Tema 2: Modelización estadística

Introducción a la modelización y el aprendizaje estadísticos

Departamento Matemática Aplicada

Universidad de Málaga

Curso 2020-2021

Índice

Índice

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, x_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 frecuencias 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

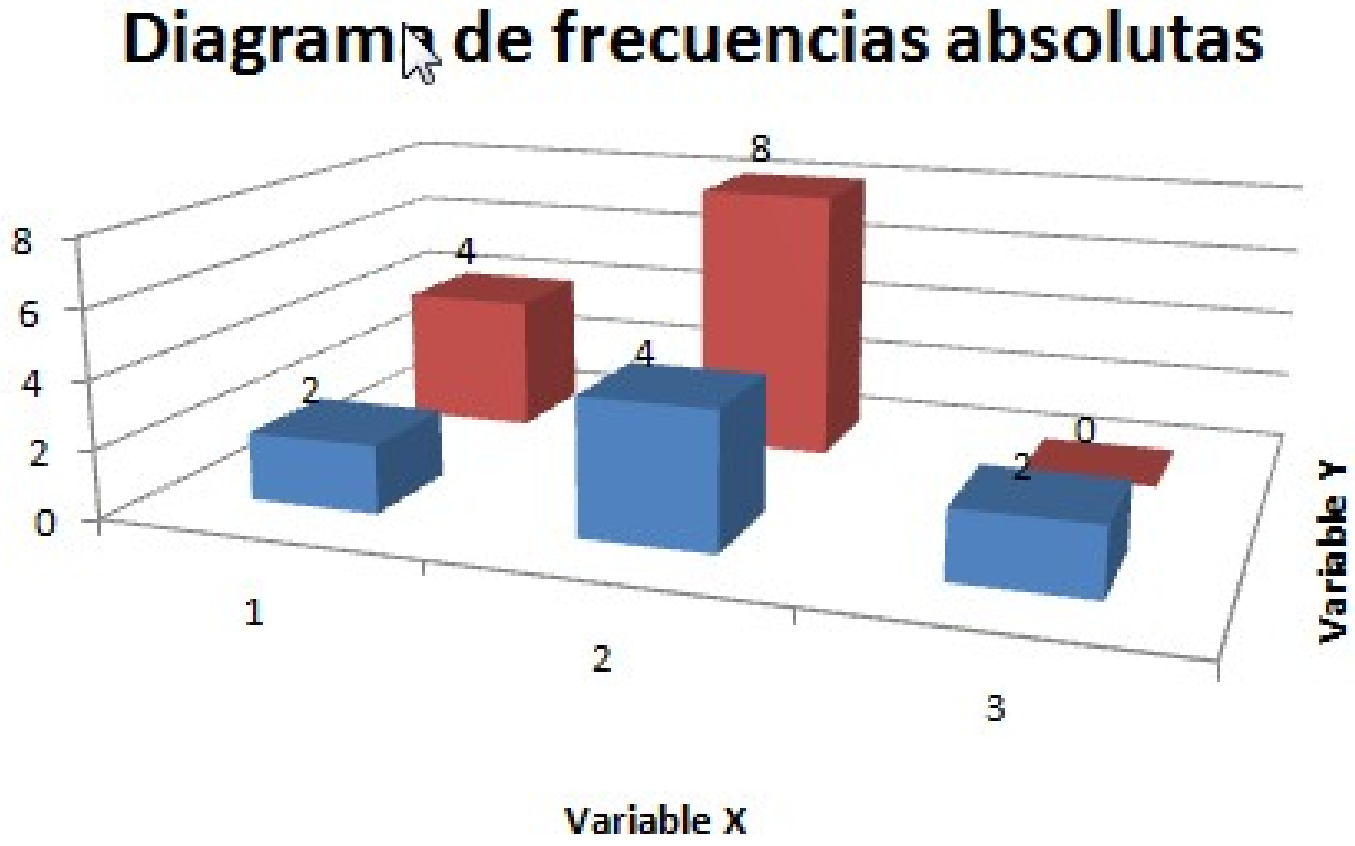
$\Longrightarrow$

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

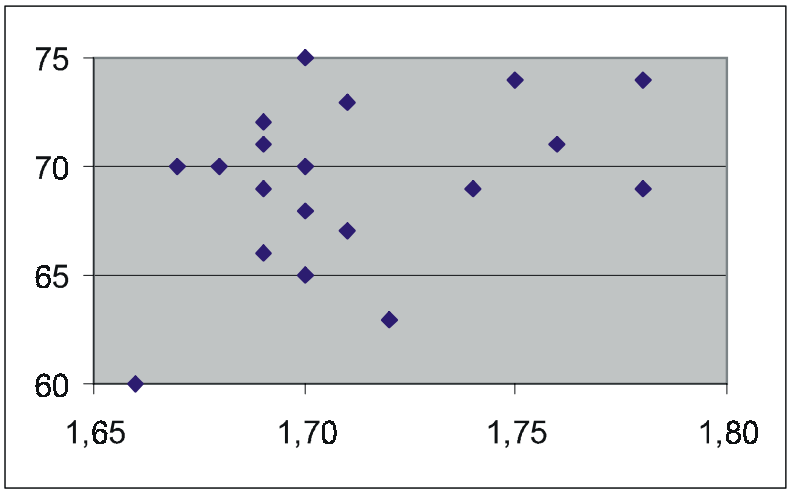
Representaciones gráficas

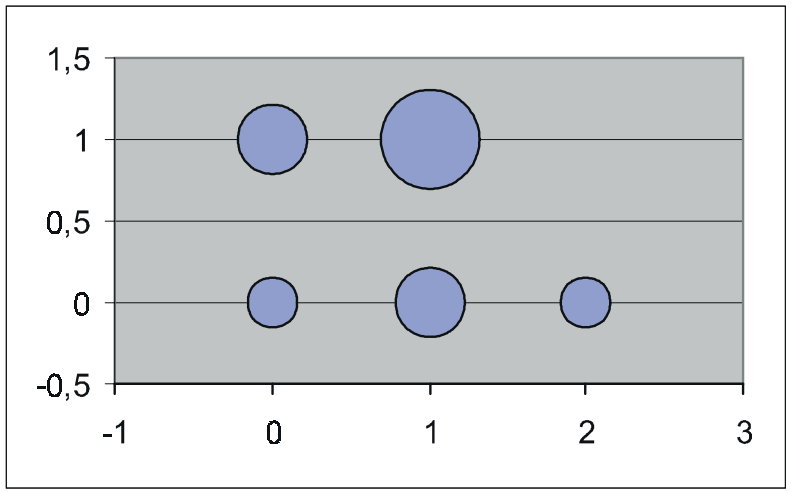
Diagrama de frecuencias. Caso discreto. Similar al diagrama de barras unidimensional. Es una representación tridimensional en la que el plano base representa los valores de las variables y la altura las frecuencias.



Representaciones gráficas-2

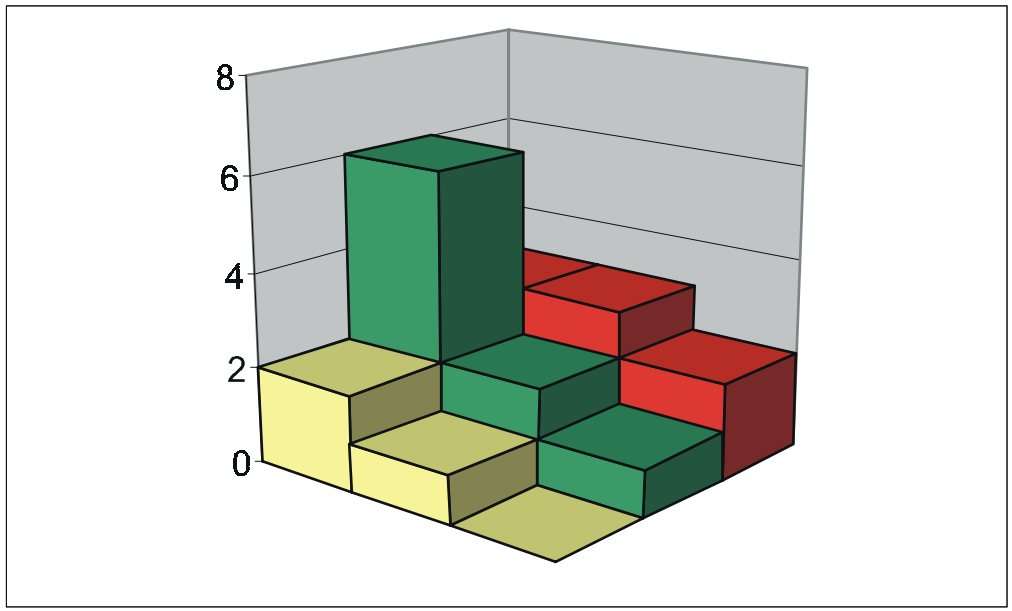
Diagrama de dispersión. Representamos los distintos pares de valores sobre unos ejes cartesianos, obteniéndose una nube de puntos. La frecuencia de cada par de puntos se puede representar usando distintos tamaños de puntos.





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 $h_{i,j} = \frac{n_{i,j}}{S_{i,j}}$, donde $S_{i,j}$ 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 queremos estudiar una de las variables de forma aislada, tenemos que separar la información relativa a dicha variable. Si X tiene modalidades $x_1, x_2, \ldots, x_k$ e Y modalidades $y_1, y_2, \ldots, y_p$ obtenemos 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.

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

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

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

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: $\Rightarrow$

Y	$n_{3\cdot}$
-2	12
-1	31
0	23
1	8
2	3
	77

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

X	$n_{\cdot 2}$
0	7
1	14
2	31
3	18
	70

Momentos

Consideramos los datos agrupados en una tabla bidimensional.

Definición

Llamamos **momento de orden** (r, s) **respecto al punto** (a, b) *a:*

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

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:

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

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

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

Centrales:

$$\mu_{0,0} = 1 \qquad \mu_{1,0} = 0 \qquad \mu_{0,1} = 0$$
$$\mu_{2,0} = \frac{1}{N}\sum_i n_{i\cdot}(x_i - \bar{x})^2 = \mathrm{s}_{\mathrm{N}}^2(\mathrm{X}) = \mathrm{m}_{2,0} - \bar{\mathrm{x}}^2$$
$$\mu_{0,2} = \frac{1}{N}\sum_j n_{\cdot j}(y_j - \bar{y})^2 = \mathrm{s}_{\mathrm{N}}^2(\mathrm{Y}) = \mathrm{m}_{0,2} - \bar{\mathrm{y}}^2$$
$$\mu_{1,1} = \frac{1}{N}\sum_i \sum_j n_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \mathrm{s}_{\mathrm{XY}} = \mathrm{Cov}(\mathrm{X}, \mathrm{Y}) = \mathrm{m}_{1,1} - \bar{\mathrm{x}}\bar{\mathrm{y}}$$

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 $\bar{t}$, $\bar{d}$, $V(t) = s_N^2(t)$, $V(d) = s_N^2(d)$, $Cov(t, d) = s_{td}$ 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
Más de 10 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

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

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

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

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

Ejemplo momentos-4

	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

$\sum_i \sum_j n_{ij} t_i d_j =$

$1(0.5)(3) + 1(1.5)(12) + 1(3)(15) + 1(6)(10) + 1(10)(22) + 3.5(0.5)(9) + 3.5(1.5)(8) +$
$3.5(3)(5) + 3.5(6)(9) + 3.5(10)(12) + 7.5(0.5)(7) + 7.5(1.5)(5) + 7.5(3)(3) + 7.5(6)(8) +$
$7.5(10)(8) + 12.5(0.5)(11) + 12.5(1.5)(7) + 12.5(3)(8) + 12.5(6)(8) + 12.5(10)(10) = 4523.75$

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

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

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

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

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

Índice

Relación entre variables

- **Independencia:** No hay relación alguna entre las variables, ninguna proporciona información 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: $f_{ij} = f_{i\cdot} f_{\cdot j} \forall i, j$

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	$3/60$	$9/60$	$6/60$	$12/60$	$3/6$
x_3	$2/60$	$6/60$	$4/60$	$8/60$	$2/6$
	$1/10$	$3/10$	$2/10$	$4/10$	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 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).

Índice

Modelizar el mundo

- Las matemáticas son la herramienta más importante para poder estudiar el funcionamiento de nuestro mundo. Son la base del resto de ciencias, como la física o la biología.
- La manera en la que aplicamos las matemáticas al mundo es mediante la creación de modelos matemáticos de la realidad.
- En los modelos matemáticos hacemos corresponder los objetos y fenómenos del mundo real (R-mundo) con objetos matemáticos (M-mundo).
- Para completar el modelo establecemos reglas y ecuaciones matemáticas que relacionan estos objetos.

Modelización determinista

- A menudo establecemos modelos matemáticos que establecen relaciones deterministas entre nuestros objetos. Estos son frecuentes en física, por ejemplo.
- Esto se hace normalmente mediante ecuaciones diferenciales o ecuaciones en derivadas parciales.
- Por ejemplo, se usan modelos deterministas de la Atmósfera para predecir el tiempo.
- O utilizamos modelos matemáticos de reacciones químicas o incluso de procesos biológicos.

Modelización probabilística

- Pero a menudo necesitamos modelizar procesos del mundo de los que no disponemos toda la información o incluso en los que podría haber una aleatoriedad intrínseca, como puede que así sea en el mundo cuántico.
- Entonces tenemos que establecer relaciones probabilísticas entre variables o utilizar procesos estocásticos que predigan la evolución de otras variables.
- La realidad es que casi nunca tenemos toda la información disponible y esto hace que la modelización probabilística sea cada vez más aplicada en todo tipo de entornos.
- Incluso en Física, hay una tendencia imparable a utilizar modelos estadísticos de la realidad.

Modelización probabilística. Ejemplos

- La predicción del boson de Higgs se basó en la creación de modelos estadísticos.
- A veces, se puede complementar con la modelización determinista. Por ejemplo, en Meteorología utilizamos modelos deterministas de la atmósfera y también utilizamos modelos estadísticos que predicen fenómenos meteorológicos.
- En ciencias sociales y sanitarias practicamente todos los modelos matemáticos que se utilizan son modelos estadísticos.
- Hoy en día, que estamos en la era de los datos, los modelos estadísticos se han convertido en la base del estudio científico de todo lo que nos rodea, cobrando una importancia suprema en la modelización matemática de nuestro mundo.

Variables de un modelo

En general, en un modelo, lo que buscamos es predecir una o más variables en base a los valores que toman otras variables. Por ejemplo, puedo querer predecir la nota de unos alumnos en base a una serie de variables como las horas que estudian, la dedicación que tienen en el campus virtual, etc. O puedo querer predecir el peso de una persona en base a su altura.
Tenemos dos tipos de variables:

- **Variables dependientes, explicadas o de respuesta**. Son las variables que queremos predecir. En el ámbito específico del Machine Learning, las llamaremos **targets**.
- **Variables independientes, explicatorias, predictoras o regresoras**. Las variables que utilizamos para predecir las variables independientes. En el ámbito específico del Machine Learning, las llamaremos **features**.

Origen de los modelos

- Podemos llegar a formular un modelo determinista o probabilista en base a diversas razones.
- Por ejemplo, en Física a menudo establecemos modelos mediante el estudio de las bases físicas del fenómeno que estemos estudiando.
- Es el caso del lanzamiento de una moneda. En base a la geometría simétrica de la moneda, establecemos un modelo probabilístico que predice una probabilidad igual para ambos resultados.
- Sin embargo, hoy en día, gracias a que contamos con cantidades enormes de datos, ha tomado un enorme peso que construyamos nuestros modelos probabilísticos en base al comportamiento estadístico previo del fenómeno, sin tener en cuenta consideraciones teóricas de su comportamiento. Esto es lo que llamaremos **Aprendizaje Estadístico**.

Índice

Del Aprendizaje humano al Aprendizaje automático

- Las personas a menudo aprendemos a base de ver muchos casos y reconocer patrones de manera intuitiva y subconsciente.
- Por ejemplo, aprendemos qué es una mesa a base de ver muchas mesas diferentes, sin que necesitemos estudiar su definición.
- Cuando observamos una mesa observamos muchas de sus características, como el número de patas o el color de la mesa.
- Al observar más mesas aprendemos sin darnos cuenta que el color no tiene nada que ver con el hecho de que sea una mesa, pero su número de patas sí.

Machine Learning

- El Aprendizaje automático o Aprendizaje estadístico, normalmente conocido como Machine Learning, en su modalidad de aprendizaje supervisado, consiste en observar una gran cantidad de individuos o casos en los que se relacionan variables y aprender de ellos.
- Cada individuo contará con variables independientes llamadas características (**features**) y variables que han de predecirse (**targets**).
- Los diferentes algoritmos de Machine Learning aprenden de los individuos y posteriormente son capaces de predecir el valor de los targets a partir de nuevos individuos en los que sólo se dispone las features.

Ejemplo Machine Learning

A continuación vemos un ejemplo dónde los individuos son estudiantes de Matemáticas brasileños.

- La base de datos original nos proporciona una serie de features para cada alumno junto con 3 variables target: las notas de los tres trimestres.
- A partir de esta información un algoritmo de Machine learning aprende y es capaz de decirnos que features son significativas a la hora de predecir los targets.
- Además, cuando proporcionemos nuevos alumnos nos podrá predecir su nota.
- Todo esto nos podrá servir para actuar sobre el sistema y mejorarlo.

Ejemplo: Descripción variables (features - target)

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

Ejemplo Machine Learning. Datos

	scho	sex	age	addr	fams	Pstat	Medu	Fedu	Mjob	Fjob	reason	guardian	trave	stud	failu	scho	fams	paid	activ	nurs	high	inter	roma	famr	freet	goou	Dalc	Walc	heal	abse	G1	G2	G3
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	6	5	6	6
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	4	5	5	6
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no	yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	yes	yes	yes	yes	yes	yes	3	2	2	1	1	5	2	15	14	15
5	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	yes	no	yes	yes	no	no	4	3	2	1	2	5	4	6	10	10
6	GP	M	16	U	LE3	T	4	3	services	other	reputatior	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	4	2	1	2	5	10	15	15	15
7	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11
8	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	6	6	5	6
9	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	yes	no	yes	yes	yes	no	4	2	2	1	1	1	0	16	18	19
10	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	5	1	1	1	5	0	14	15	15
11	GP	F	15	U	GT3	T	4	4	teacher	health	reputatior	mother	1	2	0	no	yes	yes	no	yes	yes	yes	no	3	3	3	1	2	2	0	10	8	9
12	GP	F	15	U	GT3	T	2	1	services	other	reputatior	father	3	3	0	no	yes	no	yes	yes	yes	yes	no	5	2	2	1	1	4	4	10	12	12
13	GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	0	no	yes	yes	yes	yes	yes	yes	no	4	3	3	1	3	5	2	14	14	14
14	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2	0	no	yes	yes	no	yes	yes	yes	no	5	4	3	1	2	3	2	10	10	11
15	GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no	yes	no	no	yes	yes	yes	yes	4	5	2	1	1	3	0	14	16	16
16	GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1	0	no	yes	no	no	yes	yes	yes	no	4	4	4	1	2	2	4	14	14	14
17	GP	F	16	U	GT3	T	4	4	services	services	reputatior	mother	1	3	0	no	yes	yes	yes	yes	yes	yes	no	3	2	3	1	2	2	6	13	14	14
18	GP	F	16	U	GT3	T	3	3	other	other	reputatior	mother	3	2	0	yes	yes	no	yes	yes	yes	no	no	5	3	2	1	1	4	4	8	10	10
19	GP	M	17	U	GT3	T	3	2	services	services	course	mother	1	1	3	no	yes	no	yes	yes	yes	yes	no	5	5	5	2	4	5	16	6	5	5
20	GP	M	16	U	LE3	T	4	3	health	other	home	father	1	1	0	no	no	yes	yes	yes	yes	yes	no	3	1	3	1	3	5	4	8	10	10

Machine Learning. Otros ejemplos.

- Predecir la calidad del vino a partir de parámetros anuales de lluvia, tipo de uva, etc.
- Predecir la probabilidad de que un determinado individuo abra un email de publicidad que se le envía.
- Predecir que película o canción te puede interesar a partir de las películas o canciones que has visto/escuchado anteriormente.

Algoritmos de Machine Learning.

- Regresión lineal: El algoritmo de Machine Learning más clásico (y probablemente el más usado). Puede funcionar con cantidades reducidas de datos.
- Random forest: Son árboles de decisión.
- Deep Learning: Redes neuronales. El ideal para reconocimiento de imágenes, traducción de textos, etc. Sólo funciona si se dispone de una enorme cantidad de datos.
- Hay decenas de otros algoritmos de Machine Learning, incluidos montones de variaciones sobre estos tres.

Machine Learning: Regresión lineal

- En esta asignatura vamos a estudiar el principal algoritmo de Machine Learning: La regresión lineal.
- Puede funcionar con cantidades reducidas de datos.
- Puede utilizar variables independientes cualitativas mediante la técnica «One Hot Encoding».
- Podemos predecir variables cualitativas mediante funciones logísticas.

Índice

Regresión y correlación

Definición

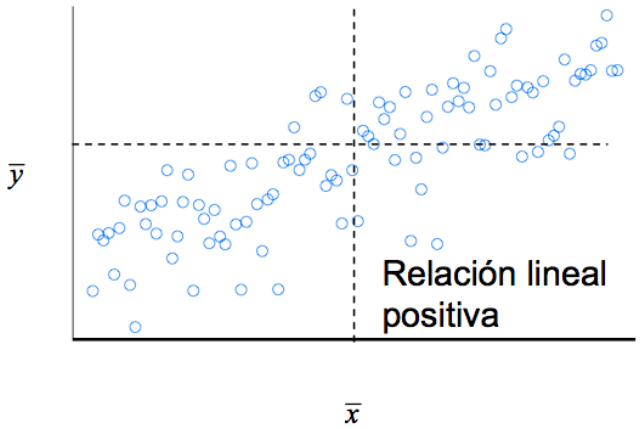
Correlación *es una medida del grado de dependencia entre las variables. La* **regresión** *pretende encontrar un modelo aproximado de la dependencia entre las variables.*

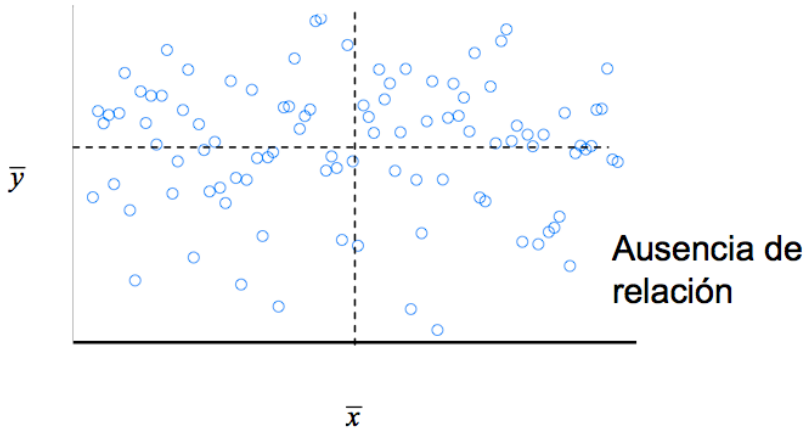
Representando los datos de la muestra de la variable bidimensional obtenemos una nube de puntos. Se llama *línea o curva de regresión* a la función que mejor se ajusta a esa nube de puntos.

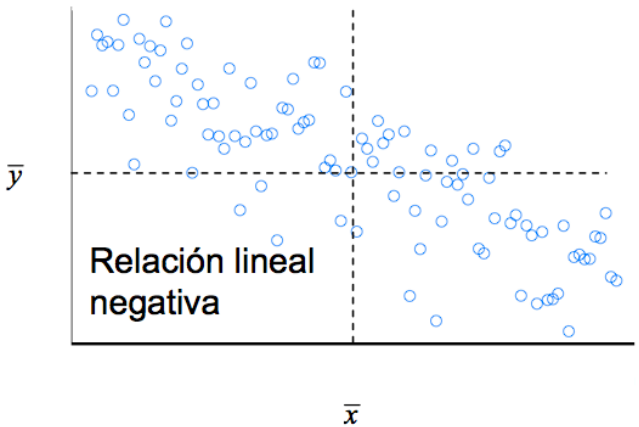
Si todos los valores de la variable satisfacen la ecuación calculada, se dice que las variables están perfectamente correladas. La ecuación de la curva de regresión nos permite predecir valores desconocidos.

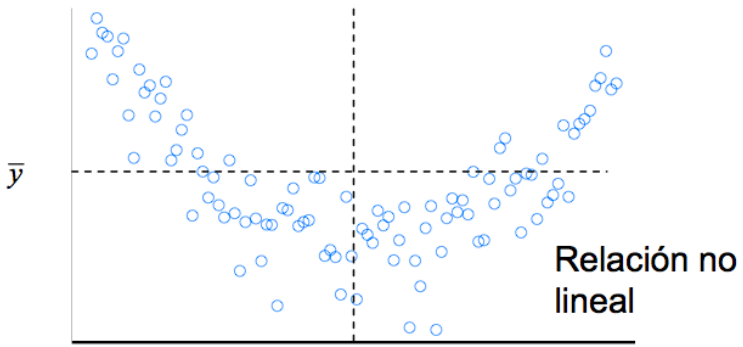
Regresión y correlación

El diagrama de dispersión muestra el tipo de relación existente:



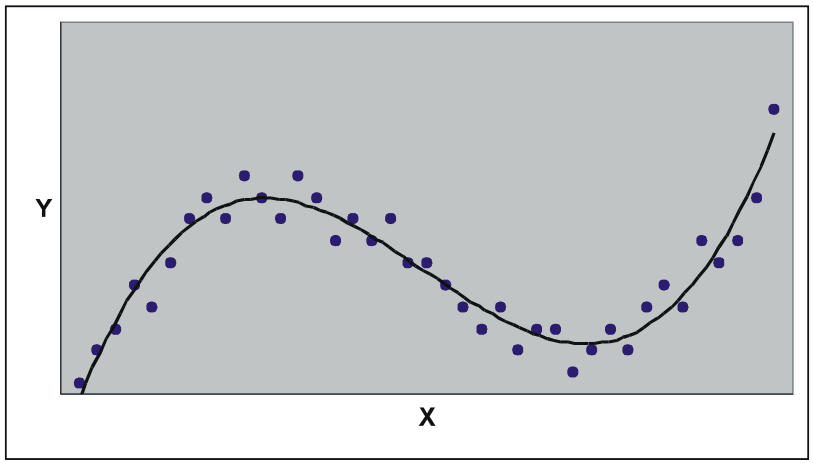






Curva de regresión

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



Al representar la curva de regresión y la nube de puntos conjuntamente, se puede observar la bondad del ajuste.

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 encontrar 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:**

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

$y_i^{est} = f(x_i)$ 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 $\left(ECM = \frac{\sum_i e_i^2}{N}\right)$.

Tipos de ajuste

El tipo de ajuste mínimos cuadrados está determinado por el tipo de función $y = f(x)$ elegido. El más utilizado es el ajuste lineal, pero a veces usamos otros tipos de ajustes.

- **Ajuste lineal:** $y = f(x) = a + bx$ (parámetros a y b).
- **Ajuste parabólico:** $y = a + bx + cx^2$ (parámetros a, b y c).
- **Ajuste hiperbólico:** $y = \frac{1}{a+bx}$ (parámetros a y b).
- **Exponencial:** $y = ae^{bx}$ (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: $F(a, b, \ldots) = \sum_i (y_i - f(x_i))^2 = \sum_i e_i^2$. 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 $Y/_{x_i}$.

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 $y = a + bx$ que mejor se ajuste a esos datos en el sentido 'mínimos cuadrados', es decir, que minimice la función:

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

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

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

Ecuaciones normales recta regresión Y/X:

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

Ajuste de la recta X/Y

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

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

Ahora los parámetros a' y b' deberan satisfacer las ecuaciones:

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

Ecuaciones normales recta regresión X/Y:

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

Ajuste lineal forma matricial

Los sistemas de ecuaciones normales, en forma matricial, para el caso de la regresión lineal son:

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

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

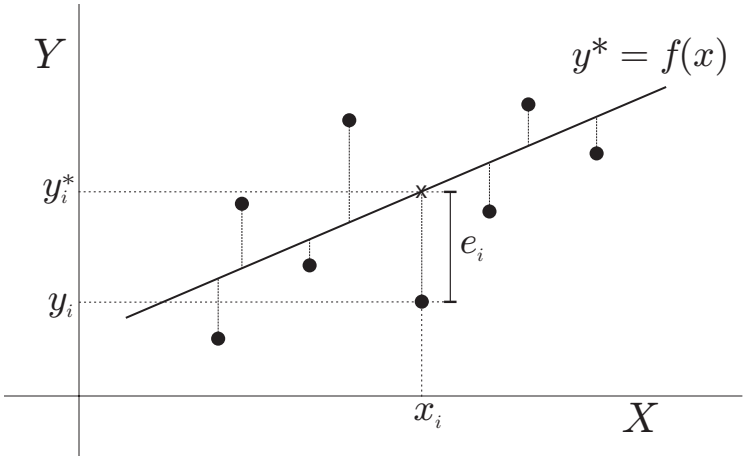
Recta de X sobre Y: (x=a'+b'y)

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

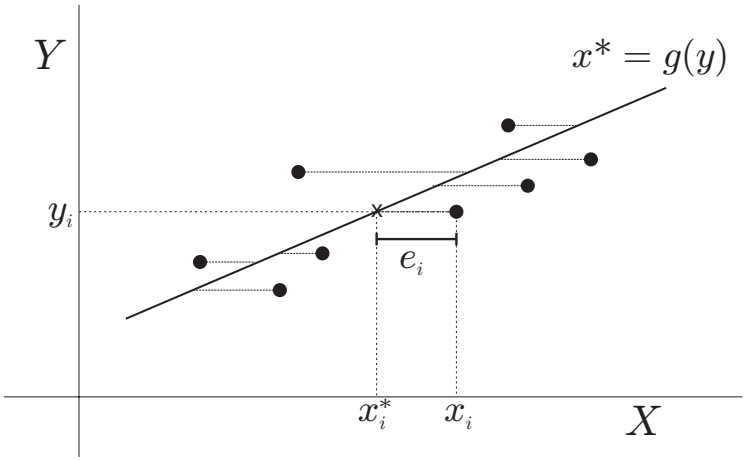
NOTA: Las ecuaciones normales pueden ser fácilmente adaptadas para los casos de disponer de datos con frecuencias.

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 $G = \sum_i (e_i')^2 = \sum_i (x_i - x_i^*)^2$



Ajuste lineal. Desarrollo

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** $G = (\bar{x}, \bar{y})$ **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:

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

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

Ajuste lineal. Ecuaciones

Así que las ecuaciones de la regresión lineal quedan así:

RECTA Y/X

$$y = \bar{y} + \frac{Cov(x,y)}{s_N^2(X)}(x - \bar{x})$$

RECTA X/Y

$$x = \bar{x} + \frac{Cov(x,y)}{s_N^2(Y)}(y - \bar{y})$$

Coeficiente de correlación lineal de Pearson:

Definición

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

$$\rho = r = \frac{Cov(x, y)}{s_N(X)s_N(Y)} \qquad (-1 \leq r \leq 1)$$

Significado: La correlación mide la magnitud y la dirección de la dependencia lineal.

- **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).

Error cuadrático medio. 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 **ECM: Error Cuadrático Medio** *a la media de los residuos al cuadrado. También se lo conoce como* **Varianza residual** *o* **MSE: Mean of Square Errors**.

$$\mathrm{MSE} = \frac{\sum_i e_i^2}{N}$$

Definición

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

$$R^2 = 1 - \frac{\mathrm{MSE}}{V(y)}$$

Discusión

El coeficiente de determinación R^2, en el caso de que se haya efectuado un ajuste por mínimos cuadrados verifica $0 \leq R^2 \leq 1$.

Definición

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

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

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

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:

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

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

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

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}.

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

Para comprobarlo los pasos son:

- $\bar{y_i^{est}} = \bar{y}$, pues ambos valen $a + b\bar{x}$
- $\sum_i (y_i - \bar{y})(y_i^{est} - \bar{y}) = 0$ haciendo uso de $\sum_i e_i = 0$ y $\sum_i e_i x_i = 0$

Simplificación varianza residual caso lineal

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

$$\mathrm{V_e} = b^2 s_N^2(X) = \left(r \frac{s_N(Y)}{s_N(X)} \right)^2 s_N^2(X) = \mathrm{r^2 V(y)}$$

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

- $\mathrm{R^2} = \mathrm{r^2}$
- $\mathrm{MSE} = (1 - \mathrm{r^2})\mathrm{V(y)}$

Índice

Planteamiento

Ahora nos vamos a ocupar de calcular la mejor función lineal para predecir una variable a partir de varias variables predictoras $X_1, \ldots, X_k$.
Supongamos que tenemos un conjunto de datos

$$\{(x_{i1}, \ldots, x_{ik}, y_i) : i = 1 \ldots N\}$$

Nuestro objetivo será encontrar la mejor función lineal de la forma:

$$y^* = a_0 + a_1 x_1 + \ldots + a_k x_k$$

que aproxime nuestros datos de la mejor manera. Al igual que en el caso lineal simple, buscaremos la mejor función lineal que minimice el error cuadrático medio.
Dado un dato $(x_{i1}, \ldots, x_{in}, y_i)$, buscamos aproximar $y_i \approx y_i^*$, donde y_i^* se calcula como $y_i^* = a_0 + a_1 x_{i1} + \ldots + a_n x_{ik}$. Definimos el **error residual** E como el error cometido por el ajuste planteado:

$$E = Y - Y^* = Y - a_0 - a_1 x_{i1} - \ldots - a_n x_{ik},$$

Minimización del vector residuo

El problema queda resuelto al encontrar los coeficientes de la regresión

$$A = (a_0, a_1, \ldots, a_k)^t .$$

que minimizan el error cuadrático medio (MSE) del vector residuo E. Podemos ver este MSE como una función de A:

$$f(A) = MSE = \frac{1}{N}\sum_{i=1}^{N} e_i^2, \qquad e_i = y_i - a_0 - a_1 x_{i1} - \ldots - a_n x_{ik}$$

Nos ocuparemos entonces, de minimizar la función $f(A)$, es decir, de encontrar los coeficientes (a_j) óptimos, que minimizan el error cuadrático medio de la variable E.

Notación

En primer lugar, vamos a organizar los datos planteando el problema en forma matricial. Dado el conjunto de datos $\{(x_{i1}, \ldots, x_{ik}, y_i)\}_{i=1}^{N}$:

$$M = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{N1} & \ldots & x_{Nk} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

Minimización del vector residuo

Podemos escribir entonces la ecuación como:

$$Y^* = M \cdot A$$

donde

$$Y^* = \begin{pmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_N^* \end{pmatrix}, \quad A = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix}$$

Minimización del vector residuo

Con esta notación, vamos entonces a minimizar el error cuadrático medio del vector residuo. Es decir, buscamos el mínimo:

$$\min_A [f(A)] = \min_A \left[(Y - Y^*)^t (Y - Y^*) \right] =$$

$$= \min_A \left[(Y - M \cdot A)^t (Y - M \cdot A) \right] =$$

$$= \min_A \left[Y^t Y - 2 (M \cdot A)^t + (M \cdot A)^t M \cdot A \right].$$

Sea $g(A) = Y^t Y - 2 (M \cdot A)^t + (M \cdot A)^t M \cdot A$,

$$\frac{\partial g}{\partial A}(A) = 0 \rightarrow -2M^t \cdot Y + 2M^t \cdot M \cdot A = 0$$

Regresión por mínimos cuadrados

De este modo, llegamos a las ecuaciones normales y a la solución de mínimos cuadrados de la regresión:

Ecuaciones Normales

$$\left(M^t \cdot M\right) \cdot A = M^t \cdot Y$$

Solución

$$A = \left(M^t \cdot M\right)^{-1} \cdot M^t \cdot Y$$

Puede demostrarse que la matriz $\left(M^t \cdot M\right)$ es siempre simétrica.

Índice

Ajuste exponencial $y = ae^{bx}$

$y = ae^{bx}$ (introducimos logaritmos) $\Rightarrow$

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

Llamando: $Y = ln(y)$, $A = ln(a)$ obtenemos: $Y = A + bx$. Podemos ajustar una recta a $\{(ln(y_i), x_i)\}$ obteniendo $A = ln(a)$, $(a = e^A)$ y b que sustituiremos en $y = ae^{bx}$

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: Varianza residual y coeficiente de determinación.

Ejemplo ajuste exponencial

x_i	0	1	2	3	6	12
y_i	7	5	4	3.5	3	22.5
$Y_i = ln(y_i)$	1.9459	1.6094	1.3863	1.2528	1.0986	7.293
x_i^2	0	1	4	9	36	50
$x_i Y_i$	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: $\begin{cases} 7.293 = 5A + 12b \\ 14.732 = 12A + 50b \end{cases} \Rightarrow \begin{matrix} A = 1.7723 \\ b = -0.1307 \end{matrix}$

$\Rightarrow a = e^{1.7723} = 5.8846$ Obtenemos: $y = 5.8846e^{-0.1307x}$

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

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

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

$y = \frac{1}{a+bx}$ (inviertiendo) $\Rightarrow \frac{1}{y} = a + bx$ Llamando: $Y = \frac{1}{y}$, obtenemos: $Y = a + bx$.

Podemos ajustar una recta a $\{(\frac{1}{y_i}, x_i)\}$ obteniendo a, y b que sustituiremos en $y = \frac{1}{a+bx}$

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: Varianza residual y coeficiente de determinación.
¿Qué ajuste es mejor el exponencial, el hiperbólico o el lineal?

Ejemplo ajuste hiperbólico

x_i	0	1	2	3	6	12
y_i	7	5	4	3.5	3	22.5
$Y_i = \frac{1}{y_i}$	0.1429	0.2	0.25	0.2857	0.3333	1.2119
x_i^2	0	1	4	9	36	50
$x_i Y_i$	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

Ajustamos: $\begin{cases} 1.2119 = 5a + 12b \\ 3.5571 = 12a + 50b \end{cases} \Rightarrow \begin{matrix} a = 0.1690 \\ b = -0.0306 \end{matrix} \Rightarrow \mathrm{y} = \frac{1}{0.169 - 0.0306\mathrm{x}}$

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

$\mathrm{V(x)} = 4.24, \quad \mathrm{cov} = \frac{41.5}{5} - \frac{12}{5}\frac{22.5}{5} = -2.5 \Rightarrow r = \frac{-2.5}{\sqrt{2}\sqrt{4.24}} = -0.8585 \Rightarrow \mathrm{r^2} = 0.7370$

Luego el mejor ajuste es el hiperbólico que explica el 86.34 % de la varianza de y.

NOTA: También se usa como criterio $SSE = \sum_i e_i^2$ que para el exponencial, hiperbólico y lineal dan respectivamente: $SSE_e = 1.8775$, $SSE_h = 1.4276$ y $SSE_L = 2.63$.

Índice

Variables cualitativas dicotómicas

- A menudo tenemos variables cualitativas que influyen en el resultado de la variable explicada.
- Pueden ser, por ejemplo, variables dicotómicas como el sexo de la persona o variables que toman más de dos valores como la comunidad autónoma a la que pertenece.
- En el caso de una variable dicotómica, basta con asignar el valor 0 a uno de los valores y 1 al otro para que podamos usarla como una variable más en la regresión lineal múltiple (es indiferente a cual le asignemos un valor u otro).
- En el caso de variables que toman más de dos valores, si es una variable ordinal (cuyos valores tienen un orden) podríamos tomarla directamente como una variable más, pero normalmente no es una buena idea porque la asignación de unos números y no otros es algo arbitrario.

Variables cualitativas ordinales

- Por ejemplo, una variable que sea el nivel de estudios: 0 sin estudios, 1: primaria, 2: secundaria, 3: bachillerato, 4: universidad.
- Si la tomamos directamente como variable explicada con esos valores, estamos suponiendo que su efecto va a ser el doble si la persona ha estudiado la universidad que secundaria y el doble si ha estudiado secundaria que primaria.
- ¿Porqué tendrían que ser esas relaciones a priori?. Es algo arbitrario.
- De hecho, podríamos hacer otra asignación: 0 sin estudios, 2: primaria, 3: secundaria, 4: bachillerato, 7: universidad.
- No hay ninguna razón para que una sea mejor que otra y nos ofrecerían modelos diferentes.
- Por esa razón, a no ser que tengamos una buena razón para mantener su orden, es mejor idea olvidarnos de él y tratarlas como variables cualitativas ordinarias.

One Hot Encoding. Variables dummy

- Cuando tenemos una variable cualitativa que toma más de dos valores, la solución es crear una variable **dummy** para cada uno de los posibles valores que puede tomar, excepto para uno, que tomamos como base.
- Esta técnica es la que conocemos como **One Hot Encoding**
- En nuestro anterior ejemplo, crearíamos 4 variables $D1$, $D2$, $D3$ y $D4$ que tomarían los siguientes valores:

	D1	D2	D3	D4
sin estudios	0	0	0	0
primaria	1	0	0	0
secundaria	0	1	0	0
bachillerato	0	0	1	0
universidad	0	0	0	1

Variables dummy. Valor base.

- Si asignáramos una variable dummy a cada uno de los posibles valores de la variable cualitativa, sin usar ninguno como valor base, el problema que tendríamos es que tendríamos una dependencia absoluta entre las variables y la matriz de regresión sería singular.
- Realmente es indiferente cuál de los valores tomemos como base, por que al final lo que nos da la correlación son una serie de hiperplanos paralelos, distanciados por las diferencias entre los coeficientes de las variables dummy.
- A pesar de ello, si suprimimos el término independiente, entonces sí podemos asignar una variable dummy a cada valor.
- Pero normalmente preferimos mantener el término independiente y asignar un valor base.

Índice

Otros algoritmos de Machine Learning.

- Random forest: Son árboles de decisión.
- Deep Learning: Redes neuronales. El ideal para reconocimiento de imágenes, traducción de textos, etc. Sólo funciona si se dispone de una enorme cantidad de datos.
- Hay decenas de otros algoritmos de Machine Learning, incluidos montones de variaciones sobre estos tres.

Índice

Selección de modelos

- Podemos obtener diferentes modelos que modelizan una variable. Por ejemplo, podemos obtenerlos por métodos estadísticos: regresión lineal, random forest, deep learning, etc. O incluso podemos generar modelos teóricos basados en características del fenómeno que estudiamos.
- Podemos querer comparar modelos en los que una o varias variables han sido transformadas mediante logaritmos, exponenciales, etc.
- A menudo querremos poder comparar modelos que involucren a un número diferente de variables, para elegir qué variables realmente son predictivas y cuales no.
- ¿Cómo compararlos y saber cuál es el mejor modelo?
- En general, la idea es poder medir la diferencia entre el valor real obtenido y el valor predicho y quedarnos con el modelo que de alguna manera minimice dicha diferencia.

El vector residuo y su norma

- Dada una nube de puntos $\{(x_i, y_i) : 1 \leq i \leq n\}$, llamamos **vector residuo** $\vec{e} = (e_i : 1 \leq i \leq n)$ con $e_i = y_i - y_i^{est}$. Es decir, e_i es el error cometido por el ajuste para la i-ésima observación.
- Para poder medir la calidad de un modelo necesitamos utilizar una norma que asigne un número real al vector residuo.
- Así, elegiremos el modelo que tenga una menor norma de su vector residuo.
- Dependiendo de lo que nos interese optimizar podríamos utilizar diferentes tipos de norma. Como una suma de valores absolutos, el máximo de los residuos, etc.
- Sin embargo, lo normal es utilizar normas equivalentes a la norma euclidiana como la media de los cuadrados o su raiz.

Suma cuadrática de errores, Error cuadrático medio y Raíz del error cuadrático medio

- Llamamos **Suma cuadrática de errores** a la suma de los residuos al cuadrado (**SSE: Sum of Squared Errors**).

$$SSE = \sum_{i=1}^{n} e_i^2$$

- Llamamos **ECM: Error Cuadrático Medio** a la media de los residuos al cuadrado (**MSE: Mean of Squared Errors**).

$$MSE = \frac{1}{n} SSE = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

- Llamamos **Raíz del error cuadrático medio** a la raíz del error cuadrático medio (**RMSE: Root of Mean of Square Roots**).

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^2}$$

Uso del error cuadrático medio y del coeficiente de determinación

- Si queremos comparar la calidad de dos modelos sobre el mismo conjunto de datos es idéntico comparar el SSE, MSE o el RMSE.
- Es decir, que el modelo que tenga el menor de cualquiera de esos estadísticos será el mismo modelo y será el que debemos elegir.
- Sin embargo, si queremos comparar dos modelos probados sobre conjuntos de datos de diferente tamaño, no podremos utilizar el SSE, porque éste depende del tamaño del conjunto de test.
- El RMSE, además, tiene las unidades de lo que estemos midiendo, por lo que su uso es más intuitivo.
- Si queremos tener una medida estándar de la calidad de un modelo, que incluso pueda comparar la calidad en poblaciones diferentes necesitaremos del **Coeficiente de Determinación**.

El RMSE es insuficiente como medida general.

- Si tomamos como predicción la media de la variable, de manera constante, para todas las muestras, su RMSE coincidirá con la desviación típica de la variable, o el MSE con la varianza.
- Eso sería algo así como el modelo mínimo que podemos hacer sin complicarnos la vida.
- Vemos que una manera de medir la calidad de un modelo es medir cuanto somos capaces de mejorar ese modelo mínimo. Es decir, cuanto somos de hacer bajar el MSE respecto de la varianza.
- El RMSE en sí no nos da una idea real de la calidad de un modelo, a no ser que lo comparemos con la desviación típica.
- Por ejemplo, si tenemos un modelo para predecir la altura de una población y su RMSE es de 10 cm, en una población donde la desviación es de 100 cm, estamos haciendo una muy buena predicción, pero si ese modelo es en una población con una desviación de 12 cm, el modelo claramente no es un buen modelo.

El Coeficiente de determinación.

Definición

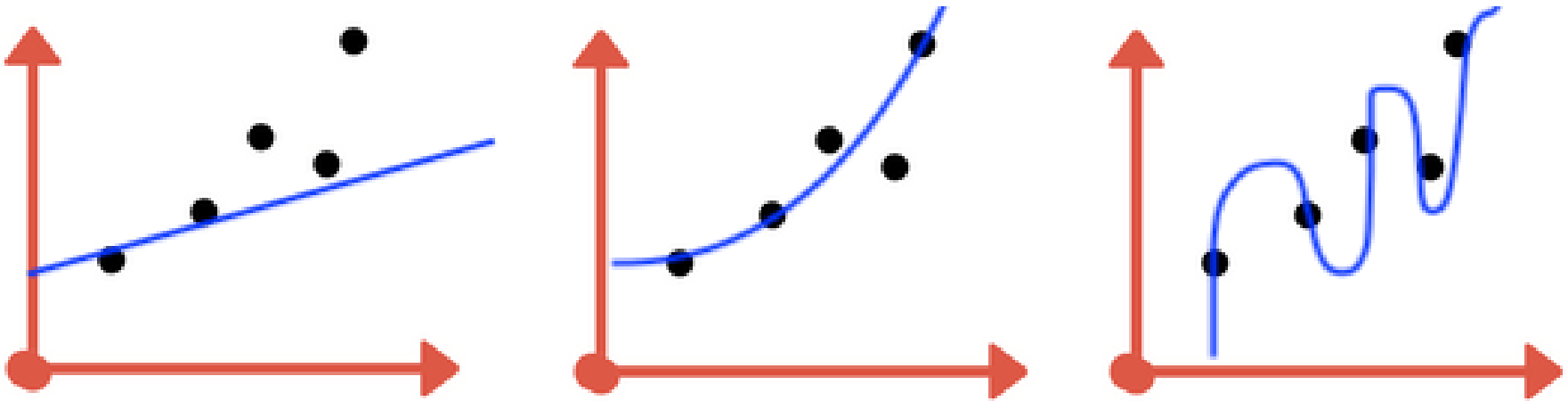
Llamamos **coeficiente de determinación** *a:*

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

- El coeficiente de determinación es una medida general de la calidad de predicción de un modelo.
- Vale 1 si la predicción es perfecta.
- Vale 0 para una predicción constante consistente en dar siempre la media de la variable.
- Si toma un valor negativo, es que el modelo es peor aún que ese modelo constante básico.
- En el caso de una regresión lineal ordinaria, coincide con el cuadrado del coeficiente de regresión de Pearson.

El problema del sobreajuste (Overfitting)

Ante una nube de puntos podemos preparar un modelo que pase muy exactamente por los puntos, pero normalmente con eso lo que haremos será introducir en el modelo fluctuaciones estadísticas, no estaremos recogiendo el modelo auténtico subyacente. Esto es lo que llamamos sobreajuste del modelo (overfitting). Si el número de parámetros del modelo se aproxima al número de muestras de las que disponemos, es muy probable que estemos generando sobreajuste.



Separación de conjuntos de entrenamiento y test

- La principal manera en la que podemos evitar el sobreajuste es entrenando con un **conjunto de entrenamiento** y testeando la calidad del modelo con un **conjunto de test**.
- Típicamente, podemos separar una muestra aleatoria de un 20 % o un 30 % de los datos, para testear y dejamos como conjunto de entrenamiento el resto.
- Por ejemplo, si nuestro modelo es una regresión lineal, eso significa que obtendremos los coeficientes de la regresión lineal con el conjunto de entrenamiento y calcularemos el coeficiente de determinación utilizando sólo el conjunto de test.

Selección de variables explicativas

- En principio, cuando añadimos más variables independientes a un modelo (lineal o no) normalmente siempre mejoraremos, aunque sea sólo un poco, la calidad del modelo y, en concreto, su R^2.
- Sin embargo, si la variable que estamos añadiendo mejora muy poco la capacidad predictiva, es mejor eliminarla, ya que estaremos complicando el modelo innecesariamente y generando probablemente sobreajuste.
- Para poder solucionar este problema utilizaremos el coeficiente de determinación ajustado $\bar{R}^2$, que es una variación del coeficiente de determinación, que tiene en cuenta el número de predictores que estamos utilzando.
- Así que si tengo muchas variables independientes y tengo que decidir cuales uso como variables predictoras, podré elegir quedarme con la combinación de ellas que genere un mayor valor de $\bar{R}^2$.

Coeficiente de determinación ajustado

Definición

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

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

donde n es el número de muestras y k es el número de predictores que estamos utilizando.

- El coeficiente de determinación ajustado siempre es menor que el coeficiente de determinación.
- Podemos utilizarlo tanto en el contexto de la regresión lineal, como en el contexto de la evaluación de cualquier otro tipo de modelos.
- Es fundamental usarlo (u otras alternativas) cuando se trata de elegir entre modelos con un número diferente de variables explicatorias.

Procedimiento general de selección de modelos

Resumiendo, el procedimiento general de selección de modelos es:

- Separar los datos en un conjunto de entrenamiento y otro de test.
- Entrenar diferentes modelos con el conjunto de entrenamiento.
- Calcular el coeficiente de determinación, el coeficiente de determinación ajustado u otras alternativas para cada modelo sobre el conjunto de test.
- Elegir el modelo con mejor resultado.

Esta manera de trabajar nos permite reducir enormemente la posibilidad de sobreajuste en él cálculo de los parámetros de los modelos. Evitando que las fluctuaciones estadísticas del conjunto de entrenamiento se incluyan en el modelo. Sin embargo, si evaluamos muchos modelos, es todavía posible que el modelo seleccionado lo sea debido a fluctuaciones estadísticas del conjunto de test.

Conjunto de validación para el modelo seleccionado

- Para medir de manera más objetiva la calidad del modelo resultante, puede ser conveniente separar de los datos originales un tercer conjunto de datos, que llamaremos **conjunto de validación**.
- Una vez realizado el procedimiento general descrito anteriormente, mediremos un resultado final de calidad (por ejemplo, con el coeficiente de determinación) utilizando el conjunto de validación.
- Por ejemplo, podemos repartir los datos originales entre los conjuntos de entrenamiento, test y validación en una proporción $50\,\% - 25\,\% - 25\,\%$, $60\,\% - 20\,\% - 20\,\%$ o cantidades similares.

Otras técnicas de comparación de modelos

Hay otros dos importantes enfoques a la hora de comparar modelos:

- Realizar contrastes de hipótesis sobre que unos modelos sean mejores que otros o sobre que ciertas variables sean significativas.
- Medidas de calidad similares al coeficiente de información ajustado, pero basados en la teoría de la información. Son algo más complejos de calcular, pero de hecho se suelen considerar más adecuados. Fundamentalmente, hay tres posibilidades:
  - El Criterio de información de Akaike (AIC).
  - El Criterio de información bayesiano (BIC).
  - El Criterio de Schwarz (SBIC, porque es una ampliación de BIC).