CAPÍTULO 9
ESTADÍSTICA DESCRIPTIVA BIDIMENSIONAL

1. INTERROGANTES CENTRALES DEL CAPÍTULO

a) Cuando sobre cada individuo se observan simultáneamente dos características cuantitativas ¿cómo se organizan y representan gráficamente esos datos bidimensionales?

b) ¿Cómo se puede saber si dos variables estadísticas están relacionadas de forma lineal, exponencial, potencial o parabólica?

c) ¿Se puede predecir el valor de una variable sabiendo el valor de otra variable que está relacionada con ella de forma lineal, exponencial, potencial o parabólica?

2. CONTENIDOS FUNDAMENTALES DEL CAPÍTULO

2.1. Tabulación de los datos

Cuando sobre cada individuo de una población se observan simultáneamente dos características cuantitativas, que unidimensionalmente podríamos representar separadamente por las variables X e Y, entonces se dice que se está observando una *variable estadística bidimensional* y se representa por (X, Y).

El conjunto de valores bidimensionales de la variable junto con sus frecuencias asociadas dará lugar a la correspondiente *distribución bidimensional de frecuencias*.

En el caso de variables bidimensionales podemos distinguir dos tipos principales de tablas:

a) Tabulación en dos columnas (o en dos filas)

Si el número de datos bidimensionales es pequeño, los datos se disponen en dos columnas (o en dos filas) sobre las que se emparejan los correspondientes valores unidimensionales de una misma realización de la variable bidimensional, como se expresa en la tabla siguiente:

variable X	variable Y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

X \ Y	B_1	B_2	$\cdots$	B_j	$\cdots$	B_k	suma
A_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1k}	f_{1*}
A_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2k}	f_{2*}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{ik}	f_{i*}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	f_{r1}	f_{r2}	$\cdots$	f_{rj}	$\cdots$	f_{rk}	f_{r*}
suma	f_{*1}	f_{*2}	$\cdots$	f_{*j}	$\cdots$	f_{*k}	n

Tabla 9.1: Tabla de doble entrada o de contingencia

b) Tabla de doble entrada o de contingencia

Si el número de observaciones bidimensionales es grande, clasificamos los n individuos de la muestra en r clases $(A_1, \ldots, A_r)$ respecto de la variable X, y en k clases $(B_1, \ldots, B_k)$ respecto de la variable Y. Entonces los datos suelen organizarse en una tabla como la Tabla 9.1, que se denomina *tabla de doble entrada* o de *contingencia*.

En la Tabla 9.1, f_{ij} es el número de individuos que pertenecen a la clase A_i de la variable X y a la clase B_j de la variable Y y se llama *frecuencia absoluta conjunta* de la clase $A_i \times B_j$ de la variable bidimensional (X, Y).

La *frecuencia relativa conjunta* de la clase bidimensional $A_i \times B_j$ es igual a:

$$h_{ij} = \frac{f_{ij}}{n}. \tag{9.1}$$

2.2. Distribuciones marginales y condicionadas. Independencia de variables

Supongamos que tenemos los datos bidimensionales organizados en una tabla de doble entrada como la Tabla 9.1.

La suma de las frecuencias absolutas conjuntas de la fila $i-$ésima, f_{i*}, es igual al número de individuos en la clase A_i de la variable X, independientemente del valor de Y, y se llama *frecuencia absoluta marginal* de la clase A_i de la variable X:

$$f_{i*} = f_{i1} + f_{i2} + \cdots + f_{ik}.$$

La *frecuencia relativa marginal* de la clase unidimensional A_i es igual a:

$$h_{i*} = \frac{f_{i*}}{n}. \tag{9.2}$$

Análogamente, la suma de las frecuencias absolutas conjuntas de la columna $j-$ésima, f_{*j}, es igual al número de individuos en la categoría B_j de la variable Y, y se llama *frecuencia absoluta marginal* de la clase B_j de la variable Y:

$$f_{*j} = f_{1j} + f_{2j} + \cdots + f_{rj}.$$

La *frecuencia relativa marginal* de la clase unidimensional B_j es igual a:

$$h_{*j} = \frac{f_{*j}}{n}. \tag{9.3}$$

Si de la Tabla 9.1 consideramos la primera y la última columna obtenemos la *distribución marginal de frecuencias absolutas de la variable* X:

X	f_{i*}
A_1	f_{1*}
$\vdots$	$\vdots$
A_i	f_{i*}
$\vdots$	$\vdots$
A_r	f_{r*}
suma	n

Análogamente, si consideramos la primera y la última fila de la Tabla 9.1, obtenemos la *distribución marginal de frecuencias absolutas de la variable* Y:

Y	f_{*j}
B_1	f_{*1}
$\vdots$	$\vdots$
B_j	f_{*j}
$\vdots$	$\vdots$
B_k	f_{*k}
suma	n

Denotaremos por X/y_j a la variable X condicionada a que Y tome el valor y_j. La *distribución de frecuencias absolutas condicionadas* de X/y_j se obtiene de la Tabla 9.1 considerando la primera columna y la columna de la clase B_j; es decir:

X/y_j	f_{ij}
A_1	f_{1j}
$\vdots$	$\vdots$
A_i	f_{ij}
$\vdots$	$\vdots$
A_r	f_{rj}
suma	f_{*j}

Por tanto, la *frecuencia relativa* de $X \in A_i$ *condicionada* a que Y tome el valor y_j es:

$$h_{i/j} = \frac{f_{ij}}{f_{*j}}. \tag{9.4}$$

Análogamente, denotaremos por Y/x_i a la variable Y condicionada a que X tome el valor x_i. La *distribución de frecuencias absolutas condicionadas* de Y/x_i se obtiene de la Tabla 9.1 considerando la primera fila y la fila de la clase A_i; es decir:

Y/x_i	f_{ij}
B_1	f_{i1}
$\vdots$	$\vdots$
B_j	f_{ij}
$\vdots$	$\vdots$
B_k	f_{ik}
suma	f_{i*}

En consecuencia, la *frecuencia relativa* de $Y \in B_j$ *condicionada* a que X tome el valor x_i es:

$$h_{j/i} = \frac{f_{ij}}{f_{i*}}. \tag{9.5}$$

Teniendo en cuenta las fórmulas 9.4 y 9.5 se obtiene:

$$f_{ij} = h_{i/j} f_{*j} = h_{j/i} f_{i*}.$$

Dividiendo por n tenemos:

$$\frac{f_{ij}}{n} = h_{i/j} \frac{f_{*j}}{n} = h_{j/i} \frac{f_{i*}}{n}.$$

Y teniendo en cuenta 9.1, 9.2 y 9.3 se tiene:

$$h_{ij} = h_{i/j} h_{*j} = h_{j/i} h_{i*}.$$

La variable X es *independiente* de la variable Y si las distribuciones de frecuencias relativas de X condicionada a cualquier valor de Y son todas idénticas; es decir, no dependen del valor que tome la variable condicionante Y; es decir:

$$h_{i/1} = h_{i/2} = \cdots = h_{i/k} \qquad \forall i \ ,$$

lo que es equivalente a:

$$\frac{f_{i1}}{f_{*1}} = \frac{f_{i2}}{f_{*2}} = \cdots = \frac{f_{ij}}{f_{*j}} = \cdots = \frac{f_{ik}}{f_{*k}} \qquad \forall i \ ,$$

y por tanto:

$$\frac{f_{ij}}{f_{*j}} = \frac{f_{i1} + f_{i2} + \cdots + f_{ij} + \cdots + f_{ik}}{f_{*1} + f_{*2} + \cdots + f_{*j} + \cdots + f_{*k}} \qquad \forall i, j \ ,$$

lo que también se puede escribir como:

$$\frac{f_{ij}}{f_{*j}} = \frac{f_{i*}}{n} \qquad \forall i, j \ .$$

En consecuencia, la definición de independencia entre las variables X e Y es equivalente a la siguiente propiedad:

$$f_{ij} = \frac{f_{i*} \, f_{*j}}{n} \qquad \forall i, j \ ,$$

o su equivalente:

$$h_{ij} = h_{i*} h_{*j} \ , \qquad \forall i, j \ ,$$

es decir, las frecuencias relativas conjuntas son iguales al producto de las correspondientes frecuencias relativas marginales.

2.3. Representaciones gráficas

Los métodos para determinar la existencia y el grado de relación entre dos variables cuantitativas deben ser capaces también de discriminar entre los tipos generales de relación que hay:

a) Se dice que dos variables cuantitativas X e Y mantienen una relación *directa* cuando los valores altos en Y tienden a emparejarse con valores altos en X, los valores intermedios en Y tienden a emparejarse con valores intermedios en X, y los valores bajos en Y tienden a emparejarse con valores bajos en X.

b) Se dice que dos variables cuantitativas X e Y mantienen una relación *inversa* cuando los valores altos en Y tienden a emparejarse con valores bajos en X, los valores intermedios en Y tienden a emparejarse con valores intermedios en X, y los valores bajos en Y tienden a emparejarse con valores altos en X.

c) Se dice que no hay relación entre dos variables cuantitativas cuando no existe un emparejamiento sistemático entre ellas en función de sus valores.

En una buena representación gráfica conjunta de dos variables estadísticas cuantitativas debe apreciarse fácilmente si existe relación entre las variables y de qué tipo es. Una representación gráfica que cumple esta condición es el *diagrama de dispersión*, que también se puede llamar *nube de puntos*.

• Si los datos no están agrupados en intervalos (como en la tabla siguiente), entonces el diagrama de dispersión se hace como se muestra en la Figura 9.1.

x_i	61	118	57	123	125	122	122	85	85	85	83	78	76	76	73	70	97	107
y_i	15	28	15	30	31	30	30	23	22	22	23	23	23	21	21	21	25	29

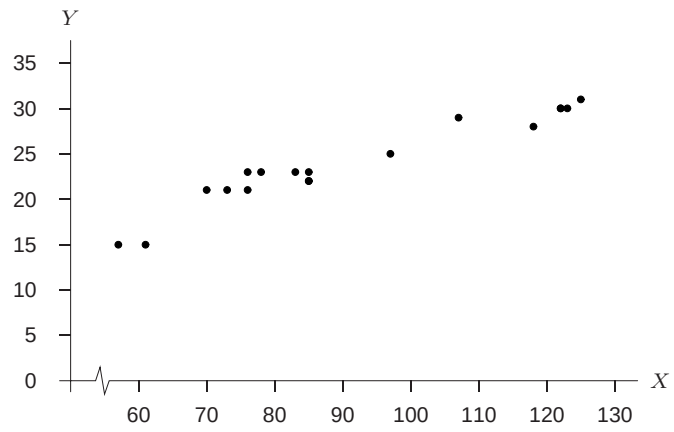

Figura 9.1: Diagrama de dispersión para datos no agrupados en intervalos

• Si los datos están agrupados en intervalos (como en la tabla siguiente), entonces el diagrama de dispersión se hace como se muestra en la Figura 9.2.

X \ Y	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]	suma
(25,75]	13	3				16
(75,125]	4	9	5	1		19
(125,175]		11	16	4		31
(175,225]		2	11	9		22
(225,275]		1	5	3	1	10
(275,325]			1	5		6
suma	17	26	38	22	1	104

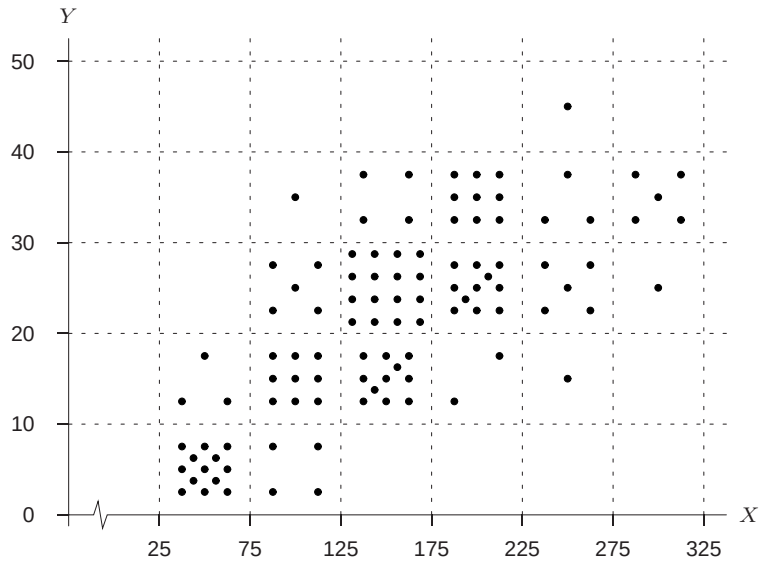

Figura 9.2: Diagrama de dispersión para datos agrupados en intervalos

2.4. Covarianza

A partir de las distribuciones marginales de X y de Y se pueden calcular las medidas descriptivas de las variables X e Y.

De entre las medidas descriptivas bidimensionales, la más utilizada es la *Covarianza* entre X e Y que se calcula de la siguiente forma:

1) Si los datos se tabulan en dos columnas (o dos filas), la covarianza entre X e Y es:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} = \frac{\sum_{i=1}^{n} x_i\, y_i}{n} \; - \; \overline{x}\,\overline{y}\,.$$

2) Si los datos se organizan en una tabla de doble entrada como la Tabla 9.1, la covarianza entre X e Y es:

$$s_{xy} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{k}(x_i - \overline{x})(y_j - \overline{y}) f_{ij}}{n} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{k} x_i\, y_j\, f_{ij}}{n} \; - \; \overline{x}\,\overline{y},$$

donde x_i es la marca de clase de la clase A_i, y_j es la marca de clase de la clase B_j y f_{ij} es la frecuencia absoluta conjunta de la clase bidimensional $A_i \times B_j$.

Si en lugar de dividir por n dividimos por $(n-1)$ tenemos la la *Cuasicovarianza* o *Covarianza modificada o corregida* entre X e Y; cuya definición, por tanto, es la siguiente:

1) Si los datos se tabulan en dos columnas (o dos filas), la cuasicovarianza entre X e Y es:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}\,.$$

2) Si los datos se organizan en una tabla de doble entrada como la Tabla 9.1, la cuasicovarianza entre X e Y es:

$$S_{xy} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{k}(x_i - \overline{x})(y_j - \overline{y})f_{ij}}{n-1},$$

donde x_i es la marca de clase de la clase A_i, y_j es la marca de clase de la clase B_j y f_{ij} es la frecuencia absoluta conjunta de la clase bidimensional $A_i \times B_j$.

En consecuencia, la covarianza y la cuasicovarianza están relacionadas de la siguiente forma:

$$(n-1)S_{xy} = ns_{xy}.$$

Por tanto, se puede calcular una de ellas a partir de la otra.

La covarianza (y, por tanto, la cuasicovarianza) es capaz de discriminar entre los dos tipos de relación lineal pues:

a) si $s_{xy} > 0$ entonces hay relación lineal directa entre X e Y,

b) si $s_{xy} < 0$ entonces hay relación lineal inversa entre X e Y, y

c) si $s_{xy} = 0$ entonces no hay relación lineal entre X e Y.

2.5. Regresión y correlación. Coeficiente de determinación

La *regresión* consiste en sustituir la nube de puntos correspondiente a una distribución bidimensional por la función matemática que mejor se ajuste a la nube de puntos. La *correlación* estima la "fuerza" con que las variables están relacionadas.

La *curva de regresión* es la curva ideal hacia la que tienden los puntos del diagrama de dispersión.

El *ajuste por el método de mínimos cuadrados* consiste en lo siguiente:

Si tenemos una nube de puntos $\{(x_i, y_i), i = 1, 2, \ldots, n\}$ y queremos ajustarle una curva cualquiera $y = f(x, a, b, \ldots)$ con parámetros a, b, $\ldots$, la determinación de éstos se hace minimizando la siguiente expresión:

$$D = \sum_{i=1}^{n}[y_i - f(x_i, a, b, \ldots)]^2$$

Para saber si la curva $y = f(x, a, b, \ldots)$ se ajusta a los puntos $\{(x_i, y_i), i = 1, 2, \ldots, n\}$ calculamos el *coeficiente de determinación*:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}[y_i - f(x_i)]^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Este coeficiente verifica:

1) $0 \leqslant R^2 \leqslant 1$.

2) Si $R^2 = 1$, entonces el ajuste es perfecto.

3) Si $R^2 = 0$, entonces la función $y = f(x)$ no se ajusta en absoluto a los puntos.

4) Cuanto más se aproxime R^2 a 1, mejor es el ajuste.

2.6. Regresión y correlación lineal

2.6.1. *Coeficiente de correlación lineal de Pearson*

La covarianza carece de unos valores máximo y mínimo estables, comunes a todos los casos, que permitan su interpretación directa. La solución a este problema consiste en dividir la covarianza por el producto de las desviaciones típicas marginales. Este índice se conoce con el nombre de *coeficiente de correlación lineal de Pearson*, y se denota por la letra r; o sea:

$$r = \frac{s_{xy}}{s_x\, s_y}\,, \tag{9.6}$$

donde s_x es la desviación típica de la variable X y s_y es la desviación típica de la variable Y.

Si la tabulación de los datos se ha hecho en dos columnas, entonces una fórmula alternativa equivalente a la expresión 9.6 es la siguiente:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}\,.$$

La razón principal por la que la covarianza no puede considerarse un índice de dependencia lineal entre dos variables es la dificultad de su valoración dado que carece de un máximo y un mínimo estables. Pero el coeficiente de correlación lineal no tiene esa dificultad ya que este índice no puede valer más de 1 ni menos de -1, es decir:

$$-1 \leqslant r \leqslant 1\,.$$

Además, la interpretación descriptiva de r es la siguiente:

a) Si $r = 1$ entonces existe una dependencia lineal directa exacta entre las variables X e Y. Los puntos del diagrama de dispersión están sobre una línea recta de pendiente positiva.

b) Si $r = -1$ entonces existe dependencia lineal inversa exacta entre X e Y. Los puntos del diagrama de dispersión están sobre una línea recta de pendiente negativa.

c) Si $r = 0$ entonces no existe dependencia lineal entre X e Y.

d) Cuanto más se aproxime r a -1 o a 1, más dependencia lineal existe entre X e Y. Cuando esto ocurra, el diagrama de dispersión se aproxima a una línea recta.

e) Cuanto más se aproxime r a 0, más independencia lineal existe entre X e Y. Cuando esto ocurra, el diagrama de dispersión no se aproxima a una recta.

f) Si r es positivo, entonces al aumentar el valor de la variable X, aumenta el valor de la variable Y.

g) Si r es negativo, entonces al aumentar el valor de la variable X, disminuye el valor de la variable Y.

2.6.2. *Recta de regresión mínimo cuadrática*

La recta de regresión mínimo cuadrática de Y sobre X es la recta $\hat{Y} = A + BX$ que mejor se ajusta (por el método de mínimos cuadrados) a los puntos del diagrama de dispersión $\{(x_i, y_i), i = 1, 2, \ldots, n\}$. Esta recta nos permitirá predecir Y a partir de los valores de X.

Tenemos que minimizar la expresión:

$$D(A,B) = \sum_{i=1}^{n} [y_i - (A + Bx_i)]^2 \ . \tag{9.7}$$

Igualando a cero las derivadas parciales de D respecto de A y de B obtenemos las siguientes *ecuaciones normales*:

$$\begin{cases} \sum y_i = B \sum x_i + nA \\ \sum x_i y_i = B \sum x_i^2 + A \sum x_i \end{cases}$$

Si los datos están tabulados en dos columnas, las fórmulas de los coeficientes A y B que hacen mínima la expresión 9.7 son las siguientes:

$$B = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2},$$

$$A = \overline{y} - B\,\overline{x}\,.$$

Estas fórmulas son equivalentes a las siguientes:

$$B = \frac{s_{xy}}{s_x^2} = r\,\frac{s_y}{s_x}\,,$$

$$A = \overline{y} - B\,\overline{x}\,.$$

Estas últimas fórmulas se pueden aplicar tanto si los datos están organizados en una tabla de dos columnas como si lo están en una tabla de doble entrada.

Análogamente, la ecuación de la recta de regresión mínimo cuadrática de X sobre Y es:

$$\hat{X} = A^* + B^* Y, \qquad \text{donde} \qquad B^* = \frac{s_{xy}}{s_y^2} = r\,\frac{s_x}{s_y}, \qquad A^* = \overline{x} - B^* \overline{y}\,.$$

2.6.3. *Coeficiente de determinación y coeficiente de correlación lineal. Predicción*

En el caso del ajuste lineal (ajuste a una recta), el coeficiente de determinación es igual a:

$$R^2 = \frac{s_{xy}^2}{s_x^2\, s_y^2}.$$

Por tanto, (sólo en el caso del ajuste lineal) se cumple que el coeficiente de determinación es igual al cuadrado del coeficiente de correlación lineal ($R^2 = r^2$).

Si el coeficiente de correlación lineal está próximo a 1 o a -1 sabemos que existe bastante relación lineal entre las variables X e Y y por tanto los puntos del diagrama de dispersión están próximos a la recta de regresión mínimo cuadrática. En este caso, a partir de la ecuación de la recta de regresión de Y sobre X se puede calcular, de forma aproximada, el valor de la variable Y cuando se conoce el valor de la variable X. Esta aproximación se conoce también por el nombre de *estimación*, *predicción* o *pronóstico*. Similarmente, a partir de la ecuación de la recta de regresión de X sobre Y se pueden predecir los valores de la variable X cuando se conocen los valores de la variable Y.

Si el coeficiente de correlación lineal no está próximo a 1 o a -1, las ecuaciones de las rectas de regresión no nos sirven para predecir los valores de una de las variables cuando se conocen los valores de la otra, pues los puntos del diagrama de dispersión no están próximos a la recta de regresión mínimo cuadrática.

2.7. Regresión exponencial

Ajuste a la curva $Y = Ae^{BX}$ por el método de mínimos cuadrados.

Tomando logaritmos neperianos: $\ln Y = \ln A + BX$.

Se hace el cambio: $Y' = \ln Y\ ,\ A' = \ln A$. Entonces $Y' = A' + BX$, con lo que se reduce a un ajuste lineal entre las variables Y' y X (se pueden utilizar las ecuaciones normales). La bondad del ajuste nos lo da el coeficiente de determinación, que coincide con el cuadrado del coeficiente de correlación lineal entre Y' y X.

2.8. Regresión potencial

Ajuste a la curva $Y = AX^B$ por el método de mínimos cuadrados.

Tomando logaritmos decimales: $\log Y = \log A + B\ \log X$.

Se hace el cambio: $Y' = \log Y\ ,\ A' = \log A\ ,\ X' = \log X$. Entonces $Y' = A' + BX'$, con lo que se reduce a un ajuste lineal entre Y' y X' (se pueden utilizar las ecuaciones normales). La bondad del ajuste nos lo da el coeficiente de determinación, que coincide con el cuadrado del coeficiente de correlación lineal entre Y' y X'.

2.9. Regresión parabólica

Ajuste a la curva $Y = A + BX + CX^2$ por el método de mínimos cuadrados.

Minimizar $D(A, B, C) = \sum[y_i - (A + Bx_i + Cx_i^2)]^2$.

Simplificando se obtiene:

$$\begin{cases} \sum y_i = An + B\sum x_i + C\sum x_i^2 \\ \sum x_i y_i = A\sum x_i + B\sum x_i^2 + C\sum x_i^3 \\ \sum x_i^2 y_i = A\sum x_i^2 + B\sum x_i^3 + C\sum x_i^4 \end{cases}$$

La bondad del ajuste a la curva $Y = A + BX + CX^2$ nos lo da el coeficiente de determinación:

$$R^2 = 1 - \frac{\sum[y_i - (A + Bx_i + Cx_i^2)]^2}{\sum(y_i - \overline{y})^2}.$$

3. ACTIVIDADES DE APLICACIÓN DE LOS CONOCIMIENTOS

A.9.1. Se está estudiando la relación existente entre los años de estudios realizados por los padres (X) y los años de estudios realizados por los hijos (Y). En una muestra de tamaño 7 se obtienen los siguientes resultados:

x_i	y_i
12	12
10	8
6	6
16	11
8	10
9	8
12	11

Dibujar el diagrama de dispersión o nube de puntos. Hallar la covarianza, s_{xy}, entre las dos variables. Hallar el coeficiente de correlación lineal r. Hallar la ecuación de la recta de regresión mínimo cuadrática de Y sobre X. Predecir el número de años de estudio de un hijo cuyo padre ha estudiado 14 años. Decir si esta predicción es fiable. Hallar la ecuación de la recta de regresión mínimo cuadrática de X sobre Y. Predecir el número de años de estudio de un padre cuyo hijo ha estudiado 15 años.

A.9.2. Determinar el grado de dependencia existente entre los años de estudio completados (X) y las faltas de ortografía cometidas en un dictado (Y) tal y como se encontró en la siguiente muestra de 10 entrevistados.

x_i	10	3	12	11	6	8	14	9	10	2
y_i	1	7	2	3	5	4	1	2	3	10

¿Cuántas faltas ortográficas tendría un entrevistado que hubiese completado 13 años de estudio? ¿Es fiable esta predicción?

A.9.3. Una factoría de una cierta marca de refrescos ha tomado al azar 18 semanas de un año, observando la temperatura media, en grados centígrados (X) correspondiente a cada una de ellas y la cantidad de refrescos pedidos durante cada uno de dichos períodos, en miles (Y). La información obtenida es la siguiente:

x_i	10	28	12	31	30	19	24	5	9	15
y_i	21	65	19	72	75	39	67	11	12	24

Dibujar el diagrama de dispersión. Hallar el coeficiente de correlación lineal r. Predecir la temperatura media de un semana en la que se hubiesen pedido 50.000 refrescos. Decir si esta predicción es fiable. Predecir el número de refrescos pedidos en una semana en la que la temperatura media fuese de 20 grados centígrados.

A.9.4. Se está estudiando la relación existente entre la edad de los hombres (X) y de las mujeres (Y) a la hora de contraer matrimonio. Se recogen los datos del año 1971 en la tabla siguiente:

X \ Y	[10,20]	(20,25]	(25,30]	(30,35]	(35,40]	(40,50]	(50,60]	(60,80]
[10,20]	4.187	16.272	7.401	864	175	127	5	
(20,25]	1.125	55.505	69.151	8.138	1.358	354	26	2
(25,30]	134	8.731	37.480	11.668	2.715	779	64	10
(30,35]	16	485	2.845	4.142	2.602	1.153	120	21
(35,40]	3	104	517	1.110	1.886	1.871	266	57
(40,50]	5	31	142	327	730	2.265	1.176	410
(50,60]		4	12	32	56	314	867	792
(60,80]			1	2	6	33	151	828

¿Existe una dependencia lineal fuerte entre la edad de los hombres y la edad de las mujeres a la hora de contraer matrimonio? Hacer una predicción de la edad de la esposa cuyo esposo tiene 25 años. ¿Es fiable esta predicción? Hacer el diagrama de dispersión.

A.9.5. El precio, en pesetas, (X) y el número de páginas (Y) de los libros contenidos en un catálogo vienen dados por:

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
19.950	496	27.500	392	21.000	240	12.000	342
9.950	208	12.500	200	15.000	278	21.000	340
17.500	300	25.000	280	27.500	420	17.000	207
15.000	448	8.000	120	10.500	128	35.000	440
12.000	200	5.950	220	9.950	249	7.500	88
30.000	288	9.950	200	35.000	392	21.000	351
32.500	324	32.500	468	20.000	400	25.000	292
35.000	525	24.000	539	27.500	300	37.250	464
37.500	384	30.000	400	15.000	240	24.000	344
25.000	250	30.000	320	16.000	230	12.500	130
18.000	200	35.000	736	12.000	144	22.500	382
15.000	224	22.000	516	38.000	336	25.000	403
30.000	384	37.500	700	37.750	550	20.000	249
25.000	256	20.000	400	30.000	478	18.000	182
17.500	215	30.000	656	17.250	437	38.500	458
17.000	278	9.500	191	9.950	288	3.500	63
20.000	376	19.500	464	18.500	496	30.000	400
22.500	421	20.500	348	18.000	236	21.500	278
32.500	450	30.000	352	12.000	143	27.500	508
30.000	243	32.500	598	17.000	284	16.000	256
12.000	202	27.500	392	28.000	520	30.500	368
15.000	251	24.000	472	38.500	758	15.000	275
21.000	320	36.500	591	25.000	413	12.500	112
35.000	460	14.500	282	27.500	394	38.000	458
12.000	342	21.000	340	17.000	207	7.500	83

Agrupar los datos de ambas variables en intervalos de clase. Determinar la distribución bidimensional de frecuencias, así como las distribuciones marginales de X y de Y. Hallar el coeficiente de correlación lineal. Predecir el precio de un libro que tuviera 205 páginas. Decir si esta predicción es fiable.

A.9.6. Las calificaciones obtenidas por un grupo de alumnos en Biología y Física son:

Biología	3	4	6	7	5	8	7	3	5	4	8	5	5	8	8	8	5
Física	5	5	8	7	7	9	10	4	7	4	10	5	7	9	10	5	7

a) Escribir la tabla de doble entrada de frecuencias absolutas.

b) Hallar las distribuciones marginales, así como la media y la varianza de dichas distribuciones unidimensionales.

c) ¿Existe relación lineal entre las calificaciones de Biología y Física?

A.9.7. Se han tomado cinco muestras de glucógeno, de una cantidad fija cada una. Se les ha aplicado una cantidad X de glucogensa (en milimoles/litro) anotando en cada caso la velocidad de reacción Y, medida en micromoles/minuto, obteniéndose los siguientes datos:

X	1	2	3	0'2	0'5
Y	18	35	60	8	10

a) ¿Se puede deducir que la velocidad de reacción aumenta con la concentración de glucogensa? Justificar la respuesta.

b) Si a una de las muestras le hubiésemos aplicado una concentración de glucogensa de $2'5$ milimoles/litro ¿cuál hubiese sido la velocidad de reacción? ¿Con qué grado de predicción?

A.9.8. Un psicólogo afirma en base a los datos obtenidos, que a medida que un niño crece, menor es el número de respuestas inadecuadas que da. Los datos son:

X	2	3	4	4	5	5	6	7	7	9	9	10	11	11	12
Y	11	12	10	13	11	9	10	7	12	8	7	3	6	5	5

donde X representa la edad en años, e Y representa el número de respuestas inadecuadas.

a) Determinar la validez de esta conclusión.

b) Si Alberto, de diez años y medio, participa en el experimento ¿cuál será el número de respuestas inadecuadas que dará?

A.9.9. Dada una variable bidimensional (X, Y), cuya tabla de frecuencias relativas es:

X \ Y	4	7	10	13	16	17
1	$0'03$	$0'04$	$0'03$	0	0	0
2	0	$0'07$	$0'09$	$0'04$	0	0
3	0	0	$0'04$	$0'12$	$0'04$	0
4	0	0	$0'04$	$0'12$	$0'04$	0
5	0	$0'07$	$0'09$	$0'04$	0	0
6	$0'03$	$0'04$	$0'03$	0	0	0

Calcular:

a) Distribuciones marginales de frecuencias absolutas.

b) Medias y varianzas marginales.

c) Recta de regresión de Y sobre X.

d) Coeficiente de correlación lineal.

A.9.10. Los datos de la tabla siguiente representan el resultado de un experimento consistente en exponer bacterias, en períodos de 1 a 15 intervalos de 6 minutos, a la radiación de rayos X a 200 kilovoltios y contabilizar el número de bacterias supervivientes. (X representa el número de intervalos de 6 minutos, e Y representa los cientos de bacterias supervivientes).

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y	355	211	197	166	142	106	104	60	56	38	36	32	21	19	15

Ajustar a los datos una curva exponencial; representar gráficamente el resultado y comprobar la bondad del ajuste.

A.9.11. Los datos de la tabla siguiente son el resultado de un estudio del efecto de la temperatura de cristalización primaria (medida en grados centígrados) de una solución, x_i, sobre el contenido en fósforo (medido en gramos por litro), y_i .

x_i	y_i
25	$10'9$
20	$9'3$
15	$8'2$
12	$7'5$
9	$6'2$
6	$5'8$
3	$4'2$
0	$3'9$
-3	$2'8$
-6	$2'0$

a) Representar gráficamente la nube de puntos. Determinar el modelo de curva adecuado para representar la relación entre las variables y encontrar, por el método de mínimos cuadrados, los parámetros de la curva.

b) ¿El ajuste anterior es bueno?

A.9.12. Los datos de la tabla siguiente pertenecen a la medida de la temperatura (X) y la presión (Y) en diferentes lugares del Himalaya.

x_i	y_i	x_i	y_i	x_i	y_i
$29'211$	$210'8$	$20'480$	$193'4$	$16'959$	$184'1$
$28'559$	$210'2$	$20'212$	$193'6$	$16'881$	$184'6$
$27'972$	$208'4$	$19'758$	$191'4$	$16'817$	$184'1$
$24'697$	$202'5$	$19'490$	$191'1$	$16'385$	$183'2$
$23'726$	$200'6$	$19'386$	$190'6$	$16'235$	$182'4$
$23'369$	$200'1$	$18'869$	$189'5$	$16'106$	$181'9$
$23'030$	$199'5$	$18'356$	$188'8$	$15'928$	$181'9$
$21'892$	$197'0$	$18'507$	$188'5$	$15'919$	$181'0$
$21'928$	$196'4$	$17'267$	$185'7$	$15'376$	$180'6$
$21'654$	$196'3$	$17'221$	$186'0$		
$21'605$	$195'6$	$17'062$	$185'6$		

Explicar la presión en función de la temperatura mediante una parábola (por el método de mínimos cuadrados).

4. ACTIVIDADES PRÁCTICAS DEL CAPÍTULO

4.1. Diagrama de dispersión

Para una variable estadística bidimensional, el gráfico más utilizado es el *diagrama de dispersión*. El programa dibuja estos diagramas si seleccionamos las opciones `Statistics|Summary Statistics|Scatter Plot`. Entonces nos aparece una ventana como en la Figura 9.3 donde debemos seleccionar las variables implicadas (`X-Axis Variable` e `Y-Axis variable`). Si deseamos agrupar los datos en intervalos (de una variable o de las dos) entonces debemos rellenar los recuadros `X-Axis (Optional)` o `Y-Axis (Optional)`, según la variable que queramos agrupar.

Ejercicio. Dibuja el diagrama de dispersión de la variable estadística bidimensional (`PESO`,`ALTURA`).

4.2. Covarianza y coeficiente de correlación lineal

Para una variable bidimensional es interesante hallar la matriz de varianzas y covarianzas corregidas. Dicha matriz es la siguiente:

$$\begin{pmatrix} S_x^2 & S_{xy} \\ S_{xy} & S_y^2 \end{pmatrix},$$

donde S_x^2 denota la cuasivarianza de X, S_y^2 indica la cuasivarianza de Y, y S_{xy} representa la cuasicovarianza entre X e Y. La matriz de varianzas y covarianzas corregidas se puede obtener seleccionando las opciones `Statistics|Linear Models|Variance-Covariance`. Entonces aparece una ventana (ver Figura 9.4) en la que debemos seleccionar las variables estadísticas de las cuales queremos calcular su matriz de covarianzas corregidas (en el recuadro `Var-Covar Variables`).

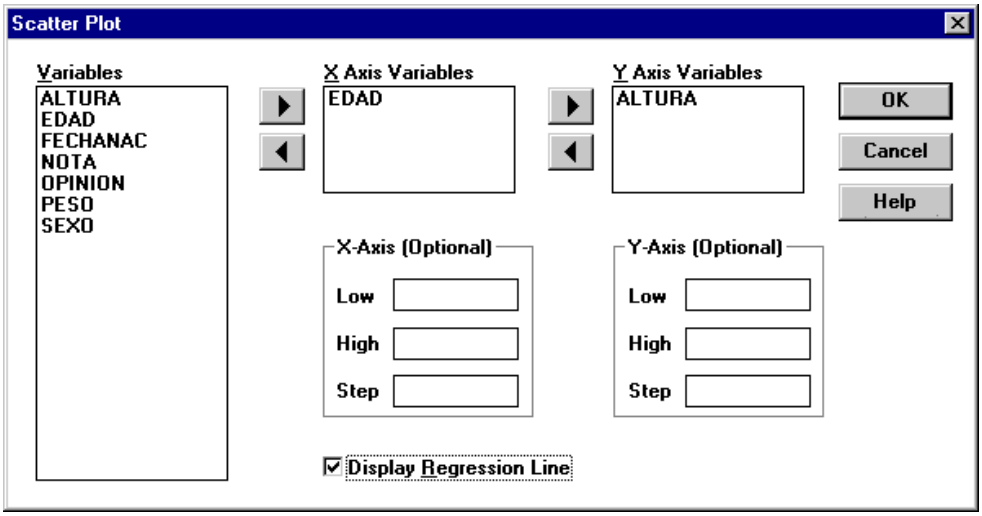

Figura 9.3: Pantalla del programa que permite seleccionar las variables para las que vamos a dibujar el diagrama de dispersión.

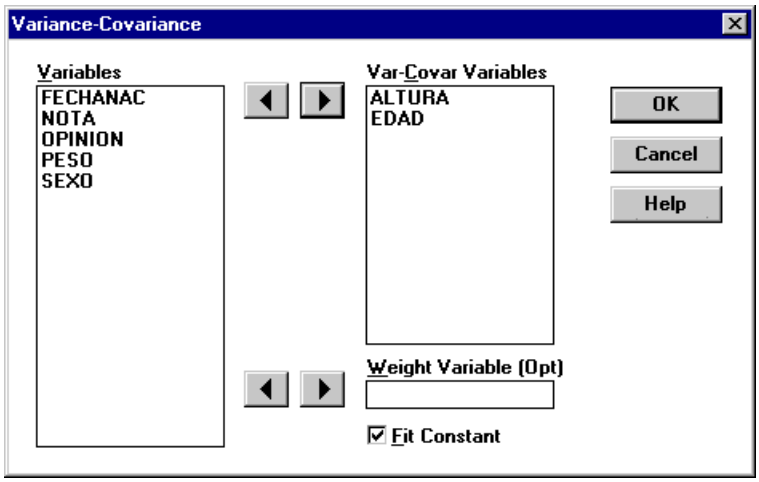

Figura 9.4: Pantalla del programa que permite seleccionar las variables para las que vamos a calcular la matriz de covarianzas.

Ejercicio. Calcula la matriz de varianzas y covarianzas corregidas de la variable bidimensional (`PESO`, `ALTURA`).

Para calcular el coeficiente de correlación lineal de Pearson entre dos variables estadísticas debemos seleccionar las opciones `Statistics|Linear Models|Correlations (Pearson)` y nos aparece una ventana como en la Figura 9.5. Tras pulsar el botón `OK` surge la ventana de resultados, con los coeficientes de correlación lineal entre todas las variables seleccionadas. En nuestro ejemplo (ver Figura 9.6) el coeficiente entre las variables `PESO` y `ALTURA` es de $0'9384$, lo que significa que existe una dependencia lineal fuerte.

Ejercicio. Determina el coeficiente de correlación lineal de Pearson entre las variables `ALTURA` y `PESO`.

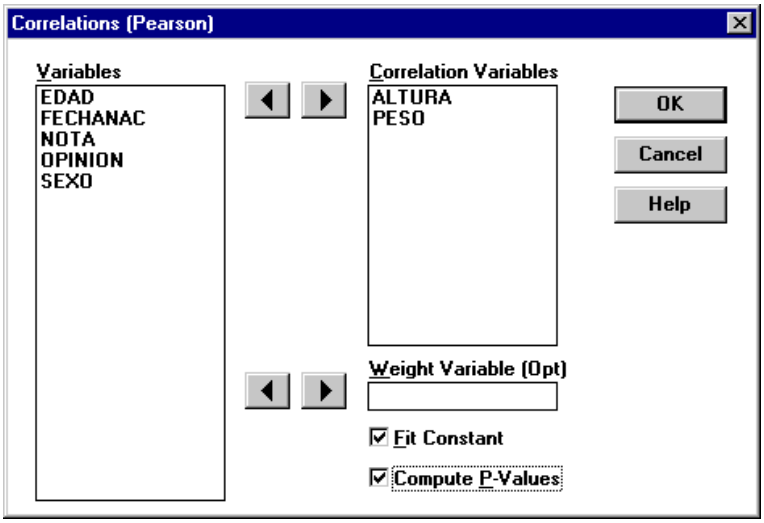

Figura 9.5: Pantalla del programa que permite calcular el coeficiente de correlación lineal de Pearson.

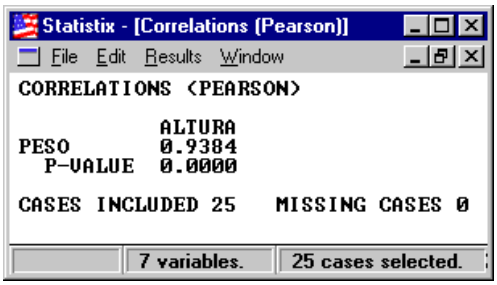

Figura 9.6: Pantalla con los resultados para la correlación entre el PESO y la ALTURA.

4.3. Recta de regresión. Predicción

La recta de regresión permite estimar el valor de una variable estadística conocido el valor de otra variable, siempre que entre las dos variables estadísticas exista dependencia lineal. Cuanto mayor sea esta dependencia lineal, mejor será la aproximación que nos da la recta de regresión. Para calcular la ecuación de la recta de regresión mínimo cuadrática debemos seleccionar las opciones `Statistics|Linear Models|Linear Regression` y nos aparece una ventana como en la Figura 9.7. En el recuadro `Dependent Variable` debemos poner la variable dependiente (la que está representada en el eje vertical) y en el recuadro `Independent Variables` la variable independiente (la que está representada en el eje horizontal).

Por ejemplo, si en la variable dependiente ponemos `ALTURA` y en la variable independiente colocamos `EDAD`, entonces se supone que queremos predecir la altura de un individuo conociendo su edad. La ventana de resultados se muestra en la Figura 9.8. En dicha figura aparecen muchos valores que no estamos en condiciones de explicar en este momento, ya que pertenecen a la parte de Estadística Inferencial. Si la ecuación de la recta de regresión de `ALTURA` sobre `EDAD` es `ALTURA = A + B EDAD`, entonces $A = 1'40089$ y $B = 0'01306$.

Para hacer una predicción debemos seleccionar las opciones `Results|Prediction`, escribiendo en el recuadro `Predictor Values` el valor de la variable independiente para el cual queremos estimar el correspondiente valor de la variable dependiente. En la casilla `Specification Method` debemos seleccionar la opción `Valued Method` (ver Figura 9.9). Así, para una edad de `24.5` años el valor correspondiente de la altura es de `1.7207` metros.

Una vez hemos calculado la recta de regresión podemos realizar una representación gráfica de la misma. Para

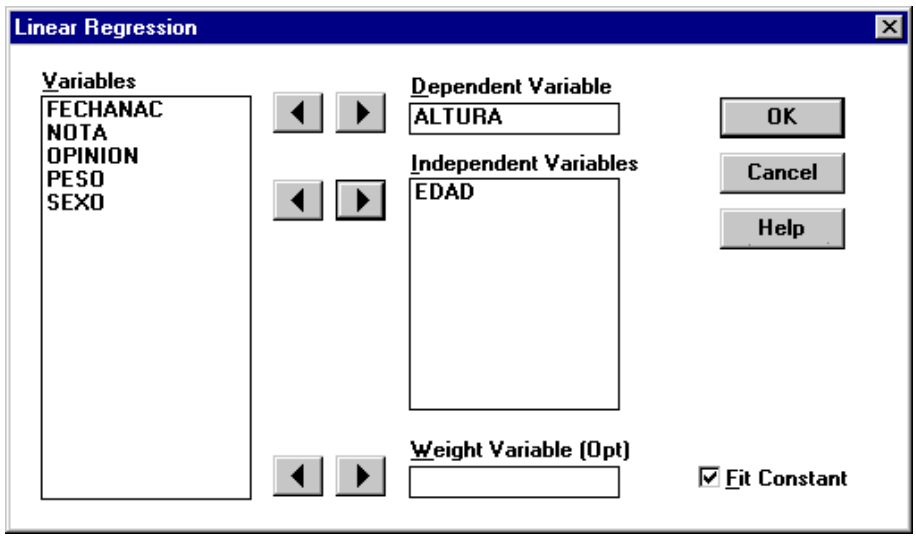

Figura 9.7: Pantalla del programa que permite seleccionar las variables para las que vamos a calcular la recta de regresión.

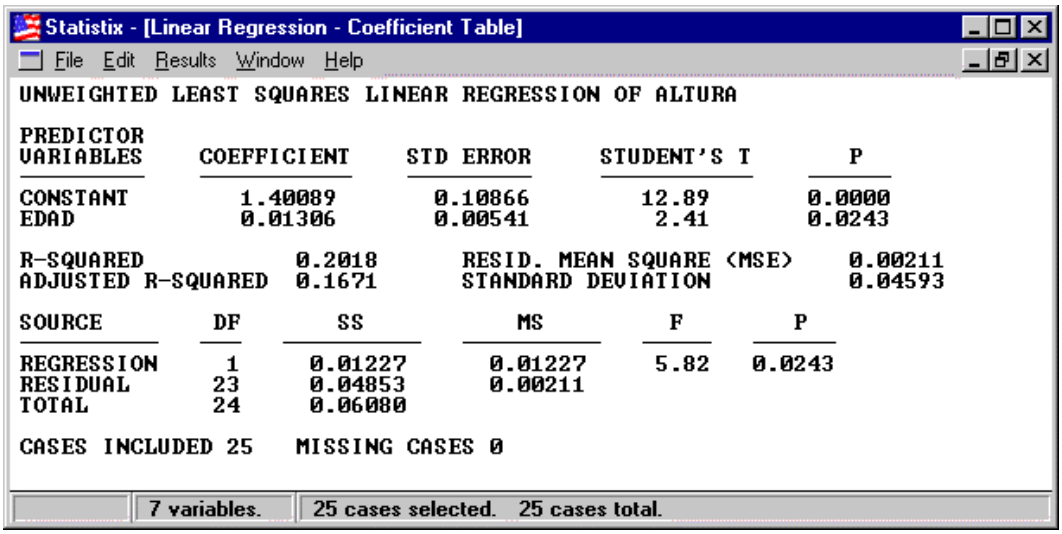

Figura 9.8: Pantalla del programa que muestra la recta de regresión de `ALTURA` sobre `EDAD`.

ello seleccionamos las opciones `Results|Plots|Simple Regression Plot`. En el gráfico resultante aparece el diagrama de dispersión (con cruces), la recta de regresión (en color azul) y dos curvas (en color rojo), una por cada lado de la recta de regresión, que delimitan una zona de confianza para los valores de la variable dependiente (ver Figura 9.10).

Ejercicio. Halla la ecuación de la recta de regresión mínimo cuadrática de `ALTURA` sobre `PESO`. Represéntala gráficamente y predice la altura de un alumno que pesa 60 kilogramos.

5. BIBLIOGRAFÍA DEL CAPÍTULO

CANDEL, J.; MARIN, A. y RUIZ, J.M. *Estadística aplicada I: Estadística descriptiva*. Barcelona: DM–PPU, 1991. Secciones 3.1, 3.3, 4.1, 4.2, 4.3, 4.4, 4.6.

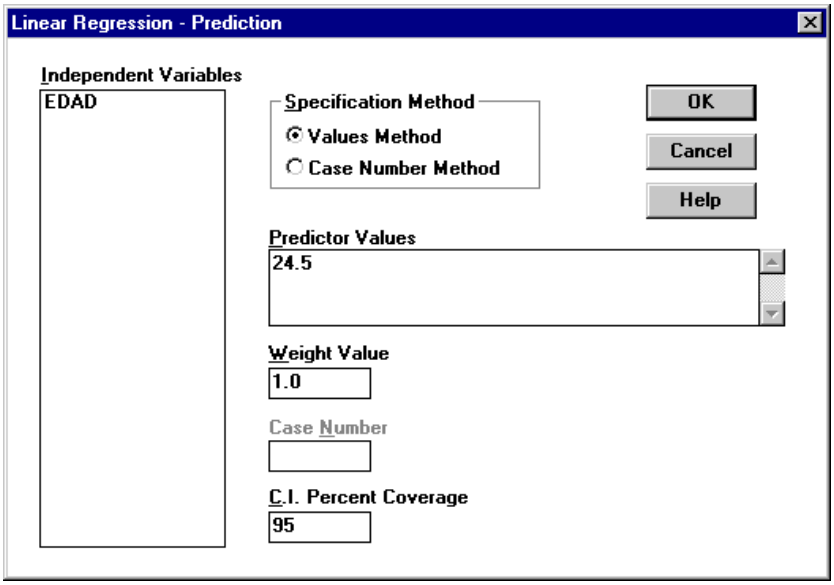

Figura 9.9: Pantalla del programa que permite estimar el valor de ALTURA para un valor de EDAD.

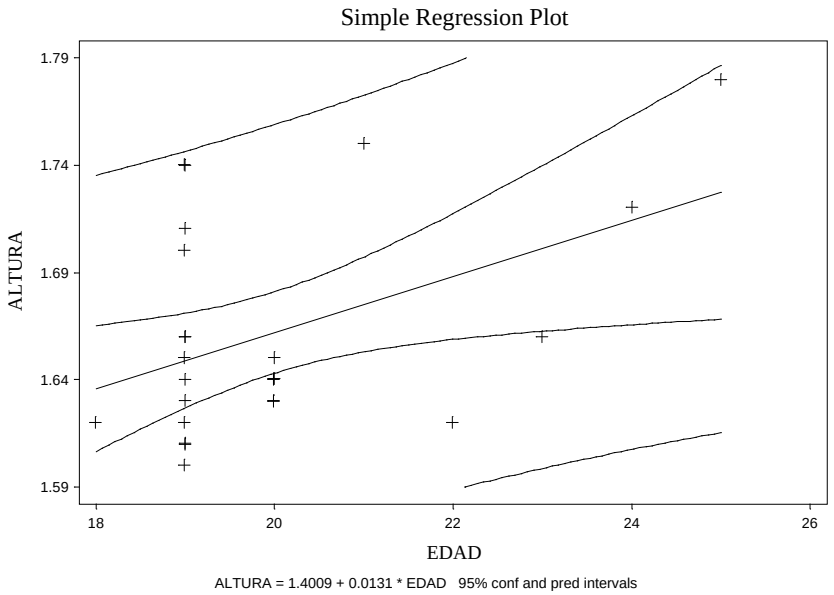

Figura 9.10: Pantalla del programa que representa la recta de regresión de ALTURA sobre EDAD.

6. PREGUNTAS DE EVALUACIÓN

E.9.1. En un grupo de alumnos de la Universidad de Murcia, se estudia el número de asignaturas aprobadas en Junio (X) y el número de horas semanales dedicadas al estudio (Y). La información obtenida es la siguiente:

X \ Y	$(0,10]$	$(10,20]$	$(20,30]$	$(30,40]$
0	6	2	0	0
1	3	6	2	1
2	1	10	8	3
3	0	10	12	8
4	1	5	10	15
5	0	2	16	10

a) Obtener la distribución marginal de frecuencias absolutas de Y. ¿Cuál es el tiempo medio semanal dedicado al estudio? ¿Cuál es el tiempo semanal de horas de estudio que deja por debajo el 50 por ciento de los tiempos semanales y por encima el 50 por ciento restante?

b) Obtener la distribución marginal de frecuencias absolutas de X. ¿Cuál es el número más habitual de asignaturas aprobadas por los alumnos? ¿Cuál es el número medio de asignaturas aprobadas? Calcular e interpretar la varianza de X. Dibujar un polígono acumulativo para la variable X y calcular su recorrido intercuartílico.

c) Obtener la distribución de frecuencias absolutas de Y condicionada a X$=1$. ¿Cuál es el tiempo medio semanal dedicado al estudio por los alumnos que han aprobado una asignatura? ¿Cuál es el tiempo semanal más habitual en los alumnos que han aprobado una asignatura? Calcular la mediana de esta distribución.

E.9.2. En una determinada empresa se ha realizado un estudio para determinar si la edad de los empleados está relacionada con el número de días de ausencia en el trabajo. Estos son los resultados:

	Edad (X)				
Días de ausencia (Y)	(20,29]	(29,38]	(38,47]	(47,56]	(56,65]
(44,50]	0	1	8	7	16
(50,56]	2	6	10	2	4
(56,62]	5	9	5	0	1
(62,68]	14	6	2	2	0

a) Obtener la distribución marginal de frecuencias absolutas y frecuencias acumuladas absolutas de X. Hallar la mediana, la media y la desviación típica de X.

b) Obtener la distribución marginal de frecuencias absolutas de Y. Hallar la moda, la media y la desviación típica de Y.

c) Determinar el coeficiente de correlación lineal entre X e Y. ¿Existe una fuerte dependencia lineal entre X e Y?

E.9.3. Una empresa nacional dedicada a la producción de videojuegos pretende sacar al mercado dos nuevos productos: uno para el segmento de 13 a 15 años y otro para el segmento de 16 a 18 años. Antes de fijar el precio, la empresa contacta con un centro de estudios sociológicos para conocer la asignación semanal de los jóvenes. Para ello, el centro extrae una muestra de 10 jóvenes y, entre otros datos, se les pregunta la edad (X, en años) y su asignación semanal (Y, en miles de pesetas), obteniendo los siguientes datos:

Edad (X)	17	16	16	15	14	13	16	18	17	13
Asignación (Y)	3	4	3	4	1	2	2	5	4	0

a) Calcular recorrido intercuartílico y coeficiente de variación de la asignación semanal.

b) ¿Cuál es la asignación semanal estimada para un joven de 16 años? ¿Es fiable dicha predicción?

E.9.4. Un equipo investigador está analizando el comportamiento de los jóvenes españoles respecto del matrimonio. Para ello extrae una muestra de 10 jóvenes parejas y les pasa un cuestionario. Entre las muchas preguntas del cuestionario figura la edad a la que contrajeron matrimonio, obteniéndose los siguientes datos:

X: Edad de la mujer	26	25	25	24	23	22	25	27	26	22
Y: Edad del hombre	26	27	26	27	24	25	25	28	27	23

a) Calcular los cuartiles Q_1, Q_2 y Q_3 de la variable X.

b) Calcular el coeficiente de variación de la variable Y.

c) Calcular la recta de regresión de Y sobre X. Si en una pareja de jóvenes, la mujer tenía 28 años cuando contrajo matrimonio, ¿cuál es la edad estimada del hombre? ¿es fiable esta predicción?

E.9.5. Una editorial está interesada en conocer los hábitos de lectura de los españoles y determinar si existe alguna relación con otras variables (nivel cultural, nivel económico, edad, etc.). Para ello se extrae una muestra de 10 personas y, entre otros datos, se les pregunta por el número de años de estudio (X) y por el número de libros que suelen comprar cada trimestre (Y). Los datos son los que recoge la siguiente tabla:

X	11	10	10	9	8	7	10	12	11	7
Y	3	4	3	4	1	2	2	5	4	0

a) Calcular la media y la mediana de Y.

b) Calcular la desviación media y la desviación mediana de Y.

c) Si una persona suele comprar 2 libros al trimestre, calcular una estimación para el número de años de estudio. ¿Es fiable dicha predicción?