CAPITULO 3.- ANÁLISIS CONJUNTO DE DOS VARIABLES

3.1 Presentación de los datos. Tablas de doble entrada.

En el capítulo anterior nos hemos interesado por el análisis y descripción de una sola variable. Para ello hemos definido un proceso de reducción de la información inicialmente disponible. Esta reducción ha dado como resultado la construcción de una tabla estadística donde se daba la distribución de frecuencias de la variable. Posteriormente se ha analizado la forma, se han definido medidas de tendencia central, medidas de dispersión, de simetría y curtosis. También se ha estudiado el problema de la concentración. Pero este análisis es de tipo unidimensional, pues de todos los caracteres de los elementos de una población solo nos ha preocupado observar un de ellos que, por lo regular, siempre ha sido de tipo cuantitativo. Pero que duda cabe que los elementos de una población cualquiera gozan de más de un carácter susceptible de ser observado. En este sentido, imaginemos que los elementos observados son las empresas. En ellas se pude observar de forma conjunta los beneficios y los costes de las mismas o cualquier otro par de caracteres. Así podríamos pensar en los gastos en publicidad y sus beneficios, o los costes y el número de empleados. El número de ejemplos que podríamos dar es tan amplio que no merece la pena seguir mencionándolos.

El objetivo de este capitulo será similar al del anterior, pero ahora buscando el análisis conjunto de dos variables o análisis bidimensional. Para ello se procederá a la observación de dos características de todos los elementos de una población. Inicialmente supondremos que esas características son de naturaleza cuantitativa. El resultado de esa observación conjunta será la definición de dos variables a las que llamaremos X e Y, las cuales pueden ser discretas o continuas, y nuestra primera preocupación será la de presentar de forma conjunta las frecuencias de los pares de valores de esas variables (x_i, y_j). El instrumento que se utiliza para alcanzar ese objetivo es lo que se conoce como tabla de doble entrada, tabla de correlaciones o tabla de contingencia. Esta última denominación se reserva especialmente para los casos de caracteres cualitativos. De todas las denominaciones que hemos señalado, usaremos la de tabla de doble entrada, pues la denominación de tabla de correlaciones tiene un significado que va más allá de la mera representación numérica de las la distribución conjunta de frecuencias.

Una tabla de doble entrada no es más que la representación de (x_i, y_j, n_{ij}) en la forma que se muestra en la Tabla 1.

Tabla 1. Distribución conjunta de dos variables

		Y						
		y_1	y_2	.....	y_j	.....	y_k	$n_{i.}$
X	x_1	n_{11}	n_{12}	.....	n_{1j}	.....	n_{1k}	$n_{1.}$
	x_2	n_{21}	n_{22}	.....	n_{2j}	.....	n_{2k}	$n_{2.}$
	. . .	. . .	. . .	. ...	. . .	. ...	. . .	. . .
	x_i	n_{i1}	n_{i2}		n_{ij}		n_{ik}	$n_{i.}$
	. . .	. . .	. . .	. ...		. ...	. . .	. . .
	x_h	n_{h1}	n_{h2}	.....	n_{hi}	.....	n_{hk}	$n_{h.}$
	$n_{.j}$	$n_{.1}$	$n_{.2}$	.....	$n_{.j}$	.....	$n_{.k}$	N

La lectura del contenido de esta tabla sería el siguiente. El valor n_{ij} nos da la frecuencia conjunta con la que se presentan el valor x_i de X y el valor y_j de Y. A su vez n_{i1} da la frecuencia conjunta de x_i y de y_1. De forma similar habría que leer e interpretar el resto de las frecuencias conjuntas que son las que están dentro del cuerpo central de la tabla, es decir, las que llevan un doble subíndice alfanumérico.

Mención aparte merecen la última fila y la última columna. A esa fila y a esa columna se les conoce como distribuciones marginales de Y y de X, respectivamente. Se trata de la distribución de frecuencias de cada una de las variables tomadas por separado. Así pues la distribución marginal de X vendría dada por los pares (x_i, $n_{i.}$), mientras que la marginal de Y vendría dada por los pares (y_j, $n_{.j}$), es decir:

Tabla 2. Distribuciones marginales de *X* y de *Y*

X	$n_{i.}$
x_1	$n_{1.}$
x_2	$n_{2.}$
.	.
.	.
x_i	$n_{i.}$
.	.
.	.
x_h	$n_{h.}$

Y	$n_{.j}$
y_1	$n_{.1}$
y_2	$n_{.2}$
.	.
.	.
y_j	$n_{.j}$
.	.
.	.
y_k	$n_{.k}$

donde

$$n_{i.} = n_{i1} + n_{i2} + ... + n_{ij} + ...+ n_{ik} = \sum_j n_{ij} \qquad j = 1, 2, ..., k \qquad (3.1)$$

$$n_{.j} = n_{1j} + n_{2j} + ... + n_{ij} + ...+ n_{hj} = \sum_i n_{ij} \qquad i = 1, 2, ..., h \qquad (3.2)$$

Finalmente se cumple que

$$\sum_i n_{i.} = \sum_j n_{.j} = \sum_i \sum_j n_{ij} = N \qquad (3.3)$$

Además de las marginales, para una tabla de doble entrada, se pueden definir también las distribuciones condicionadas, que también son de tipo unidimensional. Estas hay que darlas en términos de una condición previa. En este sentido se tendría la distribución de los valores de la variable *X* condicionada a que la variable *Y* tome un valor concreto. De igual forma se podría hacer para la variable *Y* con respecto a los posibles valores de *X*. Si se define la condicionada de *X*, entonces los valores que puede tomar esta variable son los mismos que los de su marginal. Lo único que varía son sus frecuencias, que se representarán por $n_{i/j}$. A su vez, si de lo que se trata es de la condicionada de *Y*, los

valores de esta distribución son los de la marginal de *Y*, pero las frecuencias son distintas y se representa por $n_{j/i}$. Estas nuevas distribuciones aparecen en la Tabla 3.

Tabla 3. Distribuciones condicionales de *X* y de *Y*

X/y_j	$n_{i/j}$
x_1	n_{1j}
x_2	n_{2j}
.	.
.	.
x_i	n_{ij}
.	.
.	.
x_h	n_{hj}

Y/x_i	$n_{j/i}$
y_1	n_{i1}
y_2	n_{i2}
.	.
.	.
y_j	n_{ij}
.	.
.	.
y_k	n_{ik}

La distribución condicional no es única, al el contrario de lo que ocurre con la marginal. Habrá tantas como valores pueda tomar la variable condicionante. Así, para variables continuas el número de distribuciones condicionales será infinito.

Todas y cada una de esta nuevas distribuciones univariantes que se han definido es posible tratarlas con los instrumentos de análisis definidos en las lecciones anteriores. Además, aunque la tabla de doble entrada que se ha diseñado antes lo es para variables de tipo cuantitativo, también es posible hablar de tablas de doble entrada para variables de tipo cualitativo o mixto, en cuyo caso se les conoce como tablas de contingencia. Por otro lado, en la Tabla 1 se recogen dos variables discretas con frecuencias unitarias o mayores que la unidad. Sin embargo ese diseño de tabla de doble entrada es también válido para el caso de variables continuas. Bastaría con sustituir los valores puntuales de cada variable por intervalos.

A continuación vamos a dar un ejemplo que permita aclarar todos estos conceptos.

Ejemplo 1. *Para un conjunto de 2005 empresas de menos de 9 empleados se han observado dos caracteres de las mismas. El número de sus empleados (X) y el número de días perdidos por bajas (Y) en esas empresas. Los resultados son los que se dan en la siguiente tabla de doble entrada:*

		Y									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	$\mathbf{n_{i.}}$
X	**1**	50	45	40	30	20	15	10	5	5	220
	2	40	50	45	40	30	20	15	10	5	255
	3	20	40	50	40	35	25	20	15	10	255
	4	15	30	30	50	40	30	25	20	15	255
	5	10	20	20	40	50	40	40	35	30	285
	6	5	10	15	30	40	50	45	40	35	270
	7	5	5	10	20	30	40	50	45	40	245
	8	5	5	5	10	20	30	45	50	50	220
	$\mathbf{n_{.j}}$	150	205	215	260	255	250	250	220	200	2005

A partir de esos datos, obtenga

a) La marginal de X y la condicional de X/y=5.

b) La marginal de Y y la condicional de Y/x=3

a)

Marginal de X		*Condicional de X/y=5*	
x_i	$n_{i.}$	$x_i/y=5$	$n_{i/y=5}$
1	220	1	15
2	255	2	20
3	255	3	25
4	255	4	30
5	285	5	40
6	270	6	50
7	245	7	40
8	220	8	30

b)

Marginal de Y		*Condicional de Y/x=3*	
y	$n_{.j}$	$y_j/x=3$	$n_{j/x=3}$
0	150	0	20
1	205	1	40
2	215	2	50
3	260	3	40
4	255	4	35
5	250	5	25
6	250	6	20
7	220	7	15
8	200	8	10

Ejemplo 2. *En la tabla siguiente se recoge la distribución de los asalariados fijos en explotaciones agrarias según edad y tamaño de las mismas en Andalucía para el año 1997.*

	Menos de 1 Ha.	De 1 a 2 Ha.	De 2 a 5 Ha.	De 5 a 10 Ha	De 10 a 20 Ha.	De 20 a 30 Ha.	De 30 a 50 Ha.	De 50 a 100 Ha.	Más de 100 Ha.	Total
< 35 años	336	195	1203	1145	671	577	234	518	2400	7279
35 a 44 años	409	468	788	452	592	448	349	852	4256	8613
45 a 54 años	231	144	657	581	751	341	418	801	5152	9076
55 a 64 años	62	601	559	212	1008	231	225	835	3260	6992
> 65 años	2	0	33	208	71	160	35	231	569	1309
Total	1041	1407	3239	2598	3094	1758	1261	3237	15637	33270

Fuente: Web INE.

A partir de esos datos, obtenga: a) la distribución marginal de la variable "edad de los asalariados fijos"; b) la distribución condicional de la variable "edad de los asalariados fijos" para explotaciones de más de 100 Ha.

a) *La distribución marginal que se nos pide viene dada por la primera y última columna de la tabla anterior. En concreto, sería la que aparece en la tabla siguiente:*

Edad	Asalariados
< 35 años	7279
35 a 44 años	8613
45 a 54 años	9076
55 a 64 años	6992
> 65 años	1309
Total	33270

b) *La distribución condicional de la variable "edad de los asalariados fijos" para explotaciones de más de 100 Ha. es la que se recoge en la tabla siguiente:*

Edad	Asalariados en explotaciones de más de 100 Ha.
< 35 años	2400
35 a 44 años	4256
45 a 54 años	5152
55 a 64 años	3260
> 65 años	569
Total	15637

3.2 La covariación.

En el apartado anterior hemos presentado una distribución frecuencias conjunta para dos variables. En ese apartado se ha señalado que tipo de distribuciones unidimensionales o univariantes se pueden definir a partir de la bivariante, y se ha indicado que las mismas podían ser tratadas con los instrumentos definidos en lecciones anteriores. Sin embargo, el interés de este capitulo no es precisamente el de realizar un análisis individualizado de todas y cada una de las distintas distribuciones univariantes que se pueda definir a partir de una distribución bivariante. Ahora, nuestro objetivo es el análisis conjunto de las dos variables que se definen en tabla de doble entrada.

Ya no se trata de estudiar solo los promedios y las medidas de dispersión de cada una de esas variables. El siguiente paso que se pretende dar con este capítulo es el análisis de la relación o dependencia que pueda existir entre dos variables. A esa relación la vamos a denominar covariación o variación conjunta**.**

La covariación es un fenómeno bastante habitual entre variables de carácter económico y de otra naturaleza. La covariación que puede darse entre dos variables *X* e *Y* cualesquiera puede ser de distinto tipo. Así puede hablarse de:

1º *Dependencia causal unilateral*. Este tipo de covariación se da cuando una variable influye en la otra y no al contrario. Es decir las variaciones de una variable pueden explicarse por las variaciones de otra, pero no a la inversa.

En este tipo de análisis, a la variable que ejerce influencia en la otra se le llama variable independiente, explicativa, variable causa o exógena. A la otra variable se le llama dependiente, explicada, variable efecto o endógena. Generalmente a la independiente se le suele representar por la letra *X* , mientras que a la dependiente se le representa por la letra *Y*.

A título de ejemplo se puede señalar los siguientes pares de variables: los impuestos y la renta, los benéficos empresariales y el volumen de ventas, los salarios y la cualificación profesional, etc.

2º *Interdependencia*. Esta situación se da cuando la influencia es recíproca entre las dos variables. En este caso se habla de una relación causal bilateral o interdependencia.

Un ejemplo muy claro en Economía de este tipo de relación se encuentra entre precio y producción de un bien. Es bien conocido que, en un sistema de mercado en régimen de competencia perfecta, estas dos variables están interrelacionadas.

3º *Dependencia indirecta*. Este tipo de covariación se da cuando existe una tercera variable que influye simultáneamente sobre *X* e *Y*. En estos casos no existe una relación de causalidad entre esas variables. Sin embargo, la presencia de una tercera que influye en ambas hace que ellas se muevan de forma sincronizada. Pensemos en la superficie quemada por incendios forestales y el número de viajeros en zonas turísticas. Estas dos variables se comportan a lo largo del año de una forma parecida. Pero no puede hablarse de una relación causa efecto entre ellas. En realidad es la variable temperatura climatológica la que condiciona su evolución paralela.

4º *Concordancia*. A veces se sabe que las variables *X* e *Y* son por naturaleza independientes. Sin embargo puede que muestren un movimiento sincronizado, lo que nos llevaría a pensar en un cierta dependencia. Tal podría ser el caso el resultado de las opiniones de un panel de expertos relativas a expectativas de crecimiento de la economía de un conjunto de países.

5º *Covariación casual o espúrea*. Ocurre cuando dos variable se mueven de forma sincronizada pero sin que exista una relación de causalidad entre ellas.

Es conveniente señalar que el tipo de relación que pueda existir entre dos variables no se puede determinar fácilmente mediante instrumentos estadísticos, por lo que ese tipo de covariación habrá que buscarla en el conocimiento previo que se tenga de esas variables. Lo que si puede hacer la Estadística, en cualquier caso, es cuantificar y formalizar matemáticamente la relación o covariación previamente señalada, con el fin de confirmar tal relación y utilizarla luego para *describir* el fenómeno, para *explicarlo* y para realizar *predicciones*.

La forma más sencilla, desde un punto de vista estadístico, de iniciar el estudio de la covariación entre dos variables es mediante un análisis gráfico. Ahora, como tenemos dos variables, recurriremos al eje de abscisas para representar los valores de la variable *X* y al de ordenadas para situar los valores de *Y*. Si en este diagrama bidimensional llevamos las parejas de valores de *X* e *Y*, el resultado es lo que se llama un diagrama de dispersión o nube de puntos. En este tipo de diagramas se representan parejas de valores con frecuencia unitarias. Si las frecuencias fueran mayores que uno, entonces habría que recurrir a un tercer eje donde llevaríamos las frecuencias de cada una de esas parejas de valores.

Admitiendo que trabajamos con parejas de valores con frecuencias unitarias, los diagramas de dispersión mas habituales serían los siguientes:

Figura 1. Covariación directa

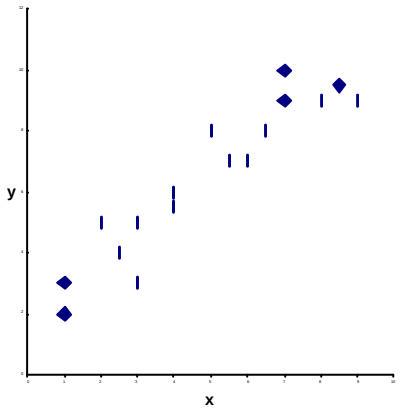


Figura 2. Covariación inversa

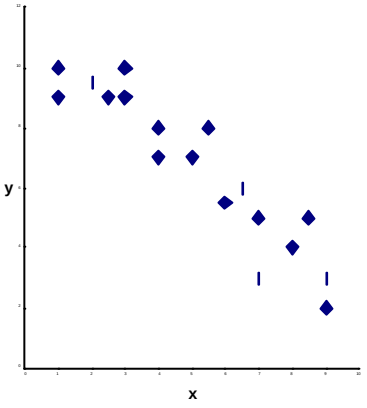


Figura 3. Ausencia de covariación

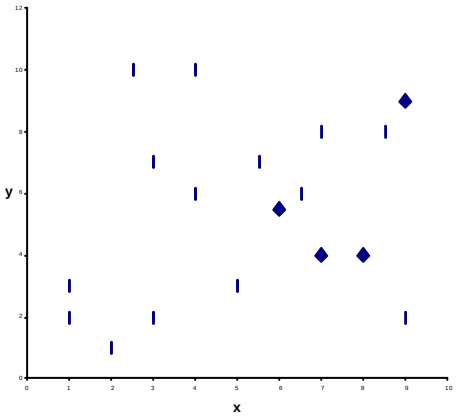
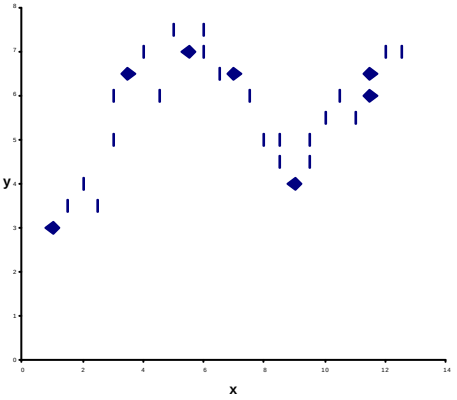


Figura 4. Covariación curvilinea



Mediante este método gráfico lo que se consigue es descubrir la posible relación que existe entre las variables. Esto representa un paso importante para un instrumento tan sencillo como es un simple gráfico.

En la Figura 1, denominada como covariación directa, se detecta una relación lineal positiva o directa. La Figura 2 nos advierte de una relación lineal negativa o inversa; la tercera nos indica que entre las variables *X* e *Y* no existe relación evidente de tipo alguno; finalmente, la última gráfica nos pone de manifiesto una relación que no es lineal.

Que duda cabe que estos cuatro modelos de diagramas de dispersión no son los únicos, pero si los más representativos.

Una vez agotada la vía gráfica para el estudio de la covariación, hay que recurrir a otros procedimientos que nos permitan cuantificar la covariación. Los dos procedimientos más utilizados son la correlación y la regresión.

Antes de finalizar este epígrafe sería conveniente resaltar que para los distintos tipos de covariación que hemos definido hay un concepto que aparece de forma recurrente. Se trata de la independencia o dependencia entre variables. Para definir este concepto en términos estadísticos haremos uso a la tabla de doble entrada que se vio en el apartado anterior. Con la terminología utilizada en esa tabla, se dice que dos variables *X* e *Y* son estadísticamente independientes si se cumple la siguiente relación:

$$n_{ij}/N = (n_{i.}/N).(n_{.j}/N) \qquad (3.4)$$

es decir, que la frecuencia relativa conjunta sea igual al producto de las frecuencias relativas marginales.

Otra forma de dar el concepto de independencia estadística es haciendo uso de las distribuciones condicionales. En este caso se dice que dos variables son estadísticamente independientes si las frecuencias relativas condicionales son iguales a sus correspondientes frecuencias relativas marginales.

$$f_{i/j} = n_{ij}/n_{.j} = (n_{i.}/N) = f_{i.} \quad (3.5)$$

$$f_{j/i} = n_{ij}/n_{i.} = (n_{.j}/N) = f_{.j} \quad (3.6)$$

Ejemplo 3. *Estudie si las variables del ejemplo 1 son o no independientes.*

En este caso, como en otros de naturaleza similar, para determinar si esas dos variables son o no independientes se procederá a aplicar alguna de las condiciones de independencia dadas con anterioridad. Para ello nos centraremos en un punto del espacio de X e Y, por ejemplo el par de valores (x=3, y=6). En este caso se tiene que

$$20/2005 = f_{36} \neq (f_{3.})(f_{.6}) = (255/2005)(250/2005)$$

$$(255/2005) = (f_{3.}) \neq (f_{3/y}) = 25/250$$

Lo anterior nos lleva a concluir que esas variables no son independientes. La selección del par (x, y) es indiferente, pues basta que para un par no se cumpla la condición de independencia para que se pueda concluir que las variables no son independientes.

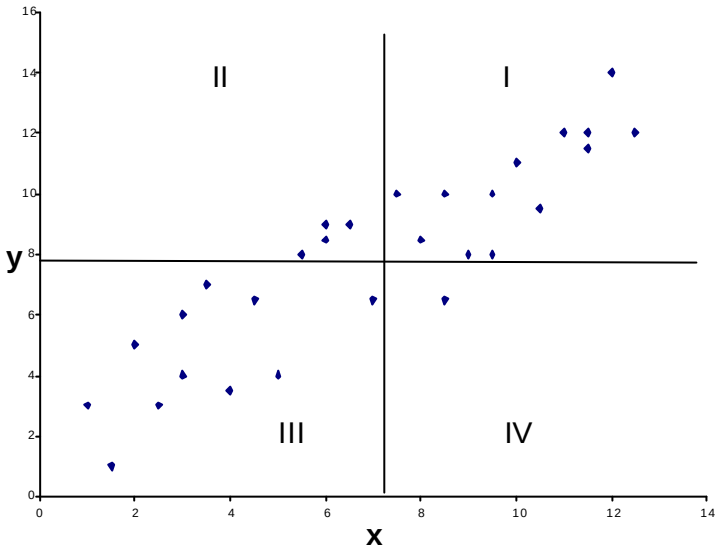
A la independencia estadística definida de esta forma se le llama determinista, frente a la estocástica.

3.3 Correlación: covariancia y coeficiente de correlación lineal.

De los distintos diagramas de dispersión que hemos mostrado en el epígrafe anterior, dos de ellos implicaban una covariación de tipo lineal, en un caso directa y en el otro indirecta o inversa. También se dijo anteriormente que una forma de cuantificar la covariación entre dos variables es mediante el análisis de la correlación. Pues bien, en lo que sigue vamos a definir un instrumento que nos va a permitir cuantificar el grado de covariación lineal entre dos variables. Se trata del coeficiente de correlación lineal.

Para deducir su expresión de cálculo y su significado haremos uso del diagrama de dispersión representado en la Figura 5.

Figua 5. Diagrama de dispersión



En este diagrama se ha realizado un cambio de origen en los ejes de forma que el nuevo origen se sitúa en el punto correspondiente a las medias para las variables *X* e *Y*. Ahora el diagrama de dispersión o nube de puntos está repartido en cuatro cuadrantes. El primer cuadrante se corresponde con los valores de *X* e *Y* mayores que sus medias. El segundo se corresponde con los valores de *X* menores que su media y con los de *Y* mayores que la suya. En el tercero están situados los valores de *X* e *Y* menores que sus medias respectivas. Finalmente, en el cuarto aparecen los valores de *X* mayores que su media y los de *Y* menores que la suya.

De acuerdo con esta distribución de puntos del diagrama, si ahora definimos las desviaciones de *X* e *Y* como $x_i = X_i - \overline{X}$ e $y_i = Y_i - \overline{Y}$, resulta que

a) el producto $x_i y_i$ del primer cuadrante será positivo
b) el producto $x_i y_i$ del segundo cuadrante será negativo

c) el producto $x_i y_i$ del tercer cuadrante será positivo
d) el producto $x_i y_i$ del cuarto cuadrante será negativo.

Teniendo en cuata esos resultados, resulta que $\Sigma_i x_i y_i$ sirve como medida de covariación entre X e Y. Esto es así porque si esa suma es positiva, la mayor parte de los puntos estarán en los cuadrantes I y III, con lo que la relación será directa. Por el contrario, si la mayoría de los puntos están en los cuadrantes II y IV, la suma será negativa y la relación será inversa. En cambio si los puntos están muy repartidos entre los cuatro cuadrantes, la suma será pequeña, tendente a cero, lo que nos informará de que no hay relación lineal alguna.

Pero ese indicador del grado de asociación lineal entre dos variables adolece de dos defectos. Por un lado bastaría con cambiar el número de pares de valores de X e Y para que el mismo fuera distinto. Por otro, el mismo viene influido por las unidades de medida de *X* e *Y*. La forma de corregir estos inconvenientes es promediar la suma (se elimina el primer problema) y expresarla en términos de la desviación estándar de *X* y de *Y*. El resultado es

$$r = \frac{\dfrac{\sum_i x_i y_i}{N}}{S_X S_Y} = \frac{S_{XY}}{S_X S_Y} \qquad (3.7)$$

que se conoce como coeficiente de correlación lineal.

Al numerador del coeficiente de correlación se le llama covariancia (S_{XY}), siendo S_X la desviación estándar de *X* y S_Y la de *Y*. Como las expresiones de cálculo de las desviaciones estándares las conocemos, habrá que dar ahora la correspondiente a la covariancia.

$$S_{XY} = \frac{\sum_i x_i y_i n_i}{N} = \frac{\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) n_i}{N} = \frac{\sum_i X_i Y_i n_i}{N} - \frac{\sum_i X_i n_i}{N}\frac{\sum_i Y_i n_i}{N} \qquad (3.8)$$

Mediante el coeficiente de correlación lineal lo que se busca es un número que indique, de forma objetiva, el grado de variación lineal conjunta entre las dos variables. El signo de este coeficiente puede ser positivo o negativo, según cual sea el de la covariancia. Los valores de este coeficiente oscilan entre menos uno y más uno. La forma de interpretar el significado de esos valores es la siguiente:

a) Si $r = 1$, la correlación lineal es perfecta y directa, o sea, la nube de puntos se sitúa sobre una línea recta creciente.

b) Si $r = -1$, la correlación lineal es perfecta y inversa, o sea, la nube de puntos se sitúa sobre una linera recta decreciente.

c) Si $r = 0$, no existe relación lineal, bien porque no exista covariación entre las variables o porque ésta no sea lineal. En este caso decimos que las variables están incorrelacionadas linealmente, lo que no significa que necesariamente sean independientes. Si el coeficiente de correlación lineal es cero, entonces las variables puede que sean independientes o bien que no lo sean y que presenten otro tipo de covariación distinto al lineal. En cambio si las variables son independientes, entonces el coeficiente de correlación lineal será siempre cero.

d) En los demás casos se puede hablar de una correlación débil o fuerte según que el valor de r esté próximo a 0 o a ± 1.

En cuanto a las propiedades del coeficiente de correlación lineal, hay que indicar que el mismo es invariante frente a cambios de origen y de escala. Para probar que esta afirmación es cierta se estudiará el comportamiento de la covariancia frente a cambios de origen y de escala en las variables X e Y, pues ya se sabe cual es la respuesta de la desviación estándar frente a este tipo de cambios. Supóngase que se definen las siguientes variables: $X' = h + kX$ e $Y' = f + gY$. Entonces:

$$S'_{xy} = \frac{\sum_i \left(x'_i - \overline{x'}\right)\left(y'_i - \overline{y'}\right)}{N} = \frac{\sum_i \left(h + kx_i - h - k\overline{x}\right)\left(f + gy_i - f - g\overline{y}\right)}{N} =$$

$$= \frac{kg\sum_i \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{N} = kgS_{xy}$$

Así pues se comprueba que, al igual que la variancia, la covariancia solo se ve afectada por cambios de escala, por lo que el coeficiente de correlación resulta invariante a los cambios de origen y de escala.

Para finalizar esta exposición sobre el coeficiente de correlación lineal, hay que señalar que para el cálculo del mismo no se asume ningún tipo de relación de causalidad. Por otro lado debe quedar claro que este coeficiente lo que mide es la intensidad de la relación lineal entre variables. Así un coeficiente *r = 0,8* indica una covariación más fuerte que *r= 0,4*, pero ello no implica que la covariación lineal en el primer caso sea doble que en el segundo.

Ejemplo 4 *Obtener el coeficiente de correlación lineal entre las variables X e Y si los valores observado de las mismas son los siguientes:*

X_i	Y_i
3	3
10	9
9	10
1	4
2	1
4	2
6	5
5	6
7	7
7	9

Para calcular el coeficiente de correlación pedido es aconsejable ampliar la tabla anterior con tres columnas adicionales como las que aparecen en la siguiente tabla:

x_i	y_i	x_i^2	y_i^2	x_iy_i
3	3	9	9	9
10	9	100	81	90
9	10	81	100	90
1	4	1	16	4
2	1	4	1	2
4	2	16	4	8
6	5	36	25	30
5	6	25	36	30
7	7	49	49	49
7	9	49	81	63
54	56	370	402	375

$$S_{XY} = \frac{\sum_i x_i y_i n_i}{N} - \frac{\sum_i x_i n_i}{N}\frac{\sum_i y_i n_i}{N} = \frac{375}{10} - \frac{54}{10}\frac{56}{10} = 7{,}26$$

$$S_x = \sqrt{\frac{\sum_i x_i^2 n_i}{N} - \left(\overline{x}\right)^2} = \sqrt{\frac{370}{10} - \left(\frac{54}{10}\right)^2} = 2{,}8$$

$$S_y = \sqrt{\frac{\sum_i y_i^2 n_i}{N} - \left(\overline{y}\right)^2} = \sqrt{\frac{402}{10} - \left(\frac{56}{10}\right)^2} = 2{,}97$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{7{,}26}{(2{,}8)(2{,}97)} = 0{,}873$$

Ejemplo 5. *Obtenga el coeficiente de correlación lineal para las variables que se recogen en la tabla siguiente.*

		Y			
		1	**2**	**3**	**4**
X	**1**	10	8	5	3
	2	7	12	6	3
	3	6	8	8	4
	4	1	4	5	10

En este caso, se trata de obtener el coeficiente de correlación cuando las frecuencias de los distintos pares de valores de las variables no son unitarias y, además, todos esos pares tienen frecuencias distintas de cero, cosa que no ocurría en el Ejemplo 4. Para calcular la correlación existente entre X e Y, es aconsejable, cuando se tiene una distribución de frecuencias como la presente, determinar previamente las marginales y después dar esa tabla de doble entrada en forma de pares de valores. Todo ello nos lleva a que:

x_i	n_i	$x_i n_i$	$x_i^2 n_i$
1	*26*	*26*	*26*
2	*28*	*56*	*112*
3	*26*	*78*	*234*
4	*20*	*80*	*320*
Total	*100*	*240*	*692*

y_i	n_i	$y_i n_i$	$y_i^2 n_i$
1	*24*	*24*	*24*
2	*32*	*64*	*128*
3	*24*	*72*	*216*
4	*20*	*80*	*320*
Total	*100*	*240*	*688*

$$S_x = \sqrt{\frac{\sum_i x_i^2 n_i}{N} - \left(\bar{x}\right)^2} = \sqrt{\frac{692}{100} - \left(\frac{240}{100}\right)^2} = 1{,}077$$

$$S_y = \sqrt{\frac{\sum_i y_i^2 n_i}{N} - \left(\bar{y}\right)^2} = \sqrt{\frac{688}{100} - \left(\frac{240}{100}\right)^2} = 1{,}058$$

x_i	y_i	n_{ij}	$x_i y_i n_i$
1	1	10	10
1	2	8	16
1	3	5	15
1	4	3	12
2	1	7	14
2	2	12	48
2	3	6	36
2	4	3	24
3	1	6	18
3	2	8	48
3	3	8	72
3	4	4	48
4	1	1	4
4	2	4	32
4	3	5	60
4	4	10	160
Total		100	617

$$S_{XY} = \frac{\sum_i x_i y_i n_i}{N} - \frac{\sum_i x_i n_i}{N}\frac{\sum_i y_i n_i}{N} = \frac{617}{100} - \frac{240}{100}\frac{240}{100} = 0{,}41$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{0{,}41}{(1{,}077)(1{,}058)} = 0{,}36$$

3.4 Regresión

El segundo procedimiento que se indicó que podía utilizarse para cuantificar la covariación entre dos variables era el análisis de regresión. La aplicación del mismo se limitará, inicialmente, al caso de dependencia causal unilateral entre dos variables.

El objetivo que se busca con el análisis de la regresión es determinar una función del tipo *y=f(x)* que relacione a estas dos variables y nos indique la forma en varían conjuntamente. Pero esta función, que se intenta cuantificar mediante el análisis de la regresión, será una

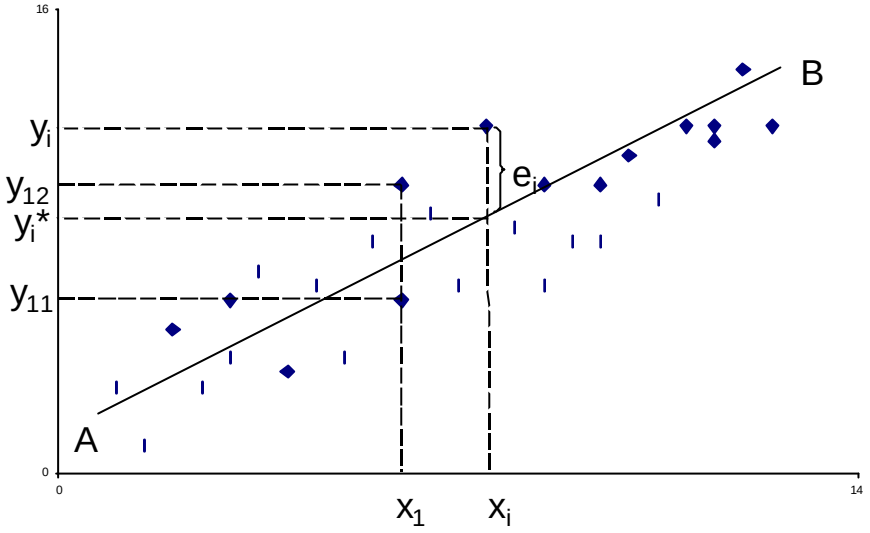
línea que intentará resumir toda la nube de puntos del diagrama de dispersión. Como tal tendrá un carácter de línea media, y esta línea nos medirá la dependencia estadística existente entre las variables. Este tipo de dependencia es distinta a la dependencia funcional o exacta. La diferencia entre las mismas radica en que en el primer caso, aunque las variables estén fuertemente relacionadas, las observaciones suelen tener una componente aleatoria que les impide que la nube de puntos aparezca exactamente distribuida a lo largo de una línea. Pero esa falta de alineación perfecta no impide que esos puntos tiendan a agruparse con mayor o menor intensidad en torno a esa línea "ideal" o media de la que se ha hablado.

Pues bien, el análisis de regresión consiste en obtener esa línea "ideal" o media, línea de regresión, hacia la cual tienden los puntos de un diagrama de dispersión. De lo que se trata, en realidad, es de determinar la dependencia exacta que se haya contenida en la dependencia estadística observada mediante la eliminación de los factores aleatorios.

Para centrar un poco estas ideas se hará uso de la Figura 6. Admitamos de entrada que esa línea media es conocida y que es la que se ha representado en el mismo como AB[1]. En ese gráfico podemos comprobar como para un determinado valor de X (x_1) observado, la variable Y puede tomar, y de hecho los toma en este caso, más de una valor (y_{11} e y_{12}), mientras que por la línea de regresión le correspondería solo uno (y^*_1). Este paso de la dependencia estadística a la dependencia exacta implica que a cada valor de la variable independiente le asignemos uno solo de la variable dependiente. Ese valor de la variable dependiente, dado por la línea de regresión, tiene categoría de valor medio, pues como ya hemos indicado, la línea de regresión tiene ese carácter de línea media.

[1] Pese a que en el gráfico la línea media o línea de regresión se ha representado como una recta, la misma puede ser una curva cualquiera.

Figura 6. Diagrama de dispersión



Mediante este gráfico también es posible comprobar como cada valor de y_i observado se puede descomponer en dos partes. Una de ellas viene dada por el valor de la línea de regresión, $y_i^*=f(x_i)$, y la otra sería la diferencia entre el valor observado y el asignado por nuestra relación funcional exacta a la que llamaremos error o residuo, e_i. Formalmente tendríamos:

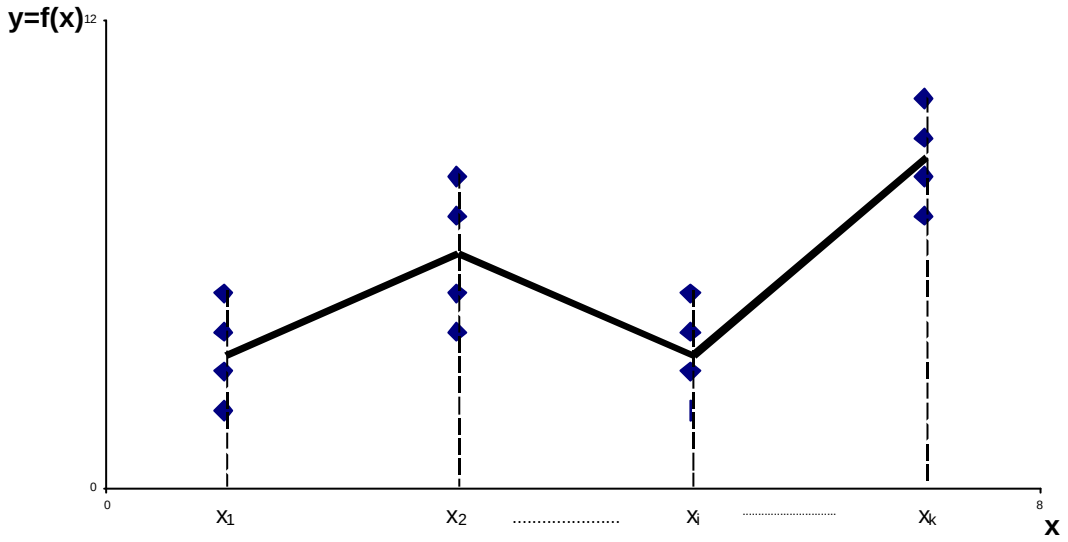
$$y_i = f(x_i) + e_i = y_i^* + e_i \,. \qquad (3.9)$$

En consecuencia el análisis de regresión lo que persigue es obtener los valores medios y^*_i de la variable dependiente que corresponden a los valores x_i observados.

El siguiente paso en el análisis de la regresión es definir los procedimientos que nos permitan obtener esa línea media que es la línea de regresión. No vamos a entrar a describir todos los posibles métodos que existen para determinar esa línea de regresión. Solo vamos a mencionar tres. El primero es el más sencillo y consiste en trazar la línea que más se ajuste a la nube de puntos. Este procedimiento gráfico, frente a su sencillez, tiene en su contra la falta de rigor.

Un segundo procedimiento consiste en sustituir todos los valores de *Y*, para un valor dado de X (x_i), por su media. Se trataría de una media condicional, $\left(x_i ; \overline{y} / x_i\right)$ y habría tantas medias como valores tome la variable independiente. Con la unión de esos valores medios se tendría la línea de regresión.

Figura 7. Línea de regresión



El tercer método es aquel que hace uso de una función matemática para explicar la dependencia exacta existente de forma implícita entre las dos variables observadas. Haciendo uso de los símbolos ya utilizados, esa función, que es la línea de regresión, es

$$y^*_i = f(x_i) \qquad (3.10)$$

Esta relación, a parte de decirnos que la variable *Y* depende de la variable *X*, y de servir para describir la relación causal exacta, permite realizar predicciones de la variable dependiente conocidos los valores de la variable independiente. Pero esas predicciones tienen un carácter de valores medios, pues los errores o residuos son impredecibles.

El siguiente paso supone la elección de la función matemática *f(x)* que ha de ser nuestra línea de regresión. Se trata pues de elegir aquella función *f(x)* que describa de la forma

más adecuada la dependencia entre las variables. A esas funciones se les denomina genéricamente como modelos. Los modelos más sencillos son los siguientes:

a) Modelo lineal : $y^* = a + bx$.
b) Modelo parabólico de segundo grado: $y^* = a + bx + cx^2$.
c) Modelo potencial: $y^* = Ax^b$.
d) Modelo exponencial: $y^* = AB^x$.
e) Modelo hiperbólico: $y^* = a/x$

En cada caso, la elección de uno de ellos dependerá de lo que digan los datos (análisis empírico) o de lo que indique la teoría. En todos esos modelos hemos introducido, además de las variables independiente y dependiente, los símbolos *a, b, c, A y B*. Los mismos reciben el nombre de coeficientes ó parámetros. Una vez seleccionado el modelo que se ajusta a la línea de puntos o que responde a una teoría existente, lo que debemos realizar a continuación es definir un método que nos permita cuantificar esos coeficientes a partir de los datos observados. El procedimiento más utilizado es el denominado método de los mínimos cuadrados ordinarios (MCO).

3.4.1 Método de los mínimos cuadrados.

Este método consiste en determinar unos valores para los coeficientes ó parámetros de la función seleccionada, $y_i^* = f(x_i)$, con la condición de que haga mínima la suma de los errores al cuadrado, conforme se definieron en (3.9) ($e_i = (y_i - y_i^*)$), es decir:

$$\sum_i e_i^2 = \sum_i \left(y_i - y_i^*\right)^2 = mínimo \qquad (3.11)$$

De forma más general esta expresión se puede dar como

$$\sum_i e_i^2 n_i = \sum_i \left(y_i - y_i^*\right)^2 = mínimo \qquad (3.12)$$

es decir, admitir frecuencias mayores que uno para los distintos pares de valores de X e Y. Con este planteamiento, y para el caso de un modelo lineal, la función a minimizar sería

$$j(a,b) = \sum_i e_i^2 n_i = \sum_i \left(y_i - y_i^*\right)^2 n_i = \sum_i \left(y_i - a - bx_i\right)^2 n_i \qquad (3.13)$$

Si a esta función le aplicamos la primera condición de mínimo se llega al siguiente sistema:

$$\frac{\partial j(a,b)}{\partial a} = -2\sum_i \left(y_i - a - bx_i\right) n_i = 0 \qquad (3.14)$$

$$\frac{\partial j(a,b)}{\partial b} = -2\sum_i \left(y_i - a - bx_i\right) x_i n_i = 0 \qquad (3.15)$$

Estas dos ecuaciones se pueden expresar como:

$$\sum_i e_i n_i = 0 \qquad (3.16)$$

$$\sum_i e_i x_i n_i = 0 \qquad (3.17)$$

A partir de este sistema de ecuaciones se llega a otro donde se aprecia, de forma más clara, que las soluciones del mismo son funciones de los valores observados de las variables:

$$\sum_i y_i n_i = Na + b\sum_i x_i n_i \qquad (3.18)$$

$$\sum_i y_i x_i n_i = a\sum_i x_i n_i + b\sum_i x_i^2 n_i \qquad (3.19)$$

A estas ecuaciones se les conoce con el nombre de ecuaciones normales. A partir de ellas se pueden obtener unas fórmulas de cálculo de los parámetros de la recta. Esas fórmulas son:

$$b = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{S_{xy}}{S_x^2} \tag{3.20}$$

$$a = \bar{y} - b\bar{x} \tag{3.21}$$

Al parámetro *b* del modelo lineal se le llama también coeficiente de regresión. Este coeficiente mide la tangente o pendiente de la recta. Su signo será el de la covariancia, que a su vez le daba el signo al coeficiente de correlación. Así pues cuando la relación es directa, el signo será positivo y cuando al relación es negativa o inversa el signo será negativo. Este coeficiente mide la variación de la variable *Y* frente un cambio unitario de la variable *X*. También se puede interpretar como la razón de una progresión aritmética.

Al parámetro *a* del modelo lineal se le conoce como término independiente u ordenada en el origen. Nos daría el valor de la variable dependiente cuando la independiente valiese cero.

De igual forma que se han obtenido las ecuaciones normales para este modelo lineal de dos variables (*Y*, *X*), también es posible llegar a un sistema de ecuaciones normales para cada uno de los modelos indicados anteriormente o para un modelo lineal donde el número de variables explicativas sea mayor que uno. Cada sistema tendrá tantas ecuaciones como parámetros existan en la línea de regresión. Para el caso del modelo parabólico las ecuaciones normales son:

$$\sum_i y_i n_i = Na + b\sum_i x_i n_i + c\sum_i x_i^2 n_i$$

$$\sum_i y_i x_i n_i = a\sum_i x_i n_i + b\sum_i x_i^2 n_i + c\sum_i x_i^3 n_i$$

$$\sum_i y_i x_i^2 n_i = a\sum_i x_i^2 n_i + b\sum_i x_i^3 n_i + c\sum_i x_i^4 n_i$$

A este modelo se le puede considerar como un modelo lineal de tres variables, una dependiente (*Y*) y dos explicativas o independientes (*X* , X^2), pues el mismo se podría haber formulado como: $y_i^* = a + bx_1 + cx_2$, donde $x_1=x$ y $x_2=x^2$. Tanto a este modelo, como a los otros señalados, se les conoce como modelos no lineales pero linealizables mediante las transformaciones adecuadas. Así, si en el modelo potencial, que como se ha visto viene dado por $y^* = Ax^b$, se toman logaritmos, entonces el nuevo modelo es lineal, y vendría dado por: $lny^* = lnA + blnx = a' + bx'$. En este modelo, el parámetro *b* es la elasticidad de *Y* con respeto de *X*, que es, además, constante, pues:

$$e = \frac{\partial y}{\partial x}\frac{x}{y} = Abx^{b-1}\frac{x}{Ax^b} = b \qquad (3.22).$$

De forma similar se podría proceder con el modelo exponencial $y^*=AB^x$. En este caso la linealización del mismo se obtendría, también, mediante el uso de logaritmos: $lny^*=lnA+xlnB=a'+b'x$. Ahora en este modelo, el coeficiente *b* es la razón de una progresión geométrica y a partir del mismo puede obtenerse la tasa de crecimiento de la variable *Y*.

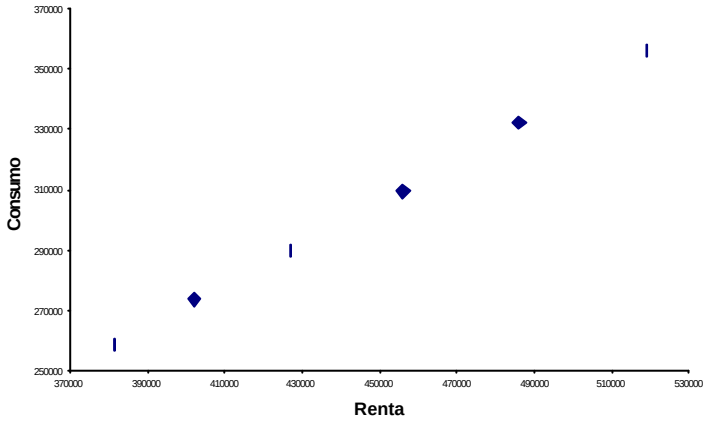
Ejemplo 6 *En la tabla siguiente se recoge la evolución, para el periodo 1999-2000, de dos de las principales macromagnitudes de la Economía de España. La Renta Nacional Disponible Neta a precios corrientes así como el Gasto en Consumo Final de los Hogares, expresadas las dos en miles millones de euros. Con estos datos: a) ajuste la función agregada de consumo, dando una interpretación del significado de los coeficientes obtenidos; b) obtenga la elasticidad del consumo para una renta de 450 miles de millones de euros así como la propensión media al consumo para esa renta.*

	Renta	*Consumo*
1995	*381,5*	*258,6*
1996	*402,3*	*273,6*
1997	*426,9*	*289,7*
1998	*456,1*	*309,3*
1999	*486,0*	*331,8*
2000	*519,0*	*356,2*

Fuente: Web INE

a) Antes de realizar ajuste alguno hay que preguntarse por la relación funcional de la función agregada de consumo, pues la Teoría Económica, de entrada, solo dice que el consumo de las familias es una función del la renta, sin especificar mucho más. Sin embargo, la teoría keynesiana señala que la relación entre ambas variables es de tipo lineal. En cualquier caso parece aconsejable realizar un análisis gráfico exploratorio previo que confirme o desmienta ese planteamiento teórico. A tal efecto se ha realizado la Figura 8 donde se aprecia que, al menos a corto plazo, el planteamiento keynesiano no es del todo erróneo, pues las seis parejas de valores están casi alineadas. Esto nos permite ensayar un ajuste lineal. Este resultado se puede confirmar calculando el coeficiente de correlación lineal.

Figura 8. Relación entre la renta disponible y el consumo de los hogares en España. Periodo 1995-2000



Los cálculos necesarios para su realización son los que aparecen a continuación.

	Renta (X)	Consumo (Y)	X^2	XY	Y^2
1995	381,534	258,647	145568,2	98682,6	66898,3
1996	402,283	273,561	161831,6	110048,9	74835,6
1997	426,908	289,675	182250,4	123664,6	83911,6
1998	456,102	309,279	208029,0	141062,8	95653,5
1999	486,049	331,825	236243,6	161283,2	110107,8
2000	518,999	356,225	269360,0	184880,4	126896,3
Total	**2671,875**	**1819,212**	**1203282,9**	**819622,5**	**558303,1**

$$S_{xy} = \frac{\sum_i x_i y_i}{N} - \frac{\sum_i x_i}{N}\frac{\sum_i y_i}{N} = \frac{819622,5}{6} - \frac{2671,9}{6}\frac{1819,2}{6} = 1584,1$$

$$S_x = \sqrt{\frac{\sum_i x_i^2}{N} - \left(\bar{x}\right)^2} = \sqrt{\frac{12032829}{6} - \left(\frac{2671,9}{6}\right)^2} = 47,4$$

$$S_y = \sqrt{\frac{\sum_i y_i^2}{N} - \left(\bar{y}\right)^2} = \sqrt{\frac{5583031}{6} - \left(\frac{1819,2}{6}\right)^2} = 33,5$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{1584,1}{(47,4)(33,5)} = 0,99966898$$

Como puede observarse, la relación lineal entre estas dos variables es muy fuerte.

$$b = \frac{Cov(x,y)}{Var(x)} = \frac{S_{xy}}{S_x^2} = \frac{1584,1}{2243,9} = 0,706$$

$$a = \bar{y} - b\bar{x} = 303,2 - (0,706)(445,3) = -11,2$$

Todos estos cálculos nos llevan a que, finalmente, la función lineal ajustada pueda expresarse como:

$$y_i = y_i^* + e_i = -11,2 + 0,706x_i + e_i$$

En este ajuste, donde la variable dependiente es el consumo agregado y la independiente la renta disponible, a la pendiente de la recta se le conoce como propensión marginal al consumo, indicando la proporción que de cada unidad monetaria de renta se dedica al consumo. En este contexto se habla de proporción porque se asume la identidad fundamental de la renta, según la cual una unidad de renta cualquiera solo puede destinarse al consumo o al ahorro. Así, para esta función agregada de consumo se tiene que la diferencia entre la unidad y la propensión marginal al consumo es la propensión marginal al ahorro.

A la ordenada en el origen o término independiente de esta función ajustada se le conoce como consumo autónomo. Es decir, se trataría de un consumo mínimo o de subsistencia correspondiente a niveles de renta nulos. En nuestro caso el signo de este coeficiente es negativo, lo que económicamente carece de sentido, pues el consumo deberá ser siempre mayor o igual que cero. La razón de este "absurdo" económico hay que buscarla en que en este ejemplo la ordena en el origen, desde el punto de vista estadístico, es una predicción, pues nos da el valor del consumo bajo el supuesto de que la renta fuera nula. Pero esos valores de la renta no han sido observados. Esto nos lleva a plantearnos nuevas cuestiones tales como la calidad del ajuste y la fiabilidad de las posibles predicciones que se puedan realizar con un modelo de regresión. Pero esto son cuestiones que se irán respondiendo en los siguientes epígrafes de es capitulo.

b) Sabemos que la elasticidad se define como

$$\boldsymbol{e} = \frac{\partial y}{\partial x}\frac{x}{y}$$

En nuestro caso tendremos que:

$$\boldsymbol{e} = \frac{\partial y}{\partial x}\frac{x}{y} = (0{,}706)\frac{450}{306{,}5} = 1{,}0365$$

lo que significa que cuando la renta varía en 1%, en un entorno próximo a los 450 miles de millones de euros, el consumo cambia en 1,0365%.

A su vez la propensión media al consumo se define como:

$$PMC = \frac{y^*}{x} = \frac{306{,}5}{450} = 0{,}6811$$

3.4.2. Regresión multivariante.

En los epígrafes anteriores se ha hablado de distribuciones bivariantes, en las que se estudiaba la distribución conjunta de dos variables, las cuales eran el resultado de observar dos caracteres de una población. Pero los caracteres observables de una población no tienen porqué limitarse a solo dos. Mas bien al contrario. Lo normal es que puedan observarse más de dos. En estos casos se tendría, también, una distribución de frecuencias conjuntas cuya representación mediante una tabla de doble entrada se hace difícil (caso de tres variables) o no puede realizarse (para más de tres variables). Pero estas dificultades que plantea su tabulación no impiden que la técnica del análisis de regresión vista anteriormente no le sea aplicable. Por el contrario, es precisamente para este tipo de situaciones donde la regresión se convierte en un instrumental de análisis realmente potente. Pero cuando lo que se pretende ajustar es una función a una nube de puntos de más de dos dimensiones, entonces el procedimiento descrito anteriormente para determinar los coeficientes de la línea de regresión, basado en la obtención de un sistema de ecuaciones (ecuaciones normales), resulta poco operativo, pues la resolución del mismo, cuando se tienen muchas incógnitas (muchos coeficientes de regresión), se hace tedioso por el elevado número de ecuaciones del que puede constar. En estos casos lo que se hace es recurrir al álgebra matricial. Para centrar un poco las ideas, supongamos que tenemos una variable *Y* que depende de *k* variables explicativas *X*. Para estas *k+1* variables (las *k* independientes más ordenada en el origen) se realizan *N* observaciones. Si admitimos que la relación funcional entre la variable dependiente y las independientes o explicativas es de tipo lineal, entonces tendríamos la función:

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots\ldots\ldots\ldots + b_k x_{ki} + e_i = y_i^* + e_i \qquad (3.23)$$

Pero este modelo, en términos matriciales, y para todas las observaciones realizadas se puede expresar como:

$$\mathbf{y = Xb + e = y^* + e} \qquad (3.24)$$

donde **y** es un vector de dimensión *Nx1,* correspondiente a las *N* observaciones de la variable *Y*, **X** es una matriz *Nx(k+1)* correspondiente a las *N* observaciones de las *k+1* variables (las *k* variables explicativas más un vector columna de unos correspondiente al

términos independiente *a*), **b** es un vector *(k+1)x1* de coeficientes y **e** es el vector *Nx1* de errores. De forma expandida el modelo quería como:

$$y_1 = a + b_1 x_{11} + b_2 x_{21} + \ldots\ldots\ldots + b_k x_{k1} + e_1$$
$$y_2 = a + b_1 x_{12} + b_2 x_{22} + \ldots\ldots\ldots + b_k x_{k2} + e_2$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$y_N = a + b_1 x_{1N} + b_2 x_{2N} + \ldots\ldots\ldots + b_k x_{kN} + e_N$$

O bien como:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots\ldots & x_{k1} \\ 1 & x_{12} & \ldots\ldots & x_{k2} \\ & \ldots\ldots\ldots\ldots\ldots & & \\ 1 & x_{1N} & \ldots\ldots & x_{kN} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ \ldots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ e_3 \end{bmatrix} = \mathbf{Xb} + \mathbf{e} \qquad (3.25)$$

Ahora, si lo que se pretende obtener es el vector de coeficientes **b** por mínimos cuadrados, entonces hay seleccionar aquel **b** que minimice la suma de cuadrados de los errores dada por:

$$\mathbf{e'e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y}-\mathbf{Xb}) = \mathbf{y'y} - 2\mathbf{b'X'y} + \mathbf{b'X'Xb} \qquad (3.26)$$

Si esta suma de cuadrados la derivamos respecto del vector **b** se obtiene la condición necesaria para que esa función sea mínima:

$$\partial(\mathbf{e'e})/\partial\mathbf{b} = -2\mathbf{X'y} + 2\mathbf{X'Xb} \qquad (3.27)$$

De este resultado se deduce que la condición necesaria de mínimo buscada es:

$$\mathbf{X'Xb} = \mathbf{X'y} \qquad (3.28)$$

Por lo que el vector de coeficientes **b** vendrá dado por:

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y} \qquad (3.29)$$

De la condición necesaria para minimizar la suma de cuadrados se deduce que:

$$\mathbf{X'Xb} = \mathbf{X'y} = \mathbf{X'Xb} + \mathbf{X'e} \qquad (3.30)$$

Es decir, que $\mathbf{X'e} = \mathbf{0}$. Este es el mismo resultado que se obtuvo en el caso de dos variables, donde se vio que $\sum_i e_i n_i = 0$ y que $\sum_i e_i x_i n_i = 0$. Estos resultados nos llevan también a que: $\sum_i y_i = \sum_i y_i^*$ en el caso multivariante, pues la suma de los errores es cero, y, en consecuencia, la media de los valores observados de *y* es igual a la media de los valores ajustados *y**.

Ejemplo 7. *Para un conjunto de empresas, de características similares, se han obtenido los siguientes datos de producción y costes medios totales. Ajuste el modelo más adecuado a esos datos en que los costes sean función de la producción.*

Producción	*Costes medios*
2	*7*
7	*3*
5	*4*
6	*3.5*
4	*5.5*
8	*3.5*
3	*5*
10	*4*
12	*5*
15	*6*

En primer lugar hay que seleccionar la forma funcional que relaciona los costes medios de una empresa con sus niveles de producción. Para ello se puede recurrir a dos vías distintas. Por un lado está el análisis gráfico y por otro está la información que pueda suministrar la teoría (en nuestro caso la teoría económica). En general, será siempre preferible recurrir a la segunda fuente de información, pues el método gráfico nos puede llevar a cuantificar relaciones funcionales que se ajusten muy bien a los datos observados cuando, en realidad, se está trabajando con variables que no están relacionadas entre si. Se trataría del caso de covariación casual o espúrea. Al recurso gráfico hay que acudir

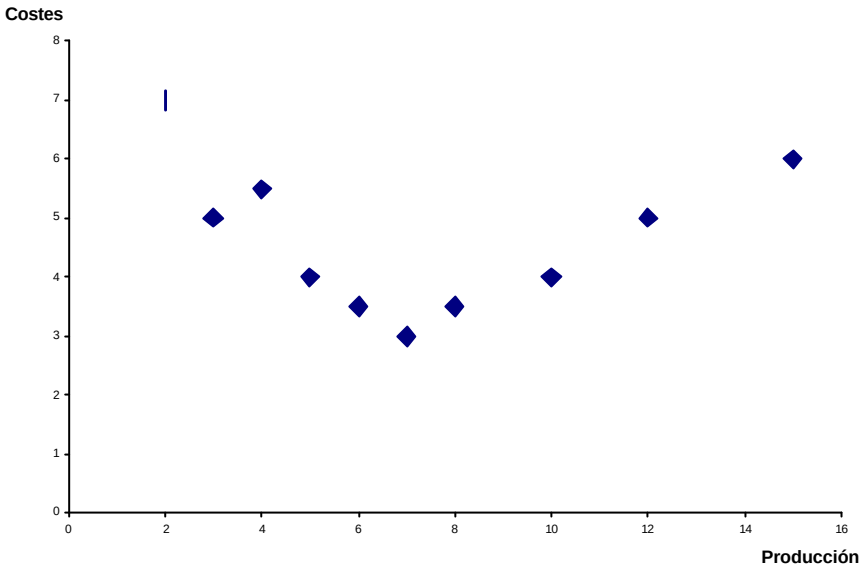
cuando no existe teoría alguna que nos informe sobre la posible relación existente entre esas variables.

Según los criterios establecidos en el párrafo anterior, lo primero que habría que hacer es indagar, desde la microeconomía, que tipo de relación hay entre los costes medios de una empresa y su producción. En este caso, la teoría nos dice que los costes medios decrecen inicialmente (debido a la caída de los costes medios fijos) hasta que se alcanza un cierto nivel de producción y después crecen. Así pues sería poco razonable pensar en un modelo lineal como el más adecuado para este ejemplo. Para confirmar lo que nos dice la teoría se puede recurrir, de forma complementaria, al análisis gráfico. En este caso, los datos de esas empresas está representados en la Figura 9.

Como puede comprobarse, y según indicaba la teoría, la relación no es lineal. Sin embargo, y a titulo de mero ejercicio, se va a proceder al ajuste lineal de estos datos mediante mínimos cuadrados. En primer lugar se calculará el coeficiente de correlación lineal, el cual nos indicará si la covariación lineal entre esas variables es fuerte o débil. Para obtener este coeficiente, y los demás que iremos obteniendo, resulta aconsejable construir la tabla siguiente:

xy	x^2y	x^2	x^3	x^4
14	28	4	8	16
21	147	49	343	2401
20	100	25	125	625
21	126	36	216	1296
22	88	16	64	256
28	224	64	512	4096
15	45	9	27	81
40	400	100	1000	10000
60	720	144	1728	20736
90	1350	225	3375	50625
331	**3228**	**672**	**7398**	**90132**

Figura 9. Relación costes-producción



En la misma aparece recogida toda la información necesaria para la resolución de este ejemplo.

El coeficiente de correlación lineal, como sabemos, es igual al cociente de la covariancia entre el producto de las desviaciones estándares.

$$S_{xy} = \frac{\sum_i x_i y_i n_i}{N} - \frac{\sum_i x_i n_i}{N}\frac{\sum_i y_i n_i}{N} = \frac{331}{10} - \frac{72}{10}\frac{46{,}5}{10} = -0{,}38$$

$$S_x = \sqrt{\frac{\sum_i x_i^2 n_i}{N} - \left(\overline{x}\right)^2} = \sqrt{\frac{672}{10} - \left(\frac{72}{10}\right)^2} = 3{,}919$$

$$S_y = \sqrt{\frac{\sum_i y_i^2 n_i}{N} - \left(\overline{y}\right)^2} = \sqrt{\frac{230{,}75}{10} - \left(\frac{46{,}5}{10}\right)^2} = 1{,}205$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{-0{,}38}{(3{,}919)(1{,}205)} = -0{,}08$$

Como puede verse, la relación lineal entre estas dos variables es muy débil y el signo del coeficiente de correlación es intranscendente en este caso, pues la curva tiene dos ramas, una decreciente y la otra creciente.

Si a pesar de todos los indicios detectados hasta ahora contra la relación lineal se siguiera adelante con este tipo de ajuste, los resultados serían los siguientes:

$$b = \frac{Cov(x,y)}{Var(x)} = \frac{S_{xy}}{S_x^2} = \frac{-0{,}38}{15{,}36} = -0{,}0247$$

$$a = \overline{y} - b\overline{x} = 4{,}65 - (-0{,}0247)(7{,}2) = 4{,}828$$

$$y_i = y_i^* + e_i = 4{,}828 - 0{,}0247x_i + e_i$$

Este ajuste lineal, como puede comprobarse por el signo negativo del coeficiente de regresión b, solo recoge el tramo decreciente de la curva, aquel donde los costes fijos medios prevalecen sobre los costes variables medios. En cambio no es capaz de detectar la rama creciente de la curva.

Si en lugar de ajustar una función lineal se trabajara con una función parabólica, entonces el sistema de ecuaciones normales asociado a ese modelo sería el siguiente:

$$\begin{aligned} 46{,}5 &= 10a + 72b + 672c \\ 331 &= 72a + 672b + 7398c \\ 3228 &= 672a + 7398b + 90132c \end{aligned}$$

con lo que los parámetros o coeficientes calculados son:

$$a = 8{,}616 \quad b = -1{,}22 \quad c = 0{,}07177$$

de forma que el modelo ajustado quedaría:

$$y_i = y_i^* + e_i = 8{,}616 - 1{,}22\, x_i + 0{,}07177\, x_i^2 + e_i$$

Al mismo resultado se habría llegado si en lugar de resolver ese sistema de ecuaciones se hubiera trabajado en términos matriciales. En este caso la matriz ***X*** *y el vector* ***y*** *vienen dados por:*

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 7 & 49 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 4 & 16 \\ 1 & 8 & 64 \\ 1 & 3 & 9 \\ 1 & 10 & 100 \\ 1 & 12 & 14 \\ 1 & 15 & 225 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 7 \\ 3 \\ 4 \\ 3{,}5 \\ 5{,}5 \\ 3{,}5 \\ 5 \\ 4 \\ 5 \\ 6 \end{bmatrix}$$

y a partir de ellas se obtiene que:

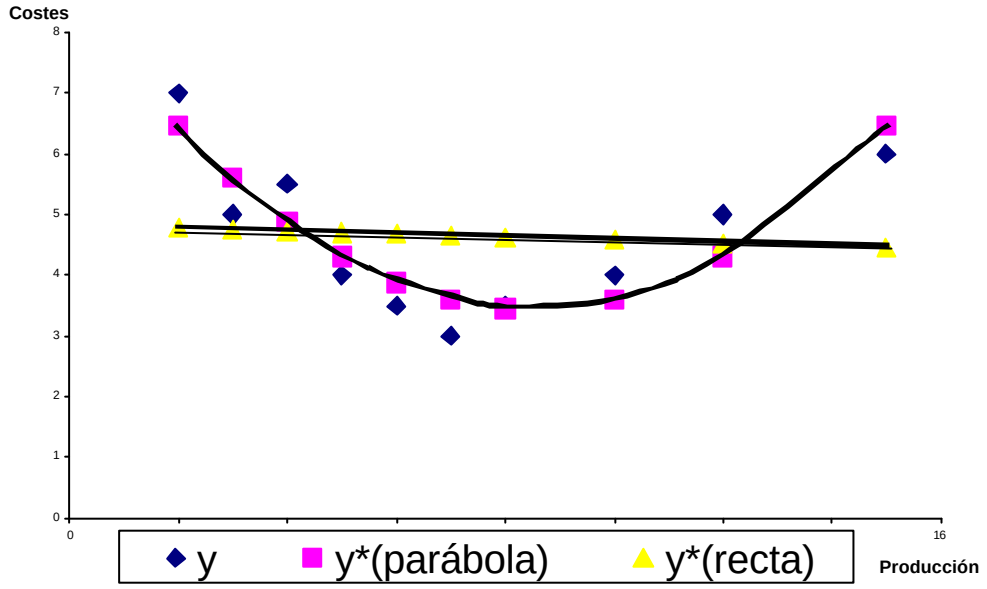
$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} n & \sum_i x_{1i} & \sum_i x_{2i} \\ \sum_i x_{1i} & \sum_i x_{1i}^2 & \sum_i x_{1i} x_{2i} \\ \sum_i x_{2i} & \sum_i x_{1i} x_{2i} & \sum_i x_{2i}^2 \end{bmatrix} = \begin{bmatrix} 10 & 72 & 672 \\ 72 & 672 & 7398 \\ 672 & 7398 & 90132 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_i y_i \\ \sum_i y_i x_{1i} \\ \sum_i y_i x_{2i} \end{bmatrix} = \begin{bmatrix} 46{,}5 \\ 331 \\ 3228 \end{bmatrix}$$

donde x_1 son los costes y x_2 son los costes al cuadrado.

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1{,}63816 & -0{,}42595 & 0{,}02275 \\ -0{,}42595 & 0{,}12619 & -0{,}00718 \\ 0{,}02275 & -0{,}00718 & 0{,}00043 \end{pmatrix}\begin{pmatrix} 46{,}5 \\ 331 \\ 3228 \end{pmatrix} = \begin{pmatrix} 8{,}6162 \\ -1{,}2207 \\ 0{,}0718 \end{pmatrix}$$

Figura 10. Valores observados y ajustados



3.4.3 Variancia residual y coeficiente de determinación.

Si ahora quisiéramos analizar cual de los dos modelos se ajusta mejor a los datos de partida se podría proceder a la inspección gráfica (Figura 10) o a calcular los errores que se cometen con los dos modelos.

Como puede apreciarse con estos resultados, tanto gráfico como numérico, la parábola representa un ajuste mejor que la línea recta, pues los errores que se comenten con la recta son mayores que con la parábola. Pero hay que fijarse no en los errores de cada uno de los modelos, pues como sabemos la suma de ellos es siempre cero de acuerdo con la primera de las ecuaciones normales, sino en el cuadrado de los mismos que es lo que se pretende hacer mínimo. En este sentido diremos que un modelo es tanto mejor

cuanto menor sea la suma de los cuadrados de sus errores, y más concretamente la media de esa suma. Este es precisamente el criterio que habrá que usar para medir la bondad o representatividad de un modelo.

		Modelo lineal			Modelo parabólico		
x_i	y_i	y*	e_i	e_i^2	y*	e_i	e_i^2
2	7	4,7786	2,2214	4,9344	6,4619	0,5381	0,2896
3	5	4,7539	0,2461	0,0606	5,6000	-0,6000	0,3600
4	5,5	4,7292	0,7708	0,5942	4,8817	0,6183	0,3823
5	4	4,7044	-0,7044	0,4962	4,3069	-0,3069	0,0942
6	3,5	4,6797	-1,1797	1,3917	3,8756	-0,3756	0,1411
7	3	4,6549	-1,6549	2,7389	3,5879	-0,5879	0,3457
8	3,5	4,6302	-1,1302	1,2774	3,4438	0,0562	0,0032
10	4	4,5807	-0,5807	0,3372	3,5861	0,4139	0,1714
12	5	4,5313	0,4688	0,2197	4,3025	0,6975	0,4865
15	6	4,4570	1,5430	2,3808	6,4537	-0,4537	0,2058
			0,0000	**14,4310**		**0,0000**	**2,4797**

Esta forma de proceder descansa en dos argumentos. El primero es el indicado antes, es decir, la selección del modelo adecuado ha de realizarse en función de que la media del cuadrado de sus errores sea mínima. En segundo lugar hay que recordar que la línea de regresión (y^*) se ha definido como una línea media que trata de resumir toda la nube de puntos. Pues bien, como tal media es aconsejable acompañarla de una indicador que mida su representatividad o bondad de ajuste, al igual que se hacía con la media aritmética y la variancia. La mayor o menor bondad dependerá de que las desviaciones de los valores observados de Y con respecto a los que se obtienen mediante la línea de regresión sean pequeñas o grandes. Si esas desviaciones son pequeñas la bondad será alta. Por el contrario, si las desviaciones son grandes la bondad será pequeña. Lo ideal es que esas desviaciones fueran siempre nulas. Pero como la suma de esas desviaciones es siempre nula (la suma de esas desviaciones es simplemente la suma de los errores) entonces la media de esa suma no nos sirve. Por esa razón, y de forma similar a como se procedió con la media y la variancia, el indicador de la bondad del ajuste se basará en la media del cuadrado de los errores. Este promedio se define como la media cuadrática de las desviaciones de los valores observados de Y respecto de sus valores "medios" Y^*. Esta media cuadrática, que sería una variancia, quedaría como:

$$S_e^2 = \frac{\sum_i \left(y_i - y_i^*\right)^2}{N} = \frac{\sum_i e_i^2}{N} \qquad (3.31)$$

A esta media cuadrática se le llama variancia residual. Se le llama así porque a los errores e_i se les conoce también como residuos. Esta variancia no es la misma que la ya definida para Y (S_y^2), pues en un caso las diferencias se toman con respecto a $\overline{Y}$ y no con respecto a Y*.

La anterior definición de variancia residual es válida para cualquier ajuste, sin importar el tipo de modelo, lineal o no. Ahora bien, el cálculo de la misma dependerá del modelo con el que se esté trabajando. En el caso lineal para dos variables la variancia residual es:

$$S_e^2 = \frac{\sum_i \left(y_i - y_i^*\right)^2}{N} = \frac{\sum_i e_i^2}{N} = \frac{\sum_i e_i\left(y_i - a - bx_i\right)}{N} = \frac{\sum_i e_i y_i - a\sum_i e_i - b\sum_i e_i x_i}{N} = \frac{\sum_i e_i y_i}{N} =$$

$$= \frac{\sum_i \left(y_i - a - bx_i\right) y_i}{N} = \frac{\sum_i y_i^2 - a\sum_i y_i - b\sum_i x_i y_i}{N} \qquad (3.32)$$

De forma similar se llega a que la variancia residual para el caso de la parábola es:

$$S_e^2 = \frac{\sum_i y_i^2 - a\sum_i y_i - b\sum_i x_i y_i - c\sum_i x_i^2 y_i}{N} \qquad (3.33)$$

Finalmente, esa variancia, para el caso lineal multivariante, viene dada por:

$$S_e^2 = \frac{\mathbf{e'e}}{N} = \frac{\mathbf{y'y} - 2\mathbf{b'X'y} + \mathbf{b'X'Xb}}{N} = \frac{\mathbf{y'y} - 2\mathbf{b'X'y} + \mathbf{b'X'X}\left(\mathbf{X'X}\right)^{-1}\mathbf{X'y}}{N} =$$
$$= \frac{\mathbf{y'y} - 2\mathbf{b'X'y} + \mathbf{b'X'y}}{N} = \frac{\mathbf{y'y} - \mathbf{b'X'y}}{N} \qquad (3.34)$$

A la raíz cuadrada de esta variancia se le conoce como error estándar del ajuste y es el equivalente a la desviación estándar ya definida previamente. Como tal, nos da el tamaño medio de los errores del ajuste. Las unidades de medida de este error estándar son las de la variable *Y*. Pero el hecho de que este indicador de los errores del ajuste no sea adimensional impide realizar comparaciones cuando se trabaja con variables dependientes de distinta naturaleza. Este problema se resolvió haciendo uso del coeficiente de variación cuando se trabajaba con una sola variable. Ahora, en el contexto de la regresión, el problema se resuelve recurriendo a lo que se conoce como coeficiente de determinación**.** Este coeficiente se utiliza para estudiar la representatividad de la línea de regresión o bondad del ajuste.

Debe recordarse que el objetivo en el análisis de la regresión es explicar las variaciones observadas en la variable *Y* mediante las variaciones de la variable explicativa *X*. Como ya se ha podido comprobar, la variable *X* no es capaz, por si sola, de explicar todas las variaciones de la *Y*, por lo que se admite la posibilidad de cometer un error, de manera que los valores de *Y,* como se indicó en (3.9) y (3.24), se pueden descomponer en dos términos:

$$y_i = y^*_i + e_i. \qquad (3.35)$$

Partiendo de esa relación se va a demostrar que se cumple la siguiente igualdad:

$$S_y^2 = S^2_{y^*} + S_e^2 \qquad (3.36)$$

La demostración es como sigue. En (3.35) se le resta la media de *Y* en ambos lados de la igualdad:

$$(y_i - \bar{y}) = (y^*_i - \bar{y}) + e_i \qquad (3.37)$$

A continuación se eleva al cuadrado (3.37) y se suma para las *N* observaciones resultado que:

$$\sum_i \left(y_i - \bar{y}\right)^2 = \sum_i \left(\left(y_i^* - \bar{y}\right) + e_i\right)^2 = \sum_i \left(y_i^* - \bar{y}\right)^2 + \sum_i e_i^2 + 2\sum_i \left(y_i^* - \bar{y}\right)e_i =$$
$$= \sum_i \left(y_i^* - \bar{y}\right)^2 + \sum_i e_i^2 \qquad (3.38)$$

pues $2\sum_i \left(y_i^* - \bar{y}\right)e_i$ vale cero. Si ahora dividimos todo por *N* se llega a:

$$S_y^2 = \frac{\sum_i \left(y_i - \bar{y}\right)^2}{N} = \frac{\sum_i \left(y_i^* - \bar{y}\right)^2}{N} + \frac{\sum_i e_i^2}{N} = S_{y^*}^2 + S_e^2 \qquad (3.39)$$

y este es precisamente el resultado al que se quería llegar.

Esta descomposición dada en (3.39) se puede interpretar como que la variancia total (S_y^2) es la suma de la variancia explicada por el modelo ($S_{y^*}^2$) más la variancia residual (S_e^2). A partir de esta relación entre variancias, y dado que todas ellas serán siempre no negativas, si dividimos ambos miembros de la igualdad por al variancia total y ordenamos términos, tendremos que:

$$\frac{S_{y^*}^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2} \qquad (3.40)$$

es decir, que la proporción de variancia explicada por el modelo respecto del total será igual a la unidad menos la proporción no explicada por el modelo. Pues bien, a esa proporción explicada por el modelo se le conoce como coeficiente de determinación (R^2), es decir:

$$R^2 = \frac{S_{y^*}^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2} \qquad (3.41)$$

Como se ha indicado antes, al ser la variancias magnitudes no negativas, el coeficiente de determinación será siempre mayor o igual que cero. Por otro lado, como la variancia de *Y**

es una parte de la variancia total, resultará que *los valores del coeficiente de determinación estarán comprendidos siempre entre cero y uno.* El valor cero lo tomará cuando la variancia explicada por el modelo sea cero, en cuyo caso diremos que el modelo seleccionado para el ajuste no es el adecuado, pues las variaciones de *X* no explican ninguna de las variaciones de *Y*. Por el contrario, cuando toma el valor uno ello implica que la variancia residual es nula. En tal caso la dependencia estadística se convierte en dependencia exacta. Esta es la situación que se da cuando todas las variaciones de la variable dependiente quedan perfectamente explicadas por las variaciones de la variable explicativa.

Todo lo anterior nos lleva a decir que el coeficiente de determinación da la proporción de la variación total de la variable dependiente que viene explicada por el modelo o la variable explicativa. Así cuando el coeficiente de determinación tome valores próximos a la unidad diremos que la bondad del ajuste es muy buena o que el modelo seleccionado para el ajuste es representativo, ya que explica una elevada proporción de las variaciones de la variable dependiente. Por todo ello resulta aconsejable que cada vez que realicemos un ajuste lo acompañemos de una medida de su bondad, y esta medida es el coeficiente de determinación.

El coeficiente de determinación definido para el caso lineal de dos variables se puede generalizar al modelo de *k* variables. También en este caso la variancia total se puede descomponer en la suma de la variancia explicada por el modelo más la variancia residual. Concretamente, el coeficiente de determinación vendrá dado por:

$$R^2 = 1 - \frac{S_e^2}{S_y^2} = 1 - \frac{\mathbf{y'\,y} - \mathbf{b'\,X'\,y}}{\mathbf{y'\,y} - N\overline{y}^2} = \frac{\mathbf{b'\,X'\,y} - N\overline{y}^2}{\mathbf{y'\,y} - N\overline{y}^2} \qquad (3.42)$$

Ejemplo 8. *Obtener el coeficiente de determinación para los ajustes realizados en el Ejemplo 7.*

a) Para el caso de la recta:

$$S_e^2 = \frac{\sum_i y_i^2 - a\sum_i y_i - b\sum_i x_i y_i}{N} = \frac{230,75 - 4,828\,(46,5) + 0,0247\,(331)}{10} = 1,44237$$

$$S_y^2 = 1,452499$$

$$R^2 = 1 - \frac{S_e^2}{S_y^2} = 0,00697$$

$y^* = 4,828 - 0,0247x;$ $\qquad R^2 = 0,00697$

b) Para el caso de la parábola el resultado es

$$R^2 = 1 - \frac{S_e^2}{S_y^2} = 0,829$$

$y^* = 8,616 - 1,22x + 0,07177x^2;$ $\qquad R^2 = 0,829$

Como se puede comprobar el modelo lineal es una mala elección, pues no llega a explicar ni siquiera el 1% de las variaciones de la variable dependiente. En cambio la parábola explica más del 82%, lo que nos permite calificarlo como un modelo bastante adecuado. Es decir, en este caso la bondad del ajuste es bastante buena, aunque mejorable, pues hay más de un 17% de variaciones no explicadas por el modelo.

A continuación vamos a mostrar una relación muy interesante que existe entre el coeficiente de determinación y el de correlación para el caso lineal. En esta situación se cumple que el coeficiente de determinación es igual al de correlación al cuadrado, es decir, $R^2 = r^2$. Para comprobar que esto es cierto debemos recordar que:

$$R^2 = \frac{S_{y^*}^2}{S_y^2} = \frac{\dfrac{\sum_i \left(y_i^* - \bar{y}\right)^2}{N}}{S_y^2} = \frac{\dfrac{\sum_i \left(a + bx_i - a - b\bar{x}\right)^2}{N}}{S_y^2} = \frac{\dfrac{b^2 \sum_i \left(x_i - \bar{x}\right)^2}{N}}{S_y^2} = \frac{b^2 S_x^2}{S_y^2} = \frac{S_{xy}^2 S_x^2}{\left(S_x^2\right)^2 S_y^2} =$$

$$= \frac{S_{xy}^2}{S_x^2 S_y^2} = r^2. \qquad (3.43)$$

Ejemplo 9. *Para el modelo de regresión lineal $y_i = f(x_i) + e_i = y_i^* + e_i = a + bx_i + e_i$, demuestre que:*

a) $e_i = y_i - \bar{y} - b(x_i - \bar{x})$

b) $\sum_i e_i = 0$

c) $\sum_i e_i^2 = \sum_i \left(y_i - \bar{y}\right)^2 - b^2 \sum_i \left(x_i - \bar{x}\right)^2$

d) $y_i^* - \bar{y} = b\left(x_i - \bar{x}\right)$

e) $\sum_i e_i \left(x_i - \bar{x}\right) = 0$

Demostración:

a) $e_i = y_i - y_i^* = y_i - \left(a + bx_i\right) = y_i - \left(\bar{y} - b\bar{x} + bx_i\right) = y_i - \bar{y} - b\left(x_i - \bar{x}\right)$

b) $\sum_i e_i = \sum_i y_i - N\bar{y} - b\sum_i \left(x_i - \bar{x}\right) = \sum_i y_i - N\bar{y} = \sum_i y_i - \sum_i y_i = 0$

c) A partir del resultado obtenido en (3.38) y de uno de los pasos intermedios de (3.43) queda demostrada la igualdad planteada en este apartado.

d) $y_i^* - \bar{y} = a + bx_i - \bar{y} = \bar{y} - b\bar{x} + bx_i - \bar{y} = b\left(x_i - \bar{x}\right)$

e) Este resultado es una consecuencia inmediata de (3.16) y (3.17), pues la suma dada en este apartado se puede poner como:

$$\sum_i e_i \left(x_i - \bar{x}\right) = \sum_i e_i x_i - \bar{x}\sum_i e_i = 0$$

Ejemplo 10. *Analice la veracidad de los siguientes resultados cuando se trabaja con un modelo de regresión lineal de dos variables.*

a) $r_{xy}= 0,5 \quad b=0,7 \quad R^2=0,9$

b) $b=0,7 \quad S_{xy} = -10$

c) $S_y^2 = 15 \quad S^2_{y^*} = 20$

d) $S_y^2 = 15 \quad S^2_{y^*} = 10 \quad S_e^2 = 5$

a) Estos resultados no pueden ser ciertos pues, aunque el signo del coeficiente de correlación lineal y el de la pendiente de la recta coinciden, dado que tienen el mismo numerador (la covariancia) y sus denominadores serán siempre no negativos, sin embargo no se cumple la relación que existe entre el coeficiente de correlación y el de determinación, que, como se ha demostrado, es: $R^2 = r^2$.

b) Este resultado tampoco es posible. El signo de b y el de S_{xy} *debe ser siempre el mismo.*

c) Teniendo en cuenta que en un modelo de regresión lineal de dos variables se cumple (3.36) y que la variancia no puede ser negativa, resulta que este resultado es imposible. Además la variancia explicada por el modelo nunca podrá ser mayor que la total.

d) Este resultado si es cierto, pues satisface la relación (3.36).

Ejemplo 11. *A veces, la relación entre las variables es de tal naturaleza que la línea de regresión debe pasar por el origen. En tales casos lo que se hace es imponer que la ordenada en el origen sea nula. Esto es equivale a introducir una restricción al modelo lineal de dos variables. Cuando se procede de esta forma algunos de los resultados obtenidos anteriormente dejan ser válidos. Veamos a continuación como afecta esta restricción a los resultados obtenidos en epígrafes anteriores.*

Para empezar hay que señalar que nuestro modelo es ahora el siguiente:

$$y_i = y^*_i + e_i = bx_i + e_i.$$

Como en este modelo solo hay un parámetro desconocido, entonces la aplicación del método de los mínimos cuadrados nos llevará a una sola ecuación normal

$$\boldsymbol{j}(b) = \sum_i e_i^2 = \sum_i \left(y_i - y_i^*\right)^2 = \sum_i \left(y_i - bx_i\right)^2$$

$$\frac{\partial \boldsymbol{j}(b)}{\partial b} = -2\sum_i \left(y_i - bx_i\right)x_i = 0$$

De estos resultados se deduce que ahora:

$$b = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

Además, como solo se trabaja con una ecuación normal, la segunda, resulta que la primera ni siquiera se cumple. Es decir, la suma de los errores ya no es cero, como comprobaremos a continuación.

$$\sum_i e_i = \sum_i \left(y_i - y^*\right) = \sum_i y_i - b\sum_i x_i$$

Pero para que esta expresión fuera cero es necesario que se cumpla que

$$b = \frac{\sum_i y_i}{\sum_i x_i}$$

lo cual no es cierto, como se ha visto en líneas anteriores. Así pues, la suma de los errores en este tipo de modelo lineal que estamos analizando no es cero. Una consecuencia de este resultado es que la media de los errores no es nula y además la media de la variable dependiente no será igual a la media de la línea de regresión. Es decir:

$$\bar{e} \neq 0 \qquad\qquad \overline{\mathrm{y}} \neq \overline{y^*}$$

Además, todo lo anterior lleva a que (3.36) tampoco se cumpla. En este caso solo es cierto que:

$$\sum_i y_i^2 = \sum_i y_i^{*2} + \sum_i e_i^2$$

Para comprobar que esto es cierto vamos a desarrollar el último sumando.

$$\sum_i e_i^2 = \sum_i \left(y_i - y_i^*\right)^2 = \sum_i y_i^2 + \sum_i y_i^{*2} - 2\sum_1 y_i y_i^* = \sum_i y_i^2 + b^2 \sum_i x_i^2 - 2b\sum_1 y_i x_i =$$
$$= \sum_i y_i^2 + b^2 \sum_i x_i^2 - 2b^2 \sum_1 x_i^2 = \sum_i y_i^2 - b^2 \sum_i x_i^2 = \sum_i y_i^2 - \sum_i \left(bx_i\right)^2 =$$
$$= \sum_i y_i^2 - \sum_i y_i^{*2}$$

Para llegar a este resultado se ha hecho uso de que $y^*_i = bx_i$ *y de que* $b = \dfrac{\sum_i y_i x_i}{\sum_i x_i^2}$.

Además, como (3.36) no se cumple, ahora no es aplicable la definición del coeficiente de determinación dada en (3.41) ni su relación con el coeficiente de correlación lineal que se mostró en (3.43).

3.4.4. Cambios de origen y de escala.

En este apartado se analizará como afectan los cambios de origen y de escala sobre las variables dependiente e independiente a los distintos coeficientes que se han definido y estudiado en este capítulo.

1º Efectos de cambios de origen y de escala en los parámetros de la línea de regresión, es decir, en la ordenada en el origen y en la pendiente.

a) Si se realiza un cambio de origen en la variable X, la ordenada el origen se ve afectada por ese cambio, pero no la pendiente.

Si se define $X' = X + k$, entonces se tiene que: $X = X' - k$, por lo que el modelo inicial $y_i^* = a + bx_i$ se transforma en: $y_i^* = a + bx_i = a + b\left(x' - k\right) = a - bk + bx' = a' - bx'$

b) Si se realiza un cambio de escala en la variable *X*, cambia la pendiente pero no la ordenada el origen.

Si se define *X' = kX*, entonces se tiene que: *X = X'/k*, por lo que el modelo inicial $y_i^* = a + bx_i$ se transforma en $y_i^* = a + bx_i = a + b(x'/k) = a + \left(\frac{b}{k}\right)x' = a - b'x'$

c) Si se realiza un cambio de origen en la variable *Y*, la ordenada el origen se ve afectada por ese cambio, pero no la pendiente.

Si se define *Y' = Y + k*, entonces el modelo inicial $y_i = y_i^* + e_i = a + bx_i + e_i$ se transforma en: $y_i' = y_i + k = a + bx_i + e_i + k = a' + bx_i + e_i$

d) Si se realiza un cambio de escala en la variable *Y*, cambia la pendiente y la ordenada el origen.

Si se define *Y' = kY*, entonces el modelo inicial $y_i = y_i^* + e_i = a + bx_i + e_i$ se transforma en: $y_i' = ky_i^* + ke_i = ka + kbx_i + ke_i = a' + b'x_i + e_i'$

Los resultados obtenidos anteriormente se podrían haber alcanzado también si se hubiera trabajado directamente con las expresiones de cálculo de la pendiente (3.20) y la ordenada en el origen (3.21).

2º Efectos de cambios de origen y de escala en las variables *X* e *Y* sobre los coeficientes de correlación y de determinación.

Estos coeficientes son invariantes frente a cambios de origen y de escala. Esta afirmación no es necesario comprobarla, pues para el coeficiente de correlación ya se hizo en el apartado 5.4, y para el de determinación resulta innecesario, dada la relación existente entre ambos.

Ejemplo 12. *Obtenga el valor de los coeficientes del ejemplo 6 si el consumo y la renta se expresaran en pesetas.*

En este caso se trata de un cambio de escala realizado simultáneamente sobre X e Y. Sabemos que tanto X como Y vienen expresadas en €. Para expresarlas en pesetas tendríamos que realizar los siguientes cambios: X' = kX. A su vez Y' = kY. Tanto en un caso como en otro sabemos que k = 166,386. Todo ello nos lleva a que de la relación inicial en euros dada por:

$$y_i = a + bx_i + e_i$$

se pase a la siguiente:

$$\frac{y_i'}{k} = a + b\frac{x_i'}{k} + e_i$$

de forma tal que la relación expresada en pesetas quedaría como:

$$y_i' = ka + bx_i' + ke_i$$

es decir, la propensión marginal al consumo no cambia, mientras que el consumo autónomo si que cambia.

3.4.5.- Predicción.

Adelantarse al futuro es y ha sido siempre un continuo deseo para el ser humano, pero a la vez inalcanzable. Muchas han sido las técnicas puestas al servicio de esa empresa. Pero ninguna de ellas consigue unos resultados aceptables sin un mínimo de "arte", porque la predicción es, en buena medida, un arte, incluso aunque las técnicas que se utilicen para la realización de esas predicciones sean, en términos estadísticos, robustas y potentes.

Como ya se ha indicado, el análisis de la regresión es una de esas técnicas que sirve para describir el comportamiento conjunto de dos variables y también para realizar predicciones. Pero lo que se predice con una línea de regresión ajustada son los valores medios de la variable dependiente, pues la componente errática de la dependencia

estadística no es predecible. Es, como se ha señalado en otro lugar, un símbolo de la ignorancia residual del estadístico.

Aún teniendo en cuenta estas limitaciones, los resultados de estos ejercicios de prospectiva deben tomarse con mucha cautela y la validez de los mismos estará sujeta a una serie de requisitos previos que han de tenerse en cuenta.

Las predicciones que se realizan a partir de la línea de regresión están condicionadas a los valores de la variable independiente. Esto hace que las mismas puedan agruparse en dos categorías distintas. Por un lado están las interpolaciones y por otro las extrapolaciones. Las primeras se corresponden con valores de la variable independiente pertenecientes al recorrido de valores observados de esta variable. En cambio, las segundas son las que se realizan cuando a X se le asignan valores fuera de su recorrido observado.

Para que la validez del primer tipo de predicciones sea aceptable será requisito necesario que el ajuste realizado sea bueno. Esto se puede medir en términos del coeficiente de determinación. Solo se debería depositar confianza en esos resultados cuando el valor de R^2 sea suficientemente alto.

Por lo que se refiere al segundo tipo de predicciones, la validez de las mismas estará condicionada a que el ajuste sea bueno (R^2 alto) y a que la relación cuantificada entre las variables X e Y en el recorrido de los valores observados de las mismas se mantenga incluso para aquellos que se alejen. Esta segunda condición es fundamental, pues la bondad del ajuste no es en absoluto suficiente para realizar pronósticos cuando nos alejamos de los valores observados.

Ejemplo 13. *A lo largo de 26 días se han observado los precios de venta (x) y las cantidades demandas (Y) de cierto producto. Los resultados de esas observaciones son los siguientes:*

Precio	*Cantidad*	*Precio*	*Cantidad*
3,5	*20*	*6,3*	*17*
3,8	*19*	*6,5*	*16,5*
4	*18*	*7,3*	*16*
4,9	*19*	*7,2*	*16*
4,5	*19*	*7*	*16,8*
5	*18*	*7,1*	*16,5*
5,2	*19*	*8*	*16*
5	*18*	*8,6*	*16*
5,3	*18*	*8,7*	*15,3*
5,7	*18*	*8*	*15,2*
6	*17*	*9*	*15,5*
6,2	*17*	*9,1*	*15*
6,7	*16*	*9*	*14,5*

Con esta información:

a) *Ajustar una función de demanda lineal.*
b) *Interpretar el significado de los coeficientes.*
c) *Obtener la elasticidad demanda-precio media.*
d) *Predecir la cantidades demandadas cuando los precios son 5,5 y 20.*

Ejemplo 14. *En unos grandes almacenes se ha observado que las compras de los clientes (expresadas en euros) de productos de marca blanca dependen de forma lineal del total de compras de los mismos. Con la información de 200 compras realizadas en un día se obtuvieron los siguientes resultados:*

b = 0,1023; r = 0,9886; Media de X = 181,5 €; Media de Y = 18,2 €; V(Y) = 95,76 €^{2}

1) *Obtener la recta de regresión.*
2) *Dar una interpretación del significado de los coeficientes obtenidos.*
3) *Estudiar la bondad del ajuste y cuantificar la variancia de la variable de dependiente no explicada por el modelo.*
4) *Obtener la elasticidad media.*

5) *Si admitimos que la relación ajustada se mantiene para cualquier volumen de compras, determine las ventas de productos de marca blanca que cabría esperar en un cliente que realiza una compra por valor de 400€.*

Ejemplo 15. *La dirección de un restaurante ha observado que el número de botellas (Y) de vino gran reserva que se sirven en una noche depende linealmente del gasto medio por persona expresado en euros (X). Para ver hasta que punto es cierto que hay ese tipo de relación se anotaron a lo largo de diez semanas el número de botellas diarias vendidas en las cenas así como el coste medio por persona de esas cenas. Los resultados de esos 70 días, de forma resumida, son los siguientes:*

Media de X = 50 ; Media de Y = 10; V(Y) = 20,25; V(X) = 325; COV(X,Y) = 80

Con estos datos:

a) *Hallar la recta de regresión.*
b) *Analizar la bondad del ajuste.*
c) *¿Cuál sería la demanda esperada de botellas si el coste medio por persona de la cena en una noche se eleva a 70 euros?*

Ejemplo 16. *En unos grandes almacenes se ha realizado una campaña publicitaria orientada a conseguir una mayor demanda por parte de sus clientes de productos de marcas blancas. Finalizada esta campaña, se observó, durante veinte días consecutivos, el volumen de comparas, medido en euros, que realizaron sus clientes de ese tipo de productos. Los resultados de esta observación son los que se dan en la tabla adjunta. Con estos datos realice un análisis del el impacto de la campaña publicitaria sobre la venta de productos de marca blanca, seleccionando para ello el modelo que considere más adecuado, estudiando la bondad del ajuste del mismo y valorando la capacidad predictiva de ese modelo.*

Días	Ventas
1	12000
2	11000
3	10500
4	9800
5	8000
6	8100
7	8000
8	7500
9	7200
10	6800
11	7000
12	7100
13	6800
14	6500
15	6200
16	6500
17	6100
18	6000
19	6200
20	6100