4. Estadística descriptiva bidimensional. Regresión lineal

4.1. Tablas de doble entrada

De la misma manera que a partir de una población se puede tomar una muestra de tamaño n para observar una variable, es posible observar varias variables para cada individuo de la muestra. Si observamos r variables $(X_1, ..., X_r)$, obtendremos para cada individuo un vector de r valores correspondientes a los valores de cada una de las variables para ese individuo. A $\overrightarrow{X} := (X_1, ..., X_r)$ se le llama **vector** o **variable estadística r-dimensional**. A la variable (unidimensional) X_k se la denomina la **componente k-ésima** del vector. Así, nuestros datos ahora serán de la forma $(x_{11}, ..., x_{1r}), ..., (x_{n1}, ..., x_{nr})$, donde x_{ij} denota el valor de la variable X_j para el individuo i-ésimo de la muestra. Si tenemos solamente dos variables, es más usual la representación (X, Y) en lugar de (X_1, X_2). En este capítulo trataremos el caso bidimensional, pero el estudio para tres o más variables se realiza análogamente.

El interés de estudiar varias variables simultáneamente radica en que es posible que los valores de las variables estén relacionados entre sí, de forma que los valores de una variable nos den información sobre los valores de otra u otras variables.

Ejemplo 44.

Consideremos nuevamente el ejemplo de las crías de conejo del capítulo anterior. En dicho capítulo habíamos estudiado el número de crías de 35 camadas. Supongamos ahora que observamos también el número de crías vivas después de dos meses. En este caso tenemos dos variables: $X \equiv$ número de crías, e $Y \equiv$ número de crías vivas dos meses

https://dx.doi.org/10.5209/docm.003.04

después. Para la camada i-ésima tenemos entonces dos valores (x_i, y_i). *Supongamos que los valores que tenemos son los que aparecen en la tabla 4.1.*

Camada	1	2	3	4	5	6	7	8	9	10	11	12
X	0	1	2	3	4	6	0	1	2	3	4	6
Y	0	0	1	2	4	4	0	1	1	3	3	5
Camada	13	14	15	16	17	18	19	20	21	22	23	24
X	1	2	3	4	6	2	3	4	6	2	3	4
Y	0	0	2	3	4	2	3	3	6	1	1	2
Camada	25	26	27	28	29	30	31	32	33	34	35	
X	6	2	3	2	3	2	3	2	3	2	3	
Y	4	1	3	1	2	0	3	1	2	0	3	

Tabla 4.1. Datos correspondientes al ejemplo 44.

Nótese que en el ejemplo anterior no tenemos 70 datos, sino que tenemos 35 pares de datos, que son los 35 datos correspondientes al número de crías inicialmente y los 35 datos correspondientes al número de crías vivas dos meses después; en definitiva, siempre el número de datos es n (número de individuos) y no $2n$.

El primer problema que nos planteamos es el de obtener una representación tabular de los datos. Podríamos considerar los distintos pares que aparecen y actuar como en el caso unidimensional, pero esto trae varios problemas:

- Las modalidades no pueden ordenarse de menor a mayor de manera natural y esto hace que la representación no sea tan clara, incluso aunque las dos componentes sean cuantitativas. Así, si tenemos los resultados (0,1) y (1,0), no tenemos ninguna razón para suponer que el primero de ellos es menor que el segundo. En consecuencia, habría que tratar la variable bidimensional como una variable cualitativa. Sin embargo, en muchas ocasiones nos va a interesar estudiar valores numéricos de cada componente, y esta forma de actuar complicaría el tratamiento.

- En muchas ocasiones, interesa estudiar las variables por separado, además de las relaciones que existen entre ellas, y esto no puede hacerse de forma eficaz a partir de la tabla de frecuencias.

Por ello, se plantea otra representación tabular más apropiada, la **tabla de doble entrada**. Esta tabla se construye de la siguiente manera: Se hace un cuadro en el que en la primera columna y la primera fila (que serán la columna y la fila 0) se dan las modalidades de las variables X e Y respectivamente, en orden creciente si es posible. Si la variable X tiene r modalidades diferentes y la variable Y tiene s modalidades diferentes, se obtiene una tabla de r filas y s columnas. En el cuadro intersección de cada fila y cada columna se da la frecuencia (absoluta o relativa) del par correspondiente; si es el valor correspondiente a la modalidad i-ésima de la variable X y la modalidad j-ésima de la variable Y, se denota este valor por n_{ij} en el caso de frecuencias absolutas y f_{ij} para las frecuencias relativas. Dicho de otra manera, el número o proporción de individuos de la muestra para los que simultáneamente X vale x_i e Y vale y_j aparece en la posición (i, j) de la tabla. Tendremos entonces $r \times s$ valores.

Una vez obtenida esta tabla, se añaden una última fila y una última columna. En ellas se dan los valores suma de todos los valores de la columna o fila correspondiente. Denotaremos esos valores por $n_{1.}, ..., n_{r.}$ para las filas y por $n_{.1}, ..., n_{.s}$ para las columnas. Por ejemplo, si consideramos frecuencias absolutas, en el primer valor de la última columna, se da el valor

$$n_{1.} = n_{11} + n_{12} + ... + n_{1s}.$$

De la misma manera, en el primer valor de la última fila aparece el valor

$$n_{.1} = n_{11} + n_{21} + ... + n_{r1}.$$

Si utilizamos frecuencias relativas en lugar de absolutas, los valores anteriores se denotan por $f_{1.}$ y $f_{.1}$, respectivamente. En la posición intersección de estas dos columnas se da el valor n (también denotado por $n_{..}$) si hemos utilizado frecuencias absolutas, o 1 si hemos utilizado frecuencias relativas. Nótese que

$$n = n_{1\cdot} + ... + n_{r\cdot} = n_{\cdot 1} + ... + n_{\cdot s},$$

es decir, que este valor sigue el mismo criterio que se utilizó en la construcción de los otros valores de esta última fila y columna.

Ejemplo 45. *(Continuación del ejemplo 44)*

En nuestro caso, la tabla de doble entrada con frecuencias absolutas se da en la tabla 4.2.

$X\backslash Y$	0	1	2	3	4	5	6	
0	2	0	0	0	0	0	0	2
1	2	1	0	0	0	0	0	3
2	3	6	1	0	0	0	0	10
3	0	1	4	5	0	0	0	10
4	0	0	1	3	1	0	0	5
6	0	0	0	0	3	1	1	5
	7	8	6	8	4	1	1	35

Tabla 4.2. Tabla de doble entrada con frecuencias absolutas para los datos del ejemplo 44.

Con frecuencias relativas la tabla de doble entrada sería la dada en la tabla 4.3.

$X\backslash Y$	0	1	2	3	4	5	6	
0	2/35	0	0	0	0	0	0	2/35
1	2/35	1/35	0	0	0	0	0	3/35
2	3/35	6/35	1/35	0	0	0	0	10/35
3	0	1/35	4/35	5/35	0	0	0	10/35
4	0	0	1/35	3/35	1/35	0	0	5/35
6	0	0	0	0	3/35	1/35	1/35	5/35
	7/35	8/35	6/35	8/35	4/35	1/35	1/35	1

Tabla 4.3. Tabla de doble entrada con frecuencias relativas para los datos del ejemplo 44.

Como se ve en las tablas anteriores, tenemos casillas con valor 0. En las tablas de doble entrada no es extraño que esto suceda. Al contrario que en el caso unidimensional en el que intentábamos evitar esta situación, esta información sí es relevante para el caso bidimensional, puesto que nos indica combinaciones de modalidades de X e Y que aunque puedan haber aparecido por separado en la muestra, no han aparecido en un mismo dato.

Nótese también que a partir de esta tabla es posible obtener conclusiones sobre posibles relaciones entre las variables. Por ejemplo, parece que cuando X es grande, hay una tendencia a que Y sea grande, porque los valores más alejados de la diagonal son nulos y casi todos los datos se identifican con modalidades que están cerca de esa diagonal.

En el caso de que alguna de las componentes del par sea continua o tome muchos valores diferentes obtendríamos, como pasaba en el caso unidimensional, una tabla muy grande en la que la mayor parte de los valores serían cero. Al igual que en el caso unidimensional, esto resta visibilidad a la tabla. Para evitar este problema, las modalidades de estas componentes se agrupan en clases, de la misma manera que en el caso unidimensional y tratando cada una de ellas por separado. Las distintas clases se tratan como si fuesen las modalidades de la variable.

Ejemplo 46.

Supongamos que estamos estudiando los individuos de una población de estudiantes. Se han seleccionado al azar 16 estudiantes sobre los que se ha medido el peso (X) y la altura (Y), obteniéndose los datos que aparecen en la tabla 4.4.

Individuo	1	2	3	4	5	6	7	8
X	68	72	69	83	60	63	90	86
Y	1.7	1.8	1.75	1.84	1.65	1.64	1.86	1.94
Individuo	9	10	11	12	13	14	15	16
X	60	67	76	74	73	86	80	63
Y	1.6	1.70	1.76	1.72	1.81	1.92	1.95	1.61

Tabla 4.4. Datos correspondientes al ejemplo 46.

En este caso, agrupando X en las clases

$$[1{,}6, 1{,}7), [1{,}7, 1{,}8), [1{,}8, 1{,}9), [1{,}9, 2)$$

e Y *en las clases*

$$[59{,}5, 69{,}5), [69{,}5, 79{,}5), [19{,}5, 89{,}5), [89{,}5, 99{,}5),$$

se obtiene la tabla de doble entrada dada en la tabla 4.5.

$X \backslash Y$	[1,60, 1,70)	[1,70, 1,80)	[1,80, 1,90)	[1,90, 2,00)	
[59,5, 69,5)	4	3	0	0	7
[69,5, 79,5)	0	2	2	0	4
[79,5, 89,5)	0	0	1	3	4
[89,5, 99,5)	0	0	1	0	1
	4	5	4	3	16

Tabla 4.5. Tabla de doble entrada con datos agrupados para ambas variables de los datos del ejemplo 46.

Nótese que se han tomado cuatro clases para cada componente y no dos clases para cada una de forma que en total haya cuatro clases. Esto es debido a que tenemos 16 datos para cada componente.

4.2. Diagramas de dispersión

Pasemos ahora a las representaciones gráficas de una variable bidimensional. De entre todas las representaciones de distribuciones bidimensionales que existen, nosotros trataremos solo los **diagramas de dispersión**, que nos serán de utilidad en la segunda parte del tema, la dedicada a regresión lineal.

Esta representación necesita que las dos componentes del par sean cuantitativas. Está especialmente pensada para cuando ambas componentes son continuas o, al menos, no haya pares repetidos. Consiste en situar en dos ejes coordenados los distintos puntos de la muestra; es decir, la observación muestral (x_i, y_i) se representa por el punto (x_i, y_i). Si tenemos muchos datos esta representación adopta la forma de una nube, de ahí que se llame también *nube de puntos*.

Ejemplo 47.

Consideremos la situación en la que se está estudiando la eficacia de dos métodos para medir el rendimiento de un medicamentos, los métodos X e Y. Para ello se selecciona al azar un grupo de individuos que están tomando el medicamento, y se mide con ambos métodos el rendimiento. Las mediciones obtenidas vienen dadas en la tabla 4.6.

Individuo	1	2	3	4	5	6	7	8	9	10
X	1.9	0.8	1.1	0.1	-0.1	4.4	4.6	1.6	5.5	3.4
Y	0.7	-1.0	-0.2	-1.2	-0.1	3.4	0.0	0.8	3.7	2

Tabla 4.6. Mediciones del rendimiento de un medicamento según los métodos X e Y para 10 individuos.

El correspondiente diagrama de dispersión viene dado en la figura 4.1.

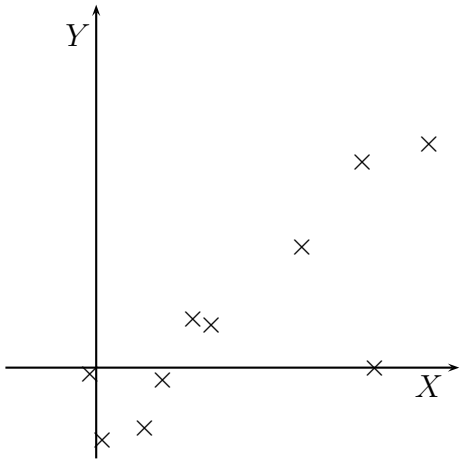

Figura 4.1. Diagrama de dispersión correspondiente a los datos del ejemplo 47.

Esta representación no solo indica el comportamiento de cada componente (sin más que ver sus valores en los ejes coordenados), sino que da una idea de posibles relaciones entre las mismas.

Ejemplo 48. *(Continuación del ejemplo 47)*

En este ejemplo, la disposición de los puntos en el diagrama de dispersión parecen indicar que hay una tendencia a que valores grandes de X conduzcan a valores grandes de Y.

Aunque es raro que aparezcan puntos repetidos, en el caso de que la observación (x, y) se repita p veces, junto al punto (x, y) se escribe en ocasiones el valor p.

Queda la situación en que alguna variable o ambas están agrupadas en clases. Supongamos en particular que ambas variables están agrupadas en clases. Entonces, si solo tenemos la tabla de doble entrada, no conocemos los valores de cada dato, sino que sabemos en qué intervalo está el valor de X y en qué intervalo está el correspondiente valor de Y. Si representamos cada combinación de clases, entonces tendremos una serie de rectángulos. Consideremos la clase i de la variable X, que viene dada por el intervalo (a, b) y la clase j de la variable Y, que vendrá dada por otro intervalo (c, d). La tabla de doble entrada nos indicará entonces que en el rectángulo de vértices $(a, c), (a, d), (b, c), (b, d)$ (esto se suele escribir matemáticamente como el rectángulo $(a, b) \times (c, d)$) hay n_{ij} datos, pero no sabemos exactamente dónde se sitúan esos puntos. Como al agrupar en clases asumimos que los valores se distribuyen de forma uniforme en cada clase, lo que se hace en esta situación es dibujar n_{ij} puntos en el rectángulo, procurando que dichos puntos estén uniformemente distribuidos en toda la superficie.

Ejemplo 49. *(Continuación del ejemplo 46)*

En este caso, el diagrama de dispersión sería el que aparece en la figura 4.2.

4.3. Distribuciones marginales y condicionadas

En muchas ocasiones no nos interesa toda la información que nos da el vector estadístico, sino solamente una parte del mismo. En esta sección veremos dos distribuciones unidimensionales que tienen gran importancia derivadas de la distribución bidimensional.

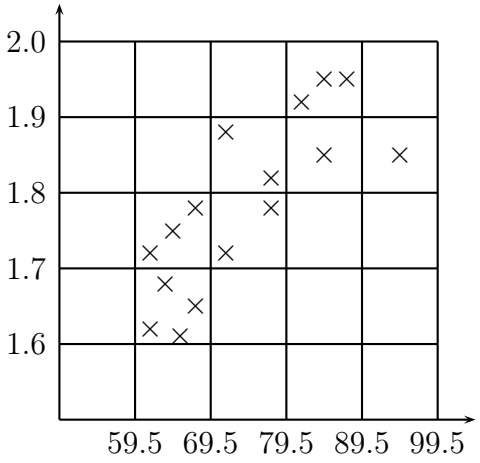

Figura 4.2. Diagrama de dispersión correspondiente a la distribución agrupada del ejemplo 46.

4.3.1. Distribuciones marginales

En este caso, suponemos que no nos interesa el resultado de una de las variables del vector, sino que solo nos interesan los resultados de uno de ellos. Por ejemplo, en el caso del número de las crías de conejo, puede interesarnos estudiar el número de crías inicialmente; en este caso, solo nos interesa la información proporcionada por la primera componente del par. Por tanto, podríamos obviar los datos referentes a la segunda componente y trabajar solamente con los datos muestrales correspondientes a la primera. Así, tendríamos la situación

$$(x_1, y_1) \to x_1, \quad (x_2, y_2) \to x_2, ...$$

Ejemplo 50. *(Continuación del ejemplo 44)*

En el caso de las crías de conejo, tenemos la situación de la tabla 4.7.

Así, tenemos la muestra

(0,0)	(1,0)	(2,1)	...
↓	↓	↓	...
0	1	2	...

Tabla 4.7. Obtención de datos cuando solo interesan los datos de X para los datos del ejemplo 44.

0 1 2 3 4 6 0 1 2 3 4 6 1 2 3 4 6 2
3 4 6 2 3 4 6 2 3 2 3 2 3 2 3 2 3

que corresponde a los valores muestrales de la primera componente.

Tendremos entonces una variable unidimensional, que se llama **distribución marginal**. En el caso de una distribución bidimensional tendremos dos distribuciones marginales (una para cada componente).

Como las distribuciones marginales son variables unidimensionales, podemos realizar el mismo estudio que se había realizado en el tema anterior. Por ejemplo, podemos construir la tabla de frecuencias. Esto puede hacerse directamente, considerando los datos muestrales, pero también puede hacerse más rápidamente a partir de la tabla de doble entrada. Consideremos la distribución marginal de X. Sea la modalidad x_i; el número de veces que X toma el valor x_i será el número de veces que X toma el valor x_i e Y toma el valor y_1, que es n_{i1}, más el número de veces que X toma el valor x_i e Y toma el valor y_2, que es n_{i2}, y seguir añadiendo las correspondientes cantidades para todas las modalidades de Y. En otras palabras, la frecuencia marginal de la modalidad x_i viene dada por

$$n_{i1} + ... + n_{is} = \sum_{j=1}^{s} n_{ij},$$

y precisamente este era el valor $n_{i.}$ que aparecía en la última columna de la tabla de doble entrada. Por tanto, esta columna nos da los valores de las frecuencias marginales de la primera componente. De la misma manera, los valores de la distribución marginal de Y vienen dados en la última fila de la tabla de doble entrada.

Ejemplo 51. *(Continuación del ejemplo 44)*

En nuestro ejemplo de las crías de conejo, las dos distribuciones marginales vienen dadas en las tablas 4.8 y 4.9.

x_i	0	1	2	3	4	6
$n_{i.}$	2	3	10	10	5	5

Tabla 4.8. Distribución marginal de X para los datos del ejemplo 44.

y_j	0	1	2	3	4	5	6
$n_{.j}$	7	8	6	8	4	1	1

Tabla 4.9. Distribución marginal de Y para los datos del ejemplo 44.

Análogamente, para cada componente se puede calcular su media, varianza... Por ejemplo, en términos de las frecuencias marginales, las fórmulas de las medias marginales vienen dadas por:

$$\overline{x} = \frac{\sum_{i=1}^{r} x_i n_{i.}}{n}, \qquad \overline{y} = \frac{\sum_{j=1}^{s} y_j n_{.j}}{n}.$$

Ejemplo 52. *(Continuación del ejemplo 47)*

En el caso del ejemplo de las mediciones de la eficacia de un medicamento se tiene

$$\overline{x} = \frac{1{,}9 + ... + 3{,}4}{10} = 2{,}33, \qquad \overline{y} = \frac{0{,}7 + ... + 2}{10} = 0{,}81.$$

$$\overline{x^2} = \frac{3{,}61 + ... + 11{,}56}{10} = 9{,}037. \Rightarrow v(X) = 9{,}037 - 2{,}33^2 = 3{,}6081,$$

$$\overline{y^2} = \frac{0{,}49 + ... + 4}{10} = 3{,}287 \Rightarrow v(Y) = 3{,}287 - 0{,}81^2 = 2{,}6309.$$

Este mismo estudio puede hacerse en el caso de que alguna de las componentes del par esté agrupada en intervalos.

4.3.2. Distribuciones condicionadas

Supongamos ahora que lo que nos interesa es cómo se comporta una de las componentes cuando el valor que toma la otra componente está fijado; por ejemplo, supongamos que lo que nos interesa es ver cómo se comporta la evolución de las crías de la camada cuando la camada inicial era de dos crías.

En situaciones como esta, no nos interesan todos los datos, sino solo aquellos que verifican la condición; en nuestro caso, solo nos interesan los datos en los que $X = 2$; es decir, los datos que nos interesan son los pares en que $x_i = 2$. Estos pares son los que aparecen en la tabla 4.10.

Camada	3	9	14	18	22	26	28	30	32	34
X	2	2	2	2	2	2	2	2	2	2
Y	1	1	0	2	1	1	1	0	1	0

Tabla 4.10. Datos que se utilizan para el ejemplo 44 cuando $X = 2$.

Más aún, lo que en realidad nos interesa de los datos seleccionados es el valor de la componente cuyo valor no está fijado (aquella que no condiciona), puesto que los valores de la otra variable ya están fijados. Nótese que entonces esta distribución es una distribución unidimensional.

Ejemplo 53. *(Continuación del ejemplo 44)*

En nuestro caso, si condicionamos por $X = 2$, tenemos que considerar solo los datos que verifican esta condición. Como el valor de X queda fijado, observamos solamente los valores de la segunda componente, con lo que la nueva muestra sería

$$1, 1, 0, 2, 1, 1, 1, 0, 1, 0.$$

Si queremos estudiar la distribución de X cuando $Y = y_j$ se dice que queremos hallar la **distribución condicionada de X por $Y = y_j$**. De la misma forma podemos tratar el caso de la distribución condicionada de Y por $X = x_i$, que es la que se consideró en el ejemplo anterior. Para determinar la distribución del primer caso, basta considerar la primera componente de los datos en que $Y = y_j$. Esto, como en el caso de las

distribuciones marginales, puede hacerse directamente a partir de los datos, tal y como se hizo en el ejemplo anterior, pero también puede hacerse a partir de la tabla de doble entrada. Para ello basta notar que los datos que consideramos ahora (en los que $Y = y_j$) son exactamente los datos que están en la columna j-ésima. Así, la distribución condicionada toma las mismas modalidades que X y tiene como frecuencia de x_i el valor n_{ij}. Nótese que en este caso el tamaño de muestra es $n_{\cdot j}$. En el caso de tener frecuencias relativas, el problema es ligeramente más complicado; en este caso la frecuencia relativa no es f_{ij} sino $\frac{f_{ij}}{f_{\cdot j}}$, pues es necesario que la suma de frecuencias relativas de la distribución condicionada sea 1. De esta manera puede construirse la tabla de frecuencias y a partir de ella podemos realizar todos los cálculos realizados en el capítulo anterior.

Ejemplo 54. *(Continuación del ejemplo 44)*

Para el ejemplo anterior, la distribución condicionada por $X = 2$ en términos de frecuencias absolutas viene dada en la tabla 4.11.

$y_{j/3}$	0	1	2	
$n_{j/3}$	3	6	1	10

Tabla 4.11. Distribución condicionada por $X = 2$ para la distribución bidimensional del ejemplo 44 en términos de frecuencias absolutas.

La distribución condicionada, en términos de frecuencias relativas viene dada en la tabla 4.12.

$y_{j/3}$	0	1	2
$f_{j/3}$	$\frac{3}{10}$=0.3	$\frac{6}{10}$= 0.6	$\frac{1}{10}$=0.1

Tabla 4.12. Distribución condicionada por $X = 2$ para la distribución bidimensional del ejemplo 44 en términos de frecuencias relativas.

En el caso en que la variable que condiciona esté agrupada en clases, se condiciona por una clase y se hace análogamente al caso anterior.

Ejemplo 55. *(Continuación del ejemplo 46)*

En el ejemplo de las alturas y pesos de los alumnos, supongamos que solo estamos interesados en estudiar las alturas de los alumnos con peso en el intervalo [59,5, 69,5)*. En este caso estamos condicionando por* $X \in [59{,}5, 69{,}5)$*, y se obtendría la distribución de alturas que aparece en la tabla 4.13.*

$y_{j/1}$	[1,6, 1,7)	[1,7, 1,8)
$n_{j/1}$	4	3

Tabla 4.13. Distribución condicionada por $Y \in [59{,}5, 69{,}5]$ para la distribución bidimensional del ejemplo 46.

Finalmente, no es necesario condicionar por un solo valor de la variable, sino que puede considerarse cualquier condición. Por ejemplo, para el ejemplo 44 podemos considerar que X sea al menos 2, o que X tome valores entre 2 y 4, ambos incluidos. En estos casos el procedimiento es igual: Se consideran los datos en los que esta condición es cierta, que serán una o varias filas de la tabla de doble entrada, y luego nos quedamos solamente con los valores de esos datos correspondientes a la segunda variable.

4.4. La covarianza

Pasemos ahora a dar un valor que mida la relación entre las componentes. La covarianza es uno de los llamados *momentos bidimensionales*. Este momento se aplicará en la parte de regresión lineal. Solo tiene sentido su aplicación en el caso de que ambas componentes del par sean cuantitativas.

Sea la variable bidimensional (X, Y). Se define la **covarianza** entre X e Y, y se denota $cov(X, Y)$ como el valor

$$\boxed{cov(X, Y) := \overline{(x - \overline{x})(y - \overline{y})},}$$

donde $\overline{x}$ e $\overline{y}$ se refieren a las medias marginales de cada componente. Veamos cómo se desarrolla esta expresión en cada caso.

Si tenemos la muestra $(x_1, y_1), ..., (x_n, y_n)$, se tiene que

$$\boxed{\begin{array}{rcl} cov(X,Y) & = & \dfrac{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n} \\ & = & \dfrac{(x_1-\overline{x})(y_1-\overline{y})+...+(x_n-\overline{x})(y_n-\overline{y})}{n}. \end{array}}$$

Esta es la expresión que usaremos con más frecuencia, puesto que es la que se usa casi siempre en regresión lineal.

Supongamos que X toma los valores $x_1, ..., x_r$ e Y toma los valores $y_1, ..., y_s$, y tal que la frecuencia absoluta del par (x_i, y_j) es n_{ij}. Es decir, supongamos que tenemos la tabla de doble entrada. Entonces, la fórmula anterior se puede escribir como

$$\boxed{\begin{array}{rcl} cov(X,Y) & = & \dfrac{\displaystyle\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i-\overline{x})(y_j-\overline{y})n_{ij}}{n} \\ & = & \dfrac{(x_1-\overline{x})(y_1-\overline{y})n_{11}+...+(x_r-\overline{x})(y_s-\overline{y})n_{rs}}{n}. \end{array}}$$

Si utilizamos las frecuencias relativas f_{ij} de cada par, entonces la expresión anterior se escribe como

$$\boxed{\begin{array}{rcl} cov(X,Y) & = & \displaystyle\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i-\overline{x})(y_j-\overline{y})f_{ij} \\ & = & (x_1-\overline{x})(y_1-\overline{y})f_{11}+...+(x_r-\overline{x})(y_s-\overline{y})f_{rs}. \end{array}}$$

Al contrario de lo que pasaba con la varianza, es interesante notar que la covarianza puede ser negativa.

Al igual que la varianza, la covarianza puede calcularse alternativamente como

$$\boxed{cov(X,Y) = \overline{xy} - \overline{x} \times \overline{y},}$$

donde

$$\overline{xy} = \frac{x_1y_1 + ... + x_ny_n}{n},$$

si tenemos los datos en forma de muestra $(x_1, y_1), ..., (x_n, y_n)$,

$$\overline{xy} = \frac{x_1 y_1 n_{11} + x_1 y_2 n_{12} + ... + x_2 y_1 n_{21} + x_2 y_2 n_{22} + ... + x_r y_s n_{rs}}{n},$$

si tenemos los datos en forma de tabla de doble entrada con frecuencias absolutas o si utilizamos frecuencias relativas,

$$\overline{xy} = x_1 y_1 f_{11} + x_1 y_2 f_{12} + ... + x_2 y_1 f_{21} + x_2 y_2 f_{22} + ... + x_r y_s f_{rs}.$$

Ejemplo 56. *(Continuación del ejemplo 47)*

En el ejemplo de las dos formas de medir la eficacia de un medicamento, se tendrían los siguientes resultados: $\overline{x} = 2{,}33, \overline{y} = 0{,}81$, *que ya fueron hallados anteriormente en el ejemplo 52. Entonces, ahora se tiene*

$$\overline{xy} = \frac{1{,}9 \cdot 0{,}7 + ... + 3{,}4 \cdot 2}{10} = \frac{43{,}59}{10} = 4{,}359.$$

Luego la covarianza vale

$$cov(x, y) = 4{,}359 - 2{,}33 \cdot 0{,}81 = 2{,}4717.$$

En el caso de que alguna de las componentes esté agrupada en clases, se utilizan las marcas de clase.

Ejemplo 57. *(Continuación del ejemplo 46)*

En este ejemplo, se tendrían los siguientes resultados:

$$\overline{x} = \frac{64{,}5 \times 7 + 74{,}5 \times 4 + 84{,}5 \times 4 + 94{,}5 \times 1}{16} = 73{,}875,$$

$$\overline{y} = \frac{1{,}65 \times 4 + 1{,}75 \times 5 + 1{,}85 \times 4 + 1{,}95 \times 3}{16} = 1{,}7875.$$

Entonces, ahora se tiene

$$\overline{xy} = \frac{64{,}5 \cdot 1{,}65 \cdot 4 + ... + 94{,}5 \cdot 1{,}85 \cdot 1}{16} = \frac{2126{,}2}{16} = 132{,}89.$$

Luego la covarianza vale

$$cov(X,Y) = 132{,}89 - 73{,}875 \cdot 1{,}7875 = 0{,}84.$$

La covarianza da una medida de la relación entre los valores de las componentes del par. Supongamos que existe una cierta tendencia a que valores grandes (consideramos grandes si son superiores a la media marginal) de la primera variable aparecen en pares en que los valores de la segunda componente son también grandes; y recíprocamente, existe una tendencia a que valores pequeños de la primera variable (menores que la media marginal) aparezcan en pares en que los valores de la segunda componente sean también pequeños; en este caso decimos que existe una **relación directa** entre las dos componentes. En este caso, si observamos la expresión de la covarianza, veremos que hay una tendencia a términos de valor positivo, en el sentido de que cuando tenemos el par (x_i, y_i) tal que x_i sea grande, el valor $(x_i - \overline{x})$ será positivo, y también lo será casi siempre $(y_i - \overline{y})$ por lo que el producto será positivo. De la misma manera, si en el par (x_i, y_i) tenemos que x_i es pequeño, el valor $(x_i - \overline{x})$ será negativo, y también lo será casi siempre $(y_i - \overline{y})$ por lo que el producto será positivo. De esta forma, casi todos los términos de la covarianza serán positivos y entonces la covarianza tomará un valor positivo.

Análogamente, si hay tendencia a que valores grandes de la primera variable aparezcan en pares en que los valores de la segunda componente son pequeños; y valores pequeños de la primera variable en pares en que los valores de la segunda componente son grandes, decimos que existe una **relación inversa** entre las dos componentes. Si observamos la expresión de la covarianza, veremos que en este caso la covarianza toma un valor negativo. Los diagramas de dispersión de las relaciones directa e inversa se pueden ver en la figura 4.3.

Ejemplo 58. *(Continuación del ejemplo 47)*

En el ejemplo de la medida de la eficacia de los medicamentos, ya habíamos visto en el diagrama de dispersión del ejemplo 47 que parecía haber una tendencia a que valores grandes de X se correspondían con valores grandes de Y. Es decir, parece existir una relación directa. Esto se comprueba con el valor de la covarianza que hemos hallado en el ejemplo 57, que es positivo.

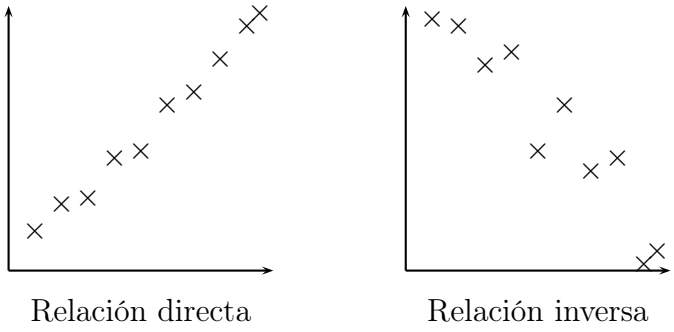

Figura 4.3. Representación gráfica de relaciones directas e inversas.

Es importante notar que si las variables no están relacionadas entre sí (son *independientes*), entonces la covarianza se anula. Sin embargo, si la covarianza se anula, esto no significa que las variables no estén relacionadas, tal y como se puede ver gráficamente en la figura 4.4.

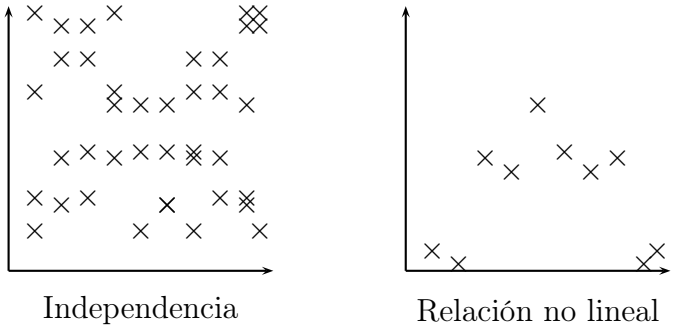

Figura 4.4. Ejemplos en los que la covarianza se anula.

4.5. Regresión lineal

Veamos ahora el modelo más sencillo de regresión. La regresión es una de las técnicas más importantes de la estadística. Se aplica en mul-

titud de situaciones prácticas y por ello es importante tener claro su funcionamiento general.

Supongamos que tenemos una población en la que se están estudiando dos características cuantitativas. Entonces nosotros tendremos una muestra de tamaño n dada por (x_1, y_1), ..., (x_n, y_n). Supongamos ahora que nosotros creemos que estas dos características pueden estar relacionadas. Por ejemplo, parece lógico suponer que una persona alta va a tener más peso que una persona baja. El objetivo de la regresión es estudiar esta relación y poder luego hacer predicciones sobre otros valores. Así, en nuestro ejemplo, podemos preguntarnos cuál sería el peso normal de una persona de 1.80 m sin tener que buscar una persona en estas condiciones. Lógicamente, estas predicciones estarán sujetas a error, ya que no todas las personas de la misma altura tienen exactamente el mismo peso. En general, en nuestras predicciones tenemos que tener en cuenta que habrá un error debido a la aleatoriedad o por deficiencias del modelo, y el valor predicho será una aproximación del valor lógico o esperado.

Ejemplo 59. *(Continuación del ejemplo 47)*

Consideremos el ejemplo de las mediciones de la eficacia de un medicamento. Parece lógico que exista una relación entre las variables X e Y, ya que si un individuo tiene una medida alta para uno de los métodos de medición, es de esperar que también tenga una medición alta para el otro método.

Hay dos aspectos que es importante tener en cuenta cuando se considera un modelo de regresión:

- *A priori* se espera que exista una relación entre las componentes, es decir, nosotros pensamos que el conocer el valor de una de las variables nos da una información sobre cómo es el valor para la otra variable. No estamos pensando en que el modelo siempre obtenga el valor correcto, pero sí a que haya una cierta tendencia. Por ejemplo, no todos los individuos altos pesan mucho, pero pensamos que hay una tendencia a que pesen más que los individuos bajos.

  Sin embargo, desde un punto de vista numérico, nosotros estamos hallando una fórmula numérica entre las dos variables. Esto

significa que se pueden encontrar relaciones muy buenas matemáticamente entre variables que no tengan ninguna relación real, simplemente porque sus valores numéricos están ordenados de manera lineal por puro azar o porque exista una tercera variable que influya sobre ambas. Por ejemplo, se puede comprobar que hay una relación numérica casi perfecta entre el número de asnos salvajes y el presupuesto del Ministerio de Educación y Ciencia, mientras que lo que ocurre realmente es que es la variable tiempo la que influye sobre el número de asnos (cada vez hay menos) y sobre el presupuesto del Ministerio de Educación y Ciencia (cada vez es mayor).

- Supongamos que hemos hallado un modelo de regresión que nos da predicciones para Y conocido el valor de X. Hay que tener en cuenta que las predicciones que se pueden realizar sobre Y deben corresponder a valores de X entre el mínimo y el máximo valor de X en la muestra. Si nos salimos de este intervalo las predicciones ya no son fiables pues es posible que la relación cambie fuera del intervalo y esto conduciría a conclusiones erróneas. Por ejemplo, si estudiamos la cantidad de luz en el mar en función de la profundidad y tenemos profundidades pequeñas, es posible que al intentar predecir la luminosidad de una profundidad grande se obtenga un valor negativo, que es imposible.

Si creemos que existe una relación entre las variables, el siguiente paso es buscar un modelo para encontrar dicha relación. Nosotros buscamos una relación matemática entre X e Y del tipo $Y \approx f(X)$. En general no vamos a tener una igualdad porque esto implicaría que hay una relación total entre las variables y entonces no tendríamos dos variables, sino solamente una escrita de dos formas diferentes. De esta forma, lo que tenemos nosotros es una fórmula del tipo

$$Y = h(X) + \epsilon(X),$$

donde ϵ se llama *error aleatorio* y es una función que mide el error que puede aparece debido a que tenemos fenómenos aleatorios. Es lo que ocurre por ejemplo entre altura y peso, donde individuos con la misma altura no tienen exactamente el mismo peso.

Hallar la función h en las condiciones anteriores no es sencillo. Por ello, lo que se hace es fijar *a priori* qué tipo de función es h. Por ejemplo, podemos suponer que h sea un polinomio de grado 3, o que h sea una función exponencial. En el caso en que tengamos dos variables, una forma gráfica de ver el tipo de relación es a partir del diagrama de dispersión.

Ejemplo 60. *(Continuación del ejemplo 47)*
Observando el diagrama de dispersión, parece que el modelo lineal, es decir, una función del tipo

$$f(X) = aX + b,$$

aproxima bastante bien esta relación.

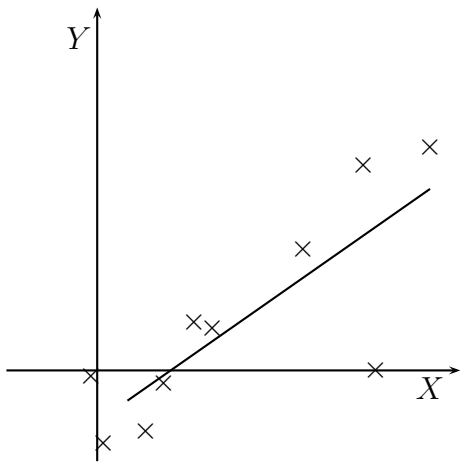

Figura 4.5. Determinación gráfica del tipo de relación entre las dos variables del ejemplo 47.

Entonces, si llamamos f a esa función del tipo fijado, esperamos que se tenga una relación del tipo

$$Y = f(X) + e(X).$$

Debe tenerse en cuenta que en general f no coincide con h ni e coincide con ϵ y, por supuesto, es posible que escojamos un modelo que

no sea el adecuado, en el sentido de que los errores que se cometen al aplicarlo sean muy grandes. En la próxima sección veremos cómo decidir si un modelo es bueno o, por el contrario, no aproxima bien la relación.

Por otra parte, es posible que tengamos varios modelos que funcionen bien para los datos que tenemos. Por ejemplo, puede ser que las relaciones

$$Y = lnX + X^{15} - tgX - e^X, \quad Y = X^4 - 2X^3 + 3X^2 + 10X - 4,$$

sean buenas aproximaciones de la relación entre X e Y. En general, nosotros queremos encontrar la relación más sencilla que nos dé buenos resultados.

En esta sección, nosotros tomaremos un modelo lineal

$$Y = aX + b,$$

que es el más sencillo y además es fácilmente interpretable. Sin embargo, todos los desarrollos que haremos a continuación valen para cualquier otro modelo, y la dificultad de otros modelos radicará en resolver los sistemas de ecuaciones que determinan el modelo.

Consideremos entonces una variable bidimensional (X, Y) de la que se tiene una muestra aleatoria simple de tamaño n que viene dada por $(x_1, y_1), ..., (x_n, y_n)$. Nosotros queremos encontrar la recta que mejor se aproxime a estos datos, es decir, la recta

$$Y^* = aX + b,$$

de forma que Y^* sea «lo más próxima posible» a Y para los elementos de la muestra. Tenemos entonces que decidir cuáles son los valores de a y b en la expresión anterior. Una vez fijados esos valores de a y b, por ejemplo $a = 3, b = 1$ podemos hacer predicciones sobre el valor de Y. Por ejemplo, si sabemos que $X = 4$, entonces nuestro modelo predice que el valor de Y será $3 \times 4 + 1 = 13$. Ahora bien, si cogemos un par de la muestra (x_i, y_i), nuestro modelo hace una predicción

$$y_i^* = ax_i + b,$$

y en realidad sabemos que para ese valor x_i el correspondiente valor de la variable Y es y_i. Entonces nosotros consideraremos que el modelo es

bueno, es decir, que hemos escogido unos buenos valores para a y b si y_i^* e y_i son valores parecidos. Aunque hay diversos criterios para medir esta proximidad entre Y e Y^*, el criterio más habitual es el conocido como «criterio de mínimos cuadrados», que consiste en medir la distancia mediante $(Y - Y^*)^2$. Una representación gráfica del criterio de mínimos cuadrados se puede ver en la figura 4.6.

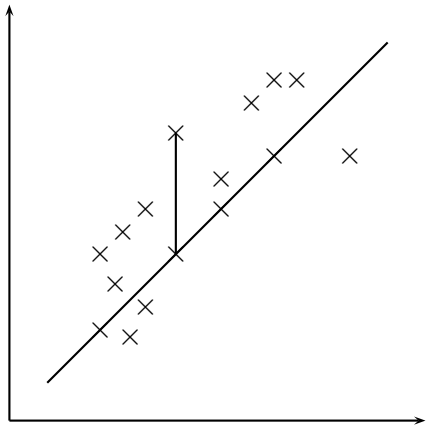

Figura 4.6. Interpretación gráfica del criterio de mínimos cuadrados.

Nótese que podríamos pensar en otros criterios alternativos a hallar las distancias en vertical, como por ejemplo considerar las distancias desde los puntos a la recta. La razón de considerar distancias entre la predicción y el valor real estriba una vez más en que nosotros no queremos que los valores de Y estén próximos a la recta del modelo, sino que la predicción sea buena, o sea, que Y e Y^* estén próximos. Y esto lo vamos a hacer para todos los puntos de la muestra. Nótese además que esta proximidad se mide con las diferencias al cuadrado para evitar compensaciones de diferencias positivas y negativas (lo mismo que ocurría por ejemplo con la definición de varianza). En definitiva, buscamos valores a, b de forma que se minimice la expresión

$$\sum_{i=1}^{n}(y_i - ax_i - b)^2.$$

Queremos entonces resolver el problema

$$\min_{a,b} \sum_{i=1}^{n}(y_i - ax_i - b)^2.$$

Si resolvemos este problema de minimización, se obtiene que la solución es

$$\boxed{a = \frac{cov(X,Y)}{v(X)}, \quad b = \overline{y} - \frac{cov(X,Y)}{v(X)}\overline{x}.}$$

Sustituyendo, la mejor recta viene dada por

$$\boxed{Y^* - \overline{y} = \frac{cov(X,Y)}{v(X)}(X - \overline{x}).}$$

Para que esta fórmula pueda aplicarse es necesario que $v(X) \neq 0$; nótese que si $v(X) = 0$, esto implica que X toma siempre el mismo valor y no tiene sentido ver cómo se comportan los valores de Y a partir de los valores de X, pues la componente X no aporta ninguna información.

Esta recta se conoce con el nombre de recta de **regresión de** Y **sobre** X. Análogamente, se puede calcular la recta de regresión de X sobre Y, en la que se intenta hacer predicciones sobre X conocido el valor de Y. En este caso tenemos que hallar los valores a, b tales que

$$\min_{a,b} \sum_{i=1}^{n}(x_i - ay_i - b)^2.$$

Ahora estamos midiendo las distancias entre x_i^* y x_i. Entonces, gráficamente estamos considerando las distancias en horizontal. Esto significa que no vamos a obtener los mismos valores que antes, puesto que la función a minimizar es diferente. Pero para hallar los valores de a, b no hace falta rehacer las cuentas, ya que basta intercambiar los valores de X e Y. De esta forma, esta recta viene dada por

$$\boxed{X^* - \overline{x} = \frac{cov(X,Y)}{v(Y)}(Y - \overline{y}).}$$

Hay que insistir, nuevamente, en que estas dos rectas no coinciden en general, pues en el primer caso estamos midiendo las diferencias «en vertical», mientras que en el segundo caso estamos midiendo las diferencias «en horizontal». La elección de una u otra recta dependerá de si deseamos predecir valores de la primera componente o de la segunda componente.

Ejemplo 61. *(Continuación del ejemplo 47)*

En el ejemplo de las medidas por dos métodos de la eficacia de un medicamento ya habíamos calculado en los ejemplos 47 y 57

$$\overline{x} = 2{,}33,\ \overline{y} = 0{,}81,\ \overline{x^2} = 9{,}064,\ v(X) = 3{,}6081,$$

$$\overline{xy} = 4{,}359,\ cov(X,Y) = 2{,}495.$$

Luego, la recta de regresión de Y sobre X es

$$Y^* - 0{,}81 = \frac{2{,}495}{3{,}6081}(X - 2{,}33) \Rightarrow Y^* - 0{,}81 = 0{,}69(X - 2{,}33).$$

4.6. Correlación

Una vez obtenidos los valores (a, b) para nuestro modelo lineal, es interesante tener una medida de lo bueno que es dicho modelo, pues no nos interesa un modelo sencillo si sus predicciones no se aproximan a la realidad. Para estudiar la bondad del modelo, hay que tener en cuenta que los errores de predicción vienen dados para los pares de la muestra por

$$(y_i - y_i^*)^2 = (y_i - ax_i - b)^2, i = 1, ..., n,$$

y entonces tiene sentido considerar como medida del error que se comete el valor

$$\sum_{i=1}^{n}(y_i - y_i^*)^2.$$

Entonces, si la expresión anterior es grande, tendremos que el modelo no es bueno, mientras que si ese valor es pequeño concluiremos que el modelo es bueno. Queda sin embargo determinar los límites para los que el valor de la expresión anterior puede considerarse grande.

Para estudiar si el modelo es bueno, utilizamos el siguiente argumento: Consideremos la variable que tiene que ser explicada (Y); los datos de la muestra de Y tienen una variación. Y las causas de esta variación son dos:

- Por una parte, existe variación porque X tiene variación, y como X influye sobre Y, esto se traduce en una variación de Y.
- Por otra parte, Y proviene de un experimento aleatorio, por lo que Y tiene una variación propia que no proviene del modelo de regresión, ya que este modelo no explica perfectamente el funcionamiento de Y; por ello, dado X no podemos predecir con exactitud el valor de Y, sino que entre Y e Y^* hay diferencias.

La variación total de Y se puede medir por

$$\sum_{i=1}^{n}(y_i - \overline{y})^2,$$

término que se conoce como **suma de cuadrados del total** (SCT). Nótese que este valor es n veces la varianza de la distribución marginal de Y. Puede demostrarse que

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - y_i^*)^2 + \sum_{i=1}^{n}(y_i^* - \overline{y})^2.$$

El término $\displaystyle\sum_{i=1}^{n}(y_i - y_i^*)^2$ mide la importancia de los errores y se llama **suma de cuadrados del error** (SCE). El término $\displaystyle\sum_{i=1}^{n}(y_i^* - \overline{y})^2$

se llama **suma de cuadrados de la regresión** (SCR) y es debida tanto a variaciones provocadas por el modelo (puesto que es n veces la varianza de las predicciones) como a errores aleatorios (puesto que X proviene de un fenómeno aleatorio). De esta forma, si dividimos la expresión anterior por $\sum_{i=1}^{n}(y_i-\overline{y})^2$, obtenemos

$$1=\frac{\sum_{i=1}^{n}(y_i-y_i^*)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}+\frac{\sum_{i=1}^{n}(y_i^*-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}.$$

Entonces, como los dos términos suman 1, el término

$$\frac{\sum_{i=1}^{n}(y_i-y_i^*)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

representa la proporción de variación de Y que no explica el modelo, y de las misma manera, el término

$$\frac{\sum_{i=1}^{n}(y_i^*-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

representa la proporción de variación de Y que sí explica el modelo.

Este último valor se llama el **coeficiente de determinación**, denotado r^2, y puede demostrarse que puede calcularse alternativamente en el caso lineal por

$$\boxed{r^2=\frac{cov(X,Y)^2}{v(X)v(Y)}.}$$

Puesto que es una proporción, este coeficiente toma valores entre 0 y 1. Veamos cómo se interpreta este coeficiente.

- Si r^2 es pequeño (cercano a 0), entonces la proporción de variación de Y que explica el modelo lineal es muy pequeña y concluimos que el modelo no explica bien la relación entre las variables. En particular, para que valga 0 es necesario que la covarianza se anule y, por tanto, la recta de regresión será una constante. Esto es lo que pasa cuando las variables son independientes (conocer una de ellas no nos da información sobre el posible valor de la otra), tal y como vimos al estudiar la covarianza, pero recordemos que hay otras situaciones en que también obtenemos $r^2 = 0$ sin que lo sean, puesto que tal y como hemos visto anteriormente, la covarianza también se puede anular aunque las variables estén relacionadas.

- Si r^2 es grande (cercano a 1), entonces la proporción de variación de Y que explica el modelo lineal es muy grande y concluimos que este modelo explica bien la relación entre las variables. En particular, cuando el coeficiente de determinación vale 1 el ajuste es perfecto y no se está cometiendo ningún error en las predicciones; esto no implica que en las predicciones para otros individuos no se cometa ningún error, pero sí es de esperar que estos errores sean muy reducidos.

En general se considera que el ajuste es bueno cuando supera el valor 0.5, aunque depende del número de datos de la muestra. Si el coeficiente de determinación vale 1, entonces puede demostrarse que las dos rectas de regresión coinciden.

Supongamos que r^2 es grande y concluimos que el modelo lineal es bueno. El coeficiente de determinación tiene el problema de que no nos da ninguna información sobre si la relación lineal entre las componentes (que ahora sabemos que existe) es directa o inversa. Puesto que es el signo de la covarianza lo que nos indica si la relación es directa o inversa, definimos el llamado **coeficiente de correlación** por

$$\boxed{r = \frac{cov(X,Y)}{d(X)d(Y)}.}$$

En realidad, el coeficiente de correlación es la raíz cuadrada del coeficiente de determinación con el signo de la covarianza. Por lo tanto toma valores entre -1 y 1. Entonces este coeficiente se interpreta de la siguiente manera:

- Un valor cercano a 0 implica una relación lineal muy mala.
- Un valor por debajo de -0.7 o por encima de 0.7 implica una relación lineal buena.
    - Si el valor es positivo, la relación entre las componentes es directa.
    - Si el valor es negativo tenemos una relación inversa.

Ejemplo 62. *(Continuación del ejemplo 47)*
En nuestro ejemplo, el coeficiente de determinación vale

$$r^2 = \frac{cov(X,Y)^2}{v(X)v(Y)} = 0{,}66.$$

Por tanto, la aproximación lineal es regular y hay que desconfiar algo de las predicciones.

4.7. Modelos derivados del modelo lineal

Todo lo visto hasta ahora se basa en que consideramos que el modelo lineal es el adecuado para modelar la relación entre las dos variables. Si suponemos otra relación funcional, como una relación de tipo cuadrático $Y = aX^2 + bX + c$, entonces las fórmulas para los coeficientes del modelo lineal ya no son válidas y es necesario volver a plantear el problema de minimización y resolver el correspondiente sistema de ecuaciones. Esto puede ser complicado para algunos modelos, incluso es posible que no haya fórmulas y tengamos que resolver el sistema a partir de los datos concretos de la muestra.

Sin embargo, hay situaciones en las que es posible aprovechar los cálculos de la parte de regresión lineal. En esta sección plantearemos otros modelos que se derivan del modelo lineal y que permiten establecer otros modelos alternativos al lineal sin tener que volver a desarrollar las fórmulas. Se resuelven mediante un proceso llamado *linealización*, que consiste en hacer una transformación que pase el modelo que nos interesa a un modelo lineal; esto se consigue haciendo un cambio de variable sobre X, Y o ambas variables. Veremos dos ejemplos de este proceso.

4.7.1. Modelo logarítmico

En este caso, el modelo que se considera adecuado es de la forma

$$Y^* = a \ln X + b.$$

Entonces tenemos que hallar los coeficientes a, b de forma que Y^* se acerque lo más posible a Y. En lugar de reproducir todo el procedimiento de mínimos cuadrados para este caso, podemos hacer $Z = \ln X$. Entonces el modelo a resolver sería $Y = aZ + b$, que es un modelo lineal para el que conocemos los valores de a y b. En definitiva, basta hallar la recta de regresión a partir de los pares

$$(\ln x_1, y_1), ..., (\ln x_n, y_n).$$

Entonces, según el modelo lineal,

$$a = \frac{cov(Y, Z)}{v(Z)} = \frac{cov(Y, lnX)}{v(lnX)}, \quad b = \overline{y} - \frac{cov(Y, lnX)}{v(lnX)}\overline{x}.$$

4.7.2. Modelo exponencial

El modelo que se considera adecuado es de la forma

$$Y^* = b \exp(aX).$$

Se trata de hallar los coeficientes a, b de forma que Y^* se acerque lo más posible a Y. Para resolver este modelo basta con tomar logaritmos de forma que se obtiene

$$\ln Y = \ln b + aX = b' + aX.$$

Y ahora basta hallar la recta que mejor explica $\ln Y$ en función de X, es decir, la recta de regresión a partir de los pares

$$(x_1, \ln y_1), ..., (x_n, \ln y_n).$$

Entonces, según el modelo lineal,

$$a = \frac{cov(lnY, X)}{v(X)}, \quad b' = \overline{lny} - \frac{cov(lnY, X)}{v(X)}\overline{x}.$$

Una vez hallados los coeficientes a, b', basta tomar $b = \exp b'$.

Ejemplo 63.

Consideremos la muestra bidimensional de la tabla 4.14.

x_i	-6	-3	0	3	6	9	12	15	20	25
y_i	2	2.8	3.9	4.2	5.8	6.2	7.5	8.2	9.3	10.9

Tabla 4.14. Datos originales para el ejemplo 63.

Vamos a hallar los coeficientes del modelo exponencial. Para ello, por lo visto anteriormente, tenemos que considerar las variables X y lnY, con lo que los datos que tenemos que trabajar son los que aparecen en la tabla 4.15.

x_i	-6	-3	0	3	6	9	12	15	20	25
$ln\, y_i$	0.69	1.03	1.36	1.44	1.76	1.83	2.02	2.10	2.23	2.39

Tabla 4.15. Datos transformados para el ejemplo 63.

Entonces tenemos que calcular la recta de regresión que explica lnY con X. Operando,

$$\overline{x} = 8{,}1, \quad \overline{lny} = 1{,}684, \quad cov(X, lnY) = 4{,}7696, \quad v(X) = 90{,}89,$$

luego $a = 0,0525, lnb = 1,259$ y los coeficientes que buscamos son

$$a = 0{,}0525, b = 3{,}522,$$

con lo que nuestro modelo final será

$$y = 3{,}522e^{0{,}0525x}.$$