	50
$x_i\hat{y}_i$	0	1.6094	2.7726	3.7583	6.5917	14.732
y_i^{est}	5.8846	5.1635	4.5308	3.9756	2.6859	
$y_i - y_i^{est}$	1.1154	−0.1635	−0.5308	−0.4756	0.3141	0.2597
$(y_i - y_i^{est})^2$	1.2442	0.0267	0.2817	0.2262	0.0987	1.8775
y_i^2	49	25	16	12.25	9	111.25

Las ecuaciones normales son: $\begin{cases} 7.293 = 5A + 12b \\ 14.732 = 12A + 50b \end{cases} \Rightarrow \begin{matrix} A = 1.7723 \\ b = -0.1307 \end{matrix}$

El ajuste buscado es $\mathbf{y = 5.8846e^{-0.1307x}}$ ya que $a = e^{1.7723} = 5.8846$.

$$V_y = \frac{111.25}{5} - \left(\frac{22.5}{5}\right)^2 = 2, \quad V_r = \frac{1.8775}{5} - \left(\frac{0.2597}{5}\right)^2 = 0.3728,$$

$$R^2 = 1 - \frac{V_r}{V(y)} = 1 - \frac{0.3728}{2} = 0.8136.$$

Ajuste hiperbólico $\mathbf{y = \frac{1}{a+bx}}$

$\mathbf{y = \frac{1}{a+bx}}$ (invirtiendo para linealizar el modelo) $\Rightarrow \frac{1}{y} = a + bx$ y entonces, llamando $\tilde{y} = \frac{1}{y}$, se tiene el modelo $\tilde{y} = a + bx$.
El ajuste de una recta a los datos $\{(x_i,\ \tilde{y}_i = \frac{1}{y_i})\}$ permite obtener los parámetros a y b del ajuste hiperbólico $y = \frac{1}{a+bx}$.

Ejemplo

Ajustar una curva del tipo $y = \frac{1}{a+bx}$ a los datos del problema anterior:

x_i	0	1	2	3	6
y_i	7	5	4	3.5	3

Hallar la varianza residual y el coeficiente de determinación.
¿Qué ajuste a esta tabla de datos es el mejor de estos tres: exponencial, hiperbólico, lineal?

Ejemplo ajuste hiperbólico

x_i	0	1	2	3	6	12
y_i	7	5	4	3.5	3	22.5
$\tilde{y}_i = 1/y_i$	0.1429	0.2	0.25	0.2857	0.3333	1.2119
x_i^2	0	1	4	9	36	50
$x_i\tilde{y}_i$	0	0.2	0.5	0.8571	2	3.5571
y_i^{est}	5.9186	5.0113	4.3451	3.8353	2.8368	
$e_i = y_i - y_i^{est}$	1.0814	−0.0113	−0.3451	−0.3353	0.1632	0.5530
$(y_i - y_i^{est})^2$	1.1693	0.0001	0.1191	0.1124	0.0266	1.4276
y_i^2	49	25	16	12.25	9	111.25
$x_i y_i$	0	5	8	10.5	18	41.5

Ajuste hiperbólico: $\begin{cases} 1.2119 = 5a + 12b \\ 3.5571 = 12a + 50b \end{cases} \Rightarrow \begin{array}{c} a = 0.1690 \\ b = 0.03059 \end{array} \Rightarrow \mathbf{y = \frac{1}{0.1690 + 0.03059x}}$

$V(y) = 2,\ V_r = \frac{1.4276}{5} - \left(\frac{0.553}{5}\right)^2 = 0.2733,\ \mathbf{R^2} = 1 - \frac{V_r}{V(y)} = 1 - \frac{0.2733}{2} = \mathbf{0.8634}$

Ajuste lineal: $\begin{cases} 22.5 = 5a + 12b \\ 41.5 = 12a + 50b \end{cases} \Rightarrow \begin{array}{c} a = 5.9151 \\ b = -0.5896 \end{array} \Rightarrow \mathbf{y = 5.9151 - 0.5896x}$

$V(x) = 4.24,\ \mu_{11} = \frac{41.5}{5} - \frac{12}{5}\frac{22.5}{5} = -2.5,\ r = \frac{-2.5}{\sqrt{2}\sqrt{4.24}} = -0.8585 \Rightarrow \mathbf{R^2 = 0.7370}$

El mejor ajuste en base a R^2 es el hiperbólico que explica el 86.34 % de $V(y)$.

NOTA: Tambien puede usarse como criterio comparativo $SSE = \sum_i e_i^2$. Compruébese que $SSE_e = 1.8775$, $SSE_h = 1.4276$ y $SSE_L = 2.63$.

Ajuste parabólico $\mathbf{y = a + bx + cx^2}$

El sistema de ecuaciones normales puede ser deducido (análogamente a como se hizo en el caso lineal) imponiendo para minimizar la función $F(a,\ b,\ c) = \sum_i (y_i - (a + bx_i + cx_i^2))^2$ que $\frac{\partial F}{\partial a} = 0$, $\frac{\partial F}{\partial b} = 0$ y $\frac{\partial F}{\partial c} = 0$.

Alternativamente, las ecuaciones normales pueden ser obtenidas imponiendo la ortogonalidad del vector de errores a las funciones básicas. Estas funciones para una función de la forma $f(x) = a \cdot 1 + b \cdot x + c \cdot x^2$ son las del conjunto $B = \{1,\ x,\ x^2\}$ (f es combinación lineal de las funciones en B), por lo que el vector error debe cumplir:

$$\left.\begin{array}{l} \langle \vec{e},\ \vec{1} \rangle = 0 \\ \langle \vec{e},\ \vec{x} \rangle = 0 \\ \langle \vec{e},\ \vec{x^2} \rangle = 0 \end{array}\right\} \Leftrightarrow \left.\begin{array}{c} \sum_i (y_i - (a + bx_i + cx_i^2)) = 0 \\ \sum_i (y_i - (a + bx_i + cx_i^2))x_i = 0 \\ \sum_i (y_i - (a + bx_i + cx_i^2))x_i^2 = 0 \end{array}\right\} \Leftrightarrow$$

$$\left\{\begin{array}{c} \sum_i y_i = Na + b\sum_i x_i + c\sum_i x_i^2 \\ \sum_i y_i x_i = a\sum_i x_i + b\sum_i x_i^2 + c\sum_i x_i^3 \\ \sum_i y_i x_i^2 = a\sum_i x_i^2 + b\sum_i x_i^3 + c\sum_i x_i^4 \end{array}\right\}$$

Otros ajustes

1) Ajustar una función del tipo $\mathbf{y = a\cos(x) + b\,sen(x)}$ a un conjunto de puntos $\{(x_i,\ y_i)\}$ dado.
En este caso $B = \{\cos(x),\ \text{sen}(x)\}$ y las ecuaciones normales son:

$$\left.\begin{array}{l} \langle \vec{e},\ \vec{\cos}(x)\rangle = 0 \\ \langle \vec{e},\ \vec{\text{sen}}(x)\rangle = 0 \end{array}\right\} \quad \left.\begin{array}{l} \sum_i (y_i - (a\cos(x_i) + b\,\text{sen}(x_i))\cos(x_i) = 0 \\ \sum_i (y_i - (a\cos(x_i) + b\,\text{sen}(x_i))\,\text{sen}(x_i) = 0 \end{array}\right\}$$

2) Ajustar una función del tipo $\mathbf{y = a + be^{-x} + c\,sen(2x)}$ a un conjunto de puntos $\{(x_i,\ y_i)\}$ dado.
En este caso $B = \{1,\ e^{-x},\ \text{sen}(2x)\}$ y las ecuaciones normales son:

$$\left.\begin{array}{c} \langle \vec{e},\ \vec{1}\rangle = 0 \\ \langle \vec{e},\ \vec{e^{-x}}\rangle = 0 \\ \langle \vec{e},\ \vec{\text{sen}}(x)\rangle = 0 \end{array}\right\} \quad \left.\begin{array}{c} \sum_i (y_i - (a + be^{-x_i} + c\,\text{sen}(2x_i)) = 0 \\ \sum_i (y_i - (a + be^{-x_i} + c\,\text{sen}(2x_i))e^{-x_i} = 0 \\ \sum_i (y_i - (a + be^{-x_i} + c\,\text{sen}(2x_i))\,\text{sen}(2x_i) = 0 \end{array}\right\}$$

Ajuste de un plano $\mathbf{z = a + bx + cy}$

Dada una nube de puntos $\{(x_i, y_i, z_i)\}_{i\in\mathcal{I}}$, podemos deducir las ecuaciones normales minimizando la función:

$$F(a,\ b,\ c) = \sum_i (z_i - (a + bx_i + cy_i))^2$$

mediante $\frac{\partial F}{\partial a} = 0$, $\frac{\partial F}{\partial b} = 0$ y $\frac{\partial F}{\partial c} = 0$. Sin embargo, alternativamente, las ecuaciones normales pueden ser obtenidas imponiendo que el vector error sea ortogonal con cualquiera de la base y, como la función de la forma $z = f(x, y) = a \cdot 1 + b \cdot x + c \cdot y$ tiene como funciones básicas las del conjunto $B = \{1,\ x,\ y\}$, se tiene que:

$$\left.\begin{array}{l} \langle \vec{e},\ \vec{1}\rangle = 0 \\ \langle \vec{e},\ \vec{x}\rangle = 0 \\ \langle \vec{e},\ \vec{y}\rangle = 0 \end{array}\right\} \Leftrightarrow \left.\begin{array}{c} \sum_i (z_i - (a + bx_i + cy_i)) = 0 \\ \sum_i (z_i - (a + bx_i + cy_i))x_i = 0 \\ \sum_i (z_i - (a + bx_i + cy_i))y_i = 0 \end{array}\right\} \Leftrightarrow$$

$$\left\{\begin{array}{c} \sum_i z_i = Na + b\sum_i x_i + c\sum_i y_i \\ \sum_i z_i x_i = a\sum_i x_i + b\sum_i x_i^2 + c\sum_i y_i x_i \\ \sum_i z_i y_i = a\sum_i y_i + b\sum_i x_i y_i + c\sum_i y_i^2 \end{array}\right\}$$

Ejemplo de ajuste de un plano

Ejemplo

Ajustar un plano $z = a + bx + cy$ a los puntos:

x_i	0	1	0	-1	0	1	1	2
y_i	0	1	1	1	3	2	2	0
z_i	2	3	4	5	7	4	5	0

Hallar la suma de los cuadrados de los errores (SSE), la varianza residual y el coeficiente de determinación.

El sistema de ecuaciones normales es:

$$\left\{ \begin{array}{c} \sum_i z_i = Na + b\sum_i x_i + c\sum_i y_i \\ \sum_i z_i x_i = a\sum_i x_i + b\sum_i x_i^2 + c\sum_i y_i x_i \\ \sum_i z_i y_i = a\sum_i y_i + b\sum_i x_i y_i + c\sum_i y_i^2 \end{array} \right\}$$

Ejemplo ajuste de un plano

x_i	y_i	z_i	x_i^2	y_i^2	z_i^2	x_iy_i	z_ix_i	z_iy_i	z_i^*	$e_i = z_i - z_i^*$	e_i^2
0	0	2	0	0	4	0	0	0	2.2045	−0.2045	0.0418
1	1	3	1	1	9	1	3	3	2.8068	0.1932	0.0373
0	1	4	0	1	16	0	0	4	3.8636	0.1364	0.0186
−1	1	5	1	1	25	−1	−5	5	4.9205	0.0795	0.0063
0	3	7	0	9	49	0	0	21	7.1818	−0.1818	0.0331
1	2	4	1	4	16	2	4	8	4.4659	−0.4659	0.2171
1	2	5	1	4	25	2	5	10	4.4659	0.5341	0.2853
2	0	0	4	0	0	0	0	0	0.0909	−0.0909	0.0083
4	10	30	8	20	144	4	7	51		0	0.6477

Las ecuaciones normales son: $(z_i^* = a + bx_i + cy_i)$

$$\begin{cases} 30 = 8a + 4b + 10c \\ 7 = 4a + 8b + 4c \\ 51 = 10a + 4b + 20c \end{cases} \Rightarrow \begin{cases} a = 2.2045 \\ b = -1.0568 \Rightarrow \mathbf{z = 2.2045 - 1.0568x + 1.6591y} \\ c = 1.6591 \end{cases}$$

$\mathbf{SSE} = \sum_i \mathbf{e_i^2} = \mathbf{0.6477}$, $\quad \mathbf{V(z)} = \frac{144}{8} - \left(\frac{30}{8}\right)^2 = \mathbf{3.9375}$,

$\mathbf{V_r} = \frac{0.6477}{8} - 0^2 = \mathbf{0.0810}$, $\quad \mathbf{R^2} = 1 - \frac{V_r}{V(z)} = 1 - \frac{0.0810}{3.9375} \approx \mathbf{0.9794}$