Estadística Descriptiva y Probabilidad

(Teoría y problemas)

3ª Edición

Autores
I. Espejo Miranda
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
A. M. Rodríguez Chía
A. Sánchez Navas
C. Valero Franco

UCA Universidad de Cádiz

Servicio de Publicaciones

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

http://www.uca.es/publicaciones

ISBN: 978-84-9828-058-6

Depósito legal:

Estadística Descriptiva y Probabilidad. Teoría y Problemas (Revisión: Febrero 2006)
I. Espejo Miranda, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, A. M. Rodríguez Chía, A. Sánchez Navas, C Valero Franco

Capítulo 3

Ajuste y regresión bidimensional

1. Introducción

Considerada una serie estadística $(x_1, y_1), \ldots, (x_n, y_n)$, procedente de una distribución (X, Y), el problema que se plantea en este capítulo consiste en encontrar alguna relación que exprese los valores de una variable en función de los de la otra. Para hacer esto, y una vez establecida cual será la variable dependiente, se tienen dos opciones: prefijar una clase funcional[1], o estimar un valor de la variable dependiente para cada valor de la variable independiente. En el primer caso, se está realizando un ajuste, y en el segundo, se tiene un problema de regresión. La regresión, viene determinada, por tanto, por un conjunto de puntos, de tal forma que uniendo los puntos contiguos por segmentos rectilíneos se obtiene la poligonal de regresión. La regresión sólo tendrá sentido cuando en la serie bidimensional haya muchos valores de la variable dependiente para cada uno de la independiente, pues en caso contrario, es casi idéntica al diagrama de dispersión y no aporta nada nuevo. Ambas técnicas, ajuste y regresión, son complementarias, siendo la poligonal de regresión una buena referencia de la clase funcional a elegir para el ajuste, a la vez que, como se verá a lo largo del capítulo, fija el techo para la bondad de éste.

[1] Por ejemplo una recta, una parábola, una función exponencial, etc....

En el caso del ajuste, la cuestión será elegir la mejor clase funcional y determinar los parámetros que identifiquen la función dentro de la clase. Esto se consigue imponiendo a dicha función que verifique condiciones de adaptación a la nube de puntos, para ello, se empleará el criterio de los mínimos cuadrados, que se desarrollará en el punto siguiente.

Los puntos de la regresión se obtienen utilizando distribuciones condicionadas. Se empleará el método de la regresión a la media, que consiste en asociar a cada valor de la variable independiente, el valor medio de la distribución de la variable dependiente condicionada a cada uno de dichos valores.

El objeto del ajuste es doble. Por una parte interpolador: puesto que se está trabajando con un cierto número de observaciones, es de esperar que si éstas son representativas del fenómeno en estudio, los elementos del colectivo se comporten de forma parecida y la función de ajuste sea también válida para ellos. Y por otra parte extrapolador: bajo la suposición de que la relación funcional entre variable dependiente e independiente permanece constante, al menos en un entorno de las observaciones, es posible hacer una previsión del valor que tomará la variable dependiente para un valor determinado de la independiente en dicho entorno. Esta característica puede ser utilizada, por ejemplo, para hacer previsiones de ventas a corto o medio plazo, estimar el volumen de cosecha en función de la lluvia caída, etc...

La elección de la familia de funciones sobre la que se hará el ajuste es uno de los problemas principales a los que se deberá hacer frente. En un principio, la observación de la nube de puntos puede dar una idea de la evolución de los valores de la variable dependiente (a partir de ahora Y) en función de los de la independiente (X). Como ya se comentó arriba, en algunos casos, puede ser de mucha utilidad el construir la poligonal de regresión.

El tema se ha dividido en dos partes, la primera dedicada al ajuste y la segunda a la regresión.

2. Ajuste. Criterio de los mínimos cuadrados

Fijada la familia de funciones que se utilizará para ajustar los valores de una serie estadística bidimensional, ésta dependerá de unos parámetros. El método que se usará para la estimación de dichos parámetros es el de los *mínimos cuadrados*, que consiste en hacer mínima la suma de las diferencias al cuadrado entre los valores observados y los correspondientes valores ajustados.

Formalmente, sean $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, los valores observados y $g(x, \alpha, \beta, \cdots, \theta)$ la función de ajuste. Los valores de los parámetros se obtienen imponiendo la condición de hacer mínima la función H, donde

$$H(\alpha, \beta, \cdots, \theta) = \sum_{i=1}^{n} [y_i - f(x_i, \alpha, \beta, \cdots, \theta)]^2.$$

Para ello, se calculan las derivadas parciales de H respecto de cada uno de los parámetros y se igualan a cero:

$$\frac{\partial H}{\partial \alpha} = 0, \frac{\partial H}{\partial \beta} = 0, \cdots, \frac{\partial H}{\partial \theta} = 0.$$

Con esto se genera un sistema de tantas ecuaciones como parámetros, llamado sistema de ecuaciones normales, que resuelto da los valores de los parámetros. A continuación, se hace un estudio más detallado para algunas funciones de uso habitual.

2.1. Caso lineal

Sean $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ los valores observados[2] y sea $f(x, a, b) = a + bx$ la recta de ajuste de los valores de Y en función de los de X. Se obtienen los valores a y b minimizando la función error cuadrático, H, dada por

$$H(a, b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \ .$$

[2]Obsérvese que no se dan las observaciones con sus respectivas frecuencias, sino que se hace por extensión para facilitar la nomenclatura.

Derivando respecto a los dos parámetros

$$\begin{aligned}\frac{\partial H(a,b)}{\partial a} &= -2\sum_{i=1}^{n}[y_i-(a+bx_i)]\\ \frac{\partial H(a,b)}{\partial b} &= -2\sum_{i=1}^{n}[y_i-(a+bx_i)]x_i\end{aligned}$$

e igualando a cero, queda el siguiente sistema de ecuaciones, que se conoce como sistema de *ecuaciones normales del modelo*:

$$\begin{aligned}\sum_{i=1}^{n} y_i &= na+b\sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n} y_i x_i &= a\sum_{i=1}^{n}x_i+b\sum_{i=1}^{n}x_i^2.\end{aligned}$$

Utilizando la notación

$$\begin{aligned}y_i^* &= f(x_i)=a+bx_i\\ e_i &= y_i-y_i^*\end{aligned}\tag{3.1}$$

el sistema de ecuaciones normales puede expresarse de la forma

$$\begin{aligned}\sum_{i=1}^{n} e_i &= 0\\ \sum_{i=1}^{n} e_i x_i &= 0\end{aligned}\tag{3.2}$$

donde e_i representa el residuo o error de la observación i-ésima.

Admitiendo que se verifican las condiciones suficientes de mínimo se pueden obtener fácilmente los valores de a y b:

$$\begin{aligned}a &= \bar{y}-b\bar{x}\\ b &= \frac{S_{xy}}{S_x^2}.\end{aligned}$$

Si se quiere obtener la línea de ajuste de X respecto a Y, llamando ahora a' y b' a los coeficientes de la recta, se obtiene

$$\begin{aligned} a' &= \bar{x} - b'\bar{y} \\ b' &= \frac{S_{xy}}{S_y^2}. \end{aligned}$$

Los valores b y b' que son las pendientes de las rectas de ajuste, reciben el nombre de *coeficientes de regresión* y representan los incrementos de las variables dependientes para aumentos unitarios de las independientes.

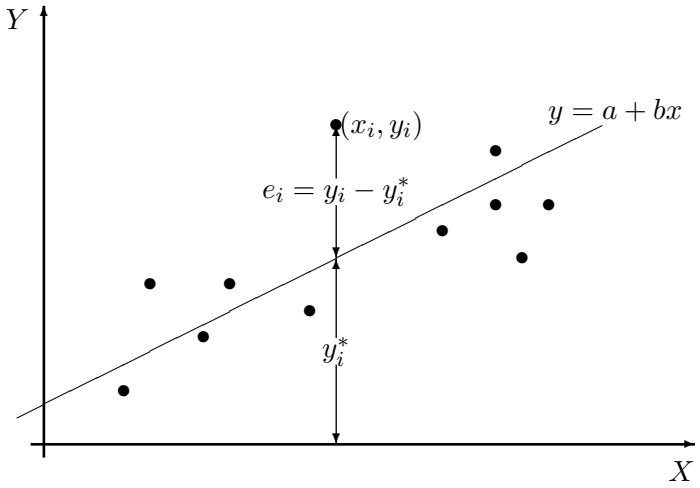

Figura 3.1: Criterio de los mínimos cuadrados

Ejemplo 3.1 *Dada la distribución bidimensional*

X	1	2	3	4	5	6
Y	2	5	9	13	17	21

Los coeficientes de la recta de ajuste de Y en función de X son

$$\begin{aligned} b &= \frac{S_{xy}}{S_x^2} = \frac{11'25}{2'91} = 3'87 \\ a &= \bar{y} - b\bar{x} = 11'166 - 3'87 \cdot 3'5 = -2'36 \end{aligned}$$

Y los coeficientes de la recta de X en función de Y son

$$\begin{aligned} b' &= \frac{S_{xy}}{S_y^2} = \frac{11'25}{43'46} = 0'26 \\ a' &= \bar{x} - b'\bar{y} = 3'5 - 0'26 \cdot 11'166 = 0'61. \end{aligned}$$

Es decir, cuando X aumenta en una unidad Y lo hace en 3'87 unidades, mientras que cuando se incrementa Y en una unidad X crece 0'26 unidades.

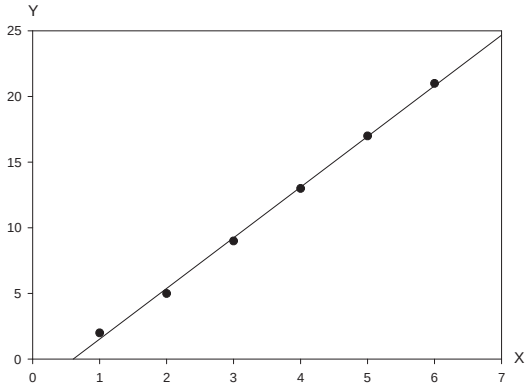

Figura 3.2: Recta de ajuste del ejemplo 3.1

Se han definido dos nuevas variables, por una parte Y^*, que representa los valores ajustados de la variable Y y que tiene por media y varianza:

$$\begin{aligned} \bar{y}^* &= \bar{y} \\ S_{y^*}^2 &= bS_{xy} \end{aligned}$$

y por otra parte, la variable residuo e, con media y varianza:

$$\begin{aligned} \bar{e} &= 0 \\ S_e^2 &= \frac{\sum_{i=1}^{n} e_i^2}{n} = \frac{\sum_{i=1}^{n} (y_i - y_i^*)^2}{n}. \end{aligned} \tag{3.3}$$

S_e^2 recibe el nombre de varianza residual, y puede expresarse también de la forma

$$S_e^2 = \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i}{n} .$$

Estas dos variables están incorreladas. En efecto, multiplicando la primera expresión de (3.2) por a y la segunda por b, se tiene

$$0 = a\sum_{i=1}^{n} e_i + b\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} e_i(a + bx_i) = \sum_{i=1}^{n} e_i y_i^*$$

y puesto que $\bar{e} = 0$, resulta

$$S_{ey^*} = \frac{\sum_{i=1}^{n} e_i y_i^*}{n} - \bar{e}\bar{y}^* = 0 \tag{3.4}$$

como se quería demostrar.

Ejercicio 3.1 *Demuestre las siguientes propiedades:*

a) *Las rectas de regresión de X sobre Y y de Y sobre X se cortan en el centro de gravedad de la distribución.*

b) *b, b', r y S_{xy} tienen siempre el mismo signo.*

c) *Las dos rectas de regresión coinciden sólo cuando $r = 1$ ó $r = -1$.*

d) *Se pueden expresar las rectas de regresión de Y sobre X y de X sobre Y, respectivamente, como:*

$$\begin{aligned} y - \bar{y} &= \frac{S_{xy}}{S_x^2}(x - \bar{x}) \\ x - \bar{x} &= \frac{S_{xy}}{S_y^2}(y - \bar{y}). \end{aligned}$$

e) *El coeficiente de correlación lineal puede obtenerse como:*

$$r = \pm\sqrt{bb'} .$$

Todo lo que se ha dicho hasta ahora es generalizable a funciones linealizables, sin más que hacer los cambios y transformaciones pertinentes, como los que se muestran en la tabla 3.1.

Función	Transformación	Cambio
$y = ax^b$	$\ln y = \ln a + b \ln x$	$\begin{cases} y' = \ln y \\ x' = \ln x \end{cases}$
$y = ab^x$	$\ln y = \ln a + x \ln b$	$\begin{cases} y' = \ln y \\ x' = x \end{cases}$
$y = a + \frac{b}{x}$	$y = a + \frac{b}{x}$	$\begin{cases} y' = y \\ x' = \frac{1}{x} \end{cases}$

Tabla 3.1: Linealización de funciones

2.2. Caso parabólico

Se considera la función de ajuste $f(x, a, b, c) = a+bx+cx^2$, parábola de segundo grado. Para obtener los valores de a, b y c se utiliza el método de los mínimos cuadrados y se minimiza la función H definida por

$$H(a, b, c) = \sum_{i=1}^{n} [y_i - (a + bx_i + cx_i^2)]^2 \ .$$

Para ello se calculan las derivadas parciales de H respecto de a, b y c y se igualan a cero:

$$\begin{aligned} \frac{\partial H}{\partial a} &= -2 \sum_{i=1}^{n} [y_i - (a + bx_i + cx_i^2)] = 0 \\ \frac{\partial H}{\partial b} &= -2 \sum_{i=1}^{n} [y_i - (a + bx_i + cx_i^2)] x_i = 0 \\ \frac{\partial H}{\partial c} &= -2 \sum_{i=1}^{n} [y_i - (a + bx_i + cx_i^2)] x_i^2 = 0. \end{aligned}$$

Despejando los términos independientes, se obtiene el sistema de ecuaciones normales:

$$\sum_{i=1}^{n} y_i = an + b \sum_{i=1}^{n} x_i + c \sum_{i=1}^{n} x_i^2$$

$$\begin{aligned}\sum_{i=1}^{n} y_i x_i &= a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} y_i x_i^2 &= a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4.\end{aligned}$$

De igual forma si se definen $y^* = a + bx_i + cx_i^2$ y $e_i = y_i - y_i^*$ el sistema de ecuaciones normales se puede expresar como

$$\begin{aligned}\sum_{i=1}^{n} e_i &= 0 \\ \sum_{i=1}^{n} e_i x_i &= 0 \\ \sum_{i=1}^{n} e_i x_i^2 &= 0.\end{aligned}$$

La varianza residual vale ahora

$$S_e^2 = \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i - c\sum_{i=1}^{n} x_i^2 y_i}{n}\ .$$

3. Análisis de la bondad del ajuste

Una vez obtenidos los valores de los parámetros, y por tanto la función de ajuste, se va a dar la medida del grado de aproximación entre los valores observados y los ajustados. Según (3.3), la varianza del error o varianza residual vale

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

que coincide con el error cuadrático medio (E.C.M.). Al venir dada dicha varianza por una suma de términos al cuadrado será siempre positiva,

salvo que todos los términos sean nulos, en cuyo caso el E.C.M. valdrá cero. Los términos serán nulos sólo si coinciden los valores observados con los ajustados, con lo cual, el ajuste será perfecto sólo si la varianza residual vale cero. Si se ajustan los mismos datos a dos modelos diferentes, el más afortunado será aquel que tenga un E.C.M. menor.

A continuación se relaciona la varianza residual, la varianza de Y y la varianza de Y^*. De la relación (3.1) se deduce que

$$y_i = y_i^* + e_i \ ,$$

además por (3.4) se tiene que y^* y e son incorreladas, es decir, la varianza de la suma es la suma de las varianzas, con lo cual se tiene que

$$S_y^2 = S_{y^*}^2 + S_e^2 \ .$$

Por tanto, la varianza de Y es igual a la varianza de los valores ajustados más la varianza residual. Este importante resultado se puede expresar también de la forma

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

sin más que multiplicar los dos miembros de la igualdad por n.

Al ser las varianzas números positivos las relaciones siguientes son evidentes:

$$\begin{aligned} 0 \leq \quad & S_e^2 \quad \leq S_y^2 \\ 0 \leq \quad & S_{y^*}^2 \quad \leq S_y^2 \end{aligned}$$

Si $S_e^2 = 0$, entonces $y_i = y_i^*$ para todo i, con lo que el modelo propuesto explica perfectamente las variaciones de la variable Y, siendo inmejorable. Si, por el contrario, $S_{y^*}^2 = 0$, $y_i^* = \bar{y}$ para todo i, con lo que el modelo de ajuste es constante y no explica, en forma alguna, las variaciones de Y. Estas son las situaciones extremas, pero el caso general se caracteriza por ser $S_e^2 > 0$ y $S_{y^*}^2 > 0$.

Evidentemente, cuanto más próximo S_e^2 esté de cero mejor será el ajuste, aunque, al ser un valor que viene acompañado de la unidad de

la variable, no se puede dar una medida exacta de la bondad de éste ni hacer comparaciones entre dos distribuciones que vengan dadas en unidades distintas. Se hace necesario, pues, construir un coeficiente abstracto, de modo que se pueda expresar la bondad del ajuste en forma de porcentaje. Con esa intención se redefine[3] el *coeficiente de determinación* de la siguiente forma:

$$R^2 = \frac{S^2_{y^*}}{S^2_y} = \frac{S^2_y - S^2_e}{S^2_y} = 1 - \frac{S^2_e}{S^2_y} \ .$$

Al ser $S^2_{y^*} < S^2_y$, R^2 varía entre 0 y 1, multiplicado por 100 expresa el porcentaje de la variabilidad de Y que queda explicado por el ajuste. Cuando $S^2_{y^*}$ vale 0, R^2 vale 0, y el ajuste explica el 0 % de la variabilidad de Y. Cuando S^2_e vale 0, R^2 vale 1 y el ajuste explica el 100 % de la variabilidad de Y. En general, y teniendo en cuenta que

$$1 = \frac{S^2_{y^*} + S^2_e}{S^2_y} = \frac{S^2_{y^*}}{S^2_y} + \frac{S^2_e}{S^2_y},$$

se puede decir que el $100\frac{S^2_{y^*}}{S^2_y}$ de la variabilidad de Y queda explicada por el ajuste, y el resto, es decir el $100\frac{S^2_e}{S^2_y}$, no se explica por éste.

Ejemplo 3.2 *El coeficiente de correlación lineal al cuadrado para la distribución del ejemplo 3.1, vienen dados por:*

$$R^2 = \frac{S^2_{xy}}{S^2_x S^2_y} = \frac{126'56}{2'914 \cdot 43'468} = 0'9991 \ .$$

Por lo que el 99'91 % de la variabilidad de Y queda explicada por X a través de la recta de ajuste y el resto, es decir el 0'09 % restante se explicaría bien por otra función de ajuste mejor o bien por otras variables distintas[4] *a X. Por último, puesto*

[3] Para el caso lineal ya se había introducido en el capítulo anterior.

[4] Observe que si para un mismo valor de X se tuvieran dos valores distintos de Y, X no podría explicar esa variabilidad, y de hecho ninguna función de X podría pasar por los dos puntos.

que los coeficientes de regresión son positivos, se tendrá que

$$r = +\sqrt{R^2} = 0'9995 \ .$$

Resultado. Cuando el ajuste realizado es lineal, R^2 coincide con el cuadrado del coeficiente de correlación lineal, es decir:

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = r^2 \ .$$

4. Regresión. Método de regresión a la media

Este método consiste en definir la línea de regresión como la poligonal que pasa por la media aritmética de los puntos de igual abscisa, (curva de regresión de Y respecto a X), ó de igual ordenada, (X respecto a Y). Si sólo hubiera un valor de Y para cada valor de X, ó un valor de X para cada uno de Y, la correspondiente curva de regresión sería la poligonal que uniera todos los puntos. La regresión tiene interés, por tanto, sólo cuando hay un gran número de observaciones y a cada valor de una de las variables le corresponden muchos valores de la otra. El nombre de regresión tiene su origen en los primeros estudios que relacionaban las estaturas de grupos de padres e hijos; se observó entonces que, en general, padres de pequeña estatura tenían hijos bajos pero no tanto como ellos, y padres de talla elevada tenían hijos altos pero no tanto como ellos, produciéndose una tendencia o "regresión" hacia los valores intermedios.

Se llama *curva de regresión* de Y respecto de X, a la función que asocia a cada x_i de X, la media condicionada de Y respecto de x_i. Es decir:

$$\phi(x_i) = \bar{y}|_{X=x_i} = \bar{y}_i$$

Ejemplo 3.3 *La regresión a la media de la siguiente distribución:*

X/Y	1	2	3	4	5	6	7	8	9
1	2	4	3	7	1	8	9	3	2
2	3	6	1	8	5	9	6	4	3
3	0	0	3	5	7	8	5	4	2
4	0	0	0	2	4	6	7	4	3
5	0	0	0	0	5	8	7	4	4
6	0	0	0	2	8	9	9	6	6
7	0	0	0	0	0	0	2	3	3

viene dada por

X	1	2	3	4	5	6	7
$\phi(x)$	$5'25$	$5'11$	$5'79$	$6'61$	$6'85$	$6'67$	$8'12$

Gráficamente la curva de regresión sería la que se muestra en la figura 3.3.

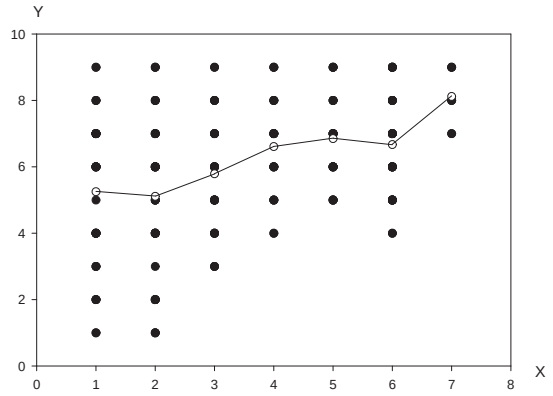

Figura 3.3: Poligonal de regresión

5. Análisis de la bondad de la regresión

A continuación se dará una medida de la proximidad de la curva de regresión a la distribución. Se considera el error cuadrático medio, que en este caso viene dado por

$$E.C.M. = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}[y_{ij} - \bar{y}_i]^2 n_{ij}}{n} = \sum_{i=1}^{r} V_i[Y]\, f_{i.} \ ,$$

donde

$$V_i[Y] = \frac{\sum_{j=1}^{s}[y_{ij} - \bar{y}_i]^2 n_{ij}}{n_{i.}}$$

es la varianza de Y condicionada a $X = x_i$.

El E.C.M. está medido en la misma unidad que la variable con los consiguientes problemas. Se hace necesario, pues, dar una medida adimensional de la bondad de la regresión. La varianza de Y se puede expresar como

$$V[Y] = \overbrace{\sum_{i=1}^{n} V_i[Y]\, f_{i.}}^{I} + \overbrace{\sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2\, f_{i.}}^{II} \ ,$$

quedando descompuesta en la suma de la media ponderada de las varianzas condicionadas (I), más la varianza ponderada de las medias condicionadas (II). Es decir, la heterogeneidad de Y, resulta de la heterogeneidad debida a las distribuciones condicionadas por cada modalidad x_i, más la heterogeneidad existente entre las distintas modalidades.

Se define la *razón de correlación* de Y con respecto a X como:

$$\eta^2_{y,x} \quad = \quad \frac{\sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2\, f_{i.}}{V[Y]}$$

$$= \frac{V[Y] - \sum_{i=1}^{n} V_i(Y)\, f_{i.}}{V[Y]}$$

$$= 1 - \frac{\sum_{i=1}^{n} V_i[Y]\, f_{i.}}{V[Y]}.$$

De forma análoga se definiría la razón de correlación de X respecto a Y.

En general las dos razones de correlación son diferentes y verifican:

$$0 \leq \eta^2_{y,x}, \eta^2_{x,y} \leq 1.$$

La razón de correlación $\eta^2_{y,x}$ vale 0 si la varianza de las medias condicionadas es nula. Es decir, si todas las medias condicionadas son idénticas, en cuyo caso la curva de regresión es paralela al eje X, y se dice que Y está incorrelada con X. El recíproco también es cierto, la ausencia de correlación de Y con X, implica que la razón de correlación vale cero.

La razón de correlación $\eta^2_{y,x}$ vale 1 cuando la media de las varianzas condicionadas $V_i(Y)$ es cero; ahora bien, una suma de términos positivos es nula sólo si todos son nulos, lo cual implica que todas las $V_i(Y)$ son cero, es decir, al valor x_i de X le corresponde un único valor de Y, y, por consiguiente, Y depende funcionalmente de X. Recíprocamente, la dependencia funcional de Y respecto a X, implica que la razón de correlación vale 1.

Los distintos casos que se pueden presentar se recogen en la tabla 3.2. De igual forma que se hacía con el coeficiente de determinación, se puede multiplicar la razón de correlación por 100 y se obtendrá, en forma de porcentaje, la parte de la variación de Y que queda explicada por la curva de regresión, que será

$$\frac{\sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2 f_{i.}}{V[Y]} 100\,\% \ ,$$

el resto hasta 100, es decir:

$$\frac{S_e^2}{V[Y]} 100\,\%$$

Razones de correlación	$\eta^2_{x,y}=0$	$0<\eta^2_{x,y}<1$	$\eta^2_{x,y}=1$
$\eta^2_{y,x}=0$	Ausencia recíproca de correlación	Ausencia de correlación de Y respecto de X	Dependencia funcional de X respecto de Y Ausencia de correlación de Y respecto de X
$0<\eta^2_{y,x}<1$	Ausencia de correlación de X respecto de Y	Caso general	Dependencia funcional no recíproca de X respecto de Y
$\eta^2_{y,x}=1$	Dependencia funcional de Y respecto de X Ausencia de correlación de X respecto de Y	Dependencia funcional no recíproca de Y respecto de X	Dependencia funcional recíproca

Tabla 3.2: Estudio de las razones de correlación

quedará sin explicar por ella.

Propiedad 3.1 *La curva de regresión es la "función" óptima para el criterio de los mínimos cuadrados.*

De este resultado se puede deducir que el E.C.M. de la regresión de la media es menor o igual que el E.C.M. del ajuste mínimo cuadrático, y como consecuencia, que el coeficiente de determinación, es siempre menor o igual que la menor de las razones de correlación, sea cual sea la función elegida. La comparación entre R^2 y mín$(\eta^2_{y,x}, \eta^2_{x,y})$ puede indicar si existe una función mejor para ajustar los datos.

6. Notas y conclusiones

1. A lo largo del capítulo se ha comentado que la regresión a la media tiene sentido cuando se cuenta con un amplio número de observaciones de la variable dependiente para un valor de la independiente. Sin embargo, en ciertos casos se puede suavizar esta condición. Así, si la variable X es continua, aún con un alto número de observaciones, es de esperar que haya pocos valores que se repitan, pero siempre se puede agrupar los datos en intervalos; éstos deben ser

lo suficientemente pequeños como para que los elementos pertenecientes a un mismo intervalo puedan considerarse, a los efectos pertinentes, iguales, y lo bastante grandes como para que caiga un número mínimo de observaciones en la mayoría de las clases.

2. Cuando la nube de puntos no proporcione una idea de la clase funcional que se debe elegir para realizar el ajuste, una posible solución puede ser el calcular la línea de regresión y representarla gráficamente.

3. Se ha utilizado una función para efectuar el ajuste, el calcular la previsión para un valor de la variable independiente se reduce a sustituir dicho valor en la función. Si debe emplearse la poligonal de regresión, la previsión para un valor x de X será el valor correspondiente a la clase a la que pertenece x. Esto hace que la previsión tenga dos matices diferentes: mientras que si se utiliza la poligonal de regresión, sólo se pueden hacer previsiones para valores encuadrados en alguna de las clases predefinidas —en realidad sería una interpolación—; cuando las previsiones se basan en una función de ajuste, no sólo pueden hacerse estimaciones para valores intermedios, sino que se pueden extrapolar y sacar conclusiones para valores exteriores; aunque, en este caso, la fiabilidad de la previsión depende de que las condiciones en que se realizó el ajuste, permanezcan constantes, lo que ocurrirá, normalmente, en un entorno de los puntos utilizados para realizarlo.

7. Ejercicios

7.1. Ejercicio resuelto

3.1 Dada la siguiente distribución bidimensional:

X	$0'7$	1	2	3	3	4	5	6	7	8
Y	$2'2$	$2'2$	$2'5$	$2'7$	$2'8$	3	$3'3$	$3'4$	4	4

a) Calcule las dos rectas de ajuste.

b) Compruebe que dichas rectas se cortan en el centro de gravedad.

c) Estime el valor de Y para X=10.
d) Interprete el significado de b y de b'.
e) Interprete el significado de $1 - R^2$.

Solución:

a) Para calcular las dos rectas se procede a obtener las ecuaciones normales, y de aquí, o bien, se utilizan las expresiones obtenidas en el capítulo para a y b, es decir, $a = \bar{y} - b\bar{x}$ y $b = \frac{S_{XY}}{S_X^2}$, o bien, se resuelve directamente el sistema. En este caso el sistema de ecuaciones normales viene dado por

$$\begin{aligned} 30'1 &= 10a + 39'7b \\ 134'14 &= 39'7a + 213'49b \ . \end{aligned}$$

Resolviendo por Cramer se tiene que

$$b = \frac{\begin{vmatrix} 10 & 30'1 \\ 39'7 & 134'14 \end{vmatrix}}{\begin{vmatrix} 10 & 39'7 \\ 39'7 & 213'49 \end{vmatrix}} = 0'262 \ ,$$

y despejando en la primera ecuación normal se obtiene $a = 1'97$. Con lo que la ecuación queda:

$$y = 1'97 + 0'262x.$$

Para obtener la otra recta se puede hacer lo mismo, no obstante, para ofrecer otra posibilidad de resolución se utiliza $b' = \frac{S_{xy}}{S_y^2}$ y $a' = \bar{x} - b'\bar{y}$. Así, puesto que $S_{xy} = 1'464$, $S_y^2 = 0'391$, $\bar{x} = 3'97$ y $\bar{y} = 3'01$, se tiene que

$$b' = \frac{1'464}{0'391} = 3'74 \quad y \quad a' = 3'97 - 3'74 \times 3'01 = -7'28$$

Quedando esta otra recta como:

$$x = -7'28 + 3'74y.$$

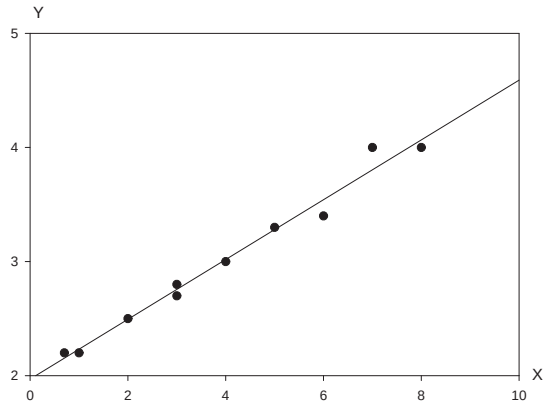

Figura 3.4: Recta de ajuste de Y en función de X

b) Para comprobar que ambas rectas se cortan en el centro de gravedad, basta con sustituir $\bar{x}$ y $\bar{y}$ en las rectas respectivas, y así:

$$\begin{aligned} y &= 1'97 + 0'262 \cdot 3'97 = 3'01 \\ x &= -7'28 + 3'74 \cdot 3'01 = 3'98. \end{aligned}$$

Como puede verse, se cometen pequeños errores de redondeo, subsanables operando con más cifras decimales.

c) Para realizar la estimación, se sustituye el valor $x = 10$ en la recta de ajuste

$$Y = 1'97 + 0'262 \times 10 = 4'59 \ .$$

La calidad de la previsión depende de que ésta se haga en un entorno próximo a los valores de la variable, en nuestro caso se puede pensar que el valor 10 está en dicho entorno, y, por otra parte, de que la recta se ajuste suficientemente bien a los puntos, para lo que se debe calcular el coeficiente de determinación R^2. Para ello, se utiliza que $R^2 = bb' = 0'262 \times 3'74 = 0'98$, es decir, la recta explica a través de X el 98 %

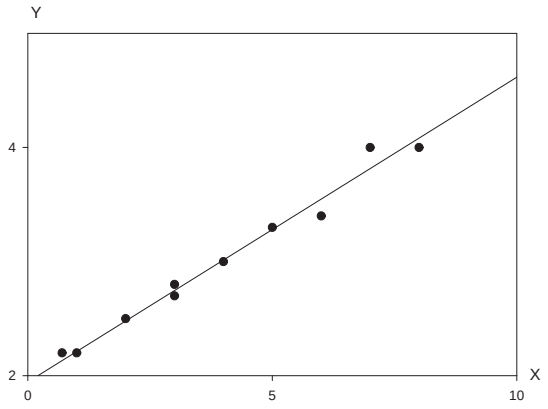

Figura 3.5: Recta de ajuste de X en función de Y

de la variabilidad de Y. A la vista de lo anterior, se puede concluir que $4'59$ es una buena previsión para Y dado que X vale 10.

d) b y b' son las pendientes de la rectas y representan los incrementos-decrementos, según que sean positivas o negativas, de la variable dependiente para incrementos unitarios de la variable independiente. En nuestro caso cuando X aumenta una unidad Y aumenta 0'262 unidades, mientras que cuando lo hace Y en una unidad X crece 3'74 unidades.

e) $1 - R^2 = 0'02$. Lo que indica que el 2 % de la variabilidad de Y, no se explica por la recta en función de X, y que puede haber una mejor función de ajuste u otras variables distintas a la X no contempladas en el modelo. En general, ocurren ambas cosas a la vez, aunque en todo caso esto habrá que discutirlo a partir de los valores de las razones de correlación.

7.2. Ejercicios propuestos

3.1. Dada la distribución:

X	1	$1'5$	2	$2'5$	3	$3'75$	$4'5$	5
Y	1	$1'5$	$2'95$	$5'65$	$8'8$	15	25	32

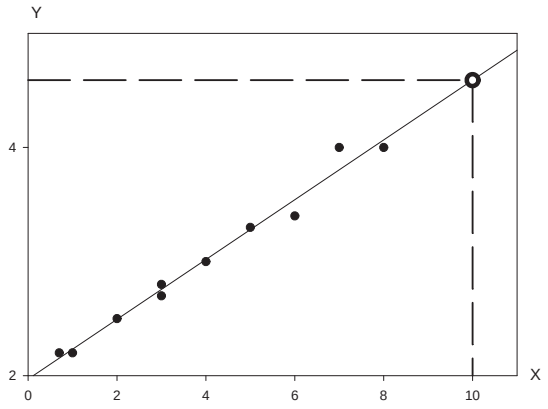

Figura 3.6: Previsión

a) Elija la mejor clase funcional para ajustar la distribución y estime sus parámetros.

b) Establezca la bondad del ajuste.

c) Calcule la previsión para Y cuando $X = 7$. Analice dicha previsión.

3.2. Dada la distribución:

Y X	1	2	3	4	5	6	7	8	9
1	2	3	3	1	1	0	1	0	0
2	1	5	6	3	1	0	0	0	0
3	1	2	7	5	2	2	0	0	0
4	0	1	7	5	3	3	0	0	0
5	0	1	5	8	5	4	3	0	0
6	0	0	4	6	7	5	4	0	0
7	0	0	2	3	8	8	7	4	0
8	0	0	1	2	5	6	7	6	1
9	0	0	1	2	3	4	6	4	5

a) Obtenga la poligonal de regresión de Y respecto a X.

b) Calcule sus razones de correlación.

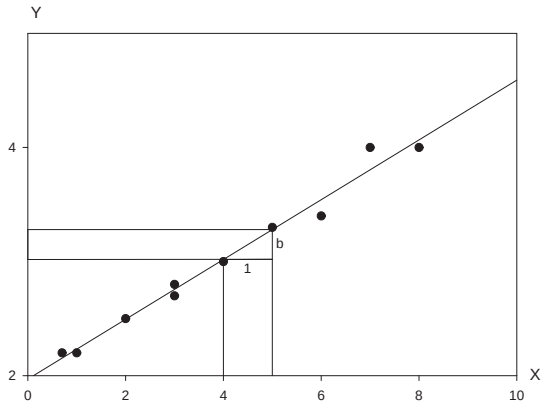

Figura 3.7: Interpretación gráfica de la pendiente de la recta

c) A la vista de la poligonal ajuste la distribución a una función, justificando dicha elección.

d) Calcule R^2, analícelo y compárelo con las razones de correlación.

3.3. Dada la distribución:

X	2′5	3′75	5	7′5	10	12′5	20
Y	8	14	23′75	40	62	90	165

a) Utilice una ecuación del tipo aX^b para ajustar la distribución.

b) Dé una medida de la bondad del ajuste.

3.4. Dada la distribución:

X	1	1′5	2	3	4	5	6	7
Y	1	1′75	2′65	4′7	7	9′5	12	15

a) Ajuste la distribución utilizando una función del tipo aX^b.

b) Analice la bondad del ajuste.

3.5. Dada la distribución:

X	5	6	8	10	13	18	20
Y	$1'5$	$1'25$	$0'93$	$0'7$	$0'46$	$0'23$	$0'15$

a) Estime los parámetros de la clase funcional $ab^{-0'2X}$ para ajustar la distribución.

b) Estudie la bondad del ajuste.

c) ¿Sería posible plantear un ajuste del tipo ab^{cX}? Justifíquelo.

3.6. Dada la distribución:

Y X	2	3	4	5	6	7
2	6	2	1	0	0	0
5	4	7	5	0	0	0
10	0	1	7	3	0	0
15	0	0	6	3	1	0
20	0	0	4	8	2	0
25	0	0	3	8	5	0
30	0	0	0	5	9	0
35	0	0	0	1	7	4
40	0	0	0	0	3	4

a) Obtenga la curva de regresión y calcule la razón de correlación de Y respecto de X.

b) Ajuste los datos a una recta y a una parábola y discuta los resultados.

3.7. Se ha obtenido utilizando el criterio de mínimos cuadrados, que la recta de ajuste de Y sobre X es $Y = 2X + 9$. Sabiendo que $\bar{x} = 5$, $\bar{y} = 10$, $S_X = 2$, $S_Y = 3$ y $S_{XY} = 0'8$. Calcule:

a) La varianza de los valores estimados en Y.

b) La recta de ajuste de Y sobre X si se le suma 5 a todos los valores de X.

c) La recta de ajuste de Y sobre X si se le suma 3 a todos los valores de Y y se multiplica por 2 todos los valores de X.

3.8. Sabiendo que las ecuaciones de las rectas de ajuste entre la temperatura (Y) y la profundidad (X) vienen dadas por

$$y^* = \frac{-x}{5} + 2 \qquad x^* = -4y + 11 \ ,$$

y además $|S_{XY}| = 8$. ¿Qué variable es más homogénea?