

Análisis predictivo. Análisis de regresión lineal

[8.1] ¿Cómo estudiar este tema?

[8.2] Ajuste de la recta

[8.3] La regresión por mínimos cuadrados

[8.4] Otros tipos de regresión

[8.5] Tipos de residuos

8

T E M A

Ideas clave

8.1. ¿Cómo estudiar este tema?

En este tema se estudiará el **análisis predictivo** a partir de la correlación entre variables. Se introducirá el concepto de **regresión** y se verá que puede haber regresión simple y múltiple. Este tema se centrará en la **regresión lineal simple** que se basa en la obtención de una ecuación de la recta a partir de una nube de puntos formada por variables que se relacionan.

Se verá cómo obtener la **ecuación de la recta** a partir de un diagrama cuando la regresión es perfecta y cuando no lo es se verá **el método de mínimos cuadrados** para el cual se introducirá el concepto de **residuos**.

Una vez obtenida la recta se estudiará su fiabilidad mediante el **error estándar de la estimación**. Así mismo, se usará la recta para **interpolar** y predecir datos y se aprenderá a calcular el **intervalo de confianza de la estimación** a determinados niveles de significancia.

Se evaluará la asociación mediante el **coeficiente de determinación** y su relación con el coeficiente de correlación.

Así mismo se darán unas pincelas a la **regresión lineal múltiple** y se verá de forma muy resumida los **diferentes tipos de residuos**.

Por último, destacar que en todos los apartados se incluye una serie de **ejemplos prácticos resueltos** que ayudarán a al alumno a comprender lo estudiado con mayor claridad.

Para estudiar este tema **deberás comprender las Ideas clave** expuestas en este documento y que han sido elaboradas por el profesor de la asignatura. Estas ideas se van a complementar con lecturas y otros documentos para que puedas ampliar los conocimientos sobre el mismo.



8.2. Ajuste de la recta

Introducción

La regresión permite conocer el valor de una variable desconocida a partir de datos de otra variable con la que se asocia en varias observaciones, por lo que permite llevar a cabo un análisis predictivo. De este modo, mediante la regresión obtenemos una ecuación de estimación que relaciona las variables y mediante el análisis de correlación se puede saber el grado en el que se asocian. La/s variable/s conocida/s se llama/n variable/ independiente/s y la/s que se desea/n predecir es/son la/s dependiente/s. Cabe destacar que la relación que se halla entre las variables es una asociación, es decir, no tiene por qué ser siempre una causa-efecto.

Hay dos tipos de regresión, la regresión lineal simple y la regresión múltiple, lineal o no, en la que varias variables influyen en una en concreto entre otras. En este tema nos centraremos en la regresión lineal simple y en el apartado 3 se introducirá la regresión lineal múltiple.

Recta de regresión

Mediante los diagramas de puntos era posible prever la asociación entre dos variables. En el caso de la asociación lineal, la unión de los puntos otorga una línea recta más o menos perfecta en función de cómo caigan los puntos en la misma. La recta puede expresarse mediante la ecuación en forma explícita, que da información de la pendiente y de la ordenada en el origen:

$$y = a + bx$$

x e y son las variables independiente y dependiente respectivamente, b la pendiente y a la ordenada en el origen.

1.

Los valores de la pendiente y la ordenada en el origen son constantes para una recta en cuestión, por lo que mediante esta ecuación es posible llevar a cabo un análisis predictivo puesto que para cualquier valor de la variable independiente se puede obtener el valor de la dependiente.

Esta recta se llama recta de ajuste a la nube de puntos y es posible obtener su ecuación a partir del diagrama cuando la asociación es perfecta tal y como se muestra a continuación.

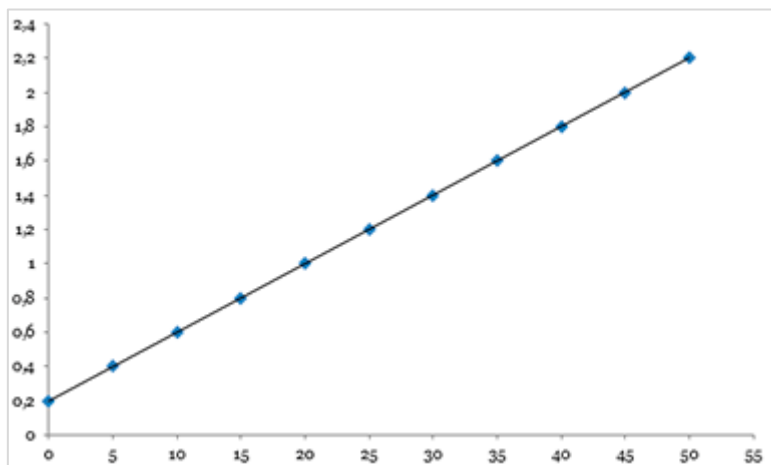


Gráfico 1. Recta de ajuste perfecta.

A partir del diagrama se puede obtener el valor de la ordenada en el origen, corte con el eje vertical, es decir cuando $x=0$, en este caso, tal valor es 0,2 en unidades de la variable dependiente y . Por otro lado, la pendiente se calcula como la tangente del ángulo que la recta de ajuste forma con el eje horizontal:

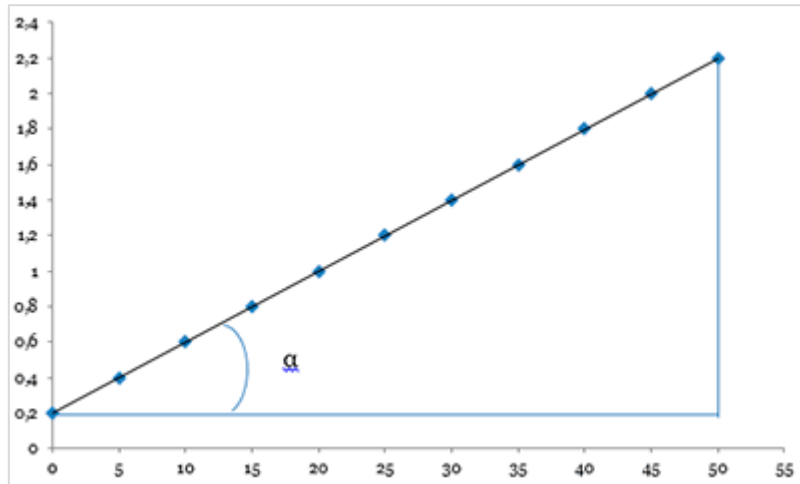


Gráfico 2. Cálculo de la pendiente.

$$b = \operatorname{tg} \alpha = \frac{\text{lado opuesto}}{\text{lado contiguo}}$$

2.

Tomando los valores de x e y , la pendiente resulta:

$$b = \frac{y_{j+n} - y_j}{x_{i+n} - x_i}$$

x_i e y_j han de ser las coordenadas de un mismo punto.

3.

Tomando las dimensiones de los lados del triángulo rectángulo en el ejemplo del gráfico 1 resulta:

$$b = \frac{y_{j+n} - y_j}{x_{i+n} - x_i} = \frac{y_{11} - y_1}{x_{11} - x_1} = \frac{2,2 - 0,2}{50 - 0} = 0,04$$

Se podría tomar cualquier otro punto ya que la relación se mantiene constante tal y como se muestra a continuación:

$$b = \frac{y_{j+n} - y_j}{x_{i+n} - x_i} = \frac{y_5 - y_4}{x_5 - x_4} = \frac{1 - 0,8}{20 - 15} = 0,04$$

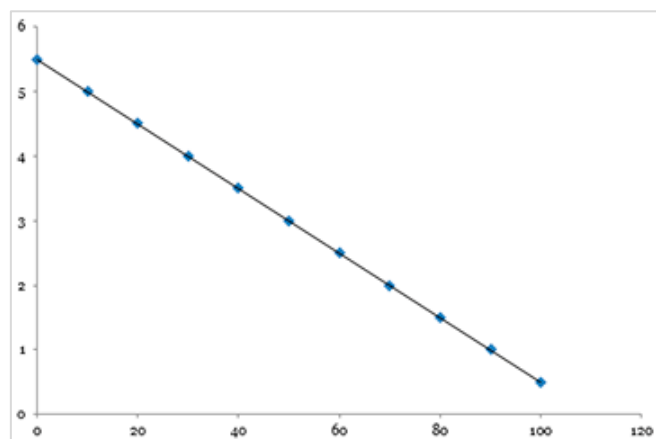
La ecuación de la recta de ajuste resulta:

$$y = 0,04x + 0,2$$

En este caso se trata de una asociación entre variables positiva, de manera que a medida que aumenta la variable independiente lo hace la dependiente debido a que la pendiente es positiva.

Ejemplo 1. Hallar la ecuación de la recta a partir del diagrama de puntos de asociación perfecta

Escribir la ecuación de la recta a partir del siguiente diagrama de puntos y predecir el valor de y cuando x vale 54,6.



La ordenada en el origen es $a=5,5$.

La pendiente se calcula:

$$b = \frac{y_{j+n} - y_j}{x_{i+n} - x_i} = \frac{y_6 - y_4}{x_6 - x_4} = \frac{3 - 4}{50 - 10} = -0,025$$

La ecuación de la recta es:

$$y = -0,025x + 5,5$$

Las variables se relacionan inversamente.

Cuando $x=54,6$ y será:

$$y = -0,025 \cdot 54,6 + 5,5 = 4,135$$

8.3. La regresión por mínimos cuadrados

Obtención de la recta de regresión

En el apartado anterior se ha visto cómo establecer una recta de ajuste a partir de un diagrama de puntos mediante un ajuste lineal perfecto. Cuando se trata de ajustar los puntos de un diagrama de dispersión se emplea el método de mínimos cuadrados que se basa en encontrar una recta que minimice el error entre los puntos dispersados y, por lo tanto, que supone el mejor ajuste entre dos variables. A continuación, se muestra un diagrama en el que la asociación entre los puntos no es una línea recta perfecta sino que hay dispersión.

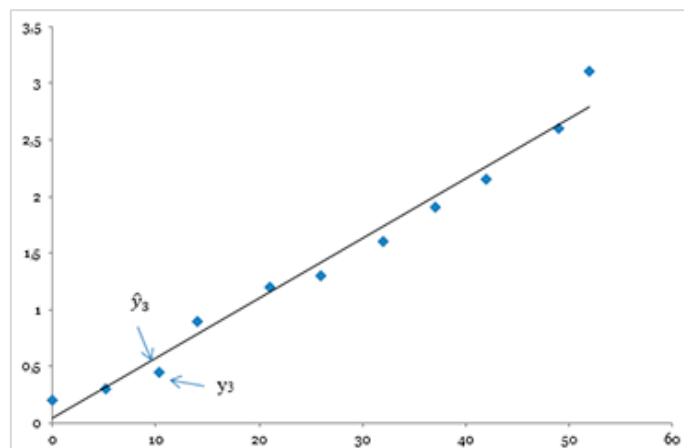


Gráfico 3. Recta de ajustada a la nube de puntos.

Es preciso establecer un nuevo símbolo para los valores de y estimados, es decir, los que caen justo en la recta de regresión. Tal valor se simboliza mediante \hat{y} .

En el gráfico 3 se puede ver que el valor de y_3 es el real y el \hat{y}_3 es el estimado por la recta de ajuste. La recta de estimación se expresa de la siguiente manera:

$$\hat{y} = a + bx$$

4.

En determinadas ocasiones no resulta sencillo dilucidar cuál es la recta que supone un mejor ajuste para una nube de puntos, principalmente en los casos en los que los mismos se hallan muy dispersos. En la siguiente se muestran dos líneas de ajuste para los mismos puntos. El error hace referencia a la diferencia entre los valores de cada punto y con los estimados por la recta, \hat{y} .

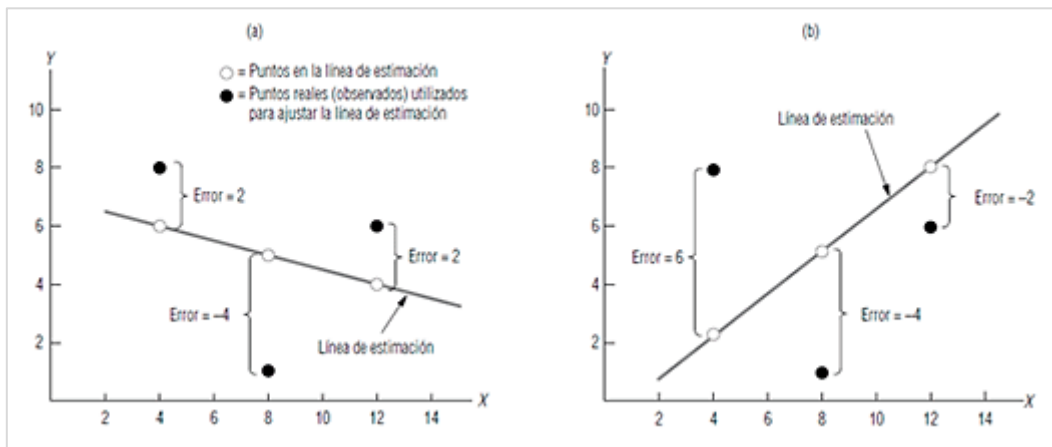


Figura 1. Errores obtenidos en dos posibles rectas de regresión para una misma nube de puntos. Fuente: *Estadística para Administración y Economía* de Levin y Rubin.

Esas diferencias entre los valores observados y los estimados por la recta se llaman residuos. En la gráfica A de la figura 1 tales residuos se han calculado de la siguiente manera:

$$y_1 - \hat{y}_1 = 8 - 6 = 2$$

$$y_2 - \hat{y}_2 = 1 - 5 = -4$$

$$y_3 - \hat{y}_3 = 6 - 4 = 2$$

En la gráfica B los residuos se han obtenido como:

$$y_1 - \hat{y}_1 = 8 - 2 = 6$$

$$y_2 - \hat{y}_2 = 1 - 5 = -4$$

$$y_3 - \hat{y}_3 = 6 - 8 = -2$$

Podría resultar lógico pensar que la suma de todos los residuos proporciona el error total y que por lo tanto, la recta que más error tenga supondrá un peor ajuste. Tal y como se puede observar, al sumar todos los residuos en ambos gráficos el error total es nulo para los dos. Este hecho significaría que ambas gráficas serían un buen ajuste ya que minimizan los errores hasta el punto de hacerlos nulos.

Ahora bien, no solo hay que tener en cuenta que el error total sea nulo, es importante que los errores individuales sean lo menor posible. En la gráfica A el valor que más se aleja un punto del estimado es en 4 unidades, mientras que en la B es 6. Se puede concluir, por lo tanto, que el procedimiento de suma de todos los residuos no es el adecuado.

Se podría suponer que una manera de solucionar este problema y conseguir saber de manera numérica cuál de las dos rectas supone un mejor ajuste es calculando los valores absolutos de los residuos. En el caso de la gráfica A la suma de todos los residuos como valores absolutos da un resultado de 8, mientras que en la gráfica B da 12. De esta manera, se predice que en la gráfica B los valores reales están más distantes de los estimados que en la gráfica A, es decir, la gráfica A supondría un mejor ajuste ya que el error absoluto total es menor.

Sin embargo, si observamos el siguiente ejemplo y calculamos los residuos y sus valores absolutos se tiene que:

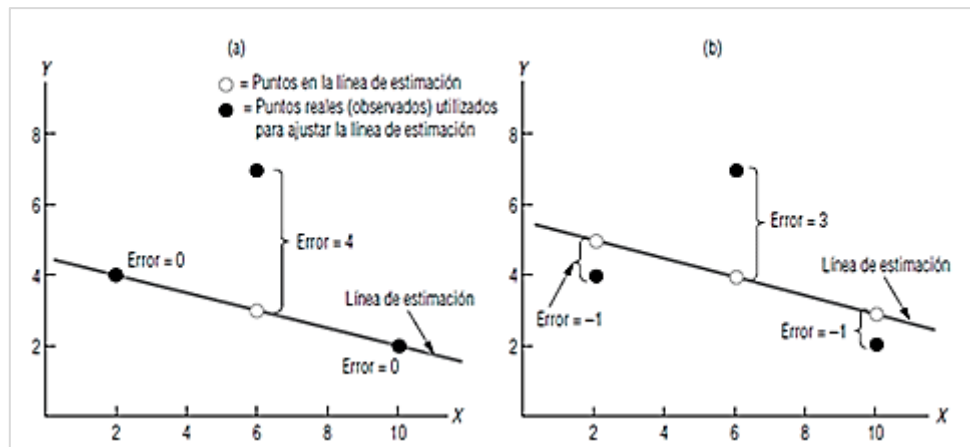


Figura 2. Rectas de estimación para corroborar que los errores absolutos no solventan el problema. Fuente: *Estadística para Administración y Economía* de Levin y Rubin.

En la gráfica A los residuos son:

$$y_1 - \hat{y}_1 = 4 - 4 = 0$$

$$y_2 - \hat{y}_2 = 7 - 3 = 4$$

$$y_3 - \hat{y}_3 = 2 - 2 = 0$$

En la gráfica B los residuos son:

$$y_1 - \hat{y}_1 = 4 - 5 = -1$$

$$y_2 - \hat{y}_2 = 7 - 4 = 3$$

$$y_3 - \hat{y}_3 = 2 - 3 = -1$$

La suma de los residuos en valor absoluto da un resultado de 4 en la gráfica A y de 5 en la gráfica B por lo que podría suponerse que la A supone un mejor ajuste. Sin embargo, a simple vista se puede observar que la gráfica A no tiene ninguna consideración hacia el punto 2 de coordenadas (6,7), mientras que en la B se ha desplazado de los puntos 1 y 3 para considerar al punto 2. La gráfica A ha hecho una recta perfecta con dos puntos, sin tener en cuenta el del medio por lo que el hecho de tomar valores absolutos tampoco soluciona el problema.

La siguiente consideración se basa en elevar al cuadrado los residuos antes de sumarlos para calcular el error total. De esta manera, las magnitudes se expresaran siempre en positivo y la suma no las anulará, como sucedía en la primera hipótesis. Por otro lado, se da más valor a los puntos que están más alejados de los estimados para que su error se intensifique y solucionar el problema que planteaba la hipótesis de los valores absolutos.

Esta nueva hipótesis da el nombre de regresión por mínimos cuadrados puesto que se pretende hallar la recta que minimice los cuadrados de los residuos. Comprobando lo que pasaría si aplicamos este nuevo planteamiento a los gráficos de la figura 2, en la gráfica A la suma de los cuadrados de los residuos es 16 y en la gráfica B es 11. Se puede demostrar que la gráfica B supone un mejor ajuste que la A, puesto que al elevar al cuadrado se ha intensificado el valor del residuo del punto 2.

La cuestión es que en la práctica las nubes de puntos no constan de un valor pequeño de los mismos, más bien la mayor parte de ajustes involucran muchos valores. Es preciso encontrar la recta que suponga un mejor ajuste para todos esos puntos y en esto se basa el método de mínimos cuadrados. Mediante este método se puede calcular el valor de la pendiente y la ordenada en el origen de la mejor estimación y para obtenerlas hay que tener en cuenta la siguiente consideración:

$$\sum_{j=1}^n (y_j - \hat{y}_j)^2 \text{ M\u00ednimo}$$

5.

Si se sustituye el valor de \hat{y} de la ecuaci\u00f3n 4 se obtiene que:

$$\sum_{j=1}^n (y_j - a - bx_i)^2 \text{ M\u00ednimo}$$

6.

Si llamamos a la función $\varphi(a, b)$ y considerando el hecho de que el error ha de ser mínimo, es preciso hacer las derivadas parciales tal y como se muestra a continuación:

$$\frac{\partial \varphi(a, b)}{\partial a} = -2 \sum_1^n (y_j - a - bx_i) = 0$$

7.

$$\frac{\partial \varphi(a, b)}{\partial b} = -2 \sum_1^n (y_j - a - bx_i)x_i = 0$$

8.

Despejando se obtiene:

$$\sum_1^n y_j = \sum_1^n a + b \sum_1^n x_i$$

9.

$$\sum_1^n y_j x_i = \sum_1^n ax_i + b \sum_1^n x_i^2$$

10.

Considerando que:

$$\sum_1^n a = na,$$

$$\sum_1^n ax_i = a \sum_1^n x_i,$$

$$\sum_1^n (x_i + y_j) = \sum_1^n x_i + \sum_1^n y_j$$

Se comprueba que:

$$\sum_1^n y_j = na + b \sum_1^n x_i$$

11.

$$\sum_1^n y_j x_i = a \sum_1^n x_i + b \sum_1^n x_i^2$$

12.

Al dividir la ecuación 11 entre n se llega a:

$$\frac{\sum_1^n y_j}{n} = a + \frac{b \sum_1^n x_i}{n} \rightarrow \bar{y} = a + b\bar{x}$$

13.

Despejando se saca el valor de la ordenada en el origen:

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

14.

Al sustituir en la ecuación 12 se obtiene que:

$$\sum_1^n y_j x_i = (\bar{y} - b\bar{x}) \sum_1^n x_i + b \sum_1^n x_i^2$$

15.

La pendiente resulta:

$$b = \frac{\sum_1^n y_j x_i - \bar{y} \sum_1^n x_i}{\sum_1^n x_i^2 - \bar{x} \sum_1^n x_i}$$

16.

Haciendo las pertinentes operaciones se llega a:

$$b = \frac{\sum_1^n (y_j - \bar{y})(x_i - \bar{x})}{\sum_1^n (x_i - \bar{x})^2}$$

17.

Al dividir arriba y abajo entre $n-1$ se puede expresar la pendiente en función de la covarianza y la varianza de la variable independiente.

$$b = \frac{S_{xy}}{S_x^2}$$

S_{xy} es la covarianza de las variables y S_x^2 es la varianza de la variable independiente.

18.

Donde S_{xy} es la covarianza de las variables y S_x^2 es la varianza de la variable independiente.

A partir de las ecuaciones 14 y 18 se obtienen los valores de la ordenada en el origen y la pendiente de la recta que supone el mejor ajuste puesto que minimiza los cuadrados de los residuos. El valor de la pendiente se le llama coeficiente de regresión de y sobre x .

La ecuación de la recta de regresión que mejor ajuste supone es:

$$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$$

19.

Donde \bar{x} e \bar{y} son los valores medios de x y de y respectivamente, S_{xy} es la covarianza y S_x^2 es la varianza de x .

Así mismo, también podría estimarse la recta de regresión de x sobre y de la siguiente manera:

$$x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y})$$

20.

La ordenada en el origen y el coeficiente de regresión lineal de x sobre y son:

$$a' = \bar{x} - b' \bar{y}$$

21.

$$b' = \frac{S_{xy}}{S_y^2}$$

22.

Ejemplo 2. Ecuación de la recta por mínimos cuadrados

Un jugador de baloncesto ha encestado las siguientes canastas desde diferentes distancias, escribir la recta de regresión obtenida por el método de mínimos cuadrados.

Distancia	Canastas
2,5	23
3,2	21
4,7	19
6,2	15
8,7	12
13,2	8
17	4
18,5	1

Se construye la siguiente tabla:

x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
2,50	-6,75	45,56	23	10,125	-68,34
3,20	-6,05	36,60	21	8,125	-49,16
4,70	-4,55	20,70	19	6,125	-27,87
6,20	-3,05	9,30	15	2,125	-6,48
8,70	-0,55	0,30	12	-0,875	0,48
13,20	3,95	15,60	8	-4,875	-19,26
17,00	7,75	60,06	4	-8,875	-68,78
18,50	9,25	85,56	1	-11,875	-109,84
$\bar{x} = 9,25$		273,70	$\bar{y} = 12,875$		-349,25

La covarianza S_{xy} es:

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{8-1} \cdot (-349,25) = -49,89$$

La varianza de x es:

$$S_x^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{273,70}{8-1} = 39,1$$

Sustituyendo en la ecuación de la recta se obtiene:

$$y - 12,87 = \frac{-49,89}{39,1}(x - 9,25)$$

Operando se saca la ecuación de la recta:

$$y = -1,28x + 24,68$$

Debido a que la covarianza es negativa la relación entre las variables es inversa y la pendiente es negativa.

Mínimos cuadrados con la calculadora

1. Borrar los datos memorizados:

Pulsar *SHIFT* + *CLR* + tecla 1 (*SCL*) = o tecla 3 (*All*).

2. Regresión lineal

Pulse *MODE* dos veces. Pulsar 2 (*REG*) y 1 (*lin*).

3. Introducir los datos:

Escribir el dato x_1 y pulsar “,” meter el dato y_1 y pulsar *M+* (*data*). Repetir para los demás datos. En *REPLAY* se pueden comprobar los datos y las frecuencias.

4. Obtener el valor de a y b:

Pulsar *SHIFT* + *S-VAR* (tecla del número 2) + *REPLAY* dos veces derecha. Ordenada en el origen tecla 1, pendiente tecla 2.

Interpolación en la recta de regresión

Tal y como se ha mencionado anteriormente, la recta de regresión se emplea para hacer predicciones. Este procedimiento se basa en interpolar valores de x tomados para sacar los valores de y . Cabe destacar que los valores de x han de estar entre el rango de los valores tomados para hacer la recta de regresión y que fuera de ese rango no podemos estar seguros de que la asociación entre las variables se mantenga.

Interpolación en la calculadora

Una vez que la calculadora tiene los datos de la pendiente y la ordenada en el origen:

1. Escribir el dato de x que se desea interpolar.
2. **SHIFT**+ **S-VAR** (tecla del número 2) y avanzar a la derecha en **REPLAY** hasta llegar a \hat{y} + tecla 2.

Error estándar de la estimación

Una vez determinada la ecuación de la recta de regresión es preciso cerciorarse de que tal estimación es buena. Tal estudio se hace mediante el error estándar de la estimación que da idea de la confiabilidad de la estimación llevada a cabo. Se trata de un parámetro de significado similar a la desviación estándar que evalúa lo que los datos se dispersan de la recta de regresión obtenida. Su cálculo se lleva a cabo a partir de la siguiente fórmula:

$$S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$$

$y - \hat{y}$ son los residuos de la variable dependiente y n el total de puntos.

23.

Cuando se eleva el error de la estimación al cuadrado se obtiene el error cuadrático medio, S_e^2 .

Ejemplo 3. Cálculo del error estándar de estimación

A partir de los datos de la tabla del ejemplo 2 calcular el error estándar de la estimación. ¿A qué distancia ha metido menos canastas de las esperadas por el análisis?

Se hace una tabla en la que se calculan los valores estimados de la variable dependiente \hat{y} , en función de los valores de x . A continuación, se calculan los residuos y, por último, se elevan al cuadrado.

x	y	$\hat{y} = -1,28x + 24,68$	$y - \hat{y}$	$(y - \hat{y})^2$
2,5	23	21,48	1,52	2,31
3,2	21	20,59	0,41	0,17
4,7	19	18,67	0,33	0,11
6,2	15	16,76	-1,76	3,10
8,7	12	13,56	-1,56	2,43
13,2	8	7,81	0,19	0,04
17	4	2,95	1,05	1,10
18,5	1	1,04	-0,04	0,002
				9,26

El error de estimación es:

$$S_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} = \sqrt{\frac{9,26}{8 - 2}} = 1,24$$

Si el error de estimación es nulo, es debido a que los residuos son nulos, ya que los datos estimados coinciden con los observados. En este caso se ha obtenido un error de estimación de 1,24 que lleva a la conclusión de la que la asociación entre las variables no es perfecta. A 6,2 metros ha encestado el menor número de canastas de las esperadas por el análisis.

El error de la estimación permite calcular los intervalos de confianza de la estimación. Si suponemos que los puntos observados siguen un modelo de distribución normal se puede suponer, del mismo modo que en el Tema 8 se vio con la desviación estándar, que el 68% de los puntos están a $\pm 1 S_e$, el 95,5% a $\pm 2 S_e$ y el 99,7 a $\pm 3 S_e$. En la siguiente figura del libro *Estadística para Administración y Economía* de Levin, R. L., Rubin, D.S., se pueden ver los *intervalos de confianza para la estimación* en función de los errores estándar de la estimación.

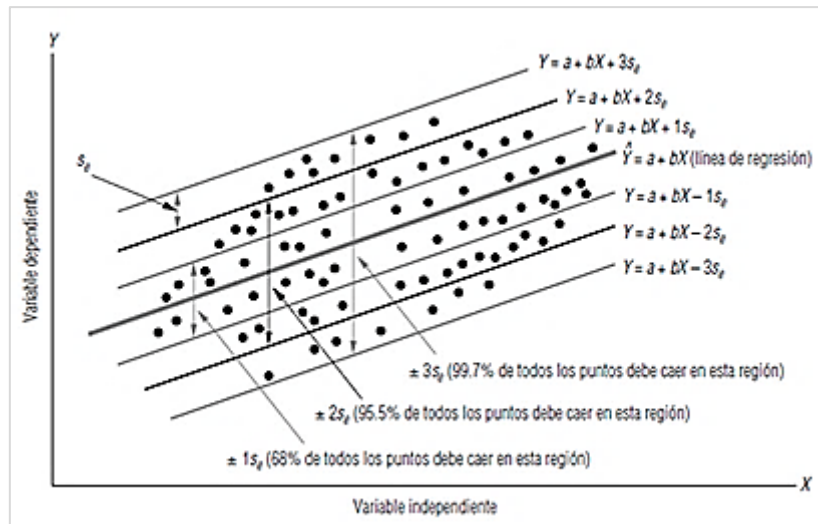


Figura 3. Distribución normal de los valores de y observados alrededor de la recta de regresión.

Fuente: *Estadística para Administración y Economía* de Levin y Rubin.

De la misma manera se puede calcular cualquier otra probabilidad mediante la estandarización con el parámetro Z.

Ejemplo 4. Cálculo de los límites de los intervalos de confianza para la estimación

Para los datos del ejemplo 2 calcular el intervalo de confianza para un valor a una distancia de 7,8 metros y a un nivel del 95,5% de significancia.

El valor estimado de y a partir de la ecuación de la recta con $x=7,8$ es:

$$\hat{y} = -1,28x + 24,68 = -1,28 \cdot 7,8 + 24,68 = 14,71$$

El 95,5% de los datos caerán a ± 2 Se por lo que los límites inferior y superior del intervalo de confianza serán:

$$\hat{y} - 2 \cdot Se = 14,71 - 2 \cdot 1,24 = 12,23$$

$$\hat{y} + 2 \cdot Se = 14,71 + 2 \cdot 1,24 = 17,19$$

A una distancia de 7,8 se espera que encesté entre 12 y 17 veces a un nivel de significancia del 95,5%.

Coeficiente de determinación

Ya hemos visto la fuerza de asociación entre variables mediante el uso de varios tipos de coeficientes en función de la naturaleza de tales variables. Ahora bien, a la hora de hacer un ajuste por regresión lineal el cálculo de un coeficiente que nos dé información sobre el grado de asociación entre las variables es de gran utilidad. Tal parámetro es el coeficiente de determinación y depende de la variación de los valores de y frente a los estimados, $(y - \hat{y})^2$ y frente a la media $(y - \bar{y})^2$. Se calcula a partir de la siguiente fórmula:

$$r^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

24.

r^2 siempre es positivo y 1 cuando se trata de correlación perfecta entre variables. Tal caso de correlación perfecta será cuando $(y - \hat{y})^2 / (y - \bar{y})^2$ sea cero, es decir, cuando los valores observados caigan sobre la recta obtenida. Por el contrario, cuando no haya correlación alguna el coeficiente de determinación será cero y tal valor se obtendrá cuando el cociente entre la diferencia de los valores observados y estimados de y la diferencia entre los observados y el valor medio sea igual.

A partir de la raíz cuadrada del coeficiente de determinación se puede obtener el valor de los coeficientes vistos en otro tema. Que sea el de correlación de Pearson o de Spearman dependerá de la naturaleza de las variables.

$$r = \sqrt{r^2}$$

25.

Ejemplo 5. Coeficiente de determinación

La siguiente tabla recoge las ganancias obtenidas por un hotel en función de los huéspedes de lunes a domingo excluyendo el miércoles.

Día de la semana	Ganancias	Número de huéspedes
Lunes	1200	12
Martes	1500	16
Jueves	3670	31
Viernes	7200	69
Sábado	10090	102
Domingo	5800	56

A. Escribir la recta de regresión y el coeficiente de correlación.

Se construye la siguiente tabla:

x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
12	-35,7	1272,3	1200	-3710	132335,7
16	-31,7	1003	1500	-3410	107994,7
31	-16,7	277,9	3670	-1240	20670,8
69	21,3	455	7200	2290	48845,7
102	54,3	2951,7	10090	5180	281429,4
56	8,3	69,4	5800	890	7413,7
$\bar{x} = 47,7$		6029,3	$\bar{y} = 4910$		598690

La covarianza S_{xy} es:

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{6-1} \cdot (598690) = 119738$$

La varianza de x es:

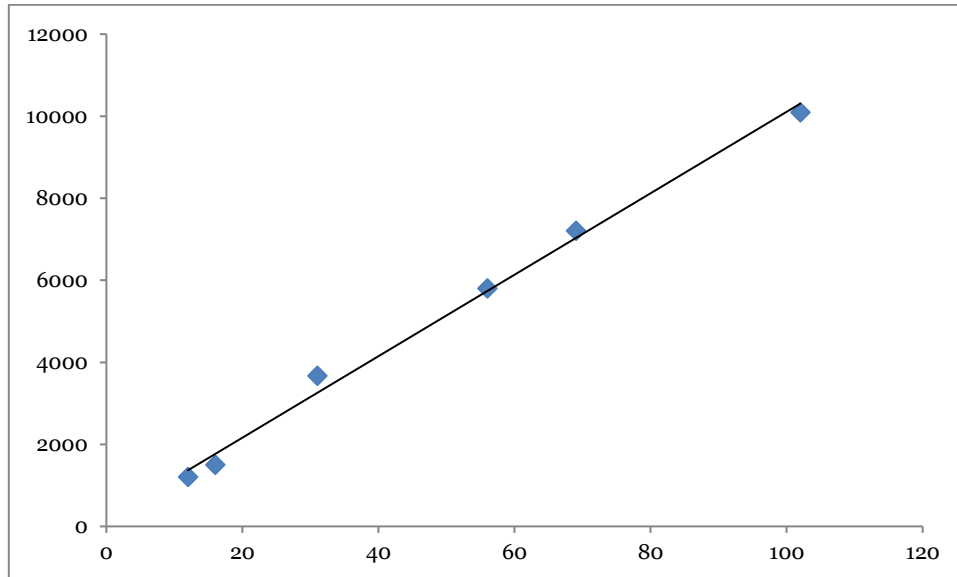
$$S_x^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{6029,3}{6-1} = 1205,9$$

Sustituyendo:

$$y - 4910 = \frac{119738}{1205,9} (x - 47,7)$$

Operando se saca la ecuación de la recta:

$$y = 99,30x + 176,9$$



Para calcular el coeficiente de determinación hacemos la siguiente tabla:

x	y	$\hat{y} = 99,30x + 176,9$	$y - \hat{y}$	$(y - \hat{y})^2$	$y - \bar{y}$	$(y - \bar{y})^2$
12	1200	1365,5	-165,5	27390,2	-3710	13764100
16	1500	1765,7	-265,7	70596,5	-3410	11628100
31	3670	3255,2	414,8	172059	-1240	1537600
69	7200	7028,6	171,4	29378	2290	5244100
102	10090	10305,5	-215,5	46440,2	5180	26832400
56	5800	5737,7	62,3	3881,3	890	792100
				345864		59798400

$$r^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{345864}{59798400} = 0,9942$$

El coeficiente de correlación es:

$$r = \sqrt{r^2} = 0,997$$

B. Calcular el error estándar de la estimación.

$$S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{345864}{6 - 2}} = 294,05$$

C. ¿Entre qué valores hubiera oscilado la ganancia si el miércoles hubiese habido 25 huéspedes a un nivel de significancia del 99,7%?

Interpolando en la recta para $x=25$ huéspedes se tiene un valor estimado de y de 2659 euros de ganancia. A un nivel de significancia del 99,7%:

$$\hat{y} - 3 \cdot Se = 2659 - 3 \cdot 294,05 = 1776,85$$

$$\hat{y} + 3 \cdot Se = 2659 + 3 \cdot 294,05 = 3541,15$$

El 99,7 % de las ganancias estarán entre 1776,85 y 3541,15 Euros.

8.4. Otros tipos de regresión

En este apartado se introducirá brevemente la regresión múltiple que se lleva a cabo cuando se emplea más de una variable independiente para estimar una variable dependiente. Tiene la ventaja de que permite usar más información disponible para estimar la variable y su procedimiento es similar a la regresión lineal simple puesto que se basa en obtener la ecuación de regresión, el error de la estimación y hacer un análisis de correlación.

Cabe destacar que la regresión múltiple sirve tanto para ajustar rectas como curvas, e incluso con la técnica de valores ficticios se pueden incluir variables cualitativas.

En este apartado haremos una breve descripción de la ecuación de regresión múltiple y el error estándar de la estimación con tan solo dos variables independientes. La tarea es harto laboriosa y no merece la pena hacer los cálculos a mano puesto que se dispone de paquetes estadísticos que son capaces de trabajar con muchas variables independientes.

La ecuación de estimación con dos variables independientes resulta:

$$\hat{y} = a + b_1x_1 + b_2x_2$$

26.

Donde x_1 y x_2 son los valores de cada uno de las dos variables independientes y b_1 y b_2 los coeficientes de regresión.

Del mismo modo que la recta de regresión simple es una línea recta, la ecuación de regresión múltiple es un plano tal y como se muestra en la siguiente figura del libro *Estadística para Administración y Economía* de Levin, R. L., Rubin, D.S.

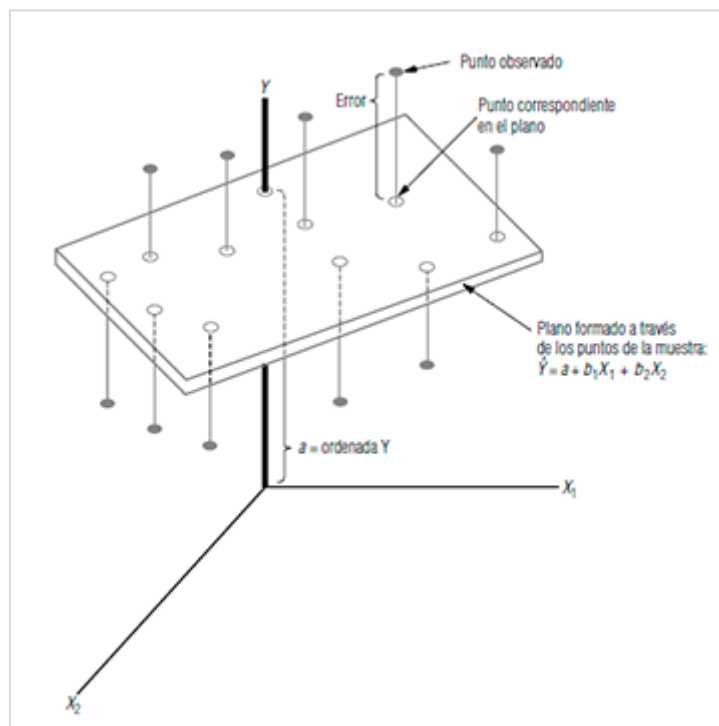


Figura 4. Plano de regresión múltiple con 10 datos. Fuente: *Estadística para Administración y Economía* de Levin y Rubin.

Tal y como se puede observar, algunos puntos caen por encima y otros por debajo del plano, del mismo modo que en la recta de regresión simple caen por arriba y por debajo de la recta. Lo que se pretende es saber cuál es el plano que mejor ajusta los puntos mediante el método de mínimos cuadrados. El mejor plano será aquel que minimice la suma de los cuadrados de los errores.

Para calcular la ecuación del plano se precisará de los datos problema y de las siguientes ecuaciones normales:

$$\sum y = na + b_1 \sum x_1 + b_2 \sum x_2$$

27

$$\sum x_1 y = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2$$

28

$$\sum x_2 y = a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2$$

29.

El error estándar de la estimación se puede calcular a partir de la siguiente fórmula:

$$S_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - k - 1}}$$

n es el número de puntos y k el número de variables independientes, en este caso dos.

30.

El coeficiente de determinación múltiple es la fracción de variación total de la variable y su fórmula es la misma que para la regresión lineal simple:

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

31.

8.5. Tipos de residuos

Tal y como se ha mencionado, un residuo es la diferencia entre un valor observado y uno estimado:

$$e = y - \hat{y}$$

32.

En el caso de los residuos ordinarios, se establece que la media de los residuos es cero:

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = 0$$

33.

La varianza se estima de manera aproximada de la siguiente manera:

$$S^2 = \frac{1}{n - k - 1} \sum_{i=1}^n (e_i - \bar{e})^2 = S_e^2$$

34.

Los residuos se pueden estandarizar de manera que tengan media cero y desviación estándar 1:

$$d_i = \frac{e_i}{\sqrt{S_e^2}}$$

35.

Se puede comprobar que:

$$S^2(e_i) = S^2(1 - h_{ii})$$

h_{ii} es el i-ésimo elemento de la matriz proyección de los residuos.

36.

El uso de residuos como estimadores de los errores requiere una mejoría de la estandarización. Puesto que el valor de h_{ii} va entre 0 y 1, el hecho de usar el error

cuadrático medio para estimar la varianza es una sobreestimación. Se recomienda trabajar con residuos estudentizados de la siguiente manera:

$$r_i = \frac{e_i}{\sqrt{Se^2(1 - h_{ii})}}$$

37.

Además, en el caso de la regresión lineal simple se cumple que el valor h_{ii} es una medida para localizar el i -ésimo punto x_i respecto al punto medio:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

38.

En el caso de la regresión lineal múltiple el valor h_{ii} se calcula como:

$$h_{ii} = \frac{1}{n} [1 + (x_i - \bar{x}) S_{xx} (x_i - \bar{x})] = \frac{1}{n} (1 - D_i^2)$$

D_i es la distancia de Mahalanobis.

39

De esta manera, la varianza de un error e_i depende de la posición a la que se halle un valor x_i . La varianza de los puntos cercanos al valor medio es mayor que la de los alejados y por lo tanto supone un pobre ajuste por mínimos cuadrados.

Es preferible trabajar con residuos estudentizados ya que con los estandarizados u ordinales el valor del residuo en sí es menor y por lo tanto más difícil de detectar en aquellos casos que se escapen del modelo.

Los residuos estudentizados son los que se emplean para el análisis de valores atípicos, los cuales tienen residuos estudentizados mayores que 2. Es preciso estudiar los valores atípicos para saber si se deben meter o no en la muestra, ya que pueden cambiar los parámetros de la regresión bruscamente. En la regresión lineal simple es bastante sencillo averiguar si hay algún valor atípico ya que se puede ver en el diagrama de dispersión pero en la regresión lineal múltiple es más complicado saberlo, por ello es recomendable estudiar aquellos casos que tengan un residuo estudentizado elevado.

Lo + recomendado

No dejes de leer...

Uso del análisis de regresión y correlación, limitaciones errores y advertencias

Levin, R. y Rubin, D. (2004). Uso del análisis de regresión y correlación: limitaciones, errores y advertencias. En *Estadística para administración y economía* (pp. 551-552).

Este apartado muestra ciertas consideraciones a tener en cuenta sobre el análisis por correlación.

Accede al artículo a través del aula virtual o desde la siguiente dirección web:

http://www.academia.edu/9701898/Estad%C3%ADstica_para_Administraci%C3%B3n_y_Econom%C3%ADa_7ma_Edici%C3%B3n_-_Richard_I._Levin_and_David_S._Rubin

La fecundidad en España, 1996-2006: mujeres de nacionalidad española frente a extranjeras

Luque, M. A. y Bueno-Cavanillas, A. (2009). La fecundidad en España, 1996-2006: mujeres de nacionalidad española frente a extranjeras. *Gaceta sanitaria*, 23 (1), 67-71.

Este artículo expone un ejemplo práctico de uso de la regresión lineal simple en el ámbito social.

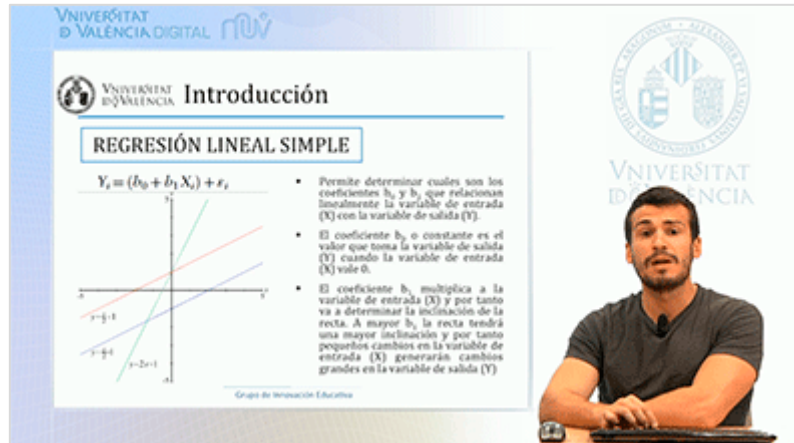
Accede al artículo a través del aula virtual o desde la siguiente dirección web:

<http://publicationslist.org/data/fmiguelangel/ref-7/MA%20Luque%20Fernandez%20FertilityinSpain,1996%E2%80%932006%20foreign%20versus%20spanish%20women%20Gac%20Sanit%202009.pdf>

No dejes de ver...

Introducción I - Regresión lineal simple y múltiple (parte 1 de 5)

En este video se recoge una introducción a la regresión lineal.



UNIVERSITAT DE VALÈNCIA DIGITAL

UNIVERSITAT DE VALÈNCIA

Introducción

REGRESIÓN LINEAL SIMPLE

$$Y_i = (b_0 + b_1 X_i) + e_i$$

- Permite determinar cuáles son los coeficientes b_0 y b_1 que relacionan linealmente la variable de entrada (X) con la variable de salida (Y).
- El coeficiente b_0 , o constante, es el valor que toma la variable de salida (Y) cuando la variable de entrada (X) vale 0.
- El coeficiente b_1 multiplica a la variable de entrada (X) y por tanto va a determinar la inclinación de la recta. A mayor b_1 , la recta tendrá una mayor inclinación y por tanto pequeños cambios en la variable de entrada (X) generarán cambios grandes en la variable de salida (Y).

Grupo de Innovación Educativa

Accede al vídeo a través del aula virtual o desde la siguiente dirección web:

http://mmedia.uv.es/buildhtml?user=carmurma&name=IntroI.mp4&path=/cream/Proyectos_Docentic/statisTIC/Xavier/S2/

+ Información

A fondo

El modelo de regresión simple: estimación y propiedades

En el apéndice 2.1 del artículo se muestra el desarrollo matemático para obtener la ecuación 17.

Accede al artículo a través del aula virtual o desde la siguiente dirección web:

<http://www.uv.es/uriel/2%20El%20modelo%20de%20regresion%20lineal%20simple%20estimacion%20y%20propiedades.pdf>

Regresión múltiple y modelado

Este tema explica de manera más detallada la regresión múltiple.

Accede al artículo a través del aula virtual o desde la siguiente dirección web:

http://www.academia.edu/9701898/Estad%C3%ADstica_para_Administraci%C3%B3n_y_Econom%C3%ADa_7ma_Edici%C3%B3n_-_Richard_I._Levin_and_David_S._Rubin

Bibliografía

García, M. (1995). *Socioestadística. Introducción a la estadística en sociología*. Madrid: Alianza Editorial.

León, O. y Montero, I. (1997). *Diseño de investigaciones*. Madrid: McGraw-Hill.

Levin, R. y Rubin, D. (2004). *Estadística para administración y economía*. México: Pearson.

Test

1. El coeficiente de determinación es:

x	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85
y	2,12	4,36	6,71	9,12	10,71	13,53	15,36	17,71

- A. 0,999.
- B. 0,998.
- C. -0,999.
- D. -0,998.

2. El coeficiente de regresión es:

x	10	15	20	22	30	32
y	4	8	12	14	16	20

- A. 0,66.
- B. 0,997.
- C. 0,995.
- D. 1,84.

3. El coeficiente de determinación es:

x	10	15	20	22	30	32
y	4	8	12	14	16	20

- A. 0,97.
- B. 0,95.
- C. 0,66.
- D. 1,84.

4. Marque la afirmación correcta:

- A. El coeficiente de determinación puede ser negativo.
- B. Regresión es lo mismo que correlación.
- C. El coeficiente de correlación lineal es la raíz cuadrada del coeficiente de determinación.
- D. B y C son verdaderas.

5. A partir de los datos de la siguiente tabla, si los valores de x aumentase diez veces y los valores de y se redujesen a la mitad, ¿cuál es la respuesta correcta?

x	10	15	20	22	30	32
y	4	8	12	14	16	20

- A. La ordenada en el origen cambiaría y la pendiente no.
- B. La pendiente cambiaría y la ordenada en el origen no.
- C. R no variaría.
- D. B y C son ciertas.

6. A partir de los siguientes datos, x es:

Concentración	Señal
0,01	12
0,02	27,3
0,05	26,5
0,1	47
0,3	61,35
x	36

- A. 38255.
- B. Ninguno de esos valores.
- C. 0,033.
- D. No se puede saber.

7. A partir de los siguientes datos:

x	y
0	0,402
10	0,584
20	0,756
30	0,934
40	1,108

- A. El coeficiente de determinación y de correlación son iguales.
- B. Para $x=25$ y es 0,84.
- C. El error de estimación es 0,0002.
- D. Todas son verdaderas.

8. A partir de los siguientes datos, el intervalo de confianza (95,5%) para $x=25$ es:

x	y
0	0,402
10	0,584
20	0,756
30	0,934
40	1,108

- A. 0,8440-0,8448.
- B. 0,8441-0,8449.
- C. -0,8440-(-0,8448).
- D. -0,8441- (-0,8449).

9. Marca la opción falsa:

- A. El cuadrado del error estándar de la estimación es el error cuadrático medio.
- B. En la regresión lineal podemos interpolar los valores de x y los de y .
- C. Los residuos se calculan como las diferencias entre los valores de x y los estimados.
- D. Todas son verdaderas.

10. Marca la afirmación verdadera sobre la regresión lineal múltiple:

- A. La ecuación es la ecuación de un plano para cualquier número de variables independientes.
- B. Se basa en ajuste por mínimos cuadrados.
- C. El coeficiente de determinación se calcula a partir de la misma ecuación de la recta que en la regresión lineal simple.
- D. Todas son ciertas.