

Apuntes de ESTADÍSTICA

Regresión y correlación



Sixto Sánchez Merino
Dpto. de Matemática Aplicada
Universidad de Málaga



Mi agradecimiento a los profesores Carlos Cerezo Casermeiro y Carlos Guerrero García, por sus correcciones y sugerencias en la elaboración de estos apuntes.


Apuntes de Estadística

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.

Usted es libre de:

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

Bajo las condiciones siguientes:

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

Capítulo 2

Regresión y correlación

En el capítulo anterior se proporcionan las herramientas para describir una población en función de los datos de una variable obtenidos en una muestra. En este capítulo se considera la observación conjunta de dos caracteres en el individuo. Los pares de datos obtenidos constituyen muestras de una variable estadística bidimensional. El objetivo del tema será describir la población a partir de las variables estudiadas, establecer la posible relación entre ellas, determinar un modelo matemático que represente dicha relación y poder cuantificar la bondad de dicho modelo.

2.1. Distribuciones bidimensionales

Para el estudio conjunto de dos caracteres de la población, consideraremos la variable X que presenta las modalidades x_1, x_2, \dots y la variable Y con modalidades y_1, y_2, \dots . Los distintos valores que podemos obtener al observar conjuntamente las dos variables constituyen una muestra de la variable bidimensional (X, Y) . La distribución de frecuencias de esta nueva variable viene determinada por las parejas (x_i, y_j) de valores observados junto a sus correspondientes frecuencias absolutas (n_{ij}) , que indican el número de veces que se repiten dichas parejas. Análogamente al caso unidimensional, se pueden definir las frecuencias relativas (f_{ij}) que indican la proporción de veces que se repite la pareja de valores (x_i, y_j) sobre el total de datos de la muestra. Si N es el tamaño de la muestra, entonces f_{ij} se calcula mediante el cociente n_{ij}/N .

Ahora mostramos distintas formas de representar la distribución de frecuencias haciendo uso de las tablas y las gráficas. La naturaleza de las variables y el tamaño o la variabilidad de los datos de la muestra determinará el procedimiento más adecuado para su representación.

2.1.1. Representación tabular

La distribución de frecuencias de una variable bidimensional se puede mostrar en forma de tabla que contiene los distintos pares de valores la variable junto a sus frecuencias. Independientemente de la naturaleza discreta o continua de las variables, consideramos tres casos en función de la cantidad y variedad de datos de la muestra.

Cuando el número de observaciones es pequeño, los valores de las variables se pueden presen-

tar en forma de *tabla simple* con dos filas (o dos columnas) conteniendo las parejas de valores. Por ejemplo, la tabla

variable X	x_1	x_2	\dots	x_N
variable Y	y_1	y_2	\dots	y_N

representa los datos de la muestra $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ de la variable (X, Y) .

Ejemplo 2.1 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

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

con los pares de datos de la muestra. □

Cuando el número de observaciones es grande, pero corresponden a pocas parejas (modalidades) distintas, los valores de las variables se pueden presentar en forma de *tabla simple* con tres filas o columnas conteniendo las parejas de valores y sus frecuencias correspondientes. Por ejemplo, la tabla de la figura 2.1 representa la distribución de frecuencias de los datos de una muestra de tamaño N que contiene k tipos de pares de datos (x_i, y_i) observados n_i veces cada uno, con $i = 1, 2, \dots, k$.

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

Figura 2.1: Tabla estadística de frecuencias

Ejemplo 2.2 Una empresa de software somete a sus programas a determinados controles para depurar errores durante su desarrollo. El número de controles efectuados disminuye los posibles errores finales pero incrementa los costes de producción. Para determinar la influencia de estas variables se observan conjuntamente el número de controles C efectuados a un software y el número de errores graves detectados D al finalizar su desarrollo obteniéndose la muestra: $(0,0)$, $(0,1)$, $(1,1)$, $(0,1)$, $(1,1)$, $(0,1)$, $(0,1)$, $(1,1)$, $(1,0)$, $(1,0)$, $(1,1)$, $(1,1)$, $(1,1)$, $(0,0)$, $(1,0)$, $(1,0)$, $(2,1)$, $(1,1)$, $(1,1)$, $(2,1)$. Utilizar una tabla estadística para representar la distribución de frecuencias.

Se ordenan los valores de la muestra y se agrupan los que corresponden al mismo par de modalidades. Después, se construye una tabla donde se representan los distintos pares de valores junto a su frecuencia absoluta y relativa.

C	D	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	1	2	0'1
		20	

Por ejemplo, la fila 4 indica que hemos observado 8 veces (frec. absoluta) que con 1 control (C) se detecta 1 error (D), y esto supone el 40 % (frec. relativa) de los casos observados. \square

Cuando hay un gran número de observaciones y de modalidades distintas, los valores de las variables se disponen en una *tabla de doble entrada*, donde los valores de cruce de cada fila y columna representan la frecuencia de la correspondiente pareja de valores. En la tabla de la figura 2.2, consideramos la variable X con k modalidades x_1, x_2, \dots, x_k y la variable Y con p modalidades y_1, y_2, \dots, y_p .

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

Figura 2.2: Tabla de doble entrada

Las distintas modalidades de las variables X e Y se ordenan en los márgenes izquierdo y superior respectivamente. La frecuencia absoluta del par (x_i, y_j) se denomina n_{ij} y se sitúa en la intersección de la fila y columna correspondiente. También se puede construir otra tabla estadística a partir de las frecuencias relativas, sin más que dividir por N las frecuencias absolutas de tal manera que

$$f_{ij} = \frac{n_{ij}}{N} \quad \text{siendo} \quad N = \sum_{i=1}^k \sum_{j=1}^p n_{ij}$$

En el margen derecho de la tabla se sitúan las frecuencias $(n_{i\cdot})$ de los valores de la variable X , que se calculan sumando por filas. En el margen inferior se localizan las frecuencias $(n_{\cdot j})$ de los valores de la variable Y , que se calculan sumando por columnas. Como veremos, los valores de las variables y sus frecuencias, representadas al margen, determinan las *distribuciones marginales*; mientras que los valores en interior de la tabla constituyen la denominada *distribución conjunta*.

Ejemplo 2.3 Representar en tablas de doble entrada las distribuciones de frecuencias absolutas y relativas para los datos del ejemplo 2.2 de la página 54.

A partir de los datos de la muestra, o de la tabla de frecuencias, construimos la tabla de doble entrada, situando en el margen izquierdo y superior las distintas modalidades de las variables X e Y respectivamente, y en el interior de la tabla, se escribe las frecuencias para cada par de valores.

C	D	n_i	f_i		n_{ij}	0	1	D		f_{ij}	0	1	D
0	0	2	0'1		0	2	4	6		0	0'1	0'2	0'3
0	1	4	0'2		1	4	8	12		1	0'2	0'4	0'6
1	0	4	0'2	\Rightarrow	2	0	2	2		2	0	0'1	0'1
1	1	8	0'4		C	6	14	20		C	0'3	0'7	1
2	1	2	0'1										
		20											

Observe la tabla estadística (izquierda), de la que se derivan las dos tablas de doble entrada, una para las frecuencias absolutas (centro) y otra para las frecuencias relativas (derecha). \square

Este tipo de representación en forma de tabla de doble entrada también se utiliza si estamos interesados en agrupar los datos en intervalos. En este caso, recuperamos los conceptos de clase, amplitud y marca, introducidos en el tema anterior.

Ejemplo 2.4 Organizar los siguientes datos de la variable (X, Y) en una tabla de doble entrada: $(1'72, 63)$, $(1'70, 75)$, $(1'70, 68)$, $(1'68, 70)$, $(1'75, 74)$, $(1'69, 72)$, $(1'71, 67)$, $(1'69, 69)$, $(1'67, 70)$, $(1'74, 74)$, $(1'76, 71)$, $(1'70, 70)$, $(1'69, 66)$, $(1'66, 60)$, $(1'78, 74)$, $(1'74, 69)$, $(1'70, 65)$, $(1'69, 71)$, $(1'71, 73)$, $(1'78, 69)$

Agrupando los valores de las variables X e Y en intervalos de amplitud 5 construimos la tabla de doble entrada

$X \backslash Y$	[60, 65]	(65, 70]	(70, 75]	
$(1'65, 1'70]$	2	6	3	11
$(1'70, 1'75]$	1	2	3	6
$(1'75, 1'80]$	0	1	2	3
	3	9	8	20

que contiene las frecuencias absolutas de los intervalos correspondientes. \square

Como hemos comentado, las tablas simples se utilizan para representar distribuciones de frecuencias con muchos datos de pocas modalidades distintas. Por el contrario, las tablas de doble entrada resultan más apropiadas para representar distribuciones de frecuencias con muchos datos pertenecientes a un gran número de modalidades distintas. Sin embargo, en cualquier caso podemos utilizar indistintamente un tipo u otro de representación tabular. Así, en los ejemplos 2.2 de la página 54 y 2.3 de la página 55 hemos representado la misma distribución de frecuencias utilizando los dos tipos de tablas. Es importante saberlas utilizar indistintamente y construir una de ellas a partir de la otra.

2.1.2. Representaciones gráficas

La representación gráfica constituye una forma ordenada de presentar la distribución de frecuencias. Las representaciones gráficas más importantes para las distribuciones bidimensionales de caracteres cuantitativos son el diagrama de dispersión, el diagrama de frecuencias y el estereograma.

Diagrama de Dispersión. Consiste en la representación de los distintos pares de valores sobre unos ejes cartesianos. De esta forma, cada par viene representado por un punto del plano XY que forman la llamada *nube de puntos*. La frecuencia de cada par de puntos puede representarse utilizando distintos tamaños de puntos.

En la figura 2.3 se muestran dos diagramas de dispersión. El primero representa la nube de puntos correspondiente a los datos (sin agrupar) del ejemplo 2.4 de la página 56 y el segundo representa los datos del ejemplo 2.2 de la página 54 donde el tamaño de los puntos es proporcional a su frecuencia.

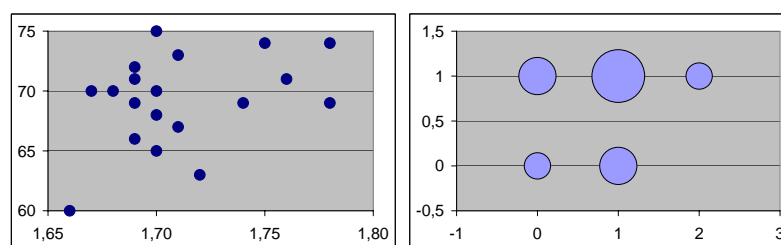


Figura 2.3: Diagramas de dispersión

Diagrama de frecuencias. Este tipo de representación está indicado para el caso discreto y es análogo a los diagramas de barras o puntos en el caso unidimensional. Consiste en una representación en tres dimensiones donde el plano base corresponde a los valores de las variables y la altura representa las frecuencias. El resultado es una serie de barras verticales apoyadas en los puntos del plano XY correspondientes a los valores (x_i, y_j) y cuya altura representa la frecuencia absoluta (n_{ij}) o relativa (f_{ij}) del par.

Este tipo de representación también se puede utilizar para representar distribuciones cuando las variables son cualitativas. En la figura 2.4 se representa mediante un diagrama de frecuencias los datos del ejemplo 2.2 de la página 54.

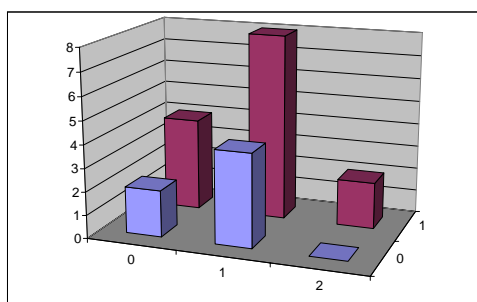


Figura 2.4: Diagrama de frecuencias

Estereograma. Se utiliza para representar aquellas distribuciones donde los datos se agrupan en intervalos y equivale al histograma para una variable. Se realiza análogamente al diagrama de frecuencias utilizando paralelepíedros, en vez de barras o puntos, cuya base son las regiones del plano correspondientes a los intervalos. En este caso, el volumen representa la frecuencia absoluta o relativa.

En la gráfica de la izquierda de la figura 2.3 de la página 57 se representaban los datos del ejemplo 2.4 de la página 56, en forma de nube de puntos. Ahora, en la figura 2.5 se muestran esos mismos datos, pero agrupados en intervalos.

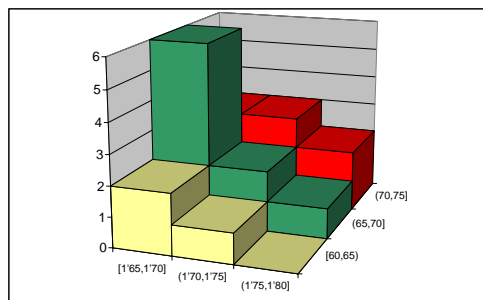


Figura 2.5: Estereograma

2.1.3. Distribuciones Marginales

La distribución de frecuencias bidimensional contiene la información conjunta de dos variables. Sin embargo, podemos estar interesados en estudiar una variable de manera aislada, sin considerar su relación con la otra. En este caso, debemos “separar” la información relativa a cada variable.

A partir de las distribuciones bidimensionales definimos las distribuciones marginales que son las distribuciones unidimensionales correspondientes a uno de los caracteres sin considerar el otro. Para obtenerlas basta con prescindir de la información de una de las variables eliminando los datos correspondientes.

Ejemplo 2.5 Calcular la distribución marginal de la variable C (número de controles efectuados a un software) del ejemplo 2.2 de la página 54 a partir de su tabla estadística de frecuencias.

Si eliminamos la columna correspondiente a la variable D y agrupamos las modalidades que sean iguales,

C	D	n_i	f_i		C	n_i	f_i		C	n_i	f_i
0	0	2	0'1		0	2	0'1		0	6	0'3
0	1	4	0'2		0	4	0'2		1	12	0'6
1	0	4	0'2	\Rightarrow	1	4	0'2	\Rightarrow	2	2	0'1
1	1	8	0'4		1	8	0'4			20	
2	1	2	0'1		2	2	0'1				
		20				20					

el resultado es la distribución de frecuencias unidimensional correspondiente a la variable C \square

Cuando la distribución se representa en una tabla de doble entrada, las distribuciones marginales aparecen “en el margen” de la tabla que contiene la suma por filas (o columnas) de los valores conjuntos de las variables. Para obtener estas distribuciones marginales sólo hay que prescindir de los valores de la variable en el interior de la tabla.

En la tabla de doble entrada de la figura 2.2 de la página 55, la distribución marginal de la variable X aparecen en el margen derecho de la tabla y cuenta con las modalidades x_1, x_2, \dots, x_k cuyas frecuencias absolutas $(n_{1\cdot}, n_{2\cdot}, \dots, n_{k\cdot})$ y relativas $(f_{1\cdot}, f_{2\cdot}, \dots, f_{k\cdot})$ correspondientes a cada modalidad se calculan sumando por filas:

$$n_{i\cdot} = \sum_{j=1}^p n_{ij} \quad \text{y} \quad f_{i\cdot} = \frac{n_{i\cdot}}{N} \quad \text{son las frecuencias marginales del valor } x_i \text{ de la variable } X.$$

Análogamente, en el margen inferior se observa la marginal de la variable Y que toma los valores y_1, y_2, \dots, y_p cuyas frecuencias absolutas $(n_{\cdot 1}, n_{\cdot 2}, \dots, n_{\cdot p})$ y relativas $(f_{\cdot 1}, f_{\cdot 2}, \dots, f_{\cdot p})$ correspondientes a cada modalidad se calculan sumando por columnas:

$$n_{\cdot j} = \sum_{i=1}^k n_{ij} \quad \text{y} \quad f_{\cdot j} = \frac{n_{\cdot j}}{N} \quad \text{son las frecuencias marginales del valor } y_j \text{ de la variable } Y.$$

Ejemplo 2.6 Calcular las distribuciones marginales de las variables C (número de controles efectuados a un software) y D (número de errores graves detectados), a partir de la tabla de doble entrada del ejemplo 2.3 de la página 55.

Para calcular la distribución marginal de la variable C se eliminan las dos columnas interiores de la tabla y permanece la columna de la derecha que contiene la suma por filas de los valores de las columnas eliminadas.

$$\begin{array}{c|cc|c} n_{ij} & 0 & 1 & D \\ \hline 0 & 2 & 4 & 6 \\ 1 & 4 & 8 & 12 \\ 2 & 0 & 2 & 2 \\ \hline C & 6 & 14 & 20 \end{array} \quad \Rightarrow \quad \begin{array}{c|c|c} C & n_i & f_i \\ \hline 0 & 6 & 0'3 \\ 1 & 12 & 0'6 \\ 2 & 2 & 0'1 \\ \hline & 20 & \end{array}$$

y para calcular la distribución marginal de la variable D se eliminan las tres filas interiores de la tabla y permanece la fila inferior que contiene la suma por columnas de los valores de las filas eliminadas.

$$\begin{array}{c|cc|c} n_{ij} & 0 & 1 & D \\ \hline 0 & 2 & 4 & 6 \\ 1 & 4 & 8 & 12 \\ 2 & 0 & 2 & 2 \\ \hline C & 6 & 14 & 20 \end{array} \quad \Rightarrow \quad \begin{array}{c|c|c} D & n_i & f_i \\ \hline 0 & 6 & 0'3 \\ 1 & 14 & 0'7 \\ \hline & 20 & \end{array}$$

En ambos casos, hemos añadido una columna correspondiente a las frecuencias relativas. \square

2.1.4. Distribuciones Condicionadas

Al igual que las marginales, las distribuciones condicionadas son también distribuciones unidimensionales. Surgen al considerar sólo aquellos valores de la muestra que presentan una determinada modalidad 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 .

En la tabla de doble entrada de la figura 2.2 de la página 55, la distribución condicionada del carácter X , respecto a la clase j del carácter Y corresponde a la columna j -ésima y cuenta con las modalidades x_1, x_2, \dots, x_k cuyas frecuencias absolutas $(n_1^j, n_2^j, \dots, n_k^j)$ aparecen directamente en la columna j -ésima $(n_{1j}, n_{2j}, \dots, n_{kj})$ de la tabla. Las frecuencias relativas $(f_1^j, f_2^j, \dots, f_k^j)$ correspondientes a cada modalidad se calculan dividiendo las absolutas entre el total de valores de X con la modalidad j , es decir, $n_{.j}$. Por tanto

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

Análogamente se puede definir la distribución condicionada del carácter Y , respecto a la modalidad i de X . Esta distribución considera los valores y_j con frecuencias:

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

Ejemplo 2.7 Determinar la distribución condicionada del carácter C respecto de la modalidad 1 del carácter D , a partir de la tabla de doble entrada del ejemplo 2.3 de la página 55.

Para determinar esta distribución condicionada, seleccionamos la segunda columna correspondiente a todos los valores de la variable C que corresponden al valor 1 de la variable D .

n_{ij}	0	1	D		C	n_i	f_i
0	2	4	6	\Rightarrow	0	4	4/14
1	4	8	12		1	8	8/14
2	0	2	2		2	2	2/14
C	6	14	20			14	

Las modalidades de C , junto a sus frecuencias correspondientes, en la columna seleccionada, constituyen la distribución de frecuencias del carácter C , respecto a la modalidad 1 del carácter D . \square

2.1.5. Distribuciones conjuntas: Momentos mixtos

En este apartado vamos a presentar algunas características de las distribuciones conjuntas y su relación con las distribuciones marginales y condicionadas.

En las tablas de doble entrada, la distribución conjunta de frecuencias se puede obtener a partir de las distribuciones de frecuencias marginales y condicionadas según las relaciones

$$f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_{i.}} \cdot \frac{n_{i.}}{N} = f_j^i \cdot f_{i.} \quad \text{o bien} \quad f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_{.j}} \cdot \frac{n_{.j}}{N} = f_i^j \cdot f_{.j}$$

A continuación vamos a definir los momentos de una distribución conjunta que se utilizan para determinar medidas de relación entre las variables. Como veremos, algunos casos particulares corresponden a las medias y varianzas de las distribuciones marginales.

Momentos mixtos. Se define el momento de orden (r, s) respecto al punto (a, b) como

$$M_{rs}(a, b) = \sum_{i=1}^N (x_i - a)^r \cdot (y_i - b)^s \cdot f_i \quad \text{o bien} \quad M_{rs}(a, b) = \sum_{i=1}^k \sum_{j=1}^p (x_i - a)^r \cdot (y_j - b)^s \cdot f_{ij}$$

según consideremos la distribución de frecuencias correspondiente a una tabla simple (figura 2.1 de la página 54) o a una tabla de doble entrada (figura 2.2 de la página 55).

Ejemplo 2.8 Calcular el momento de orden $(2, 3)$ respecto al punto $(0, 1)$ para la distribución de frecuencias del ejemplo 2.2 de la página 54.

$$\begin{aligned} M_{23}(0, 1) &= \frac{2(0-0)^2(0-1)^3 + 4(0-0)^2(1-1)^3 + 4(1-0)^2(0-1)^3 + \dots}{20} = \\ &= \frac{0 + 0 - 4 + 0 + 0}{20} = -\frac{4}{20} = -0'2 \end{aligned}$$

□

Resultan de especial interés los siguientes dos casos particulares:

Momentos mixtos ordinarios. Si $a = b = 0$ entonces el momento de orden (r, s) recibe el nombre de momento ordinario y se denota por

$$m_{rs} = \sum_{i=1}^N x_i^r \cdot y_i^s \cdot f_i \quad \text{o bien} \quad m_{rs} = \sum_{i=1}^k \sum_{j=1}^p x_i^r \cdot y_j^s \cdot f_{ij}$$

Ejemplo 2.9 Calcular el momento ordinario de orden $(2, 3)$ para la distribución de frecuencias del ejemplo 2.2 de la página 54.

$$m_{23} = \frac{2 \cdot 0^2 \cdot 0^3 + 4 \cdot 0^2 \cdot 1^3 + 4 \cdot 1^2 \cdot 0^3 + 8 \cdot 1^2 \cdot 1^3 + 2 \cdot 2^2 \cdot 1^3}{20} = \frac{0 + 0 + 0 + 8 + 8}{20} = \frac{16}{20} = 0'8$$

□

Momentos mixtos centrales. Si $a = \bar{x}$ y $b = \bar{y}$ entonces el momento de orden (r, s) recibe el nombre de momento central y se denota por

$$\mu_{rs} = \sum_{i=1}^N (x_i - \bar{x})^r \cdot (y_i - \bar{y})^s \cdot f_i \quad \text{o bien} \quad \mu_{rs} = \sum_{i=1}^k \sum_{j=1}^p (x_i - \bar{x})^r \cdot (y_j - \bar{y})^s \cdot f_{ij}$$

Ejemplo 2.10 Calcular el momento central de orden $(2, 3)$ para la distribución de frecuencias del ejemplo 2.2 de la página 54.

Para calcular el momento central es necesario disponer de la media de las distribuciones marginales:

$$\bar{c} = \frac{0 \cdot 6 + 1 \cdot 12 + 2 \cdot 2}{20} = \frac{16}{20} = 0'8 \quad \text{y} \quad \bar{d} = \frac{0 \cdot 6 + 1 \cdot 14}{20} = \frac{14}{20} = 0'7$$

Después aplicamos la fórmula del momento central

$$\begin{aligned}\mu_{23} &= \frac{2(0-0'8)^2(0-0'7)^3 + 4(0-0'8)^2(1-0'7)^3 + 4(1-0'8)^2(0-0'7)^3 + \dots}{20} = \\ &= \frac{-0'43907 + 0'06912 - 0'01372 + 0'00864 + 0'07776}{20} = -\frac{0'29727}{20} = -0'0148635\end{aligned}$$

□

Para los momentos centrales y ordinarios, destacamos las siguientes propiedades que muestran su relación con algunas medidas de posición (media) y dispersión (varianza y desviación típica) para las distribuciones marginales:

$$\begin{array}{lll}m_{00} = 1 & m_{10} = \bar{x} & m_{01} = \bar{y} \\ \mu_{00} = 1 & \mu_{10} = 0 & \mu_{01} = 0\end{array}$$

Como en el caso unidimensional, se puede establecer una relación entre los momentos centrales y ordinarios. Destacamos las siguientes propiedades que establecen fórmulas alternativas para calcular determinadas medidas:

$$\mu_{11} = m_{11} - m_{10}m_{01} \quad \mu_{20} = m_{20} - m_{10}^2 \quad \mu_{02} = m_{02} - m_{01}^2$$

Medias marginales. La media marginal de la variable X corresponde a la medida de tendencia central *media aritmética* de la distribución marginal de la variable X . Análogamente se define la media marginal de la variable Y y ambas se calculan a partir de los momentos ordinarios:

$$\bar{x} = m_{10} = \sum_{i=1}^k x_i \cdot f_i \quad \bar{y} = m_{01} = \sum_{j=1}^p y_j \cdot f_j$$

El punto (\bar{x}, \bar{y}) es el *punto medio* o *centro de gravedad* de la distribución.

Ejemplo 2.11 Calcular el centro de gravedad para la distribución de frecuencias del ejemplo 2.2 de la página 54.

$$\bar{c} = \frac{0 \cdot 6 + 1 \cdot 12 + 2 \cdot 2}{20} = \frac{16}{20} = 0'8 \quad \text{y} \quad \bar{d} = \frac{0 \cdot 6 + 1 \cdot 14}{20} = \frac{14}{20} = 0'7$$

Por lo tanto, el centro de gravedad de la distribución es el punto $(0'8, 0'7)$. □

Varianzas marginales. La varianza marginal de la variable X corresponde a la medida de dispersión *varianza* de la distribución marginal de la variable X . Análogamente se define la varianza marginal de la variable Y y ambas se calculan a partir de los momentos centrales

$$\sigma_x^2 = V(X) = \mu_{20} = \sum_{i=1}^k (x_i - \bar{x})^2 \cdot f_i \quad \sigma_y^2 = V(Y) = \mu_{02} = \sum_{j=1}^p (y_j - \bar{y})^2 \cdot f_j$$

o de los momentos ordinarios, aplicando la propiedad que los relaciona.

Las desviaciones típicas marginales se definen como la raíz cuadrada positiva de las varianzas marginales correspondientes.

Ejemplo 2.12 Calcular la desviación típica marginal de la variable C en la distribución de frecuencias del ejemplo 2.2 de la página 54.

$$\sigma_c^2 = \frac{6 \cdot 0^2 + 12 \cdot 1^2 + 2 \cdot 2^2}{20} - 0'8^2 = 1 - 0'64 = 0'36$$

□

Covarianza. La covarianza o varianza conjunta es el momento central de orden (1,1)

$$\mu_{11} = \sum_{i=1}^k \sum_{j=1}^p (x_i - \bar{x}) \cdot (y_j - \bar{y}) \cdot f_{ij} \quad \text{o bien} \quad \mu_{11} = \sum_{i=1}^k (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot f_i$$

y se denota por $Cov(X, Y)$, o bien por σ_{xy} . Las propiedades que relaciona los momentos centrales y ordinarios nos permiten obtener una nueva fórmula para calcular la covarianza: *la media de los productos menos el producto de las medias*.

$$Cov(X, Y) = \sum_{i=1}^k \sum_{j=1}^p x_i y_j f_{ij} - \bar{x} \bar{y}$$

La covarianza es una medida de la variación conjunta de las variables y forma parte en la definición de los coeficientes que miden la relación entre esas variables. La covarianza se basa en las unidades de medida originales de las dos variables X e Y . Por lo tanto, no es posible comparar la covarianza de distintas distribuciones conjuntas. Si dividimos su fórmula por el producto de las desviaciones típicas de las variables X y Y obtenemos el coeficiente de correlación lineal de Pearson

$$r = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y}$$

que es una medida adimensional que permite comparar covarianzas de distintas distribuciones conjuntas.

Ejemplo 2.13 Calcular la covarianza y el coeficiente de correlación lineal de Pearson para la distribución de frecuencias del ejemplo 2.2 de la página 54.

En primer lugar, calculamos la covarianza:

$$Cov(C, D) = \frac{2 \cdot 0 \cdot 0 + 4 \cdot 0 \cdot 1 + 4 \cdot 1 \cdot 0 + 8 \cdot 1 \cdot 1 + 2 \cdot 2 \cdot 1}{20} - 0'8 \cdot 0'7 = 0'6 - 0'56 = 0'04$$

Después calculamos las desviaciones típicas marginales

$$\sigma_c^2 = 0'36 \quad \text{y} \quad \sigma_d^2 = 0'21$$

y finalmente aplicamos la fórmula para obtener el coeficiente de correlación

$$r = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{0'04}{\sqrt{0'36} \cdot \sqrt{0'21}} \approx 0'145$$

□

2.2. Regresión y correlación

En esta sección se introducen algunas técnicas estadísticas que nos permitirán estudiar la relación entre dos variables de una misma población o muestra. El interés se centrará en aquellos casos donde intuimos que existe una relación entre las variables, pero no somos capaces de encontrar una función matemática que describa esta relación. Por ejemplo, intuimos que el peso y la altura de un individuo están relacionados, sin embargo, no existe ninguna fórmula matemática que nos permita determinar el peso exacto de una persona en función de su altura.

El objetivo es encontrar un modelo o función matemática que recoja, de la manera más acertada, la relación entre dos variables de este tipo. Además, cuando hayamos determinado el modelo, será necesario proporcionar alguna medida de la bondad de dicho modelo. Por tanto, hay que resolver dos problemas:

1. Encontrar un modelo que permita relacionar dos variables
2. Determinar el grado de relación entre esas dos variables.

La regresión estudia la naturaleza estadística de la relación entre dos variables y nos proporciona un modelo de dicha relación. El modelo consiste en una función matemática cuya gráfica se aproxima a los datos observados. La función encontrada permitirá obtener los valores aproximados de una de las variables a partir de los valores prefijados de la otra variable.

La correlación se encarga de solucionar el segundo problema estableciendo la correspondencia en las pautas de variación de dos variables. La correlación cuantifica esta dependencia entre las variables mediante el cálculo de los coeficientes de correlación.

Veamos, en primer lugar, qué tipos de relaciones pueden existir entre las variables. En segundo lugar, presentaremos algunos métodos para obtener modelos que determinan la relación entre las variables. En tercer lugar, introduciremos medidas que permitan estudiar la bondad de esos modelos. Y, por último, presentaremos, a modo de ejemplo, algunos modelos importantes, como el modelo lineal.

2.2.1. Relación entre variables

El objetivo de analizar conjuntamente dos variables diferentes en una misma población o muestra es estudiar el tipo de relación que hay entre ellas. Según el grado extremo de relación existente distinguimos tres casos: Si no hay relación alguna decimos que las variables son independientes; si, por el contrario, hay una relación total decimos que las variables dependen funcionalmente; y en los casos intermedios decimos que las variables mantienen una dependencia estadística. Desde el punto de vista estadístico, este último caso es el más interesante pues permite estudiar el grado de dependencia entre las variables, proporcionando un modelo matemático que explique la relación entre ellas.

Independencia

Cuando no existe relación alguna entre las variables, es decir, ninguna de ellas proporciona información sobre la otra, decimos que existe una independencia entre las variables. En este

caso, se dice que las variables son *independientes* una de la otra. Por ejemplo: la velocidad de un ordenador y el grosor del papel utilizado en la impresora.

Formalmente la independencia se define así:

1. Se dice que el carácter X es independiente de Y , si todas las condicionadas de X respecto a cualquier clase de Y coinciden con la marginal de X , es decir $f_i^j = f_i$ para todo j .
2. Análogamente se define la independencia de Y respecto a X si $f_j^i = f_j$ para todo i .

Se deduce que si X es independiente de Y , entonces Y es independiente de X y esto ocurre si y sólo si $f_{ij} = f_i \cdot f_j$. En este caso, es fácil determinar, a la vista de la tabla de frecuencias, la independencia de caracteres porque las columnas son proporcionales entre sí, al igual que ocurre con las filas.

Ejemplo 2.14 Comprobar que la siguiente tabla de frecuencias corresponde a dos variables independientes.

	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

Se puede observar que las columnas de la tabla son proporcionales: la segunda columna es tres veces la primera, la tercera es dos veces la primera y la cuarta es cuatro veces la primera.

Si representamos la distribución de frecuencias relativas en forma de tabla de doble entrada

	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

observamos que el producto de las frecuencias de las distribuciones marginales coincide con la frecuencia correspondiente de la distribución conjunta. Por ejemplo, $f_2 \cdot f_{\cdot 3} = f_{23}$, es decir, $3/6 \cdot 2/10 = 6/60$.

También se puede comprobar que las distribuciones de frecuencias condicionadas y marginal son iguales. Por ejemplo, en las siguientes tablas calculamos las distribución de frecuencias de la variable X condicionada a cualquier modalidad de la variable Y (izquierda) y comprobamos que todas coinciden y son iguales a la distribución de frecuencias marginal de la variable X (derecha).

x_i	f_i^j	f_i^1	f_i^2	f_i^3	f_i^4	x_i	f_i
x_1	$\frac{1}{6}$	$= \frac{1}{1+3+2}$	$= \frac{3}{3+9+6}$	$= \frac{2}{2+6+4}$	$= \frac{4}{4+12+8}$	x_1	$\frac{1}{6}$
x_2	$\frac{3}{6}$	$= \frac{3}{1+3+2}$	$= \frac{9}{3+9+6}$	$= \frac{6}{2+6+4}$	$= \frac{12}{4+12+8}$	x_2	$\frac{3}{6}$
x_3	$\frac{2}{6}$	$= \frac{2}{1+3+2}$	$= \frac{6}{3+9+6}$	$= \frac{4}{2+6+4}$	$= \frac{8}{4+12+8}$	x_3	$\frac{2}{6}$

Análogamente podíamos comprobar que se verifica para la variable Y calculando sus frecuencias condicionadas y marginal. \square

Dependencia funcional

En el estudio conjunto de dos variables puede ocurrir que la aparición de un determinado valor de una de las variables esté perfectamente determinado conociendo el valor de la otra para esa misma observación. En este caso, decimos que existe una dependencia funcional entre las variables y podemos establecer un modelo matemático que relaciona ambas variables.

Por ejemplo, si tomamos varias muestras de las longitudes de las circunferencias (L) y sus radios (R) observamos que los valores de las variables están relacionados por la fórmula: $L = 2\pi R$. Es decir, existe un modelo matemático que me permite calcular el valor que toma la variable L sin necesidad de observarlo, conociendo el valor correspondiente de la variable R .

A la vista de la tabla de frecuencias es fácil determinar la dependencia funcional. Si para cada modalidad x_i de X existe una única modalidad y_j de Y tal que $n_{ij} \neq 0$, decimos que la variable Y depende funcionalmente de la variable X . Esta relación de dependencia funcional no es recíproca, es decir, si X depende funcionalmente de Y no implica que Y dependa funcionalmente de X . Por ejemplo: $Y = a \cdot X^2$ donde Y depende de X y no al revés.

Ejemplo 2.15 *Comprobar que la siguiente tabla de frecuencias corresponde a dos variables que dependen funcionalmente. Determinar la dependencia y establecer el modelo matemático.*

	y_1	y_2	y_3	y_4	y_5
x_1	0	0	3	0	0
x_2	0	0	0	0	1
x_3	0	0	2	0	0
x_4	4	0	0	0	0

Como se observa en la tabla, para cada modalidad x_i de la variable X existe una única modalidad y_j de la variable Y cuya frecuencia conjunta es distinta de 0. En este caso, decimos que la variable Y depende funcionalmente de la variable X y se establece el siguiente modelo matemático en forma de tabla

X	x_1	x_2	x_3	x_4
Y	y_3	y_5	y_3	y_1

que permite determinar los valores de Y en función de la observación del valor de X . □

Dependencia estadística

La independencia y la dependencia funcional son dos casos extremos de la relación entre las variables cuando ésta no existe o es total. Generalmente, cuando se estudian conjuntamente dos variables para establecer la relación entre ambas surgen los casos intermedios.

Cuando una variable puede dar información sobre otra, pero la relación entre ambas no es determinista y por tanto no existe o no se conoce una expresión matemática que las relacione, se dice que existe una *dependencia aleatoria o estadística*. Por ejemplo, sabemos que el peso y la estatura de una persona son dos variables relacionadas y sin embargo no se puede establecer una fórmula matemática que determine, en todos los casos, el peso de una persona en función de su altura.

La dependencia estadística también suele considerarse en aquellos procesos o variables cuya relación es determinista pero resulta muy complejo su estudio. Por ejemplo, en el comportamiento atmosférico sólo intervienen fenómenos físicos perfectamente estudiables y sin embargo, su estudio es intratable cuando pretendemos establecer una predicción meteorológica. Igual ocurre con las placas tectónicas terrestres, aunque su movimiento se rige por leyes físicas, su complejidad impide la predicción exacta de un terremoto. En estos casos, se considera que las variables presentan una dependencia estadística y se estudia su relación a partir de muestras.

2.2.2. Regresión: Método de los mínimos cuadrados

Cuando existe una dependencia estadística entre variables, el objetivo es encontrar un modelo o función matemática que determine, de manera aproximada, la relación entre las variables

La representación de los datos obtenidos en la muestra de una variable estadística bidimensional (X, Y) sobre el plano (diagrama de dispersión) constituye 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 correlacionadas o que hay *correlación perfecta* entre ellas. En general, como se observa en la figura 2.6, se trata de una línea ideal en torno a la cual se distribuyen los puntos de la nube.

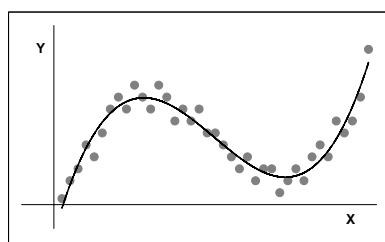


Figura 2.6: Nube de puntos y curva de regresión

En la práctica, la obtención de esta línea no es sencilla y, en general, no tiene que tener una expresión matemática en forma de ecuación. Por esta razón, la estadística se limita a calcular líneas “ideales” con expresiones matemáticas conocidas con formas rectas, parabólicas, exponenciales, logarítmicas, hiperbólicas, etc.

Cuando dispongamos de la ecuación de esta curva de regresión podemos utilizarla para estudiar las características de la relación entre las variables y predecir valores desconocidos.

El problema general de la regresión es ajustar una función o curva de ecuación conocida a la nube de puntos que representa las observaciones de una variable bidimensional (X, Y) . En primer lugar, hay que determinar qué variable es la dependiente, y cual es la independiente. Después, y a la vista de la nube de puntos, hay que elegir un tipo de modelo o función $y = f(x)$, que puede ser lineal, cuadrático, exponencial, etc., que determina la relación entre las variables.

El tipo de modelo de regresión $y = f(x)$, elegido para ajustar la nube de puntos, dependerá de una serie de coeficientes o parámetros. Los métodos de regresión nos permiten calcular los coeficientes o parámetros que determinan el modelo que mejor se ajusta a la nube de puntos.

Por ejemplo, si hemos elegido un modelo lineal de regresión del tipo $y = a + bx$, el método de regresión nos ayudará a calcular los valores de a y b que determinan la recta $y = a + bx$ que mejor se ajusta a la nube de puntos.

Para poder determinar los coeficientes de un modelo de regresión es necesario disponer de un mínimo número de puntos. En general, será necesario que haya tantos puntos como coeficientes haya que determinar en el modelo. Por ejemplo, si consideramos el modelo lineal $y = a + bx$, entonces será necesario que la nube de puntos tenga, al menos, dos puntos, pues, con un sólo punto habría infinitud de rectas que ajustasen (perfectamente) el modelo y ninguna de ellas sería mejor que las otras. O, por ejemplo, pensemos en un modelo parabólico $y = a + bx + cx^2$. En este caso, será necesario que la nube de puntos tenga más de tres elementos, pues con un número menor, por ejemplo dos, hay infinitud de parábolas que pasan por esos dos puntos, y todas ellas, se ajustan perfectamente a la nube de puntos.

Por lo tanto, y para evitar trivialidades, consideraremos que el número de observaciones de una variable bidimensional (X, Y) es mayor o igual al número de coeficientes del modelo de regresión que deseamos ajustar. Además, como veremos en esta sección, será necesario que esos puntos tengan valores distintos de la variable independientes, es decir, que el número de coeficientes del modelo debe ser menor o igual al número de observaciones con valores distintos de la variable independiente.

Método de los mínimos cuadrados

El método de los mínimos cuadrados permite ajustar modelos de regresión, y consiste en minimizar las distancias entre el modelo y los puntos correspondientes a los valores observados en la muestra. Estas distancias reciben el nombre de *errores* o *residuos*.

Consideramos una muestra de tamaño N de una variable bidimensional (X, Y) que toma los valores $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)$ con frecuencias absolutas n_1, n_2, \dots, n_k , y supongamos que hemos determinado que la variable Y depende de la variable X . Primero, elegimos el modelo o función $y = f(x)$ que depende de ciertos parámetros a_1, \dots, a_m . Después, a cada valor x_i de la variable X le asignamos un valor teórico $y_i^* = f(x_i)$ calculado a partir del modelo.

Como se ve en la figura 2.7, las diferencias entre los verdaderos valores y_i y los valores y_i^* estimados por el modelo, a partir de los correspondientes valores x_i , determinan los errores cometidos al utilizar el modelo, que se denotan por $e_i = (y_i - y_i^*)$.

El objetivo es minimizar los errores, pero hay que tener en cuenta que los valores e_i pueden ser positivos o negativos en función de la posición relativa del punto (x_i, y_i) respecto de la función $y = f(x)$. Por lo tanto, la simple suma de estos errores puede dar una visión equivocada del ajuste del modelo a la nube de puntos. Por ejemplo, si la suma de los errores es 0, puede ser que la función pase efectivamente por todos los puntos de la nube indicando un ajuste perfecto; o puede ser también que los errores de signo positivo se hayan compensado con los negativos y el ajuste no sean tan bueno como creíamos.

Utilizar los valores absolutos de los errores puede dificultar notablemente los cálculos, de manera que, para evitar estos problemas, utilizaremos los cuadrados de los errores. Y ya estamos en disposición de construir una función objetivo F , definida como la suma de los cuadrados de los errores e_i . Esta función sólo depende de los parámetros de la función $f(x)$ que hay que

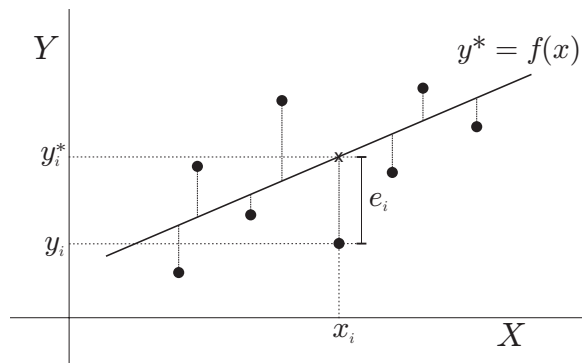


Figura 2.7: Método de los mínimos cuadrados Y/X

determinar:

$$F(a_1, \dots, a_m) = \sum_{i=1}^k e_i^2 \cdot n_i = \sum_{i=1}^k (y_i - y_i^*)^2 \cdot n_i = \sum_{i=1}^k (y_i - f(x_i))^2 \cdot n_i$$

Para calcular el valor de los parámetros que minimizan la función basta con resolver el sistema obtenido al igualar a cero las derivadas parciales de F respecto de los parámetros de los que depende $f(x)$, es decir, resolver el sistema: $\nabla F = 0$ donde ∇ es el operador gradiente. En definitiva, el método consiste en minimizar la suma de los cuadrados de los errores y de ahí su nombre.

Para explicar el método de los mínimos cuadrados hemos considerado que la variable independiente era X . En este caso, los errores se definían como las diferencias entre los valores observados de la variable Y y los valores estimados según el modelo $y = f(x)$. Si consideramos que Y es la variable independiente entonces el modelo es de la forma $x = g(y)$ y los errores se determinan como diferencias de los valores observado y los estimados para la variable X , es decir, $e_i = (x_i - x_i^*)$ donde $x_i^* = g(y_i)$ (ver figura 2.8).

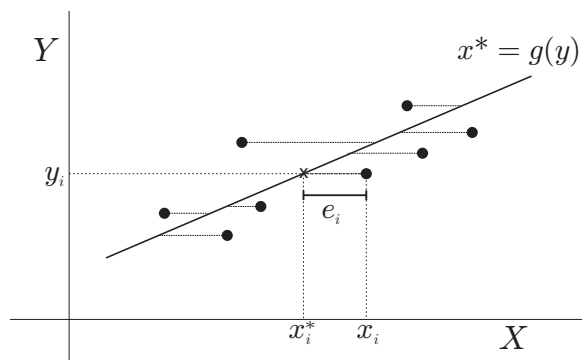


Figura 2.8: Método de los mínimos cuadrados X/Y

Curva general de regresión

La curva general de regresión es un conjunto de puntos que representa a la nube de puntos. Como veremos, ajustar un modelo de regresión a la nube de puntos, equivale a ajustarlo a la curva de regresión. Este resultado simplificará notablemente los cálculos en aquellos ejemplos cuyos datos se presentan en forma de tabla de doble entrada.

Consideramos la distribución de frecuencias de la variable (X, Y) que presenta las modalidades (x_i, y_j) con frecuencias relativas f_{ij} con $i = 1, \dots, k$ y $j = 1, \dots, p$. Se define la **curva general de regresión de Y sobre X** como la función que asigna, a cada valor x_i de la variable de X , la media \bar{y}_i de la distribución de la variable Y condicionada al valor x_i de la variable X .

Con esta definición, podemos decir que la curva de regresión está formada por los valores (x_i, \bar{y}_i) con frecuencia relativa f_i con $i = 1, \dots, k$, siendo $\bar{y}_i = \sum_{j=1}^p y_j f_{ij}$. Obsérvese que se podría definir, de manera análoga, la curva general de regresión de X sobre Y como la función que asigna, a cada valor y_j de la variable de Y , la media $\bar{x}_j = \sum_{i=1}^k x_i f_{ij}$ de la distribución de la variable X condicionada al valor y_j de la variable Y .

La importancia de estas curvas radica en la siguiente propiedad de la curva general de regresión: *El problema de ajustar un modelo de regresión Y sobre X a la nube de puntos, por el método de los mínimos cuadrados, es equivalente a ajustar dicho modelo a la curva general de regresión, por el método de los mínimos cuadrados.*

Esta propiedad tiene dos implicaciones inmediatas en el ajuste por mínimos cuadrados. Por un lado, cuando tengamos un conjunto de observaciones donde algunos puntos tienen el mismo valor de la variable independiente, podemos simplificar el conjunto de datos. En particular, cuando tengamos un problema donde la distribución de frecuencias viene expresada con una tabla de doble entrada, podemos transformarla en una tabla estadística de frecuencias. Para ello, sustituiremos los valores de la variable dependiente por las medias de las distribuciones condicionadas correspondientes.

En la figura 2.9 se muestra como se transforma la tabla de doble entrada de la distribución de frecuencias de la variable (X, Y) , en una tabla de frecuencias donde cada modalidad (x_i, y_j) ha sido sustituida por la modalidad (x_i, \bar{y}_i) , siendo \bar{y}_i la media de la variable $Y/X = x_i$, para todo $i = 1, \dots, k$.

$X \setminus Y$	y_1	y_2	\dots	y_j	\dots	y_p			x_i	y_i	n_i
x_1	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1p}	$n_{1\cdot}$		x_1	\bar{y}_1	$n_{1\cdot}$
x_2	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2p}	$n_{2\cdot}$		x_2	\bar{y}_2	$n_{2\cdot}$
\vdots	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots		\vdots	\vdots	\vdots
x_i	n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{ip}	$n_{i\cdot}$	\longrightarrow	x_i	\bar{y}_i	$n_{i\cdot}$
\vdots	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots		\vdots	\vdots	\vdots
x_k	n_{k1}	n_{k2}	\dots	n_{kj}	\dots	n_{kp}	$n_{k\cdot}$		x_k	\bar{y}_k	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	\dots	$n_{\cdot j}$	\dots	$n_{\cdot p}$	N				N

Figura 2.9: Aplicación de la propiedad de la curva general de regresión

Esta simplificación del conjunto de observaciones no tiene sentido realizarse cuando todos los puntos tiene distinto valor de la variable independiente. En tal caso, la curva de regresión coincide con la nube de puntos.

Por otro lado, veamos que la propiedad de la curva general de regresión que hemos presentado, tiene otra consecuencia inmediata, para evitar trivialidades, imponiendo una restricción al modelo de regresión.

Sabemos que el número de observaciones de una variable bidimensional (X, Y) debe ser mayor o igual que el número de coeficientes del modelo de regresión que deseamos ajustar. Pero según la propiedad de las curvas de regresión, ajustar un modelo a la nube de puntos es igual que ajustarlo a la curva de regresión. Por lo tanto, el número de coeficientes del modelo debe ser menor o igual al número de puntos de la curva de regresión, es decir, al número de observaciones con valores distintos de la variable independiente.

2.2.3. Correlación

La correlación mide el grado de relación entre las variables, a partir del modelo de regresión. Para ello, se definen medidas que determinan la bondad de dicho modelo.

La aproximación de la curva de regresión a la nube de puntos viene determinada por los residuos. Las medidas de correlación deben cuantificar la dispersión de los datos en torno al modelo, es decir, lo cerca o lejos de la curva que están los puntos. Para ello, será necesario hacer un estudio de las varianzas y de los residuos.

En las fórmulas que vamos a obtener para estas medidas, consideramos una muestra de tamaño N de una variable (X, Y) que toma los valores (x_i, y_i) , con frecuencias absolutas n_i , y relativas f_i , respectivamente para todo $i = 1, \dots, k$.

Varianzas del modelo

En el estudio del modelo general de regresión $y = f(x)$ para las variables X e Y , hemos considerado dos nuevas variables: los valores (E) de los errores o residuos y los valores (Y^*) estimados por el modelo. Para cada pareja de valores (x_i, y_i) de la variable (X, Y) hemos considerado un valor $y_i^* = f(x_i)$ de la variable Y^* y un valor $e_i = y_i - y_i^*$ de la variable E .

Vamos a considerar las varianzas de estas variables Y^* y E , cuyos valores se obtienen a partir del modelo ajustado. Ambas medidas se utilizan para determinar la bondad del ajuste y, junto a la varianza de Y , forman parte en la definición de algunos coeficientes de correlación.

Se llama **varianza explicada** a la varianza de los valores estimados y_i^* de la variable Y^*

$$\sigma_{y^*}^2 = \sum_{i=1}^k (y_i^* - \bar{y}^*)^2 \cdot f_i$$

Se llama **varianza residual** o **varianza no explicada** a la varianza de los errores e_i de la variable E

$$\sigma_e^2 = \sum_{i=1}^k (e_i - \bar{e})^2 \cdot f_i \quad \text{siendo} \quad \bar{e} = \sum_{i=1}^k e_i \cdot f_i \quad \text{y} \quad e_i = y_i - y_i^*$$

Y, se llama **varianza total** a la varianza de la variable dependiente Y .

Coefficiente de determinación

En el estudio del modelo general de regresión $y = f(x)$ para las variables X e Y , la variable E (errores o residuos) mide las diferencias entre los valores de la variable Y y los valores de la variable (Y^*) estimados por el modelo. Por lo tanto, se espera que E sea una variable cuya media debe ser 0, y cuya varianza debe ser pequeña (en comparación con la de Y).

Por esta razón, se define el *coeficiente de determinación* como 1 menos el cociente entre la varianza residual y la varianza de la variable Y , y se denota por R^2

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}$$

Si el ajuste, mediante la curva de regresión, es bueno, cabe esperar que este coeficiente tome un valor próximo a 1. De esta manera, el coeficiente de determinación mide el grado de bondad del ajuste.

Residuos

Los residuos indican la discrepancia entre el modelo y los datos. Una comparación entre estos valores para distintos modelos permite elegir el más adecuado.

En el método de los mínimos cuadrados, se definía la función suma de los cuadrados de los residuos. Esta función dependía de los coeficientes del modelo que se obtenían al ajustar la curva a la nube de puntos.

Por tanto, a partir del modelo ajustado $y = f(x)$ y de los puntos (x_i, y_i) con frecuencias absolutas n_i podemos obtener los residuos $e_i = y_i - f(x_i)$ que son los valores de la variable E correspondientes al modelo. Si calculamos la suma de los cuadrados de los residuos (sin promediarlos)

$$\text{SSE} = \sum_{i=1}^k e_i^2 \cdot n_i = \sum_{i=1}^k (y_i - y_i^*)^2 \cdot n_i$$

obtenemos un coeficiente de correlación que denotamos por SSE (Sum of Squared Errors).

Este coeficiente sirve para comparar la bondad de dos modelos que se ajustan a una misma nube de puntos. SSE determina los errores cometidos cuando se utilizan los valores estimados por el modelo en lugar de los verdaderos valores de la variable. Por tanto, el modelo que presente un menor valor de SEE corresponde al modelo que mejor se aproxima a la nube de puntos.

Veamos que la curva general de regresión que hemos presentado en la pagina 70 tiene una interesante propiedad de correlación que determina una cota inferior del error que se comete cuando se ajusta cualquier modelo de regresión.

Consideramos la distribución de frecuencias de la variable (X, Y) que presenta las modalidades (x_i, y_j) con frecuencias absolutas n_{ij} con $i = 1, \dots, k$ y $j = 1, \dots, p$, y sean (x_i, \bar{y}_i) los puntos que definen la curva general de regresión de Y sobre X . Entonces se verifica que el valor del coeficiente SSE de cualquier modelo de regresión $y = f(x)$ es mayor o igual que el valor del

coeficiente SSE de la curva general de regresión Y/X , es decir,

$$\sum_{i=1}^k \sum_{j=1}^p (y_j - \bar{y}_i)^2 n_{ij} \leq \sum_{i=1}^k \sum_{j=1}^p (y_j - f(x_i))^2 n_{ij}$$

Por lo tanto, la expresión del primer miembro de la ecuación, determina una cota inferior del error que se comete cuando se ajusta cualquier modelo de regresión. Sin embargo, si todos los valores de la variable independiente son distintos, entonces la cota que determina la curva general de regresión es trivial, pues vale 0.

2.3. El modelo lineal

Sea $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ una muestra de la variable estadística bidimensional (X, Y) . Para simplificar las fórmulas, hemos considerado que todas las modalidades presentan frecuencia absoluta igual a uno; en otro caso, los distintos valores aparecerían multiplicados por su frecuencia absoluta correspondiente.

Nuestro objetivo será encontrar un modelo lineal que se ajuste a la nube de puntos y un coeficiente que determine el grado de aproximación del modelo a los datos.

2.3.1. Regresión lineal

El modelo lineal que mejor se aproxima a la nube de puntos recibe el nombre de *recta de regresión de Y sobre X* . Este modelo de ecuación $Y = a + b \cdot X$ queda determinado conociendo los valores de los parámetros a y b . Aplicando el método de los mínimos cuadrados se obtienen fórmulas que permitan calcular estos parámetros en función de los datos de la muestra.

A cada valor x_i de la variable X le corresponde un valor y_i de la variable Y . Sin embargo, la recta de regresión le asigna a x_i el valor estimado $y_i^* = f(x_i) = a + bx_i$. Por tanto, la diferencia (también llamada error o residuo) entre el valor “teórico ajustado” y el valor “real” es

$$e_i = y_i - y_i^* = y_i - a - bx_i$$

Aplicando el método de los mínimos cuadrados, imponemos la condición de que la suma de los errores al cuadrado sea mínima. Para ello, minimizamos la función

$$F(a, b) = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - a - bx_i)^2$$

donde x_i e y_i son datos del problema.

Ahora, los puntos críticos de la función F , que resultan ser mínimos¹, se obtienen resolviendo la ecuación $\nabla F(a, b) = 0$.

¹En el ejercicio 35 de la página 96 se propone la demostración de este resultado

$$\left\{ \begin{array}{l} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \end{array} \right\} \leftrightarrow \left\{ \begin{array}{l} -2 \sum_{i=1}^N (y_i - a - bx_i) = 0 \\ -2 \sum_{i=1}^N x_i (y_i - a - bx_i) = 0 \end{array} \right\} \leftrightarrow \left\{ \begin{array}{l} \sum_{i=1}^N y_i = a \cdot N + b \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i y_i = a \sum_{i=1}^N x_i + b \sum_{i=1}^N x_i^2 \end{array} \right\}$$

El sistema anterior recibe el nombre de *sistema de ecuaciones normales* que expresado en forma matricial resulta:

$$\begin{pmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i y_i \end{pmatrix}$$

Resolviendo el sistema se obtienen los siguientes resultados:

$$b = \frac{N \sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \cdot \sum_{i=1}^N y_i}{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2} \quad y \quad a = \frac{1}{N} \left(\sum_{i=1}^N y_i - b \sum_{i=1}^N x_i \right)$$

Si se divide el numerador y el denominador de la expresión de b por N^2 y se observa la expresión obtenida para a tenemos

$$b = \frac{Cov(X, Y)}{\sigma_x^2} \quad y \quad a = \bar{y} - b\bar{x}$$

Por tanto, la ecuación de la *recta de regresión de Y sobre X* (Y/X) es

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

Si consideramos que Y es la variable independiente y X la dependiente, entonces la ecuación del modelo lineal es $X = a + bY$. Para ajustar el modelo a la nube de puntos, aplicamos el método de los mínimos cuadrados y obtenemos la recta de regresión de X sobre Y (X/Y) que es

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

Como podemos observar, las dos rectas de regresión obtenidas pasan y se cortan en el punto del plano correspondiente al centro de gravedad (\bar{x}, \bar{y}) .

En este punto, hay que hacer una observación importante sobre las rectas de regresión Y/X y X/Y . Desde el punto de vista matemático, las dos rectas son distintas pues, en general, si en

la recta de regresión Y/X despejamos la variable X en función de la variable Y , no se obtiene la recta de regresión de X/Y , y viceversa.

Los modelos de regresión permiten “predecir” los valores de la variable dependiente en función de los valores de la variable independiente. Así, la recta de regresión de Y/X determina los valores de Y en función de los valores de X , y por lo tanto, si deseamos utilizar un modelo lineal para calcular un valor de X en función de uno de Y , no podemos utilizar el modelo lineal Y/X . En este caso será necesario calcular la recta de regresión X/Y .

Ejemplo 2.16 En el ejemplo 2.2 de la página 54 se consideran las variables “número de controles efectuados” (C) y “número de errores detectados” (D) en programas de software. Determinar la variable dependiente y calcular la recta de regresión para los datos de la muestra.

Evidentemente, el número de errores detectados (variable D) depende del número de controles efectuados (variable C), y por lo tanto, consideramos el modelo lineal $D = a + bC$. Veamos tres formas de calcular los valores de a y b , que determinan el modelo.

1. Resolviendo el sistema de ecuaciones normales:

$$\begin{aligned}\sum d_i n_i &= a \cdot N + b \sum c_i n_i \\ \sum c_i d_i n_i &= a \sum c_i n_i + b \sum c_i^2 n_i\end{aligned}$$

y, para ello, resulta útil considerar la siguiente tabla estadística

c_i	d_i	n_i	$c_i n_i$	$c_i^2 n_i$	$d_i n_i$	$c_i d_i n_i$
0	0	2	0	0	0	0
0	1	4	0	0	4	0
1	0	4	4	4	0	0
1	1	8	8	8	8	8
2	1	2	4	8	2	4
		20	16	20	14	12

que determina el sistema de ecuaciones

$$14 = 20a + 16b$$

$$12 = 16a + 20b$$

cuya solución es $a = \frac{11}{18} \approx 0'611$ y $b = \frac{1}{9} \approx 0'111$.

2. Aplicando las fórmulas

$$b = \frac{\text{Cov}(X, Y)}{\sigma_x^2} \quad \text{y} \quad a = \bar{y} - b\bar{x}$$

a las variables y datos de nuestro ejemplo, siendo $X = C$ e $Y = D$,

$$b = \frac{\text{Cov}(C, D)}{\sigma_C^2} = \frac{0'04}{0'36} = \frac{1}{9} \approx 0'111$$

$$a = \bar{D} - b\bar{C} = 0'7 - \frac{1}{9} \cdot 0'8 = \frac{11}{18} \approx 0'611$$

también obtenemos esos mismos valores para los coeficientes a y b del modelo.

3. Calculamos la curva general de regresión:

n_{ij}	0	1	D		c_i	d_i	n_i
0	2	4	6	\longrightarrow	0	$2/3$	6
1	4	8	12		1	$2/3$	12
2	0	2	2		2	1	2
C	6	14	20				20

Aplicando la propiedad de la curva general de regresión, si ajustamos el modelo lineal de regresión a esta distribución de frecuencias, obtenemos la misma recta que en los casos anteriores.

Independientemente del método usado, la recta de regresión de D/C es el siguiente modelo lineal que relaciona ambas variables:

$$D = \frac{11}{18} + \frac{1}{9}C$$

En la figura 2.10 se representa la nube de puntos (puntos azules) y la curva general de regresión (cruces en rojo), cada una de ellas con un número que indica su frecuencia absoluta. Además, se representa la recta de regresión (línea discontinua) que se ajusta a estos datos.

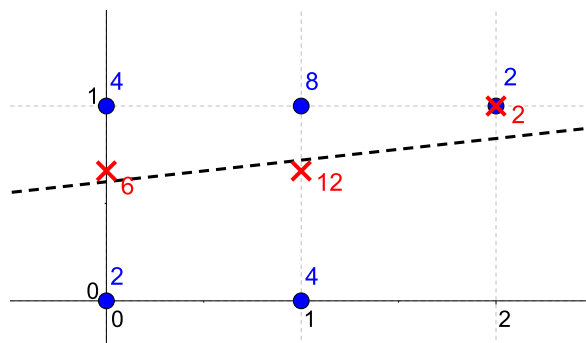


Figura 2.10: Ajuste lineal

Obsérvese que este modelo me permite “predecir” el valor de D en función del valor de C . Por ejemplo, cuando se realicen $C = 5$ controles, se espera que se detecten $D = 7/6 \approx 1'167$ errores.

Sin embargo, este modelo, no se debe utilizar para predecir el valor de C , en función de D , aunque, matemáticamente, si sea posible. En ese caso, habría que utilizar la recta de regresión C/D . \square

2.3.2. Correlación lineal

Una vez visto el problema de *regresión* o *estimación* de una variable, se verá ahora el problema de la *correlación*, o grado de interconexión entre variables. Se pretende determinar con qué precisión se describe o explica la relación entre variables en una ecuación lineal.

Coeficientes de regresión y correlación lineal

Dada una variable bidimensional (X, Y) , podemos obtener dos rectas de regresión: la de Y sobre X y la de X sobre Y . Para cada una de estas rectas definimos el *coeficiente de regresión lineal* como

$$b_{y/x} = \frac{\text{Cov}(X, Y)}{\sigma_x^2} \quad \text{y} \quad b_{x/y} = \frac{\text{Cov}(X, Y)}{\sigma_y^2}$$

siendo $b_{y/x}$ el coeficiente de regresión de la recta de regresión Y/X y $b_{x/y}$ de la recta de regresión de X/Y .

Estos coeficientes tienen el mismo signo y están estrechamente relacionados con las pendientes de las rectas. Por ello, los valores que toman determinan el crecimiento, decrecimiento, horizontalidad o verticalidad de las rectas de regresión. Por ejemplo, si $b_{y/x}$ es un número positivo, entonces la recta de regresión de Y sobre X es creciente e indica que aumenta la variable Y al aumentar la X .

Las pendientes de las rectas son:

$$m_{y/x} = b_{y/x} = \frac{\text{Cov}(X, Y)}{\sigma_x^2} \quad \text{y} \quad m_{x/y} = \frac{1}{b_{x/y}} = \frac{\sigma_y^2}{\text{Cov}(X, Y)}$$

siendo $m_{y/x}$ la pendiente de la recta de regresión Y/X y $m_{x/y}$ de la recta de regresión X/Y .

Ejemplo 2.17 Calcular los coeficientes de regresión lineal para las variables del ejemplo 2.2 de la página 54.

Sabiendo que $\text{Cov}(C, D) = 0'04$, $\sigma_c^2 = 0'36$ y $\sigma_d^2 = 0'21$, aplicamos la fórmula que se deduce de la definición:

$$b_{d/c} = \frac{\text{Cov}(C, D)}{\sigma_c^2} = \frac{0'04}{0'36} = \frac{1}{9} \approx 0'111 \quad , \quad b_{c/d} = \frac{\text{Cov}(C, D)}{\sigma_d^2} = \frac{0'04}{0'21} = \frac{4}{21} \approx 0'1905$$

□

Coeficiente de correlación lineal

Siempre que los datos tiendan a agruparse en torno a una línea recta se puede afirmar que existe *correlación lineal* o dependencia de tipo lineal. Además, distinguimos dos tipos:

- Si la recta tiene pendiente positiva, la correlación o dependencia lineal es *directa*, es decir, incrementos positivos de una variable implican aumentos en la otra.
- Si la recta tiene pendiente negativa, la correlación o dependencia lineal es *inversa*, es decir, al aumentar una variable disminuye la otra.

El *análisis de correlación* consiste en determinar un número que permita conocer cuál es el grado de asociación entre las variables y en qué sentido (directa o inversamente). Por esta razón se introduce el concepto de coeficiente de correlación lineal de Pearson.

El **coeficiente de correlación lineal de Pearson** es una medida que se sólo se define para el modelo lineal, y que determina el grado de ajuste entre una nube de puntos y la recta de

regresión correspondiente. Este coeficiente es adimensional, se denota por r o ρ y viene definido por la media geométrica de los coeficientes de regresión lineal:

$$r = \rho = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} \quad -1 \leq r \leq 1$$

Obsérvese que este coeficiente no puede calcularse si alguna de las variables es degenerada, es decir, toma un único valor. En ese caso, como la desviación típica de la variable degenerada es siempre 0, la definición anterior carece de sentido.

El coeficiente de correlación lineal resulta ser siempre un número en el intervalo $[-1, 1]$ con las siguientes interpretaciones² en función a su valor y su signo:

- El signo de este indicador va a coincidir con el de la covarianza pues las desviaciones de las variables son siempre positivas. De esta manera
 - Si $r > 0$ entonces la relación entre las variables es directa.
 - Si $r < 0$ entonces la relación entre las variables es inversa.
- El valor del coeficiente determina el grado de ajuste de la recta. De esta manera
 - Si $r = -1$ ó $r = 1$ entonces la correlación es perfecta e indica que existe una dependencia funcional entre las variables. En este caso, los datos representados en la nube de puntos están situados sobre una recta, que resulta ser la recta de regresión Y/X y que coincide con la de X/Y .
 - Si $r = 0$ entonces las rectas de regresión son paralelas a los ejes ($y = \bar{y}$ y $x = \bar{x}$), y se dice que las variables están linealmente incorreladas.
 - Los valores intermedios determinan los grados intermedios de ajustes. Cuanto más cerca de 1 ó -1 esté el valor de r la correlación será más *fuerte*, mientras que valores próximos a 0 indican una correlación *débil*.

El coeficiente r que hemos definido, resulta ser una medida objetiva de correlación lineal entre dos variables, en el sentido de que no depende de la escala de medición utilizada, es decir, es adimensional. Sin embargo, es importante tener en cuenta que sólo tiene sentido definir este coeficiente en el caso lineal.

Ejemplo 2.18 *Determinar e interpretar el valor del coeficiente de correlación lineal de Pearson para los datos del ejemplo 2.2 de la página 54.*

Sabiendo que $\text{Cov}(C, D) = 0'04$, $\sigma_c^2 = 0'36$ y $\sigma_d^2 = 0'21$, aplicamos la definición y obtenemos:

$$r = \frac{\text{Cov}(C, D)}{\sigma_c \cdot \sigma_d} = \frac{0'04}{\sqrt{0'36} \cdot \sqrt{0'21}} \approx 0'1455$$

Que el valor de r sea positivo, indica que la relación entre las variables es directa, es decir, que D aumenta, cuando aumenta C . Además, el hecho de que el valor de $|r|$ esté más próximo a 0 que a 1, indica que la correlación entre las variables es débil, es decir, que no hay mucha relación lineal entre ellas. □

²Que $r \in [-1, 1]$ es consecuencia de la expresión $\sigma_e^2 = \sigma_y^2(1 - r^2) \geq 0$ que se deduce en el caso lineal; y la interpretación de los posibles valores de r es consecuencia de la fórmula $r^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}$ que se deriva de la expresión anterior.

El coeficiente de correlación lineal permite establecer una relación entre las pendientes de la recta de regresión Y/X y la recta de regresión X/Y para un mismo conjunto de datos.

Si multiplicamos el numerador y denominador de $m_{y/x}$ por σ_y y el de $m_{x/y}$ por σ_x obtenemos

$$m_{y/x} = \frac{Cov(X, Y)}{\sigma_x^2} \cdot \frac{\sigma_y}{\sigma_y} = r \cdot \frac{\sigma_y}{\sigma_x} \quad \text{y} \quad m_{x/y} = \frac{\sigma_y^2}{Cov(X, Y)} \cdot \frac{\sigma_x}{\sigma_x} = \frac{1}{r} \cdot \frac{\sigma_y}{\sigma_x}$$

siendo r el coeficiente de correlación lineal. Como el valor de r es siempre un número comprendido entre -1 y 1, se puede establecer la siguiente relación entre las pendientes de las rectas de regresión

$$|m_{y/x}| \leq |m_{x/y}|$$

Esta relación permite determinar cuál de las dos rectas de regresión es la de Y sobre X y cuál es la de X sobre Y comparando, simplemente, sus pendientes.

Ejemplo 2.19 Sean $y = 4x - 7$ e $y = x - 1$ las rectas de regresión de las variables X e Y cuya covarianza es 9. Calcular las medias y las varianzas de las variables X e Y y determinar el coeficiente de correlación lineal de Pearson.

Como las dos rectas de regresión pasan por el punto (\bar{x}, \bar{y}) , basta resolver el sistema de ecuaciones formado por las dos rectas para obtener el valor de \bar{x} y \bar{y} que resulta ser 2 y 1, respectivamente.

Para obtener el resto de las medidas es necesario determinar cual de las dos rectas es la Y/X , y cual es la X/Y . Para ello, utilizamos la relación que se establece entre sus pendientes: $|m_{y/x}| \leq |m_{x/y}|$. Por lo tanto, $y = x - 1$ es la recta de regresión Y/X , e $y = 4x - 7$ es la recta de regresión X/Y .

Como la covarianza es 9, y sabemos que $m_{y/x} = Cov(X, Y)/\sigma_x^2 = 1$, entonces se deduce que $\sigma_x^2 = 9$. Y, análogamente, como sabemos que $m_{x/y} = \sigma_y^2/Cov(X, Y) = 4$, entonces se deduce que $\sigma_y^2 = 36$.

Finalmente, para calcular el coeficiente de correlación lineal, aplicamos su definición:

$$r = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{9}{\sqrt{9} \cdot \sqrt{36}} = \frac{9}{18} = 0.5$$

□

Descomposición de la varianza

Las características del modelo lineal permiten expresar la varianza de la variable Y (varianza total), como suma de las varianzas residual y explicada. Esta fórmula se conoce como *descomposición de la varianza*.

Consideramos una muestra de tamaño N de una variable (X, Y) que toma los valores (x_i, y_i) con frecuencias relativas f_i para todo $i = 1, \dots, k$; y consideramos el modelo lineal de regresión $y = a + bx$ que determina las variables Y^* , que toma los valores $y_i^* = a + bx_i$ con frecuencias relativas f_i para todo $i = 1, \dots, k$, y la variable E , que toma los valores $e_i = y_i - y_i^*$ con frecuencias relativas f_i para todo $i = 1, \dots, k$.

Para el modelo lineal, como consecuencia de la primera ecuación normal, se verifican las siguientes propiedades:

(p1) Las medias de las variables Y e Y^* son iguales, es decir, $\bar{y} = \bar{y}^*$.

(p2) La suma de los residuos es cero y, por lo tanto, la media es cero, es decir, $\bar{e} = \sum_{i=1}^n e_i \cdot f_i = 0$.

Con estas propiedades podemos simplificar la fórmula de la varianza residual

$$\sigma_e^2 = \sum_{i=1}^k (e_i - \bar{e})^2 \cdot f_i \stackrel{(p1)}{=} \sum_{i=1}^k e_i^2 \cdot f_i = \sum_{i=1}^k (y_i - y_i^*)^2 \cdot f_i$$

y obtener una fórmula que relaciona las varianzas del modelo con la varianza total.

$$\sigma_y^2 = \sigma_{y^*}^2 + \sigma_e^2$$

Ejemplo 2.20 Calcular las varianzas residual y explicada para el modelo lineal calculado en el ejemplo 2.16 de la página 75 y comprobar que la varianza marginal de la variable D es la suma de las varianzas del modelo.

Para calcular las varianzas del modelo, necesitamos obtener las distribuciones de frecuencias de las variables D^* , que representa los valores estimados por el modelo, y E , que representa los residuos. Las distribuciones de ambas variables se obtienen a partir del modelo ajustado:

$$D = \frac{11}{18} + \frac{1}{9}C$$

- La varianza explicada es la varianza de la variable D^* que toma los valores $d_i^* = f(c_i)$ con las frecuencias absolutas de las modalidades c_i de la variable C :

C	n_i	f_i	D^*
0	6	0'3	11/18
1	12	0'6	13/18
2	2	0'1	15/18

y, por lo tanto, la varianza explicada toma el valor $\sigma_{d^*}^2 = \frac{4}{900} \approx 0'0044$.

- La varianza residual es la varianza de la variable E que toma los valores $e_{ij} = d_j - f(c_i)$ con las frecuencias absolutas n_{ij} de las modalidades (c_i, d_j) de la variable (C, D) :

e_{ij}	$d_1=0$	$d_2=1$	con frecuencias	n_{ij}	$d_1=0$	$d_2=1$
$c_1=0$	-11/18	7/18		$c_1=0$	2	4
$c_2=1$	-13/18	5/18		$c_2=1$	4	8
$c_3=2$	-15/18	3/18		$c_3=2$	0	2

y, por lo tanto, la varianza residual toma el valor $\sigma_e^2 = \frac{185}{900} \approx 0'2056$.

Si comparamos estas dos varianzas, con la varianza de la variable E , obtenemos la siguiente relación:

$$\sigma_{e^*}^2 + \sigma_e^2 = \frac{4}{900} + \frac{185}{900} = \frac{189}{900} = \frac{21}{100} = \sigma_d^2$$

que es una característica del modelo lineal de regresión. □

La descomposición de la varianza permite, en el caso lineal, definir el coeficiente de determinación (R^2) como el cociente entre la varianza explicada y la varianza total

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2} = \frac{\sigma_y^2 - \sigma_e^2}{\sigma_y^2} = \frac{\sigma_{y*}^2}{\sigma_y^2} \quad \text{con} \quad 0 \leq R^2 \leq 1$$

Además, sólo en el caso lineal, donde tiene sentido calcular el coeficiente de correlación lineal (r), se verifica la siguiente relación entre los coeficientes de correlación y determinación

$$R^2 = r^2$$

Obsérvese que el 2 que hay sobre r indica una potencia (elevar al cuadrado), mientras que el 2 de la expresión R^2 es simplemente un símbolo (notación), pues no se define lo que significa R .

El valor de R^2 , en el caso lineal, siempre es un número en el intervalo $[0, 1]$, de manera que si R^2 está próximo a 1 significa que el ajuste es “bueno” mientras que un valor de R^2 próximo a 0 indica que el modelo no es el adecuado.

Ejemplo 2.21 *Calcular el coeficiente de determinación para el modelo lineal calculado en el ejemplo 2.16 de la página 75 y determinar la bondad del ajuste.*

Sabiendo los valores de las varianzas, $\sigma_d^2 = \frac{189}{900} = 0'21$ de la variable D , y $\sigma_e^2 = \frac{185}{900}$ de la variable E (residuos), aplicamos la fórmula

$$R^2 = 1 - \frac{185/900}{189/900} = 1 - \frac{185}{189} = \frac{4}{189} \approx 0'0212$$

y obtenemos el valor del coeficiente de determinación que, al ser próximo a 0, indica que la correlación entre las variables C y D es débil, es decir, que no hay mucha relación entre ellas.

Otra forma más sencilla de calcular este coeficiente es aplicando la fórmula que lo relaciona con el coeficiente de correlación lineal

$$R^2 = r^2 = \frac{Cov(C, D)^2}{\sigma_c^2 \cdot \sigma_d^2} = \frac{0'04^2}{0'36 \cdot 0'21} = \frac{0'0016}{0'0756} \approx 0'0212$$

□

2.4. Modelos de regresión no lineal

El modelo de regresión lineal que hemos estudiado es el más utilizado habitualmente. Sin embargo, la forma de la nube de puntos puede sugerir la consideración de otros modelos de regresión. Como veremos, en general, recurriremos al método de los mínimos cuadrados para ajustar el modelo y determinar el valor de los coeficientes. Sin embargo, hay algunos modelos que se pueden reducir al caso lineal, aplicando alguna transformación algebraica, y utilizar las fórmulas obtenidas antes.

2.4.1. Linealización de modelos

En muchos casos, los modelos de regresión utilizados pueden reducirse al caso lineal que hemos estudiado. Para ello, se realizan algunas transformaciones algebraicas y se determina un cambio de variables. Para obtener el nuevo modelo se aplican los cambios de las variables, transformando todas las modalidades.

Por ejemplo, a partir del modelo $y = a \cdot b^x$ y aplicando logaritmos neperianos

$$y = a \cdot b^x \quad \Longrightarrow \quad \ln(y) = \ln(a) + \ln(b) \cdot x \quad \Longrightarrow \quad Y = A + B \cdot X$$

y podemos considerar el modelo lineal $Y = A + B \cdot X$ donde

$$Y = \ln(y) \quad , \quad X = x \quad , \quad A = \ln(a) \quad \text{y} \quad B = \ln(b)$$

Ahora, aplicamos a y el cambio de variable, transformando todas sus modalidades. En este caso, las modalidades de la nueva variable Y se obtiene calculando el logaritmo neperiano de las modalidades de la variable y . Por último, ajustamos la nueva nube de puntos a la recta para obtener los valores de A y B y poder calcular los coeficientes a y b del modelo original

$$a = e^A \quad \text{y} \quad b = e^B$$

Obsérvese que para aplicar esta reducción al caso lineal, es necesario que todos los valores de y sean positivos, pues estamos considerando su logaritmo.

Ejemplo 2.22 Ajustar el modelo $y = a \cdot e^{bx}$ a los siguientes datos:

Variable X	1	2	3	4	5
Variable Y	4'5	6'5	10'0	15'0	22'0

Si aplicamos logaritmos neperianos al modelo $y = a \cdot e^{bx}$ obtenemos

$$y = a \cdot e^{bx} \quad \Longrightarrow \quad \ln(y) = \ln(a) + b \cdot x \quad \Longrightarrow \quad Y = A + B \cdot X$$

y podemos considerar el modelo lineal $Y = A + B \cdot X$ donde

$$Y = \ln(y) \quad , \quad X = x \quad , \quad A = \ln(a) \quad \text{y} \quad B = b$$

Si aplicamos estas transformaciones a los valores de las variables, obtenemos la siguiente tabla:

Nueva variable $X = x$	1	2	3	4	5
Nueva variable $Y = \ln y$	1'504	1'872	2'303	2'708	3'091

Para estos datos, calculamos la recta de regresión Y/X que resulta ser $y = 1'0925 + 0'401x$. Y deshaciendo los cambios de variable, obtenemos que $a = e^A = e^{1'0925} = 0'272$ y $b = B = 0'401$. Por lo tanto, el modelo ajustado es

$$y = 0'272 \cdot e^{0'401x}$$

□

2.4.2. Ajuste parabólico

Consideramos una muestra $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ de una variable bidimensional (X, Y) . Nuestro objetivo es ajustar una función del tipo

$$y = a + bx + cx^2$$

Aplicando el método de los mínimos cuadrados, obtenemos la función

$$F(a, b, c) = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - a - bx_i - cx_i^2)^2$$

La solución del problema pasa por minimizar la función $F(a, b, c)$ para determinar los valores de a , b y c . Para ellos, se resuelve el siguiente sistema de ecuaciones normales

$$\left\{ \begin{array}{l} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \\ \frac{\partial F}{\partial c} = 0 \end{array} \right\} \iff \left\{ \begin{array}{lll} \sum_{i=1}^N y_i & = aN & +b \sum_{i=1}^N x_i + c \sum_{i=1}^N x_i^2 \\ \sum_{i=1}^N x_i y_i & = a \sum_{i=1}^N x_i & +b \sum_{i=1}^N x_i^2 + c \sum_{i=1}^N x_i^3 \\ \sum_{i=1}^N x_i^2 y_i & = a \sum_{i=1}^N x_i^2 & +b \sum_{i=1}^N x_i^3 + c \sum_{i=1}^N x_i^4 \end{array} \right\}$$

que escrito en forma matricial resulta

$$\begin{pmatrix} N & \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i^3 \\ \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i^3 & \sum_{i=1}^N x_i^4 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N x_i^2 y_i \end{pmatrix}$$

Este resultado se puede generalizar (observar la estructura y disposición de los elementos de las matrices en el caso polinómico) para ajustar un modelo polinómico de cualquier grado. De manera que el sistema de ecuaciones normales, para el modelo polinómico general

$$y = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$$

expresado en forma matricial es

$$\begin{pmatrix} N & \sum x_i & \sum x_i^2 & \dots & \sum x_i^n \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \dots & \sum x_i^{n+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \dots & \sum x_i^{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_i^n & \sum x_i^{n+1} & \sum x_i^{n+2} & \dots & \sum x_i^{2n} \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \\ \vdots \\ \sum x_i^n y_i \end{pmatrix}$$

cuya solución nos permite obtener los valores de los $n + 1$ parámetros a_0, a_1, \dots, a_n .

Ejemplo 2.23 Ajustar el modelo parabólico $D = a + b \cdot C + c \cdot C^2$ a los datos del ejemplo 2.2 de la página 54.

Consideremos la tabla estadística de la distribución de frecuencias de las variables C y D , a la que hemos añadido una serie de columnas que nos resultarán útiles.

c_i	d_i	n_i	$c_i n_i$	$c_i^2 n_i$	$c_i^3 n_i$	$c_i^4 n_i$	$d_i n_i$	$c_i d_i n_i$	$c_i^2 d_i n_i$
0	0	2	0	0	0	0	0	0	0
0	1	4	0	0	0	0	4	0	0
1	0	4	4	4	4	4	0	0	0
1	1	8	8	8	8	8	8	8	8
2	1	2	4	8	16	32	2	4	8
		20	16	20	28	44	14	12	16

Con los valores de la tabla, podemos construir el siguiente sistema de ecuaciones lineales:

$$\left\{ \begin{array}{l} \sum d_i = aN + b \sum c_i + c \sum c_i^2 \\ \sum c_i d_i = a \sum c_i + b \sum c_i^2 + c \sum c_i^3 \\ \sum c_i^2 d_i = c \sum c_i^2 + b \sum c_i^3 + c \sum c_i^4 \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{l} 14 = 20a + 16b + 20c \\ 12 = 16a + 20b + 28c \\ 16 = 20a + 28b + 44c \end{array} \right\}$$

cuya solución, determina los valores de los parámetros $a = \frac{2}{3}$, $b = -\frac{1}{6}$ y $c = \frac{1}{6}$, y por tanto, el modelo que buscamos es:

$$D = \frac{2}{3} - \frac{1}{6} \cdot C + \frac{1}{6} \cdot C^2$$

En la figura 2.11 se representa la nube de puntos (puntos azules) y la curva general de regresión (cruces en rojo), cada una de ellas con un número que indica su frecuencia absoluta. Además, se representa la parábola de regresión (línea discontinua) que se ajusta a estos datos.

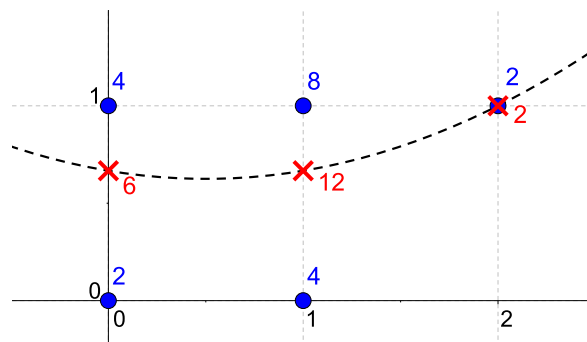


Figura 2.11: Ajuste parabólico

En este ejemplo, podíamos haber utilizado la curva general de regresión para ajustar el modelo parabólico y haber obtenido el mismo resultado. Como se observa en la figura, el modelo de regresión pasa exactamente por los puntos de la curva general de regresión, lo que supone que el ajuste es perfecto en el sentido de que los errores cometidos con cualquier otro modelo será siempre mayores.

Y todo esto ocurre porque el número de puntos distintos de la curva general de regresión (que es tres) coincide con el número de coeficientes del modelo (que también es tres por ser un modelo parabólico completo), en virtud de las propiedades de la curva general de regresión.

Si aplicamos el método de los mínimos cuadrados para ajustar cualquier otro modelo de regresión con más de tres coeficientes (por ejemplo un modelo cúbico completo) daría lugar a un sistema de ecuaciones normales compatible indeterminado, pues existirían infinitud de modelos (por ejemplo, infinitud de polinomios de grado 3) que se ajustan perfectamente a la nube de puntos, en el sentido de que, todo ellos, pasan por todos los puntos de la curva general de regresión. \square

2.4.3. Otros ajustes

En general, para ajustar un modelo de regresión, utilizaremos el método de los mínimos cuadrados descrito en la sección 2.2.2. Este método que ya hemos usado para determinar el modelo lineal y el polinómico, se resume en los siguientes pasos:

1. Consideramos el conjunto de datos $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$.
2. Representamos la nube de puntos para determinar qué modelo resulta más adecuado.
3. Si el modelo es $f(x)$ y depende de los parámetros a_1, a_2, \dots, a_n , entonces consideramos la función

$$F(a_1, a_2, \dots, a_n) = \sum_{i=1}^N (y_i - f(x_i))^2$$

4. Calculamos todas las derivadas parciales de la función F y las igualamos a 0 para obtener el sistema de ecuaciones normales.
5. Al resolver este sistema obtenemos el valor de los parámetros que determinan el modelo de regresión ajustado.

Ejemplo 2.24 *Obtener una fórmula que permita determinar el modelo de regresión $y = bx$ para el conjunto de datos $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$.*

Para aplicar el método de los mínimos cuadrados, debemos minimizar la función

$$F(b) = \sum_{i=1}^N (y_i - bx_i)^2$$

que sólo depende de un parámetro. En este caso, la derivada de F , igualada a 0, determina la ecuación normal:

$$\frac{dF}{dx}(b) = 0 \implies \sum_{i=1}^N x_i y_i - b \sum_{i=1}^N x_i^2 = 0$$

La solución de esta ecuación determina el mínimo³ de la función F que corresponde al valor del coeficiente b , calculado a partir del conjunto de puntos.

$$b = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2}$$

³El punto crítico obtenido es el mínimo de la función F pues $\frac{d^2}{dx^2} F(b) = 2 \sum x_i^2 > 0$.

OBSERVACIÓN: Aunque el modelo $y = bx$ que hemos ajustado es lineal, no debemos confundirlo con el modelo lineal general $y = a + bx$. Un error muy común, que debemos evitar, es aplicar la fórmula $Cov(X, Y)/\sigma_x^2$ del modelo lineal general para calcular el valor del parámetro b , considerando que el término independiente (a) es igual a 0. \square

En ocasiones, es posible aplicar los dos métodos (linealizar o aplicar, directamente, mínimos cuadrados) a un mismo modelo de regresión. Veamos un ejemplo.

Ejemplo 2.25 *Ajustar el modelo $y = ax + bx^3$ de dos maneras distintas (linealización del modelo y método de los mínimos cuadrados) para ajustarlo al siguiente conjunto de datos de la variable (X, Y) :*

$$\{(1, 5), (2, 8), (3, 9), (4, 8), (5, 0)\}$$

En primer lugar ajustamos el modelo reduciéndolo a un modelo lineal:

- Si dividimos por x la expresión del modelo obtenemos

$$y = ax + bx^3 \implies \frac{y}{x} = a + bx^2 \implies Y = A + B \cdot X$$

y podemos considerar el modelo lineal $Y = A + B \cdot X$ donde

$$Y = \frac{y}{x}, \quad X = x^2, \quad A = a \quad \text{y} \quad B = b$$

Si aplicamos estas transformaciones a los valores de las variables, obtenemos el siguiente conjunto de datos de las nuevas variables (X, Y) :

$$\{(1, 5), (4, 4), (9, 3), (16, 2), (25, 0)\}$$

Para estos valores, calculamos la recta de regresión Y/X que resulta ser $Y = 4'9765 - 0'1979 \cdot X$. Deshaciendo los cambios de variable, obtenemos que $a = A = 4'9765$ y que $b = B = -0'1979$. Por tanto, el modelo ajustado es

$$y = 4'9765x - 0'1979x^3$$

Obsérvese que este método no podría haberse utilizado si el valor de x de alguna de las observaciones hubiese sido 0, pues no hubiese sido posible aplicar la transformación.

Y ahora, utilizamos el método de los mínimos cuadrados para ajustar directamente el mismo modelo:

- En primer lugar, consideramos la función

$$F(a, b) = \sum (y_i - ax_i - bx_i^3)^2$$

Después, calculamos todas las derivadas parciales de la función F , y las igualamos a 0 para obtener el sistema de ecuaciones normales.

$$\left\{ \begin{array}{l} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \end{array} \right\} \iff \left\{ \begin{array}{l} \sum x_i y_i = a \sum x_i^2 + b \sum x_i^4 \\ \sum x_i^3 y_i = a \sum x_i^4 + b \sum x_i^6 \end{array} \right\}$$

que aplicado a los valores de nuestras variables, resulta ser el sistema de ecuaciones:

$$\begin{aligned} 80 &= 55a + 979b \\ 824 &= 979a + 20515b \end{aligned}$$

cuya solución, determina el valor de los parámetros $a \approx 4'912$ y $b \approx -0'194$ que determinan el modelo de regresión ajustado:

$$y = 4'912x - 0'194x^3$$

En este ejemplo, para hacer el ajuste, no tendría sentido simplificar el conjunto de observaciones utilizando la curva general de regresión, pues todos los valores de la variable independiente son distintos, y por lo tanto, la curva general de regresión coincide con la propia nube de puntos. \square

Obsérvese, en el ejemplo anterior, que aunque son muy parecidos, los coeficientes de los modelos obtenidos por cada uno de los métodos son distintos. El objetivo de los dos métodos es minimizar los errores, sin embargo, la transformación aplicada en la linealización del modelo distorsiona estos errores. Por lo tanto, el uso directo del método de los mínimos cuadrados proporciona un modelo más ajustado que el método de linealización, si bien, en muchos casos puede resultar más sencillo aplicar este último.

2.4.4. Bondad del ajuste

En los modelos que se reducen al caso lineal, se suele calcular el coeficiente de correlación lineal para el modelo transformado que es tipo lineal. Este coeficiente se puede utilizar como indicativo de la bondad del propio ajuste. Sin embargo, no debemos utilizarlo para comparar dos ajustes distintos.

En los modelos polinómicos completos (con todos sus términos) se verifica la fórmula de la descomposición de la varianza que, en general, no es cierta para cualquier modelo. Por lo tanto, en el caso polinómico resulta apropiado utilizar el coeficiente de determinación (R^2) como medida de correlación. Además, este coeficiente toma valores en el intervalo $[0, 1]$ con la misma interpretación que se le daba en el caso lineal.

Ejemplo 2.26 Utilizar el coeficiente de determinación para estudiar la bondad de los modelos lineal y parabólico ajustados a los datos del ejemplo 2.16 de la página 75.

Para el modelo lineal $D = \frac{11}{18} + \frac{1}{9}C$ podemos determinar los residuos (E)

e_{ij}	$d_1=0$	$d_2=1$		n_{ij}	$d_1=0$	$d_2=1$
$c_1=0$	$-11/18$	$7/18$	con frecuencias	$c_1=0$	2	4
$c_2=1$	$-13/18$	$5/18$		$c_2=1$	4	8
$c_3=2$	$-15/18$	$3/18$		$c_3=2$	0	2

y calcular la varianza residual $\sigma_e^2 = 0'2056$ que nos permite calcular el coeficiente de determinación para el modelo lineal

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_d^2} = 1 - \frac{0'2056}{0'21} = 0'0212$$

Para el modelo parabólico $y = \frac{2}{3} - \frac{1}{6}C + \frac{1}{6}C^2$ podemos determinar los residuos (E)

e_{ij}	$d_1=0$	$d_2=1$	con frecuencias	n_{ij}	$d_1=0$	$d_2=1$
$c_1=0$	-2/3	1/3		$c_1=0$	2	4
$c_2=1$	-2/3	1/3		$c_2=1$	4	8
$c_3=2$	-1	0		$c_3=2$	0	2

y calcular la varianza residual $\sigma_e^2 = 0'2$ que nos permite calcular el coeficiente de determinación para el modelo lineal

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_d^2} = 1 - \frac{0'2}{0'21} = 0'0476$$

En ambos casos, el coeficiente de determinación es muy próximo a cero, lo que indica que los ajustes no son apropiados. Además, de los resultados se deduce que la parábola es un modelo mejor que la recta para ajustar los datos de la muestra, pues el valor de R^2 es mayor. Esta conclusión es siempre cierta para estos dos modelos en cualquier conjunto de datos pues la expresión de la parábola $y = a + bx + cx^2$ generaliza a la de la recta $y = a + bx$, que es un caso particular que se obtiene cuando $c = 0$. \square

En general, para determinar el grado de bondad de un modelo cualquiera, se suele utilizar el coeficiente de determinación. Sin embargo, hay que tener en cuenta que, sólo en el caso polinómico completo, incluyendo el caso lineal, este coeficiente toma un valor entre 0 y 1. Por esa razón, para comparar la bondad de dos ajustes cualesquiera, a una misma nube de puntos, es preferible utilizar el coeficiente SSE que determina la suma de los cuadrados de los residuos

$$\text{SSE} = \sum_{i=1}^N e_i^2 n_i = \sum_{i=1}^N (y_i - f(x_i))^2 n_i \quad \text{o bien} \quad \text{SSE} = \sum_{i=1}^k \sum_{j=1}^p e_{ij}^2 n_{ij} = \sum_{i=1}^k \sum_{j=1}^p (y_j - f(x_i))^2 n_{ij}$$

Ejemplo 2.27 Determinar qué modelo, el lineal o el parabólico, se ajusta mejor a los datos del ejemplo 2.16 de la página 75.

Para el modelo lineal podemos determinar los residuos (ver ejemplo anterior) y calcular el valor de $\text{SSE} = 37/9 \approx 4'111$. De la misma manera, para el modelo parabólico podemos determinar los residuos (ver ejemplo anterior) y calcular el valor de $\text{SSE} = 4$.

La curva general de regresión establece una cota inferior del valor de SSE para cualquier modelo que se ajuste a este conjunto de datos.

n_{ij}	0	1	D
0	2	4	6
1	4	8	12
2	0	2	2
C	6	14	20

,

c_i	\bar{d}_i	n_i
0	2/3	6
1	2/3	12
2	1	2
		20

 $\xrightarrow{(*)}$

$$\sum_{i=1}^3 \sum_{j=1}^2 (d_j - \bar{d}_i)^2 n_{ij} = 4$$

$$(*) \sum_{i=1}^3 \sum_{j=1}^2 (d_j - \bar{d}_i)^2 n_{ij} = 2(0 - \frac{2}{3})^2 + 4(1 - \frac{2}{3})^2 + 4(0 - \frac{2}{3})^2 + 8(1 - \frac{2}{3})^2 + 2(1 - 1)^2 = 4$$

Lo que significa que cualquier modelo que se ajuste a los datos del ejemplo, por el método de los mínimos cuadrados, debe tener un valor de SSE mayor o igual a 4. El hecho de que el modelo parabólico haya sido exactamente 4, indica que este ajuste parabólico es perfecto, en el sentido de que ningún otro ajuste puede disminuir la suma de los cuadrados de los residuos. \square

2.5. Relación de problemas

1. En la elaboración de la siguiente tabla de frecuencias de la variable (X, Y) se ha cometido un error.

$Y \setminus X$	0	1	2	3	4	5	
$[0, 4]$	3	3	1	0	0	0	7
$(4, 6]$	3	4	2	0	0	0	9
$(6, 8]$	1	3	2	1	0	0	7
$(8, 12]$	0	0	1	2	3	2	9
	7	11	6	3	3	2	32

Se pide:

- Detectar y corregir la errata.
 - Representar la distribución condicionada $(Y/X = 2)$ y calcular el sesgo y la curtosis.
 - Calcular las rectas de regresión de X sobre Y y de Y sobre X .
 - Calcular el coeficiente de correlación lineal y la varianza residual del modelo lineal Y/X .
2. Demostrar la igualdad de las dos siguientes fórmulas que permiten calcular la covarianza:

$$\text{Cov}(X, Y) = \sum_{i=1}^k \sum_{j=1}^p (x_i - \bar{x}) \cdot (y_j - \bar{y}) \cdot f_{ij} = \sum_{i=1}^k \sum_{j=1}^p x_i \cdot y_j \cdot f_{ij} - \bar{x}\bar{y}$$

3. Demostrar las siguientes propiedades de los momentos ordinarios y centrales que se establecen en la sección 2.1.5 de la página 60:

$$\begin{array}{lll} m_{00} = 1 & m_{10} = \bar{x} & m_{01} = \bar{y} \\ \mu_{00} = 1 & \mu_{10} = 0 & \mu_{01} = 0 \end{array}$$

4. La siguiente tabla recoge los valores de fuerza (F) y elongación (E), registrados en 6 pruebas de tensión de acero.

F	1	2	3	4	5	6
E	15	35	41	63	77	84

Estimar el modelo lineal de regresión E/F y obtener una medida de la bondad del ajuste.

5. Representar gráficamente los datos de las muestras de las variables (X, Y_i) con $i = 1, 2, \dots, 7$ que se proporciona en la siguiente tabla:

X	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
0	0	13	4	11	0	10	2
4	2	11	3	13	2	7	5
6	3	10	8	8	4	12	3
8	4	9	6	4	3	4	8
12	6	7	7	7	7	5	4
14	7	6	13	6	6	2	4
16	8	5	2	3	8	8	10
22	11	2	11	2	11	4	12
26	13	0	0	1	13	5	6

- a) A la vista de las gráficas, elegir, en la siguiente lista, un valor para el coeficiente de correlación lineal de cada una de las muestras anteriores y justificar la elección.

$$-1 \quad , \quad -0'875 \quad , \quad -0'543 \quad , \quad 0 \quad , \quad 0'606 \quad , \quad 0'986 \quad , \quad 1$$

- b) Calcular los coeficientes de correlación lineal y comprobar que se ha elegido correctamente.

6. Representar gráficamente, calcular la recta de regresión y determinar el grado de correlación lineal de la variable X con cada una de las variables Y que se presentan en la siguiente tabla:

X	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
1	4	1	6	1	7	6	1
2	2	3	5	1	5	4	4
3	3	5	4	3	4	3	2
4	2	7	3	5	2	1	1
5	4	9	2	6	2	5	5

7. Los siguientes datos están tomados de un estudio sobre el flujo de tráfico a través de un túnel para vehículos. Las cifras son los valores promedio basados en las observaciones que se hicieron en 10 intervalos de 5 minutos.

Densidad(veh/km)	43	55	40	52	39	33	50	33	44	21
Velocidad(km/h)	27'0	23'8	30'7	24'0	34'8	41'4	27'0	40'4	31'7	51'2

Se pide:

- a) Representar el diagrama de dispersión.
- b) A la vista del diagrama, elegir el valor correcto de r entre estos tres valores: 0'968, -0'968, -0'198.
- c) Verificar la respuesta calculando r .
- d) ¿Hay alguna evidencia real de que exista asociación entre la velocidad de los vehículos y la densidad?

8. Recordando que dos variables son linealmente incorreladas si $r = 0$. Se pide

- a) Justificar que 2 variables aleatorias son linealmente incorreladas si y solo si su covarianza es 0.
- b) Dados los puntos (1,0), (2,1), (4,1) y (5,a), hallar el valor de a sabiendo que las variables X e Y son incorreladas. Determinar las rectas de regresión.

9. Veamos la importancia de la representación gráfica de los datos. Las siguientes tablas presentan tres conjuntos de datos que tienen la misma correlación y la misma recta de

regresión:

X_1	Y_1	X_2	Y_2	X_3	Y_3
10	8'04	10	9'14	10	7'70
8	6'95	8	8'14	8	6'60
13	7'58	13	8'74	13	9'60
9	8'81	9	8'77	9	7'80
11	8'33	11	9'26	11	8'70
14	9'96	14	8'10	14	9'90
6	7'24	6	6'13	6	7'96
4	4'26	4	3'10	4	5'92
12	10'84	12	9'13	12	8'80
7	4'82	7	7'26	7	6'90
5	5'68	5	4'74	5	2'62

- Calcular la recta de regresión y el coeficiente de correlación lineal de cada conjunto y comprobar que son iguales.
 - Utilizar el diagrama de dispersión para representar los conjuntos de datos junto a la recta de regresión calculada.
 - ¿En qué conjunto de datos utilizarías la recta de regresión para predecir el valor de la variable Y cuando $X = 16$?
10. Sea (X, Y) una variable estadística bidimensional. La variable X presenta las modalidades a y 0 mientras que la variable Y toma los valores $a - 1$ y 1. Además, se conoce que la proporción de datos muestrales que presentan la modalidad 0 en la variable X es 0'75 y la proporción de datos muestrales que presentan la modalidad $a - 1$ en la variable Y es 0'5. Sabiendo que la recta de regresión mínimo cuadrática de X sobre Y es $X + Y = 1$. Calcular:
- El coeficiente de correlación.
 - Estimar el valor de X para $Y = 0$ y el de Y para $X = 1$.
11. Las rectas $x - 2y = 4$ y $2x - 9y = 8$ son las rectas de regresión de una variable estadística bidimensional (X, Y) , con $N = 10$ y $\sigma_x^2 = 9$.
- Hallar el coeficiente de correlación lineal, la varianza de Y y la covarianza.
 - Si se descubre que uno de los puntos considerados, el $(2, -1)$, no debería haberse utilizado, hallar las nuevas rectas de regresión.
12. A partir de 30 observaciones de una variable estadística bidimensional (X, Y) se obtuvieron las rectas de regresión: $X = (Y - 1)/2$ e $X = Y - 1$, sabiéndose que la varianza de X es 1. Más adelante se obtuvo una nueva observación que resultó ser el punto $(0, 1)$.
- Obtener las nuevas rectas de regresión.
 - Las varianzas residuales de ambos ajustes. ¿Han aumentado o disminuido?
 - ¿Mejoran los ajustes al tomar una nueva observación?
13. Sea una regresión lineal mínimo cuadrática del tipo Y/X obtenida a partir de N observaciones de una variable estadística bidimensional (X, Y) , con centro de gravedad el origen de coordenadas.

Con objeto de obtener más información, se realiza una nueva observación, que resulta ser de nuevo el centro de gravedad.

Ante la duda de que esta información adicional que parece reiterativa, no aporte nada nuevo, se decide realizar una nueva regresión lineal del tipo Y/X con las $N+1$ observaciones.

- a) Estudiar si esta información es de utilidad, pues hace disminuir la varianza residual.
- b) Comprueba si aumenta o no, el coeficiente de correlación lineal.
- c) ¿ En qué tanto por ciento como máximo, disminuye la varianza residual con respecto a la inicial ?

14. Consideremos los siguientes modelos de regresión:

$$y = a \cdot e^{bx} \quad , \quad y = a \cdot x^b \quad , \quad y = \frac{1}{a + b \cdot x}$$

Para cada uno de ellos, se pide:

- a) Determinar los cambios de variable necesarios para reducir los siguientes modelos al caso lineal.
- b) Determinar las ecuaciones que permiten calcular los coeficientes del modelo original, a partir de los coeficientes del modelo lineal
- c) Determinar las restricciones que debe verificar el conjunto de datos para poder aplicar la reducción.

15. Ajustar el modelo $y = a \cdot b^x$ (reduciéndolo al caso lineal) a los siguientes datos:

Variable X	1	2	3	4	5
Variable Y	3'0	4'5	7'0	10'0	15'0

16. Ajustar el modelo $y = a \cdot x^b$ (reduciéndolo al caso lineal) a los siguientes datos:

Variable X	1	2	3	4	5
Variable Y	0'5	2'0	4'5	8'0	12'5

17. Ajustar el modelo $y = \frac{1}{a + b \cdot x}$ (reduciéndolo al caso lineal) a los siguientes datos:

Variable X	1	2	3	4	5
Variable Y	1'00	0'50	0'33	0'25	0'20

18. Consideramos la muestra (1,0), (2,1), (1,2), (-1,0), (2,2) de la variable (X,Y). Se pide:

- a) Ajustar un modelo del tipo $Y = a + b(1/X)$.
- b) Ajustar la recta Y/X .
- c) ¿Qué modelo resulta más apropiado?

19. Dados los puntos: (1,1) , (2,1) , (3,2) , (4,4) y (5,8), se pide:

- a) Estudiar si resultaría conveniente realizar un ajuste lineal.
- b) Ajustar una función del tipo $y = a \cdot b^x$.

- c) Utilizar los modelos para predecir y comparar los valores y para $x = 6$ y $x = 10$. A la vista de los resultados, elegir el modelo más adecuado para la predicción y justificar la respuesta.
- d) Comparar los dos modelos utilizando el coeficiente de correlación lineal y SSE.
20. Dados los puntos $(0,0'9)$, $(2,1/3)$, $(3,1/7)$, $(4,1/10)$ y $(6,1/82)$ obtener los coeficientes del ajuste por transformación al modelo lineal, para una relación entre ambas variables del tipo $y = 1/(ab^x + 1)$.
21. Se probó el desgaste (d en $mm.$) de seis moldes, probando cada uno de ellos bajo una diferente temperatura (t en unidades de $100^\circ C$) de operación controlada en un baño de aceite. Los resultados de la prueba fueron:

t	1	1,5	2	3	3,5	4
d	3'3	5'0	5'5	9'4	11'4	12'8

Puede suponerse que los valores de la temperatura no tienen error y hay bases para suponer que el desgaste y la temperatura están relacionados por una función lineal. Se pide:

- a) Obtener la ecuación del modelo lineal de regresión.
- b) Estimar el desgaste cuando la temperatura de operación es $250^\circ C$.
- c) Elegir otro modelo de regresión que resulte más apropiado y que no contemple desgaste cuando la temperatura es de 0 grados.
22. Vamos a estudiar el movimiento uniformemente acelerado de un objeto a partir de los datos del espacio (e) y del tiempo (t) recogidos en la siguiente tabla:

tiempo	1	2	3	4	5	6	7
espacio	13	41	67	119	176	245	333

- a) Ajustar mediante mínimos cuadrados la expresión del espacio en función del tiempo.
- b) Estimar el espacio inicial, la velocidad inicial y la aceleración.
- c) Predecir el espacio recorrido cuando $t = 10$.
- d) Hallar la nueva ecuación considerando el nuevo dato $e = 6$ para $t = 0$. (Observación: utilizar los cálculos anteriores).
23. El tiempo total necesario para detener un automóvil después de percibir un peligro se compone del tiempo de reacción más el tiempo de frenado. Por tanto, la velocidad del vehículo no es suficiente para calcular este tiempo total aplicando las leyes de la mecánica. Para estudiar este fenómeno se considera la siguiente tabla que contiene las distancias (d en metros) de frenada de un automóvil que marcha a la velocidad (v en Km/h) desde el instante en que se observa el peligro.

v	30	45	60	75	90	105
d	1'30	2'25	3'50	5'20	7'50	10'50

Representar gráficamente los datos, determinar un modelo que se ajuste a la nube de puntos y utilizarlo para estimar d cuando v es $80 Km/h$. (interpolación) y $120 km/h$. (extrapolación). Estudiar las limitaciones del modelo.

24. Ajustar una recta y una parábola de regresión Y/X al conjunto de puntos $\{(2, 3), (3, 4), (8, 9), (9, 8)\}$. Comprobar que la parábola se ajusta mejor a los datos y justificar por qué ocurre siempre esto, independientemente del conjunto de puntos.

25. Dada la tabla:

$X \backslash Y$	0	1	2
20	2	0	0
30	1	3	2
40	1	3	2
50	2	0	0

se pide

- Ajustar un modelo lineal de regresión.
 - Calcular el coeficiente de correlación lineal y la covarianza.
 - Estudiar la dependencia e independencia de las distribuciones.
 - Ajustar una parábola de regresión y comparar la bondad del modelo con el caso lineal.
26. **Cambio de variable.** Al analizar los datos, a veces conviene aplicar una transformación que simplifique su aspecto general. La siguiente tabla muestra el contenido de oxígeno Y a los X metros de profundidad de un lago:

X	10	20	30	40	50	60	70
Y	6'5	5'6	5'4	6'0	4'6	1'4	0'1

Dar respuesta a las siguientes cuestiones:

- Aplicar el cambio de variable $X' = (X - 40)/10$ y calcular la media de la nueva variable X' .
 - Ajustar la recta de regresión Y/X' .
 - Estudiar la correlación lineal.
 - Utilizar el modelo para predecir el contenido de oxígeno a los 65 metros.
 - Ajustar una parábola y comparar la bondad del ajuste con el modelo lineal.
27. Dada la siguiente tabla de frecuencias:

$Y \backslash X$	1	2	3	4	5
3 - 4	5	3	1		
4 - 5	1	2	1	2	
5 - 6			4	3	1
6 - 7			1	2	2
7 - 8					2

- Aplicar el cambio de variable $W = X - 3$ y $Z = Y - 5'5$.
- Calcular la recta de regresión Z/W .
- Ajustar el modelo parabólico de regresión $z = a + bw^2$.
- Ajustar el modelo de regresión $z = a + bw + cw^3$.
- Comparar la bondad de los modelos utilizando una medida de correlación apropiada.

28. El número de agricultores españoles, en millones viene dado por los puntos (1973, 9'47), (1974, 9'26), (1975, 8'86), (1976, 8'25), (1977, 7'81), (1978, 8'01), (1979, 7'55), (1980, 7'24), (1981, 7'01), (1982, 6'88) y (1983, 7'03).
- Aplicar una traslación a los años para obtener una nueva variable con media 0.
 - Predecir el número de agricultores en el año 1970 suponiendo una dependencia lineal entre las variables.
 - Hallar el coeficiente de correlación lineal.
 - Ajustar una curva del tipo $y = a \cdot b^x$ y comparar la bondad del ajuste con el modelo lineal.
29. Estudiar en qué medida le afectan los cambios de origen y de escala al coeficiente de correlación lineal.
30. Los datos que muestra el siguiente ejemplo provienen del registro del número de automóviles que salen de una población grande por la carretera principal hacia la costa en cada uno de los 10 domingos seleccionados al azar. Las observaciones se hicieron en un punto de observación sobre la carretera durante un intervalo de tiempo fijo, y para mantener los números sencillos, se expresan redondeándolos al 1000 más cercano. También se muestra la temperatura (en grados centígrados) que se registró en la población al principio del día.

t	13	16	9	10	18	23	19	27	15	10
v	18	19	9	12	21	25	26	30	24	14

Se pide:

- Representar gráficamente los datos.
 - Elegir y ajustar un modelo que permita establecer la relación que existe entre la temperatura (t) y el número de vehículos (v).
 - Justificar la elección del modelo del apartado anterior.
31. Algunas veces se requiere que la curva de regresión pase por el origen. En estos casos, elegimos modelos que no tengan término independiente, como en el siguiente ejercicio. Ajustar el modelo $E = aC$ a los siguientes datos obtenidos en un experimento para determinar la rigidez de un resorte. Se midió la extensión (E) del resorte (a partir de su longitud natural) bajo la acción de diferentes cargas (C):

Carga (Newtons)	2	4	6	8	10	12
Extensión (mm)	10	19	29	40	48	56

32. **Regresión múltiple.** En la tabla, z representa una propiedad física particular de las barras de acero forjado, y x e y son los porcentajes de elementos a y b que se encuentran presentes en la aleación. Se escogieron cuatro niveles para x y cuatro para y , lo que da 16 posibles combinaciones, y se registró experimentalmente un valor de z para barra de cada tipo. (Este es un ejemplo de lo que se conoce como diseño factorial completo).

x	5	5	5	5	10	10	10	10	15	15	15	15	20	20	20	20
y	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
z	28	30	48	74	29	50	57	42	20	24	31	47	9	18	22	31

- a) Demostrar que las ecuaciones normales para el modelo lineal de regresión múltiple $z = a + b \cdot x + c \cdot y$ en forma matricial son

$$\begin{pmatrix} n & \sum x & \sum y \\ \sum x & \sum x^2 & \sum xy \\ \sum y & \sum xy & \sum y^2 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum z \\ \sum xz \\ \sum yz \end{pmatrix}$$

- b) Resolver el sistema para deducir las fórmulas que determinan los coeficientes a , b y c en función de los momentos:

$$a = \bar{z} - b\bar{x} - c\bar{y} \quad , \quad b = \frac{\sigma_{xz}\sigma_y^2 - \sigma_{yz}\sigma_{xy}}{\sigma_x^2\sigma_y^2 - (\sigma_{xy})^2} \quad \text{y} \quad c = \frac{\sigma_x^2\sigma_{yz} - \sigma_{xy}\sigma_{xz}}{\sigma_x^2\sigma_y^2 - (\sigma_{xy})^2}$$

- c) Utilizar las fórmulas anteriores para calcular los valores de a , b y c a partir de los datos del experimento.
- d) La linealidad del modelo significa que hay una relación lineal entre z y x cuando y está fija, y entre z e y cuando x está fija. Sobre el diagrama de dispersión, representar las distintas rectas obtenidas al fijar los valores de y que se investigan (1, 2, 3 y 4). Hacer lo mismo para los valores de x que se investigan (5, 10, 15 y 20).
- e) Calcular el mayor valor de z estimado por el modelo dentro del intervalo de valores de x e y que se investigan. Justificar la respuesta.

33. Fórmulas de codificación.

- a) Obtener la fórmula lineal de codificación que transforme respectivamente los valores 5, 10, 15 y 20 de la variable x en los valores en -3, -1, 1 y 3 de la variable u .
- b) Análogamente, obtener la fórmula lineal de codificación que transforme los valores 1, 2, 3 y 4 de la variable y en los valores en -3, -1, 1 y 3 de la variable v .
- c) Aplicar estas fórmulas de codificación a los datos del ejercicio 32 de la página 95 y comprobar que el nuevo modelo de regresión lineal múltiple $z = a + bu + cv$ coincide con la ecuación que relacionaba x e y con z .

34. Consideramos los datos (1,2,1), (1,4,3), (2,2,4), (2,2,5), (2,4,3), (1,4,3) y (2,4,5) de una muestra de la variable (x, y, z) . Se pide:

- a) Ajustar un plano de regresión a la nube de puntos.
- b) Ajustar un modelo del tipo $z = a + b \cdot \ln(xy)$.
- c) Determinar el modelo de regresión más apropiado.

35. Para determinar el modelo $y = a + bx$, aplicando el método de los mínimos cuadrados, tenemos que minimizar la función $F(a, b) = \sum (y_i - a - bx_i)$. Los valores de a y b obtenidos, resolviendo la ecuación $\nabla F(a, b) = 0$, son puntos críticos de la función F , pero ¿son mínimos de la función? Para ello, es necesario aplicar algún criterio de clasificación de extremos de un campo escalar. Un criterio sencillo, que podemos aplicar aquí, consiste en calcular la matriz Hessiana de F ($\nabla^2 F(a, b)$) y comprobar que, tanto el elemento que ocupa la posición (1,1), como el determinante de la matriz, son números positivos. Se pide:

- a) Calcule la matriz Hessiana de F y úsela para determinar que el punto crítico obtenido, aplicando el método de los mínimos cuadrados, es un mínimo de la función F .
- b) Realice esta misma comprobación para el modelo parabólico de regresión.

2.6. Anexo I: Justificación de algunos resultados

En esta sección vamos a presentar la justificación de algunos de los resultados que hemos visto en este tema. Incluiremos aquellas demostraciones que utilizan resultados básicos de matemáticas o aquellas que se apoyan en los conocimientos aprendidos en otras asignaturas de matemáticas de la titulación.

Consideramos una muestra de tamaño N de una variable bidimensional (X, Y) que toma los valores (x_i, y_i) con frecuencias absolutas n_i , siendo $N = \sum n_i$, y frecuencias relativas f_i para todo $i = 1, \dots, k$.

Sea $y = a + bx$ la recta de regresión de Y sobre X con $b = \sigma_{xy}/\sigma_x^2$ y $a = \bar{y} - b\bar{x}$, y consideremos las variables Y^* (de los valores estimados) que toma los valores $y_i^* = a + bx_i$ con frecuencias f_i , y la variable E (de los residuos) que toma los valores $e_i = y_i - y_i^*$ con frecuencias f_i .

2.6.1. Descomposición de las varianzas para el modelo lineal de regresión

Vamos a demostrar que, en el caso lineal, se verifica la propiedad: $\sigma_y^2 = \sigma_{y^*}^2 + \sigma_e^2$. Para ello, vamos a demostrar el resultado equivalente $\sigma_e^2 = \sigma_y^2 - \sigma_{y^*}^2$, mediante la siguiente cadena de igualdades:

$$\begin{aligned} \sigma_e^2 &= \sum_{i=1}^N (y_i - a - bx_i)^2 f_i \stackrel{(1)}{=} \sum_{i=1}^N (y_i - \bar{y} + b\bar{x} - bx_i)^2 f_i = \sum_{i=1}^N [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 f_i = \\ &= \sum_{i=1}^N [(y_i - \bar{y})^2 - 2b(y_i - \bar{y})(x_i - \bar{x}) + b^2(x_i - \bar{x})^2] f_i = \sigma_y^2 - 2b\sigma_{xy} + b^2\sigma_x^2 \stackrel{(2)}{=} \\ &\stackrel{(2)}{=} \sigma_y^2 - 2\frac{\sigma_{xy}}{\sigma_x^2} + \frac{\sigma_{xy}^2}{\sigma_x^2} = \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2} \stackrel{(3)}{=} \sigma_y^2 - \sigma_{y^*}^2 \end{aligned}$$

donde

(1) Sustitución: $a = \bar{y} - b\bar{x}$

(2) Sustitución: $b = \frac{\sigma_{xy}}{\sigma_x^2}$

(3) Sustitución: $\sigma_{y^*}^2 = \frac{\sigma_{xy}^2}{\sigma_x^2}$ pues si $y^* = a + bx$ entonces $\sigma_{y^*}^2 = b^2\sigma_x^2 = \frac{\sigma_{xy}^2}{\sigma_x^4}\sigma_x^2 = \frac{\sigma_{xy}^2}{\sigma_x^2}$

2.6.2. El coeficiente de correlación lineal de Pearson (r) es un número comprendido entre -1 y 1

Veamos que se verifica la propiedad: $\sigma_{y^*}^2 = r^2\sigma_y^2$

$$\sigma_{y^*}^2 \stackrel{(1)}{=} b^2\sigma_x^2 \stackrel{(2)}{=} \left(\frac{\text{Cov}(X, Y)}{\sigma_x^2}\right)^2 \sigma_x^2 \stackrel{(3)}{=} \left(\frac{\text{Cov}(X, Y)}{\sigma_x}\right)^2 \frac{\sigma_y^2}{\sigma_y^2} \stackrel{(2)}{=} \left(\frac{\text{Cov}(X, Y)}{\sigma_x\sigma_y}\right)^2 \sigma_y^2 \stackrel{(4)}{=} r^2\sigma_y^2$$

donde

- (1) Propiedad de la varianza frente a la transformación afín $Y^* = a + bX$
- (2) Operar y simplificar
- (3) Multiplicar y dividir por σ_y^2
- (4) Definición del coeficiente r

Considerando el resultado anterior y la fórmula de la descomposición de la varianza obtenemos:

$$\left. \begin{array}{l} \sigma_y^2 = \sigma_{y^*}^2 + \sigma_e^2 \\ \sigma_{y^*}^2 = r^2 \sigma_y^2 \end{array} \right\} \implies \sigma_y^2 = r^2 \sigma_y^2 + \sigma_e^2 \implies \sigma_e^2 = (1 - r^2) \sigma_y^2$$

Ahora bien, como $\sigma_e^2 \geq 0$ y $\sigma_y^2 \geq 0$ (por definición de la varianza) entonces $(1 - r^2) \geq 0$ y, por lo tanto, $r^2 \leq 1$, es decir, $-1 \leq r \leq 1$.

2.7. Anexo II: Comandos de R

```

> x=c(1,2,3,4,5)
> y=c(2,4,6,8,9)
> table(x,y)           # Tabla de doble entrada
> cov(x,y)             # Covarianza muestral (dividido por N-1)
> cor(x,y)             # Coef. Correlación lineal de Pearson

###  MODELOS DE REGRESIÓN LINEAL  ###

> reg1<-lm(y ~ x)           # Regresión lineal:  $Y = a_0 + a_1 * X$ 
> reg2<-lm(y ~ x+I(x^2)+I(x^3)) # Regresión:  $Y = a_0+a_1*X+a_2*X^2+a_3*X^3$ 

###  MODELOS DE REGRESION NO LINEAL  ###

> reg3<-nls(y ~ a*exp(b*x)   # Regresión:  $Y = a * e^{bX}$ 
> reg4<-nls(y ~ a*b^x)       # Regresión:  $Y = a * b^X$ 
> reg5<-nls(y ~ a+b*x)       # Regresión:  $Y = a + b * X$ 

###  DATOS DEL MODELO: Regresión y Correlación  ###

> reg=lm(y ~ x)
> plot(x,y);abline(reg) # Representa la nube de puntos y el modelo ajustado
> summary(reg)          # Resumen datos del modelo
> names(reg)            # Datos de la Regresión lineal almacenados en "reg"
> reg$fitted.values     # Valores estimados de "y" por el modelo
> reg$residuals         # Residuos estimados
> coef(reg)             # Coeficientes del Modelo
> resid(reg)            # Residuos del Modelo
> fitted(reg)           # Valores ajustados por el modelo

# Fórmulas para definir  $R^2$  y SSE

> 1-var(resid(reg))/var(y) # Coeficiente de determinación ( $R^2$ )
> sum(resid(reg)^2)        # Suma de los cuadrados de los residuos (SSE)

#####

> reg1<-lm(y~x)

> reg1

Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
          0.4          1.8

```

```

> summary(reg1)

Call:
lm(formula = y ~ x)
Residuals:
    1      2      3      4      5 
-2.000e-01  3.193e-16  2.000e-01  4.000e-01 -4.000e-01 
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4000     0.3830   1.044 0.373021
x             1.8000     0.1155  15.588 0.000574 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3651 on 3 degrees of freedom
Multiple R-squared:  0.9878, Adjusted R-squared:  0.9837 
F-statistic:  243 on 1 and 3 DF,  p-value: 0.0005737

#####

> reg2<-nls(y~a+b*x)

> reg2

Nonlinear regression model
  model: y ~ a + b * x
 data:  parent.frame()
  a      b
0.4 1.8
residual sum-of-squares: 0.4
Number of iterations to convergence: 1
Achieved convergence tolerance: 1.251e-07

> summary(reg2)

Formula: y ~ a + b * x
Parameters:
      Estimate Std. Error t value Pr(>|t|)
a    0.4000     0.3830   1.044 0.373021
b    1.8000     0.1155  15.588 0.000574 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3651 on 3 degrees of freedom
Number of iterations to convergence: 1
Achieved convergence tolerance: 1.251e-07

```