8

ASOCIACIÓN ENTRE DOS VARIABLES

8.1 INTRODUCCIÓN

En el campo de la investigación biológica surge frecuentemente la necesidad de establecer si dos o más variables están relacionadas o asociadas. La mayoría de las veces esta correspondencia no es un hecho evidente y se hace necesario examinar el comportamiento conjunto de los valores de las variables involucradas para establecer si existe algún tipo de asociación entre ellas. Por ejemplo, es posible que en algunos estudios se tenga interés en saber si existe relación entre la altura de una planta y la concentración de algún nutriente en el suelo; la talla y el peso de los individuos de cierta especie; la frecuencia cardiaca y la concentración de una droga en la sangre; la densidad de una población y el tiempo de vida de sus miembros, etc. Son varias las técnicas estadísticas que pueden usarse para establecer la naturaleza, intensidad y tipo de asociación entre variables como las de los ejemplos mencionados anteriormente. Sin embargo, los dos métodos estadísticos de asociación de variables más utilizados en la biología son el Análisis de Correlación y el Análisis de Regresión. En éste capítulo, se estudiarán ambos métodos.

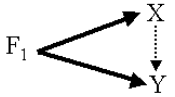
8.2 NATURALEZA DE LA ASOCIACIÓN ENTRE VARIABLES

Aunque son numerosas las razones por las cuales dos variables biológicas se relacionan, Sokal y Rolf (1980) señalan algunas formas básicas de asociación frecuentemente encontradas en los estudios biológicos.

1.- La variable X es la causa única de la variación de otra variable Y.

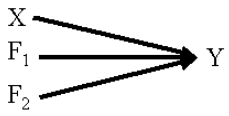
Esta situación no es frecuente, sin embargo algunos ejemplos son los siguientes: a) Edad de un pez (X) y número de anillos de crecimiento en las escamas (Y); b) Concentración de una sustancia (X) en una solución y el color desarrollado por la solución (Y), y c) Concentración de sólidos disueltos (X) y conductividad del agua (Y)

2.- Las variables X e Y están relacionadas porque dependen de un factor común (F1)



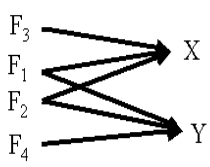
Estos casos tampoco son frecuentes. Se puede poner como ejemplo la relación que puede existir entre el peso (X) y la talla (Y) de un organismo, cuya asociación suele depender de la edad (F_1)

3.- La variable X es una de las varias causas de Y



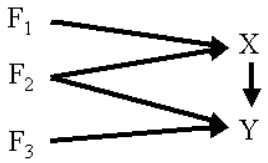
Este modelo es más frecuente que los anteriores. Por ejemplo, para muchos organismos la frecuencia cardiaca (Y) además de depender del peso (Y) depende de otros factores como la edad (F_1) y la temperatura (F_2)

4.- La relación entre la variables X e Y depende de varias causas comunes o no.



Posiblemente, estas situaciones complejas sean las más comunes en la naturaleza. Como ejemplo se puede tomar el caso de los estudios morfométricos en insectos, donde el tamaño de dos estructuras, como la cabeza (X) y la longitud del fémur de una pata (Y) depende de varios factores comunes como la edad (F_1), el peso o la talla (F_2) y de factores específicos como su función (F_3), la posición (F_4), etc.

5.- La variable X afecta directamente a la variable Y, estando ambas afectadas por una variable común.



Esta caso representa una situación de mayor complejidad. Por ejemplo, en muchos animales al aumentar el tiempo de desarrollo (X) incrementa el peso (Y), sin embargo un tercer factor como el incremento en el número de cría (F_2) disminuye el peso (Y) y aumenta el tiempo de desarrollo (X). De modo que la variable común tiende a contrarrestar la acción directa entre las variables X e Y.

8.3 FORMAS DE EVALUAR UNA ASOCIACIÓN

8.3.1 Evaluación gráfica

Una manera rápida de determinar la eventual asociación entre dos variables es representar en un eje cartesiano los valores de las dos variables estudiadas y marcar con un punto la intersección de sus coordenadas. Esto produce una dispersión de los pares de valores en el plano. La tendencia que siga todo el grupo de puntos puede indicar si existe asociación entre las variables y la naturaleza de ésta asociación. La asociación será positiva si al aumentar los valores de una variable aumentan los de la otra (Figura 8.1A y B); o será negativa si al

aumentar una variable la otra disminuye (Figura 8.1C). Si no hay asociación no se manifiesta ninguna tendencia (Figura 8.1D). Una desventaja de le evaluación gráfica es que requiere un número grande de pares de valores de las variables X e Y.

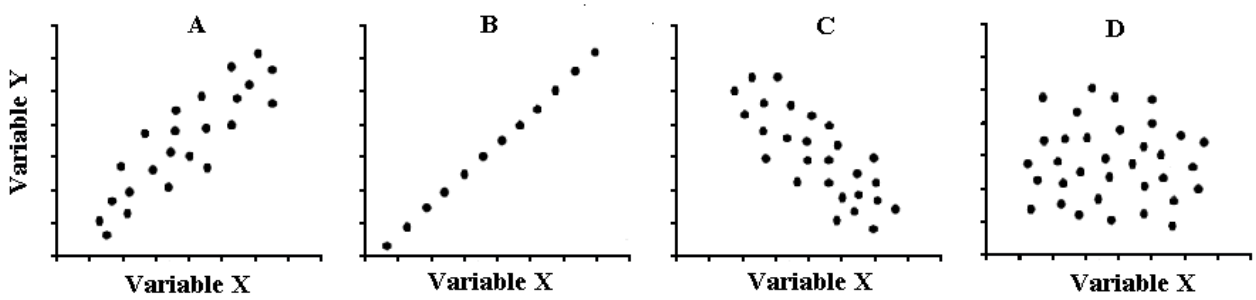


Figura 8.1. Asociación entre dos variables. A) positiva; B) positiva perfecta; C) negativa; D) sin asociación.

8.3.2 Evaluación Estadística

La evaluación estadística de una asociación entre dos variables permite dos cosas: i) decidir objetivamente si existe asociación entre dos variables, ii) determinar la naturaleza, intensidad y forma de la asociación. El Análisis de Correlación y el Análisis de Regresión son los dos los métodos estadísticos más utilizados en la biología para lograr los dos propósitos antes señalados.

8.4 ANÁLISIS DE CORRELACIÓN

El análisis de correlación sirve para demostrar si dos variables aleatorias están asociadas y cuantificar la intensidad de dicha asociación. Este método supone que la asociación entre las variables X e Y produce una distribución bivariada. Veamos como se origina este tipo de distribuciones.

Supóngase que a 10 personas se les midió la altura y que la distribución de valores es la que se muestra en la Tabla 8.1.

Tabla 8.1. Frecuencia de tallas en un grupo de 10 personas

Talla (m)	Frecuencia	Frecuencia relativa
1,68	1	0,1
1,69	2	0,2
1,70	4	0,4
1,71	2	0,2
1,72	1	0,1

La distribución de tallas de la Tabla 8.1 se puede representar mediante un histograma como el de la Figura 8.2

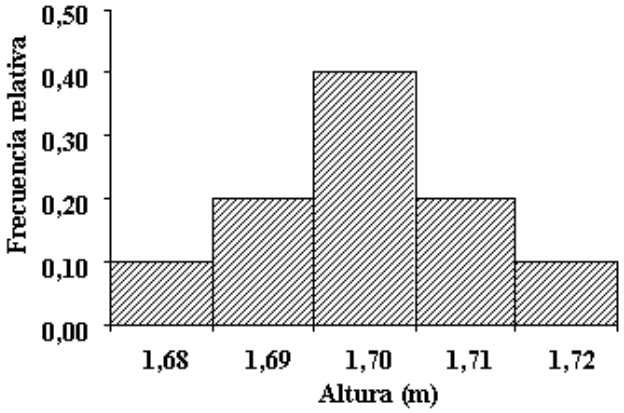


Figura 8.2. Histograma para la distribución de tallas

La distribución de frecuencias relativas de la Figura 8.2 representa la distribución de probabilidades de la variable talla. Como es obvio se trata de la distribución probabilística de una sola variable. Supóngase nuevamente que para cada una de las personas también se registró el peso en kilogramos (Tabla 8.2).

Tabla 8.2. Frecuencia de talla y peso en un grupo de 10 personas

Talla (m)	Peso	Frecuencia absoluta	Frecuencia relativa
1,68	65	1	0,1
1,69	66	2	0,2
1,70	67	4	0,4
1,71	68	2	0,2
1,72	69	1	0,1

La distribución de probabilidades de las tallas y pesos estaría formada por 25 valores, tal y como se muestra en la Tabla 8.3.

Tabla 8.3. Distribución conjunta de frecuencias relativas de la talla y el peso de 10 personas

		Talla (x)					
		1,68	1,69	1,70	1,71	1,72	fr(y)
Peso (y)	65	**0,1**	0,0	0,0	0,0	0,0	0,1
	66	0,0	**0,2**	0,0	0,0	0,0	0,2
	67	0,0	0,0	**0,4**	0,0	0,0	0,4
	68	0,0	0,0	0,0	**0,2**	0,0	0,2
	69	0,0	0,0	0,0	0,0	**0,1**	0,1
	fr(x)	0,1	0,2	0,4	0,2	0,1	1,0

En la tabla anterior se puede obtener fácilmente la probabilidad de que un determinado par de valores de talla y peso ocurran conjuntamente. Por ejemplo la probabilidad de que ocurran simultáneamente los eventos Talla = 1,69 y Peso = 66 es 0, 2. Una representación más formal es la siguiente: P(x = 1,69; y = 66) = 0,2. Otros valores de probabilidad son:

$$P(x = 1{,}70;\ y = 66) = 0{,}0;\ \ P(x = 1{,}71;\ y = 68 = 0{,}2);\ \ P(x = 1{,}72;\ y = 68) = 0{,}0$$

La distribución conjunta o bivariada del histograma de frecuencias relativas de la talla y el peso de la Tabla 8.3 se muestra en la Figura 8.3.

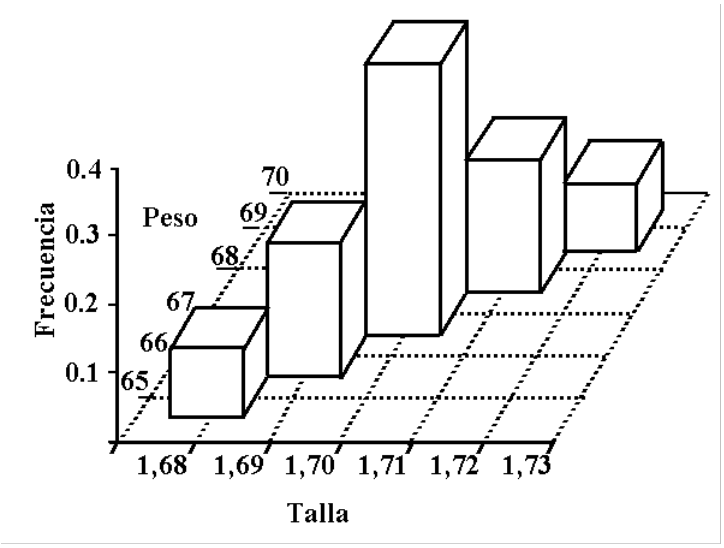


Figura 8.3.

El histograma anterior se construyó con los 16 valores de la Tabla 8.3. En este caso debido al bajo número de observaciones a cada valor de X o de Y le corresponde un solo dato. Sin embargo si se incrementa el número de datos se produce un aumento en el número de observaciones para cada valor de X o de Y. Como ejemplo en la Tabla 8.4, se muestra la talla y el peso de 111 personas donde se observa que a cada valor de X le corresponden varios valores de Y e igualmente a cada valor de Y le corresponden varios valores de X.

Tabla 8.4. Frecuencia de la talla y el peso de 111 personas

		Talla					
		1,68	**1,69**	**1,70**	**1,71**	**1,72**	**1,73**
	65	0	1	2	1	0	0
	66	2	3	4	3	2	0
	67	3	5	7	5	3	2
Peso	**68**	4	6	8	6	4	2
	69	2	4	9	4	2	0
	70	0	3	7	3	0	0
	71	0	1	2	1	0	0

El histograma que se origina de los datos de frecuencia de la tabla anterior se muestra en la Figura 8.4.

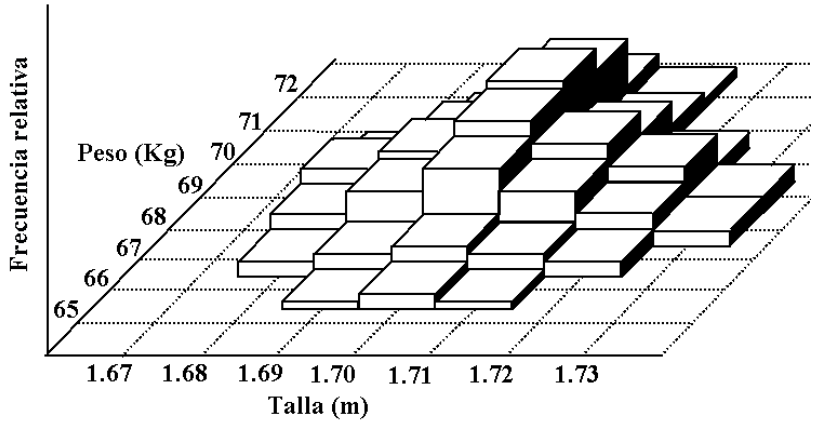


Figura 8.4.

Al seguir incrementando el número de valores para las dos variables la forma del histograma se suaviza cada vez más, hasta que finalmente cuando el número de pares de observaciones sea infinito el histograma de frecuencias tendrá una forma acampanada (Figura 8.5).

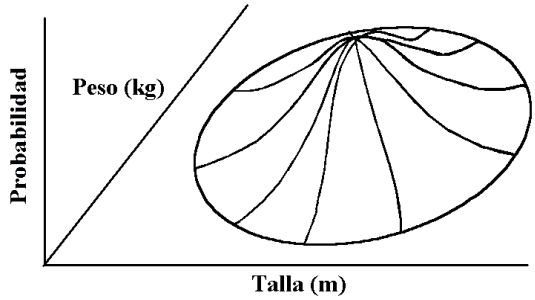


Figura 8.5.

Esta distribución bivariada se caracteriza porque para cada valor de x (talla) existe una subpoblación de valores de y (peso) que se distribuyen normalmente (Figura 8.6).

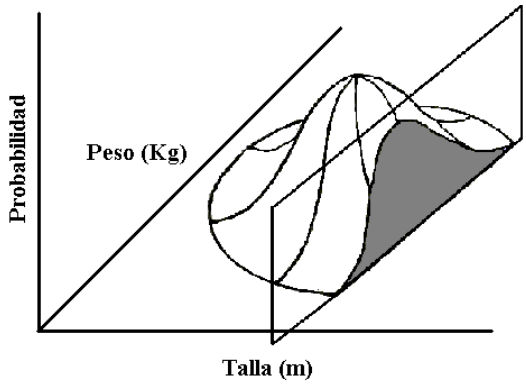


Figura 8.6.

Igualmente para cada valor de y (peso) existe una subpoblación de valores de x (talla) distribuidos normalmente (Figura 8.7).

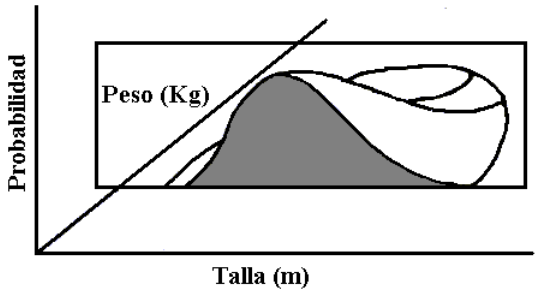


Figura 8.7.

Matemáticamente la distribución normal bivariada es descrita por la función de probabilidades siguiente:

$$f(x,y)=\frac{e^{-\frac{1}{2(1-\rho_{xy})^2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho_{xy}\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}}$$

Dicha función está caracterizada por los parámetros siguientes:

$\mu_x =$ *media de la variable aleatoria X.* $\mu_y =$ *media de la variable aleatoria Y.*

$\sigma_x =$ *desviación de la variable aleatoria X.* $\sigma_y =$ *desviación de la variable aleatoria Y.*

$\rho_{xy} =$ *coeficiente de correlación para las dos variables.* $\sigma_{xy} =$ *covarianza de x e y*

Conociendo los parámetros y usando la función anterior, es posible calcular la probabilidad de ocurrencia de un dado par de valores (x,y).

8.4.1 Coeficiente de Correlación

El parámetro ρ_{xy} se denomina coeficiente de correlación y puede considerarse como una medida del grado de relación entre X e Y. El valor del coeficiente de correlación es igual a:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

donde:

σ_{xy} = *covarianza de x e y.*

$\sigma_x =$ *desviación de la variable aleatoria X.*

$\sigma_y =$ *desviación de la variable aleatoria Y.*

El coeficiente ρ_{xy} tiene las características siguientes:

1. Su valor varía entre –1 y +1 $(-1 \le \rho_{xy} \le +1)$.
2. Cuando dos variables no están correlacionadas el coeficiente $\rho_{xy} = 0$.

3. Si las variables X e Y están correlacionadas el coeficiente $\rho_{xy} \neq 0$.
4. Si $0 \leq \rho_{xy} \leq +1$ se dice que las variables X e Y está correlacionadas positivamente.
5. Si $-1 \leq \rho_{xy} \leq 0$ se dice que las variables X e Y están correlacionadas negativamente.
6. Si $\rho_{xy} = +1$, se dice que existe una correlación positiva perfecta entre X e Y. Todos los puntos del diagrama de dispersión están ubicados sobre una línea con pendiente positiva.
7. Si $\rho_{xy} = -1$, se dice que existe una correlación negativa perfecta entre X e Y. Todos los puntos del diagrama de dispersión están ubicados sobre una línea con pendiente negativa.

8.4.2 Estimación de ρ_{xy}.

El parámetro ρ_{xy} es estimado por el estadístico,

$$r = \frac{S_{xy}}{S_x S_y}$$

denominado coeficiente de correlación muestral. S_{xy} es el estimador de la covarianza entre X e Y. Igualmente S_x y S_y son respectivamente los estimadores de las desviaciones de las variables X e Y.

La expresión $r = S_{xy} / S_x S_y$ puede simplificarse si la covarianza y desviaciones se consideran en términos de las sumas de cuadrados. Así se tiene que sí:

$$S_{xy} = \frac{SPxy}{n-1}; \qquad S_x = \sqrt{\frac{SCx}{n-1}}; \qquad S_y = \sqrt{\frac{SCy}{n-1}}$$

entonces: $$r = \frac{S_{xy}}{S_x S_y} = \frac{\frac{SP_{xy}}{n-1}}{\sqrt{\frac{SC_x}{n-1}}\sqrt{\frac{SC_y}{n-1}}} = \frac{SP_{xy}}{n-1\sqrt{\frac{SC_x SC_y}{(n-1)^2}}} = \frac{SP_{xy}}{\frac{n-1}{n-1}\sqrt{SC_x SC_y}} = \frac{SP_{xy}}{\sqrt{SC_x SC_y}}$$

En resumen:

$$r = \frac{SP_{xy}}{\sqrt{SC_x SC_y}}$$

donde:

$$SP_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}; \qquad SC_x = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}; \qquad SC_y = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

De modo que para calcular r sólo es necesario conocer los valores de $SP_{xy}; SC_x \;\; y \;\; SC_y$.

Ejemplo 8.1.
Se desea conocer el valor del coeficiente de correlación entre la Conductividad y la Dureza del agua, si los valores obtenidos para las dos variables en mediciones efectuadas en 8 ríos fueron las siguientes:

Conductividad (mμ)	**Dureza (mg/L $CaCO_3$)**
30	12,5
40	20,6
60	29,5
90	50,1
50	25,5
100	30,6
150	61,2
20	10,7

En primer lugar es necesario obtener los valores siguientes:

$$\sum_{i=1}^{n} x_i = 540\,; \quad \sum_{i=1}^{n} x_i^2 = 49600\,; \quad \sum_{i=1}^{n} y_i = 240{,}7\,; \quad \sum_{i=1}^{n} y_i^2 = 9407{,}41\,; \quad \sum_{i=1}^{n} x_i y_i = 21207{,}0$$

Luego se calculan los valores de las respectivas sumas de cuadrados: $SP_{xy}\,;\ SC_x \ \ y \ \ SC_y$

$$SP_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = 21207 - \frac{(540)(240{,}7)}{8} = 4959{,}75$$

$$SC_x = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 49600 - \frac{(540)^2}{8} = 13150{,}00$$

$$SC_y = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 9407{,}41 - \frac{(240{,}7)^2}{8} = 2165{,}3481$$

Finalmente se obtiene el valor de r.

$$r = \frac{SP_{xy}}{\sqrt{SC_x SC_y}} = \frac{4959{,}75}{\sqrt{(13150)(2165{,}3481)}} = \frac{4959{,}745}{5336{,}134} = 0{,}93$$

Dado que el valor de $r = 0{,}93$ es cercano a +1, se supone que entre la conductividad y la dureza del agua existe una correlación alta.

Ejemplo 8.2.
En los mismos ocho ríos del ejemplo anterior se quiere saber si la inclinación del cauce y la velocidad de la corriente están correlacionadas. Los datos obtenidos para las dos variables fueron los siguientes:

Pendiente (%)	**Velocidad (m/seg)**
5,0	0,55
13,5	0,65
16,0	0,59
3,5	0,60
7,0	0,30
11,0	0,40
3,0	0,40
15,0	0,69

Datos necesarios:

$$\sum_{i=1}^{n} x_i = 74{,}0\,; \quad \sum_{i=1}^{n} x_i^2 = 879{,}5\,; \quad \sum_{i=1}^{n} y_i = 4{,}18\,; \quad \sum_{i=1}^{n} y_i^2 = 2{,}319\,; \quad \sum_{i=1}^{n} x_i y_i = 41{,}1150$$

Cálculo de las sumas de cuadrados:

$$SP_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = 41{,}1150 - \frac{(74)(4{,}18)}{8} = 2{,}45$$

$$SC_x = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 879{,}5 - \frac{(74)^2}{8} = 195$$

$$SC_y = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 2{,}319 - \frac{(4{,}18)^2}{8} = 0{,}1352$$

Cálculo del valor de r:

$$r = \frac{SP_{xy}}{\sqrt{SC_x SC_y}} = \frac{2{,}45}{\sqrt{(195)(0{,}1352)}} = \frac{2{,}45}{5{,}135} = 0{,}48$$

Este valor parece indicar que la pendiente del cauce y la velocidad de la corriente están correlacionados positivamente, aunque con menos fuerza que las dos variables del ejemplo anterior cuyo $r = 0,93$.

8.4.3 Prueba de significación para *r*

Se ha comprobado empíricamente que cuando el tamaño de la muestra de pares de valores es muy grande $(n > 100)$, son independientes las dos variables estudiadas si el valor del coeficiente de correlación se encuentra en el intervalo $-3 \leq r \leq +3$. Por el contrario, valores fuera de este rango indican asociación. Sin embargo, para tamaños de muestras menores a 100 es necesario disponer de algún criterio estadístico que determine si el valor del coeficiente de correlación muestral se desvía significativamente de cero.

Puesto que r es un estadístico y por lo tanto una variable cuyo valor cambia para diferentes muestras, se puede recurrir a una prueba de hipótesis, para determinar si la desviación de r del valor cero confirma la asociación entre las variables y no es debida a la aleatoriedad del muestreo. A fin de desarrollar la prueba de hipótesis es necesario que se conozcan las propiedades de la distribución probabilística de la variable *r*.

En este sentido se sabe que cuando no existe correlación entre las dos variables estudiadas, es decir que $\rho = 0$, el estadístico r sigue una distribución de *t* con *n-2* grados de libertad (Figura 8.8), alrededor del valor cero y con una desviación igual a $S_r = \sqrt{1-r^2} \big/ \sqrt{n-2}$

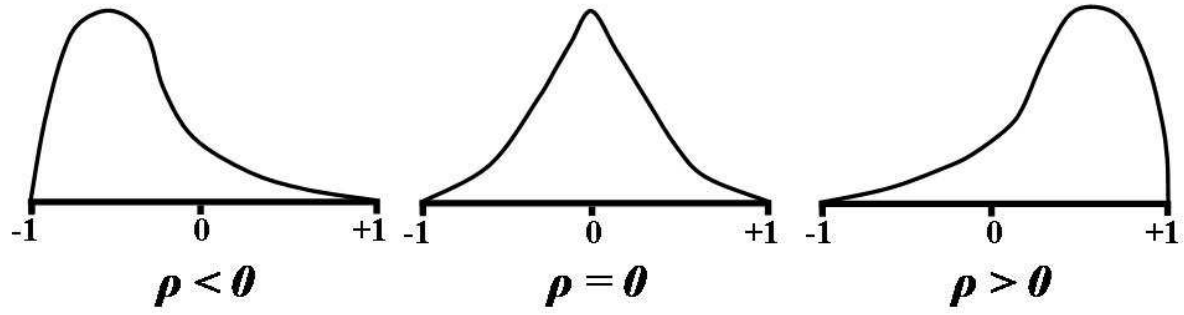


Figura 8.8. Distribución de r para diferentes valores de ρ.

Conociendo la distribución de probabilidades de r y su desviación es fácil someter a prueba la hipótesis nula de no existencia de correlación entre X e Y. El procedimiento es similar a los usados en el capítulo 6 y su manejo lo demostraremos con los datos del ejemplo 8.1, cuyo coeficiente de correlación muestral entre la conductividad y la dureza fue igual a $r = 0,93$.

8.4.3.1 Prueba de hipótesis para $\rho = 0$

a. Formular las hipótesis

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

b. Especificar un valor de probabilidad crítico o nivel de significación. Ante la ausencia de una especificación particular, se puede escoger como nivel de significación un valor de $\alpha = 0,05$.

c. Elegir el estadístico de la muestra y su distribución para someter a prueba las hipótesis. Bajo el supuesto que $\rho = 0$ la variable r se distribuye como t y el estadístico de prueba es:

$$t = \frac{r-\rho}{S_r} = \frac{r-\rho}{\frac{\sqrt{1-r^2}}{\sqrt{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

d. Establecer la zona de aceptación o rechazo para Ho. Como $H_1 : \rho \neq 0$ se trata de una prueba de dos colas. La zona de aceptación de H_0 es la siguiente:

$$ZA = \left\{T \,/\, -t_{(1-\alpha/2;n-2)} < T < +t_{(1-\alpha/2;n-2)}\right\}$$

e. Calcular el valor del estadístico de prueba y los valores críticos de la zona de aceptación.

$$t = \frac{r-\rho}{S_r} = \frac{r-\rho}{\frac{\sqrt{1-r^2}}{\sqrt{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,93\sqrt{8-2}}{\sqrt{1-(0,93)^2}} = 6,198$$

$$ZA == \left\{T \,/\, -t_{(0,975;6)} < T < +t_{(0,975;6)}\right\} = \left\{T \,/\, -2,45 < T < +2,45\right\}$$

f. Tomar la decisión. Como $t = 6,198 > t_{(0,975;6)} = 2,44$ el valor del estadístico de prueba se encuentra fuera de la zona de aceptación de Ho. Por lo tanto se concluye que los datos proporcionan suficiente evidencia para rechazar Ho y que las variables conductividad y dureza del agua están correlacionadas.

Ejemplo 8.3.
Comprobar si el valor de $r = 0,48$ calculado en el ejemplo 8.2 se diferencia significativamente del valor cero.

g. Formular las hipótesis

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

h. Especificar un valor de probabilidad crítico o nivel de significación. Ante la ausencia de una especificación particular, se puede escoger como nivel de significación $\alpha = 0,05$.

i. Elegir el estadístico de la muestra y su distribución para someter a prueba las hipótesis. Bajo el supuesto que $\rho = 0$ la variable r se distribuye como t y el estadístico de prueba es:

$$t = \frac{r-\rho}{S_r} = \frac{r-\rho}{\frac{\sqrt{1-r^2}}{\sqrt{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

j. Establecer la zona de aceptación o rechazo para H_0. Al ser $H_1 : \rho \neq 0$, se trata de una prueba de dos colas. La zona de aceptación de H_0 es la siguiente:

$$ZA = \left\{ T \,/\, -t_{(1-\alpha/2;n-2)} < T < +t_{(1-\alpha/2;n-2)} \right\}$$

k. Calcular el valor del estadístico de prueba y los valores críticos de la zona de aceptación.

$$t = \frac{r-\rho}{S_r} = \frac{r-\rho}{\frac{\sqrt{1-r^2}}{\sqrt{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}48\sqrt{8-2}}{\sqrt{1-(0{,}48)^2}} = 1{,}34$$

$$ZA = \left\{ T \,/\, -t_{(0{,}975;6)} < T < +t_{(0{,}975;6)} \right\} = \left\{ T \,/\, -2{,}45 < T < +2{,}45 \right\}$$

l. Tomar la decisión. Como $-2{,}45 < t = 1{,}34 < +2{,}45$ el valor del estadístico de prueba se encuentra dentro de la zona de aceptación de Ho. Por lo tanto los datos no proporcionan suficiente evidencia para rechazar Ho y se concluye que las variables pendiente y velocidad del agua no están correlacionadas.

Este último ejemplo demuestra que es necesario tener cuidado de no tomar ninguna decisión acerca del coeficiente de correlación mientras no se pruebe si es significativamente diferente de cero. El valor de $r = 0{,}48$ parece suficientemente grande y alejado de cero y se podría estar tentado a aceptar que existe correlación débil, entre las dos variables. Sin embargo, cuando hay muy pocos datos $(n < 20)$, si ρ es muy bajo, el coeficiente de correlación muestral *r* es muy variable, de modo que la prueba de hipótesis es muy exigente o prudente para indicar la existencia de correlación.

8.4.3.2 Prueba de hipótesis para $\rho = \rho_0$

El procedimiento de prueba de hipótesis descrito anteriormente sólo es válido bajo la suposición de no existencia de correlación entre las variables, es decir que $\rho = 0$. Si se quiere probar que *r* es igual a cualquier otro valor $\rho = \rho_0$, la distribución muestral de r no obedece la distribución de *t* y presenta un sesgo que hace la distribución asimétrica (Figura 8.8). Sin embargo, es posible usar la denominada transformación *Z* o de Fisher, la cual convierte la distribución asimétrica de *r* en una distribución aproximadamente normal. La transformación se hace de la manera siguiente:

$$Z_r = \frac{1}{2} ln \frac{1+r}{1-r}$$

Esta nueva variable Z_r se caracteriza porque el promedio o valor esperado es igual a: $E_{(Z_r)} = \frac{1}{2} ln \frac{1+\rho}{1-\rho}$ y su desviación es igual a $\sigma_{Z_r} = 1/\sqrt{n-3}$. Conociendo estos valores es posible efectuar la prueba de hipótesis para ρ.

Con el propósito de evitar el uso de los logaritmos naturales se han construido tablas donde conociendo el valor de r o ρ se encuentran directamente los valores de Z_r o de Z_ρ.

Ejemplo 8.4.
Un investigador sospecha que el coeficiente de correlación entre dos variables es $\rho = 0,6$. Para verificar esta hipótesis tomó 50 pares de observaciones y calculó el coeficiente de correlación muestral cuyo valor fue $r = 0,70$. ¿ Será el valor de *r* diferente al predicho por el investigador?

1. Hipótesis: $H_0 : \rho = 0,6$
   $H_1 : \rho \neq 0,6$

2. Nivel de significación: $\alpha = 0,05$

3. Estadístico de prueba: $Z = \frac{Z_r - Z_\rho}{S_{Z_r}}$

4. Zona de aceptación: $ZA : \{Z / -z_{(1-\alpha/2)} < Z < +z_{(1-\alpha/2)}\}$

5. Cálculos:
   $Z_r = \frac{1}{2} ln \frac{1+r}{1-r} = \frac{1}{2} ln \frac{1+0,7}{1-0,7} = 0,8673$

   $Z_\rho = \frac{1}{2} ln \frac{1+0,6}{1-0,6} = 0,6932$

   $S_{Z_r} = 1/\sqrt{n-3} = 1/\sqrt{50-3} = 0,146$

   $Z = \frac{Z_r - Z_\rho}{S_{Z_r}} = \frac{0,8673 - 0,6932}{0,146} = 1,19$

   $ZA : \{Z / -z_{(0,975)} < Z < +z_{(0,975)}\} = \{Z / -1,96 < Z < +1,96\}$

6. Decisión: Como el valor de $Z = 1,19$ se encuentra dentro de los límites críticos de la zona de aceptación no existe suficiente evidencia para rechazar $H_{0,,}$ lo que apoya la hipótesis del investigador.

La transformación de Fisher se emplea preferentemente cuando el tamaño de la muestra es superior a 25. Para muestras de tamaño $10 \le n \le 25$ se puede emplear el método de Hotteling, cuyo estadístico de prueba es:

$$Z^* = \frac{z^* - \zeta^*}{\frac{1}{\sqrt{n-1}}}$$

Donde: $z^* = z_r - \frac{3z_r + r}{4n}$ y $\zeta^* = z_\rho - \frac{3z_\rho + \rho}{4n}$

8.4.3.3 Prueba de hipótesis para dos coeficientes de correlación *($\rho_1 = \rho_2$)*

También es posible someter a prueba la diferencia entre dos coeficientes de correlación usando el estadístico de prueba:

$$Z = \frac{\left(Z_{r_2} - Z_{r_1}\right) - \left(Z_{\rho_2} - Z_{\rho_1}\right)}{\sqrt{\frac{1}{n_2 - 3} + \frac{1}{n_1 - 3}}}$$

Este estadístico sigue la distribución normal.

Ejemplo 8.5.

Se ha determinado que durante la época de sequía los valores del coeficiente de correlación (ρ) entre la alcalinidad y la dureza en los ríos andinos de páramo es igual 0,90 y en los ríos andinos de zonas bajas es igual a 0,96. Con el propósito de verificar si en la época de lluvia se mantiene la diferencia entre los coeficientes de correlación para los ríos de ambas zonas, se midieron las dos variables en 30 ríos de cada zona, encontrándose en los ríos de páramo un valor $r_1 = 0{,}76$ y en los ríos de zonas bajas un valor $r_2 = 0{,}84$ ¿Cuál es la conclusión?.

2. Hipótesis: $H_0 : \rho_2 - \rho_1 = 0{,}6$
   $H_1 : \rho_2 - \rho_1 \neq 0{,}6$

7. Nivel de significación: $\alpha = 0{,}05$

8. Estadístico de prueba: $Z = \frac{\left(Z_{r_2} - Z_{r_1}\right) - \left(Z_{\rho_2} - Z_{\rho_1}\right)}{\sqrt{\frac{1}{n_2 - 3} + \frac{1}{n_1 - 3}}}$

9. Zona de aceptación: como se trata de una prueba de dos colas, la zona de aceptación será:

$$ZA : \left\{ Z \,/\, -z_{(1-\alpha/2)} < Z < +z_{(1-\alpha/2)} \right\}$$

10. Cálculos:

$$Z_{\rho_1} = \frac{1}{2} ln \frac{1+0,90}{1-0,90} = 1,47 \qquad Z_{\rho_2} = \frac{1}{2} ln \frac{1+0,96}{1-0,96} = 1,946$$

$$Z_{r_1} = \frac{1}{2} ln \frac{1+0,76}{1-0,76} = 0,996 \qquad Z_{r_2} = \frac{1}{2} ln \frac{1+0,84}{1-0,84} = 1,22$$

$$Z = \frac{\left(Z_{r_2} - Z_{r_1}\right) - \left(Z_{\rho_2} - Z_{\rho_1}\right)}{\sqrt{\frac{1}{n_2 - 3} + \frac{1}{n_1 - 3}}} = \frac{\left(1,22 - 0,996\right) - \left(1,946 - 1,47\right)}{\sqrt{\frac{1}{30-3} + \frac{1}{30-3}}} = \frac{-0,252}{0,272} = -0,926$$

$$ZA: \left\{ Z \,/\, -z_{(0,975)} < Z < +z_{(0,975)} \right\} = \left\{ Z \,/\, -1,96 < Z < +1,96 \right\}$$

11. Decisión: Como el valor de $Z = -0,93$ se encuentra dentro de los límites críticos de la zona de aceptación no existe suficiente evidencia para rechazar H_0, y se concluye que la diferencia entre los índices de correlación de las variables dureza y alcalinidad se mantiene en ambos grupos de ríos.

8.5 ANÁLISIS DE REGRESIÓN

El propósito del análisis de regresión es el de establecer una relación funcional entre dos variables X e Y. Cuando existe esta relación una de las variables cambia en función de la otra que permanece constante. La que cambia se denomina variable dependiente y se identifica con la letra Y, mientras que a la otra se le denomina variable independiente y se identifica con la letra X. Esta última puede ser una variable aleatoria o no, pero debe ser medida con exactitud o controlada por el investigador. El valor que adquiere la variable dependiente Y en una situación determinada será función del valor de la variable independiente X. Esta relación funcional se representa en forma genérica mediante la expresión Y = f(x). La misma puede tener diversas formas específicas (Figura 8.9).

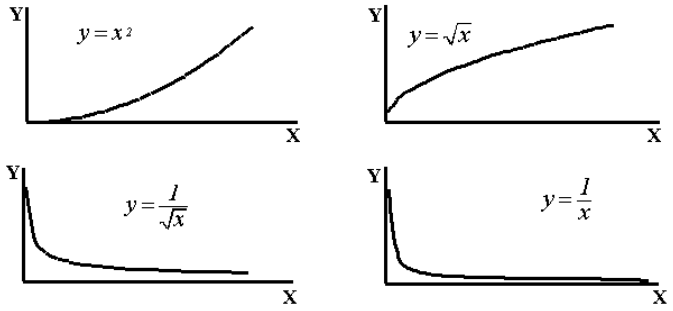


Figura 8.9. Diferentes formas de relación entre x e y

Sin embargo la función más sencilla es la que representa una recta (Figura 8.10) en la cual las variables X e Y están relacionadas linealmente mediante la ecuación $f(x) = a + bx$, caracterizada por dos cantidades constantes: b = valor de la pendiente de la recta y a = valor donde la recta intercepta el eje Y.

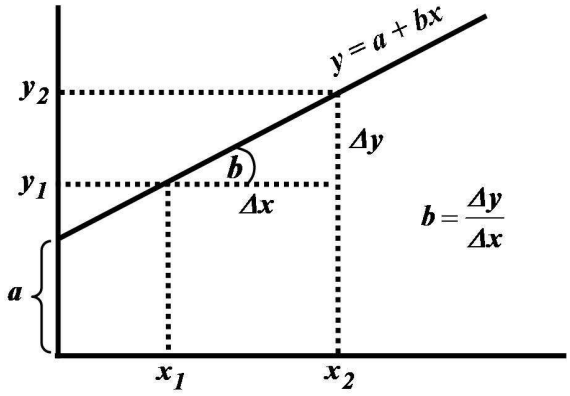


Figura 8.10

8.5.1: Buscando la mejor ecuación de una recta.

Cuando se trabaja con datos provenientes de experimentos o de observaciones de campo y se quiere conocer la función que relaciona linealmente una variable independiente con la variable dependiente, se pueden usar diferentes procedimientos.

Ejemplo 8.6.

Supongamos que en un experimento se obtuvieron los datos siguientes:

X	0	1	2	3	4	5	6	7	8	9
Y	2,0	2,5	2,9	3,6	3,95	4,4	5,1	5,6	5,9	6,0

El conjunto de pares de valores, representado por los puntos de la Figura 8.11, se encuentran alineados, por lo que usando una regla se pueden trazar "al ojo" una línea que una todos los puntos. Del propio gráfico se obtiene que el intercepto a es igual a 2 y la pendiente b es igual a 0,5, de modo que la ecuación de la línea recta es $y = 2 + 0,5x$

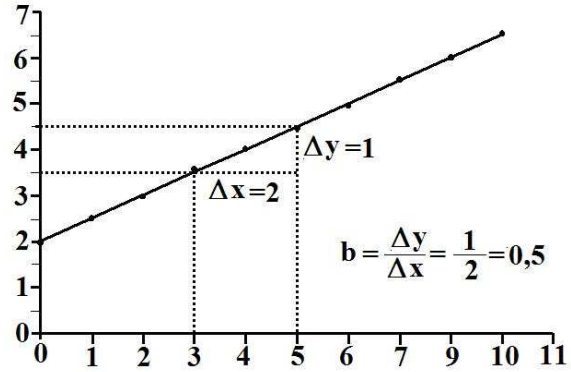


Figura 8.11

Los datos del ejemplo anterior permiten que el trazado de la recta sea muy fácil puesto que todos los puntos se encuentran sobre la misma línea. Sin embargo la mayoría de las veces, especialmente cuando se estudian fenómenos aleatorios, aunque dos variables estén relacionadas linealmente, los puntos que cuantifican la intersección entre las dos variables no se disponen exactamente sobre una línea recta.

Ejemplo 8.7.
Supóngase que un experimento produjo el conjunto de datos siguientes:

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

La dispersión de los puntos no permite delinear una recta que pase por todos ellos (Figura 8.12). Sin embargo se puede trazar "al ojo" una recta a través de la nube de puntos procurando ubicarla lo más cerca posible de cada punto. Una vez dibujada la línea se obtienen los valores del intercepto y la pendiente de la recta. En este caso, el trazado de la línea es muy subjetivo e inexacto. Por ejemplo, diferentes personas, aún usando el mismo criterio pueden perfilar líneas rectas diferentes y consecuentemente distintas ecuaciones, tal y como se muestra en la Figura 8.12. Por otra parte, cuando la nube de puntos es muy grande, no es posible trazar una recta a simple vista.

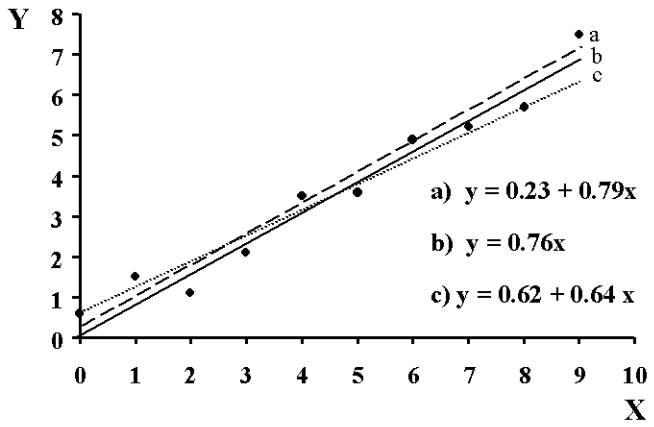


Figura 8.12.

8.5.1.1 Método de los mínimos cuadrados.
Debido a los problemas que se generan con el trazado de rectas "al ojo" es necesario disponer de métodos más objetivos de ajuste. Dentro de este contexto, se entiende por "ajuste" la disminución del error total, es decir la disminución del valor que resulta de la suma de las desviaciones de todas las observaciones a la recta. El método de los mínimos cuadrados tiene como objetivo encontrar los valores de a y b que hacen que la recta $y = a + bx$ se encuentre desviada cierta distancia de cada punto de tal forma que el resultado de la suma de los cuadrados de esas desviaciones sea el menor valor que se pueda obtener. En otras palabras, una vez establecida la ecuación de la recta por el método de los mínimos cuadrados, no existe otra recta que produzca un valor menor para la suma de cuadrados de las desviaciones.

8.5.1.1.1 Cálculo de a y b por el método de los mínimos cuadrados

Si para un dado un conjunto de pares de valores (x_i, y_i) se considera que $y = a + bx$ es una recta cualquiera trazada a través del conjunto de puntos (Figura 8.13), se puede representar la desviación de cada punto a la recta de la forma siguiente:

$$d_i = y_i - y$$

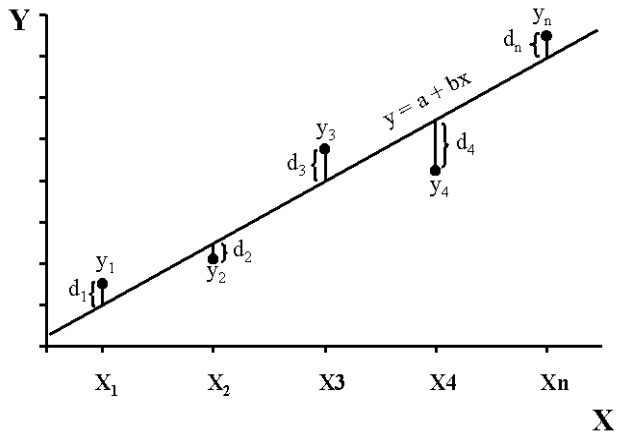


Figura 8.13

El valor d_i será pequeño cuando el punto y_i esté muy cerca de la recta y será mayor en la medida que la recta se aleje de dicho punto. De tal manera que la mejor recta que se puede trazar a través del conjunto de puntos, sería aquella que logre disminuir al mínimo cada valor d_i. Sin embargo en la práctica, debido a lo disperso de los puntos, cualquier intento de acercamiento de la recta a un determinado punto, produce simultáneamente un alejamiento de la recta de otros puntos. Por lo tanto la mejor recta será aquella cuyo trazado se haga de tal manera que la suma de todas las diferencias $d_i = y_i - y$, sea la menor que se pueda encontrar, es decir que

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (y_i - y) = \textit{valor mínimo}$$

Sin embargo como la suma algebraica de diferencias respecto a un valor promedio siempre es igual a cero y la recta es un valor promedio, se tiene que $\sum_{i=1}^{n} (y_i - y) = 0$; por tal razón es preferible trabajar con el cuadrado de las desviaciones y se dice que la mejor recta será aquella cuyo trazado se haga de tal manera que la suma del cuadrado de todas las diferencias $d_i = y_i - y$, sea la menor que se pueda encontrar, es decir que:

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - y)^2 = \textit{valor mínimo}$$

Puesto que $y = a + bx$, la expresión anterior puede cambiarse a la forma:

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \textit{valor mínimo}$$

El problema se reduce ahora a encontrar los valores de los términos *a* y *b* que hacen cumplir la igualdad anterior. Para esto se recurre al cálculo diferencial, por el cual se sabe que una función $f(x)=y$, presenta un valor mínimo para aquellos valores de *x* que hacen la primera derivada igual a cero $(dy/dx=0)$.

Como la función: $\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$, depende de *a* y *b*, su valor mínimo se puede encontrar encontrando las derivadas parciales con respecto a estos términos:

$$1)\ \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0$$

$$2)\ \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0$$

La diferenciación anterior produce el resultado siguiente:

$$1)\ \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = -2\sum_{i=1}^{n}(y_i - a - bx_i)$$

$$2)\ \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i$$

Al igualar las derivadas parciales a cero y reacomodar los términos, se obtienen las denominadas ecuaciones normales:

$$1)\ na + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

$$2)\ a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

Se puede demostrar que los valores de *a* y *b* que resuelven el anterior sistema de ecuaciones son los siguientes:

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}} = \frac{SP_{xy}}{SC_x} \qquad \text{y} \qquad a = \bar{Y} - b\bar{X}$$

Obtenidos los valores *a* y *b*, se sustituyen en $y = a + bx$ quedando así establecida la ecuación de la línea recta que mejor se ajusta al conjunto de puntos.

Ejemplo 8.8.
Establecer la ecuación de la recta que relaciona las variables X-Y del ejemplo 8.7.

Datos necesarios:

$$\sum_{i=1}^{n} x_i = 45,0 \qquad \sum_{i=1}^{n} y_i = 35,0$$

$$\sum_{i=1}^{n} x_i^2 = 285,0 \qquad \sum_{i=1}^{n} y_i^2 = 167,6$$

$$\sum_{i=1}^{n} x_i y_i = 217,5 \qquad \sum_{i=1}^{n} X_i \sum_{i=1}^{n} y_i = 1575,0$$

$$\bar{X} = 4.5 \qquad \bar{Y} = 3.5$$

$$n = 10$$

Sumas de cuadrados:

$$SP_{xy} = \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} = 217,5 - \frac{1575}{10} = 60,0$$

$$SC_x = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} = 285 - \frac{(45)^2}{10} = 82,5$$

$$SC_y = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} = 167,6 - \frac{(35)^2}{10} = 45,10$$

Cálculo de *a* y *b*.

$$b = \frac{SP_{xy}}{SC_x} = \frac{60}{82,5} = 0,727 \qquad a = \bar{Y} - b\bar{X} = 3,5 - 0,727(4,5) = 0,229$$

La ecuación de la recta resultante es $y = 0,229 + 0,727x$ (Figura 8.14). Compare esta ecuación con las obtenidas por el ajuste “al ojo” del ejemplo 8.7.

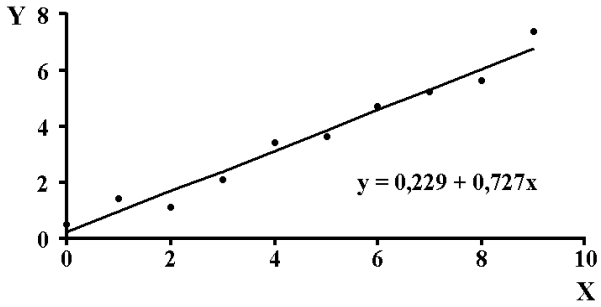


Figura 8.14

8.5.2. Modelos de regresión lineal.

Como se ha visto el método de los mínimos cuadrados sirve para encontrar la ecuación de la recta que relaciona *n* pares de valores de las variables X e Y. Ahora bien, desde el punto de vista estadístico se puede considerar que esta recta es una de las muchas rectas que se pueden generar si se repite numerosas veces el proceso de obtener un valor de Y para cada valor de X. Esta situación la ejemplifican los resultados que se muestran en la Tabla 8.5. Donde se observa que para un mismo valor de la variable peso, que es la variable independiente porque está controlada por el investigador, se midieron cinco valores distintos de la variable dependiente: concentración de glucosa.

Tabla 8.5. Niveles de glucosa en el suero sanguíneo de hombres adultos con diferente peso.

	Glucosa (mg/l)				
Peso (kg)	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
50	83	78	85	74	81
55	82	81	80	85	77
60	95	86	95	85	92
65	89	90	89	79	95
70	104	97	99	105	101
75	101	102	101	99	110
80	115	109	105	101	112
85	116	114	115	111	108
90	128	115	120	110	115
95	135	120	125	131	131

Presumiendo que entre las dos variables existe una relación de causa-efecto, se usó el método de los mínimos cuadrados a fin de calcular, para cada grupo de personas, las ecuaciones de las rectas que relacionan las dos variables estudiadas. Las fórmulas obtenidas fueron las siguientes:

Grupo 1: $y_1 = 19,558 + 1,1758x_1$

Grupo 2: $y_2 = 27,315 + 0,9915x_2$

Grupo 3: $y_3 = 31,976 + 0,9576x_3$

Grupo 4: $y_4 = 20,842 + 1,0642x_4$

Grupo 5: $y_5 = 26,976 + 1,0376x_5$

Cada una de estas ecuaciones representa una recta muestral que estiman una misma recta poblacional, que es aquella que se espera establezca la mejor relación entre las variables X-Y. Por supuesto que la recta poblacional es ideal puesto que para poder encontrarla sería necesario obtener los infinitos valores de la variable Y que se pueden generar para cada valor de la variable X. Si las condiciones del experimento o de observación se mantuvieron mas o menos constantes, el recorrido de las cinco rectas anteriores debe ser muy parecido al de la recta poblacional imaginaria (Figura 8.15). En esta figura hemos supuesto, con un sentido pedagógico, que la recta poblacional sigue el trazado de la línea cortada. Dada la similitud de las ecuaciones de las rectas no es problema escoger cualquiera de ellas y usarla para estimar la verdadera recta. Esto nos deja ver que el análisis de regresión es un método de inferencia estadística.

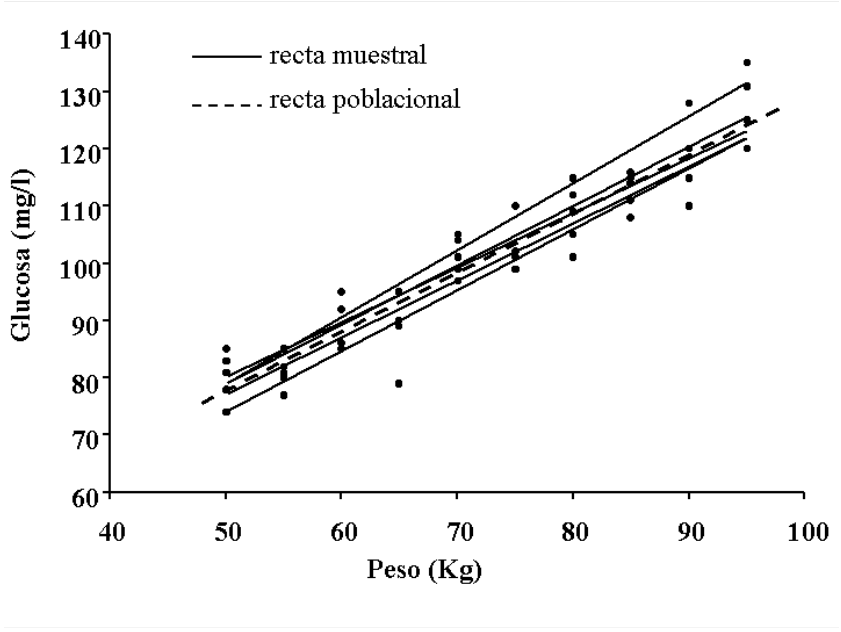


Figura 8.15.

Como se sabe todo proceso de inferencia requiere de algunos supuestos sobre las características de la población, en éste caso de la población de valores de la variable Y. Estos supuestos se encuentran establecidos en los denominados Modelos de Regresión cuyas premisas fundamentales son las siguientes:

8.5.2.1 Modelo I

1. La variable independiente X es fija, sus valores son controlados por el investigador. Esto no significa que la variable X no sea aleatoria, sino que el investigador cambia a voluntad los valores de X y mantiene las condiciones experimentales de tal forma que los cambios en cada nivel de X sean mínimos. Por el contrario la variable Y no es controlada y cambia aleatoriamente para cada valor de X.

2. Para cada valor x_i los valores y_i están distribuidos normalmente, independientemente y con la misma varianza $\sigma^2_{y/x}$ alrededor del valor promedio μ_{y/x_i}. La Figura 8.16. ilustra esta situación.

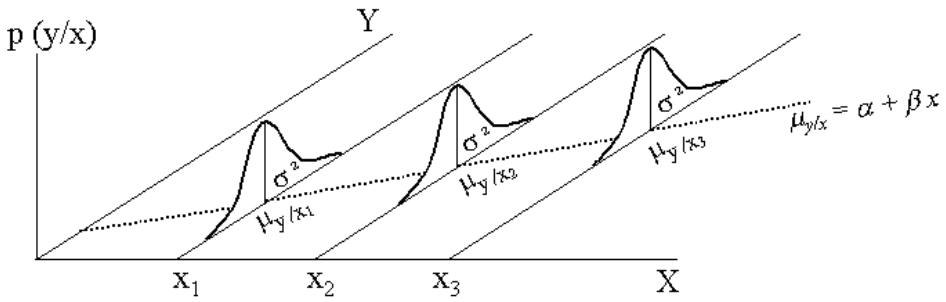


Figura 8.16.

3. Los valores promedios de Y están todos sobre la línea $\mu_{y/x} = \alpha + \beta x$, la cual se denomina recta paramétrica de regresión, donde: μ_{y/x_i} = promedio de la subpoblación de valores *y* para un dado x_i; α = intercepto de la recta con el eje Y; y β = pendiente de la recta.

Con el fin de ilustrar el proceso de estimación de la recta paramétrica bajo los postulados del modelo de regresión, supóngase que para cada uno de los tres valores de X de la Figura 8.16 se midió un valor de Y que fueron identificados como y_1, y_2 e y_3 respectivamente. A través de estos puntos se dibujo una recta $\hat{y} = \hat{\alpha} + \hat{\beta} x$ usando el método de los mínimos cuadrados (Figura 8.17). Esta recta es una estimación de la recta paramétrica $\mu_{y/x} = \alpha + \beta x$ El término $\hat{y}$ estima al valor $\mu_{y/x}$, y los términos $\hat{\alpha}$ y $\hat{\beta}$ son los estadísticos que estiman a los dos parámetros α y β.

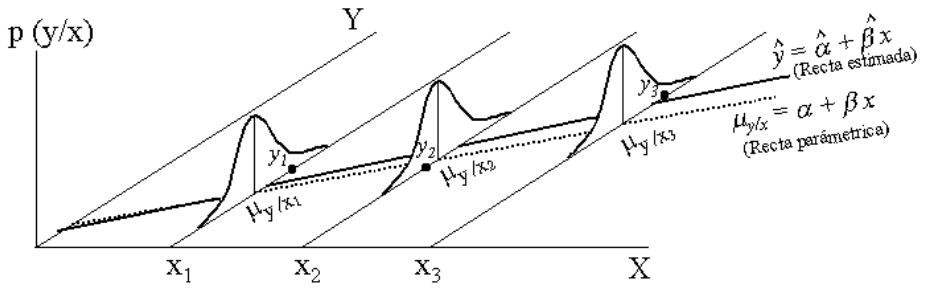


Figura 8.17

La recta $\hat{y} = \hat{\alpha} + \hat{\beta} x$ no es necesariamente la recta más próxima a la recta $\mu_{y/x} = \alpha + \beta x$ sino la mejor que se pudo construir con los puntos obtenidos. Cada muestra distinta de pares de valores X-Y producirá una recta diferente. De modo que cualquier recta $\hat{y} = \hat{\alpha} + \hat{\beta} x$ siempre será una aproximación a la recta paramétrica $\mu_{y/x} = \alpha + \beta x$.

8.5.2.2 Modelo II

El Modelo II de regresión se caracteriza porque hay dos variables dependientes. Para cada valor de X se origina una subpoblación de valores de Y; y para cada valor de Y se origina una subpoblación de valores de X. La distribución de las dos variables se denomina bivariada. Puesto que cualquiera de las dos variables se puede usar como variable independiente, es

posible obtener dos ecuaciones de regresión: la regresión de Y sobre X ó la regresión de X sobre Y.

1. Regresión de Y sobre X: $\hat{y} = \hat{\alpha} + \hat{\beta}x$.
2. Regresión de X sobre Y: $\hat{x} = \hat{\alpha}_* + \hat{\beta}_* y$

La elección de cual de las dos tipos de regresiones usar depende de cual variable se quiere usar como predictora. Por ejemplo, supóngase que a cien aves de una especie se les midió la longitud del tarso y el tamaño del cuerpo. Si se quiere predecir el tamaño de un ave a partir de la longitud del tarso, esta última variable se considera como independiente. En caso contrario, si se quiere usar el tamaño del cuerpo para predecir la longitud del tarso, la variable independiente será el tamaño del cuerpo.
También se puede aplicar el modelo tipo I usando estos dos datos. Por ejemplo se fijan previamente seis valores de la longitud del tarso, supongamos que son los siguientes 5,0; 5,5, 6,0; 6,5; 7,0; y 7,5 cm. Luego se escogen aleatoriamente entre todos los datos de tamaño del cuerpo uno para cada una de las longitudes de tarso seleccionadas y con los seis pares de valores se puede efectuar el análisis de regresión.

8.5.2.3 Modelo I vs. Modelo II

Ambos tipos de modelo tienen ventajas y desventajas. Por ejemplo con el modelo Tipo I es posible elegir previamente el intervalo de valores de X dentro del cual el investigador está interesado en efectuar predicciones o inferencias. Por el contrario, en el modelo tipo II el intervalo de valores no es controlado y lo determina el azar. Esto puede ser inconveniente si dentro del intervalo obtenido no están representados valores importantes para el investigador. El modelo tipo II, es ventajoso en situaciones donde se dificulta obtener o controlar los valores de la variable independiente. Esto sucede frecuentemente cuando el tamaño de la muestra obtenida es pequeño. Pongamos nuevamente el caso de las aves. Si en un muestreo de su población sólo se capturan ocho individuos, es difícil aplicar el modelo I de regresión. En este caso resulta complicado usar el procedimiento explicado anteriormente de fijar unos valores de tamaño de tarso y luego seleccionar aleatoriamente entre las muchas repeticiones un valor de tamaño de cuerpo. Ante la dificulta de fijar los valores de la variable independiente, es mucho más sencillo recurrir al modelo II de regresión y usar los ocho valores de longitud de tarso como variable independiente y sus respectivos tamaños de cuerpo como variable dependiente.

8.5.3. Estimación de α y β.

Como el propósito del análisis de regresión es construir a partir de la muestra de pares de valores X-Y, una recta muestral $\hat{y} = \hat{\alpha} + \hat{\beta}x$ que estima la verdadera recta paramétrica $\mu_{y/x} = \alpha + \beta x$, se puede usar el método de los mínimos cuadrados para calcular los estimadores $\hat{\alpha}$ y $\hat{\beta}$.

Ejemplo 8.9.

Con el propósito de predecir el contenido de calcio a partir de la edad de la mujer, se quiere establecer la ecuación que relaciona la edad y la concentración de calcio en los huesos de mujeres adultas. Con tal fin se determinó el contenido de calcio promedio en muestras óseas provenientes de mujeres de seis edades diferentes, encontrándose los resultados siguientes:

Edad (años)	45,0	50,0	55,0	60,0	65,0	70,0
Calcio (mg/g)	420,1	380,2	280,3	249,9	210,2	198,7

Suponga que el contenido de calcio se distribuye en forma normal.

Datos necesarios:

$$\sum_{i=1}^{n} x_i = 345,0 \qquad \sum_{i=1}^{n} y_i = 1739,4 \qquad \sum_{i=1}^{n} x_i y_i = 95897,0$$

$$\sum_{i=1}^{n} x_i^2 = 20275,0 \qquad \sum_{i=1}^{n} y_i^2 = 545719,88 \qquad \sum_{i=1}^{n} X_i \sum_{i=1}^{n} y_i = 600093,0$$

$$\overline{X} = 57,5 \qquad \overline{Y} = 289,9 \qquad n = 6$$

Sumas de cuadrados:

$$SP_{xy} = \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} = 95897,0 - \frac{600093,0}{6} = -4118,5$$

$$SC_x = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} = 20275 - \frac{(345)^2}{6} = 437,5$$

$$SC_y = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} = 545719,88 - \frac{(1739,4)^2}{6} = 41467,82$$

Cálculo de $\hat{\alpha}$ y $\hat{\beta}$.

$$\hat{\beta} = \frac{SP_{xy}}{SC_x} = \frac{-4118.5}{437.5} = -9,4137$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 289,9 - 9,4137(57,5) = 831,1886$$

La ecuación de la recta resultante es $\hat{y} = 831,1886 - 9,4137x$ (Figura 8.18).

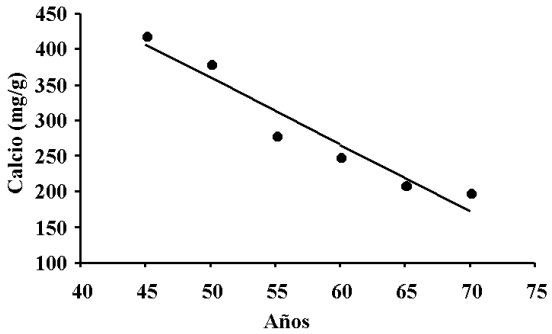


Figura 8.18

8.5.4. Evaluación de la regresión.

8.5.4.1 Prueba de hipótesis para β.

Una vez obtenida la recta de regresión $\hat{y} = \hat{\alpha} + \hat{\beta}x$ es necesario, comprobar estadísticamente si se justifica su aplicación. El grado de regresión lineal entre dos variables X-Y, es una condición que varia desde una situación donde existe una relación lineal perfecta positiva (Figura 8.19a) hasta la situación donde hay relación perfecta negativa (Figura 8.19d). Entre estos extremos existe un continuo de probables estados intermedios (Figura 8.19bd) que incluye situaciones con una relación muy débil o en las que no hay relación alguna (Figura 8.19e).

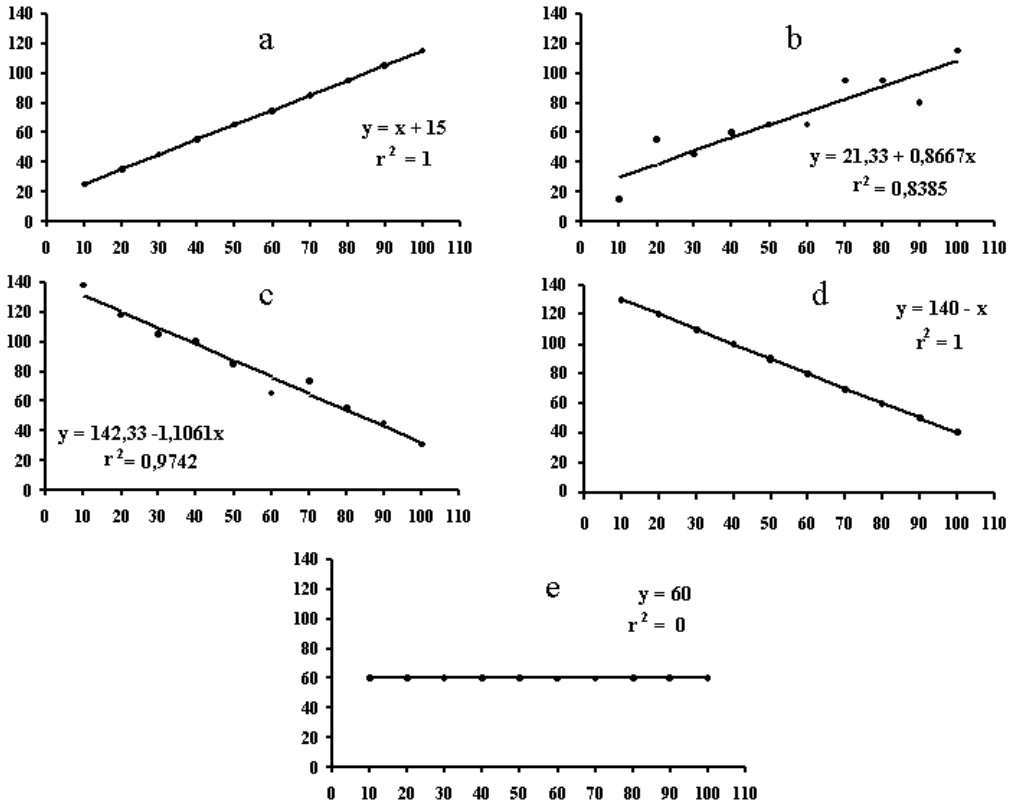


Figura 8.19.

Un buen indicador del grado de regresión lineal es la pendiente (β) de la recta, que también se denomina *coeficiente de regresión*. El valor de β no sólo cuantifica el grado de regresión sino la dirección de la relación. Si $\beta > 0$ la variable Y aumenta al incrementar X (Figura 8.19a,b); si $\beta < 0$ la variable Y disminuye al incrementar X (Figura 8.19cd); y si $\beta = 0$ la variable Y no cambia al incrementar X (Figura 8.19e).
En consecuencia, si β es diferente a cero, entonces la variable Y depende de la variable X y se justifica el uso de la ecuación de regresión lineal. Sin embargo, como se trabaja con muestras de *n* pares de valores (x_i; y_i) no se conoce el verdadero valor de β sino su valor aproximado, representado por el estimador $\hat{\beta}$, que como hemos visto antes, es una variable cuyo valor fluctúa aleatoriamente alrededor de β. Afortunadamente se conoce la distribución de probabilidades de $\hat{\beta}$, lo que posibilita el hacer inferencias estadísticas acerca de β, como sería la de someter a prueba la hipótesis que presume un $\beta \neq 0$.
En efecto, $\hat{\beta}$ se distribuye normalmente:

$$\hat{\beta} : N(\mu_\beta ; \sigma^2_{\hat{\beta}})$$

Donde: $\mu_\beta = \beta$ y $\sigma^2_{\hat{\beta}} = \sigma^2_{y/x} \Big/ SC_x$

Si se conoce σ_β, las inferencias sobre β se pueden hacer usando el estadístico:

$$z = \frac{\hat{\beta} - \beta}{\sigma_\beta}$$

Si se desconoce σ_β pero el tamaño de la muestra de n pares de valores (x_i; y_i) es grande $(n \geq 30)$, entonces se usa:

$$z = \frac{\hat{\beta} - \beta}{S_\beta}$$

Siendo $S_\beta = \sqrt{S^2_{y/x} \Big/ SC_x} = \sqrt{\left(SC_y - \beta^2 SC_x\right) \Big/ n-2} \Big/ \sqrt{SC_x}$

Si se desconoce σ_β y el tamaño de la muestra de *n* pares de valores (x_i; y_i) es pequeño ($n < 30$), entonces se usa:

$$t = \frac{\hat{\beta} - \beta}{S_\beta} \text{ (con n-2 grados de libertad)}$$

Ejemplo 8.10.
Verificar si la ecuación de regresión establecida en el ejemplo 8.9 puede utilizarse para predecir los valores del contenido de calcio en los huesos.

1. Hipótesis: $H_0 : \beta = 0$
   $H_1 : \beta \neq 0$

2. Nivel de significación: $\alpha = 0,05$

3. Estadístico de prueba:
   Como se desconoce σ_β y $n < 30$ el estadístico de prueba a usar es:
   $$t = \frac{\hat{\beta} - \beta}{S_\beta}$$

4. Zona de aceptación: como se trata de una prueba de dos colas, la zona de aceptación será:
   $$ZA: \left\{T / -t_{(1-\alpha/2;n-2)} < T < + t_{(1-\alpha/2;n-2)}\right\}$$

5. Cálculos:

   $$S_\beta = \sqrt{S^2_{y/x} \Big/ SC_x} = \sqrt{\left(SC_y - \beta^2 SC_x\right) \Big/ n-2} \Big/ \sqrt{SC_x}$$

   $$S_\beta = \sqrt{\frac{[41467,82 - (-9,4137)^2 (437,5)]}{4}} \Big/ \sqrt{437,5} = 25,969 / 20,917 = 1,24$$

   $$t = \frac{\hat{\beta} - \beta}{S_\beta} = \frac{-9,4137 - 0}{1,24} = -7,58$$

   $$ZA: \left\{T / -t_{(0,975;4)} < T < +t_{(0,975;4)}\right\} = \left\{T / -2,78 \ < \ T \ < \ +2,78\right\}$$

6. Decisión: Como el valor de $t = -7,58$ se encuentra fuera de los límites críticos de la zona de aceptación se rechaza H_0, y se concluye que los datos proporcionan suficiente evidencia para aceptar con un 95% de confianza que existe regresión lineal entre la edad y el contenido de calcio.

8.5.4.2 Análisis de varianza

La regresión lineal también se puede evaluar mediante un análisis de las varianzas. En este caso las varianzas involucradas son:

- La varianza residual originada por la dispersión de los valores y_i alrededor de $\hat{y}$.
- La varianza de la regresión, equivalente a la varianza de los tratamientos del ANDEVA. Esta varianza es causada por el efecto que tiene cada nivel x_i sobre $\hat{y}$.
- La varianza total que incluye las dos varianzas anteriores.

Dado *n* pares de valores X-Y, las fórmulas de cálculo de las tres varianzas son las siguientes:

Varianza residual o del error = $S_E^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2} = \dfrac{SCE}{n-2}$

Varianza debido al regresión = $S_R^2 = \dfrac{\sum_{i=1}^{n}(\hat{y} - \bar{Y})^2}{1} = SCR$

Varianza Total = $S_T^2 = \dfrac{\sum_{i=1}^{n}(y - \bar{Y})^2}{n-2} = \dfrac{SCT}{n-2}$

En la Figura 8.20 se observan las relaciones entre las tres varianzas anteriores. Para una recta de regresión $\hat{y} = \hat{\alpha} + \hat{\beta}x$ establecida, la desviación $y_i - \bar{Y}$ simboliza la desviación total entre un punto cualquiera $(x_i ; y_i)$ y el valor promedio $\bar{Y}$, que es la posición de referencia de los valores de Y si no hubiese regresión entre X e Y. Esta desviación total esta compuesta por dos desviaciones parciales: $\hat{y} - \bar{Y}$ que representa la desviación producida por el efecto de regresión y la diferencia $y_i - \hat{y}$ que mide la desviación del error o no controlada.

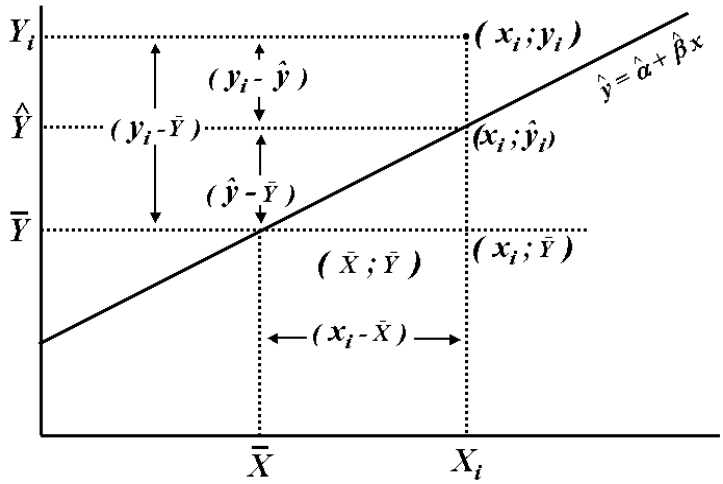


Figura 8.20.

Resulta obvio de la figura que:

Desviación total = desviación de la regresión + desviación del error

ó

$$(y - \bar{Y}) = (\hat{y} - \bar{Y}) + (y - \hat{y})$$

Por tanto:

$$\sum_{i=1}^{n}(y-\bar{Y})^2=\sum_{i=1}^{n}(\hat{y}-\bar{Y})^2+\sum_{i=1}^{n}(y-\hat{y})^2$$

De modo que conociendo dos de las sumas de cuadrados, por ejemplo la suma de cuadrados total y la suma debido a la regresión, se puede calcular por diferencia la tercera, en este caso la suma de cuadrados residual. Las fórmulas de cálculo de estas sumas de cuadrados son las siguientes:

$$\textit{Suma de cuadrados total}=SCT=\sum_{i=1}^{n}(y-\bar{Y})^2=\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}$$

$$\textit{Suma de cuadrados de la regresión}=SCR=\hat{\beta}^2\left[\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right]$$

$$\textit{Suma de cuadrados del error}=SCE=SCT-SCR$$

Conocidas las sumas de cuadrados se puede construir la tabla resumen del Análisis de Varianza para la regresión:

Tabla de Andeva

Fuente de variación	**Suma de cuadrados**	**Grados de Libertad**	**Cuadrados Medios**	**F_o**
Regresión	$\hat{\beta}^2\left[\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right]$	*1*	$\frac{SCR}{1}$	$\frac{CMR}{CME}$
Residual o Error	$SCT-SCR$	*n-2*	$\frac{SCE}{n-2}$	
Total	$\sum_{i=1}^{n}y_i^2-\left[\left(\sum_{i=1}^{n}y_i\right)^2\Big/n\right]$	*n-1*		

Ejemplo 8.11.
Verificar mediante un ANDEVA si la ecuación de regresión establecida en el ejemplo 8.9 puede utilizarse para predecir los valores del contenido de calcio en los huesos.

1. Hipótesis: $H_0 : \beta = 0$
   $H_1 : \beta \neq 0$

2. Nivel de significación: $\alpha = 0,05$

3. Estadístico de prueba: $F_o = \frac{S_R^2}{S_E^2} = \frac{CMR}{CME}$

4. Cómputos necesarios. Del ejemplo 8.9. se sabe que:

   $SC_x = 437,5$ $SC_y = 41467,82$ $\hat{\beta} = -9,4137$

   Cálculo de las sumas de cuadrados:

$$SCR = \hat{\beta}^2 \left\{ \sum_{i=1}^{n} x_i^2 - \left[\left(\sum_{i=1}^{n} x_i \right)^2 \Big/ n \right] \right\} = (-9,4137)^2 (437,5) = 38770,26$$

$$SCE = SCT - SCR = SCy - \hat{\beta}^2 (SC_x) = 41467,82 - 38770,26 = 2697,56$$

Cálculo de los Cuadrados medios

$$CMR = SCR / 1 = 38770,26 \qquad CME = \frac{SCE}{n-2} = \frac{2697,56}{4} = 674,39$$

5. Tabla resumen del análisis de varianza.

Fuente de variación	Suma de cuadrados	Grados de Libertad	Cuadrados Medios	F_o
Regresión	*38770,26*	*1*	*38770,26*	*57,48****
Residual o Error	*2697,56*	*4*	*674,39*	
Total	*41467,82*	*5*		

6. Zona de aceptación para la hipótesis de igualdad de las varianzas

$$ZA: \left\{ F \,/\, F < f_{\left[1-\alpha \,;\, 1/n-2\right]} \right\} = \left\{ F \,/\, F < f_{\left[0,95 \,;\, 1/4\right]} \right\} = \left\{ F \,/\, F < 7,70 \right\}$$

7. Decisión: Como el valor del estadístico de prueba $F_0 = 57,48$ es mucho mayor al límite crítico $(F = 7,7)$ se rechaza Ho.

8. Conclusión: se acepta con un 95% de confianza que existe regresión lineal entre la edad y el contenido de calcio en los huesos.

8.5.5 Bondad del ajuste por regresión lineal

Además de establecer la función lineal y verificar la existencia de regresión es importante saber que tan bueno ha sido el ajuste por regresión, es decir que tan cerca está la recta de regresión de los puntos usados para su construcción. En la Figura 8.21 se muestran dos rectas con coeficientes de correlación muy parecidos, sin embargo la dispersión de puntos alrededor de la recta de la izquierda (Figura 8.21a) es mayor que para la recta del lado derecho (Figura 8.21b). Se podría decir que la fuerza de la regresión, es decir el grado de dependencia de la variable Y con respecto a la variable X, es mayor para el caso b de la Figura 8.21.

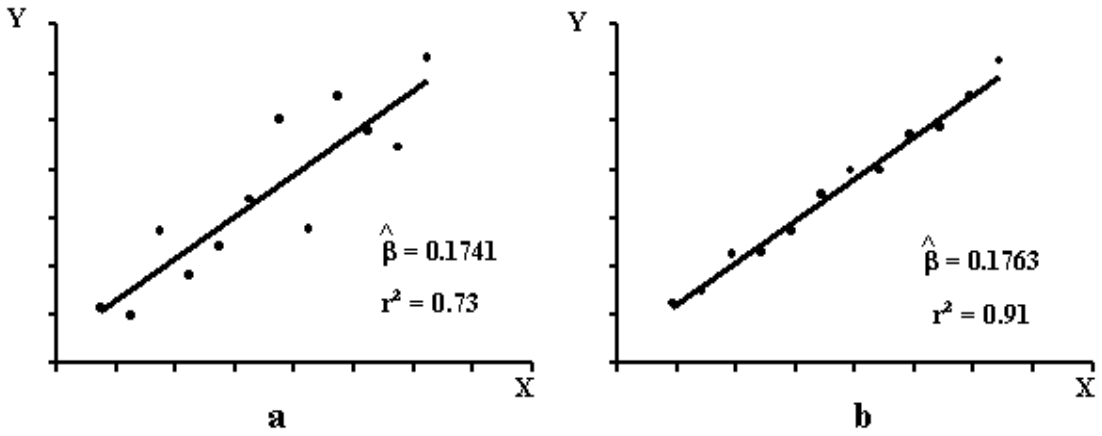


Figura 8.21.

Anteriormente se había determinado que la suma de cuadrados de los valores de Y resulta de la adición de dos sumas de cuadrados: una debido a la regresión y la otra debido al error:

$$SCT = SCR + SCE$$

ó

$$\sum_{i=1}^{n} (y - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{y} - \bar{Y})^2 + \sum_{i=1}^{n} (y - \hat{y})^2$$

De la expresión anterior y con la ayuda de la Figura 8.21 se ve que cuando la regresión entre las variables X-Y es muy fuerte, es decir que los puntos están muy cerca de la recta, la suma de cuadrados debido a la regresión es la proporción más importante de la suma de cuadrados total. De modo que con el propósito de valorar la fuerza de la regresión se ha definido el denominado *coeficiente de determinación* r^2, que simplemente es el cociente que resulta de dividir la variación debido a la regresión entre la variación total.

$$r^2 = \frac{SCR}{SCT} = \frac{\hat{\beta}^2 SC_x}{SC_y}$$

El valor de éste estadístico varía entre cero y uno. Un valor de cero indica que no existe variación debido a regresión. Un valor de uno, revela que toda la variación total es explicada por regresión. En término porcentuales un $r^2 = 0.93$, denota que el 93% de la variación total de Y es explicada por regresión. En el caso de la recta de la izquierda de la Figura 8.22 sólo el 73% de la variación es explicada por regresión, y en la recta de la derecha esta proporción es de un 91%.

Ejemplo 8.12.
Calcule el valor del coeficiente de determinación para los datos del ejemplo 8.9.

Sabiendo que:

$SC_x = 437,5$ $SC_y = 41467,82$ $\hat{\beta} = -9,4137$

Se tiene que:

$$r^2 = \frac{SCR}{SCT} = \frac{\hat{\beta}^2 SC_x}{SC_y} = \frac{(-9,4137)^2(437,5)}{41467,82} = \frac{38770,26}{41467,82} = 0,93$$

Se concluye que existe un buen ajuste por regresión entre la edad y el contenido de calcio en los huesos, puesto que el 93% de la variación del contenido de calcio depende de la edad de la persona.

8.5.6 Comprobación del modelo

Una vez aplicado el método de los mínimos cuadrados y ajustada la recta o modelo de regresión, toda la información de la variación que no puede ser explicada por el modelo, se encuentra contenida en los residuales, que no son otra cosa que la diferencia $d_i = y_i - \hat{y}$, es decir la diferencia que existe entre un valor observado (y_i) y el valor predicho $(\hat{y})$ para un dado valor x_i (Figura 8.22).

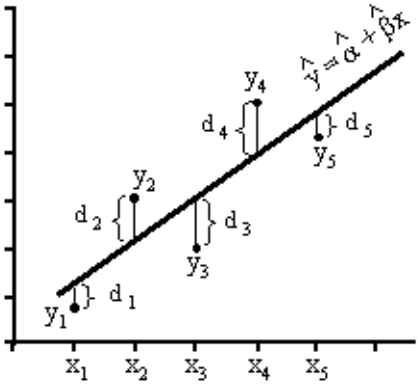


Figura 8.22.

Bajo los supuestos del modelo de regresión el valor d_i estima la varianza del error $(\sigma^2_{y/x})$, que como hemos visto se supone constante, independiente, con media cero y distribución normal. Por lo tanto, una de las maneras de evaluar el modelo de regresión, es analizar la tendencia de cambio de los residuales al variar los valores de $\hat{y}$ para comprobar la consistencia de las premisas. Las formas de dichas tendencias se pueden examinar graficando los valores de $\hat{y}$ contra los valores de d_i en un plano cartesiano (Figura 8.23).

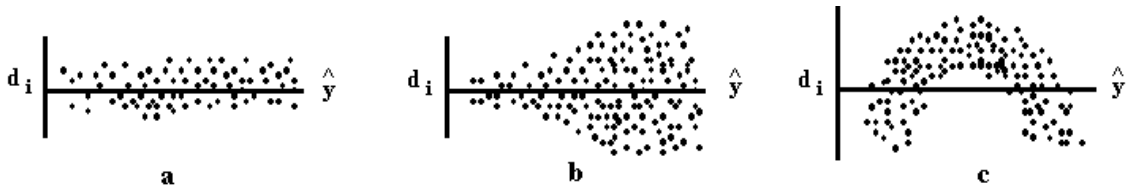


Figura 8.23

Si el modelo de regresión lineal es el adecuado para describir la relación entre las variables X e Y, la dispersión de puntos se muestra como una banda de un ancho más o menos constante alrededor de $\hat{y}$ (Figura 8.23ª). Cuando la amplitud de la banda de puntos aumenta con el valor de $\hat{y}$ (Figura 8.23b), es porque la varianza $(\sigma^2_{y/x})$ tiende a incrementar con el aumento del nivel de respuesta, lo que pone en duda la homogeneidad de la varianza. Esta situación se puede resolver, transformando los datos a fin de estabilizar la varianza. La dispersión de puntos no aleatoria como la que se muestra en la Figura 8.23c indica que el modelo usado no es el adecuado y debe usarse uno no lineal.

8.5.7 Formas de utilizar la ecuación de regresión.

Una vez establecida una ecuación de regresión lineal, la misma puede usarse de dos maneras diferentes: a) predecir el valor probable y_i para una dado valor x_i, y b) estimar el valor promedio de la subpoblación de valores de Y, es decir $\mu_{y/x}$, dado un valor x_i (Figura 8.24)

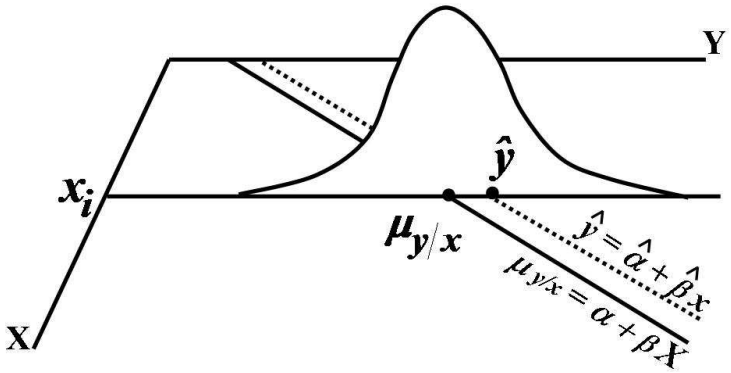


Figura 8.24.

La estimación puntual de y_i y de $\mu_{y/x}$ es la misma, porque ambas medidas se estiman usando la ecuación $\hat{y} = \hat{\alpha} + \hat{\beta}x$. Sin embargo, el intervalo de predicción para y_i es más amplio que el intervalo de confianza para $\mu_{y/x}$, no obstante que ambos se construyen a partir del mismo valor predicho $\hat{y}$. Esta diferencia se produce porque las varianzas usadas en los dos casos son diferentes.

8.5.7.1 Predicción de y para un dado valor x_i .

Puesto que se sabe que $\hat{y}$ se distribuye como *t* con *n-2* grados de libertad, se puede construir el intervalo de confianza a partir del estadístico:

$$t = \frac{\hat{y} - y_i}{S_{(\hat{y}-y)}}$$

Donde:

y_i = valor probable de Y para un dado valor de X.
$\hat{y}$ = estimación puntual de *y* para un dado x_i .

$$S_{(\hat{y}-y)} = S_{y/x}\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}} = \textit{desviación estándar de y}$$

$$S_{y/x} = \sqrt{\frac{SC_y - \beta^2 SC_x}{n-2}} = \textit{desviación del error}$$

$$\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2 = SC_x$$

El intervalo de predicción será el siguiente:

$$IP = \left[\hat{y} \pm t_{[1-\alpha/2;n-2]} S_{y/x}\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}}\right]$$

Ejemplo 8.13.

Suponga que en una investigación se quiere encontrar una ecuación que relacione linealmente una variable independiente X con la variable dependiente Y. Si una muestra de seis observaciones pareadas de las dos variables produjo los resultados siguientes:

X	2	3	4	5	6	7
Y	7	9	10	11	14	15

Haga lo siguiente: a) establezca la ecuación de regresión; b) compruebe si se justifica su uso; c) determine la bondad del ajuste por regresión, y d) encuentre el intervalo de predicción para x = 2; x = 4,5 y x = 7

Encontrando la ecuación de la recta

Datos necesarios:

$$\sum_{i=1}^{n} x_i = 27,0 \qquad \sum_{i=1}^{n} y_i = 66,0$$

$$\sum_{i=1}^{n} x_i^2 = 139,0 \qquad \sum_{i=1}^{n} y_i^2 = 772,0$$

$$\sum_{i=1}^{n} x_i y_i = 325,0 \qquad \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = 1782,0$$

$$\overline{X} = 4,5 \qquad \overline{Y} = 11,0$$

$$n = 6$$

Sumas de cuadrados:

$$SP_{xy} = \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} = 325,0 - \frac{1782,0}{6} = 28,0$$

$$SC_x = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} = 139,0 - \frac{(27)^2}{6} = 17,5$$

$$SC_y = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} = 772,0 - \frac{(66)^2}{6} = 46,0$$

Cálculo de $\hat{\alpha}$ y $\hat{\beta}$.

$$\hat{\beta} = \frac{SP_{xy}}{SC_x} = \frac{28,0}{17,5} = 1,6 \qquad \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 11 - 1,6(4,5) = 3,8$$

La ecuación de la recta resultante es $\hat{y} = 3,8 + 1,6\,x$ (Figura 8.25).

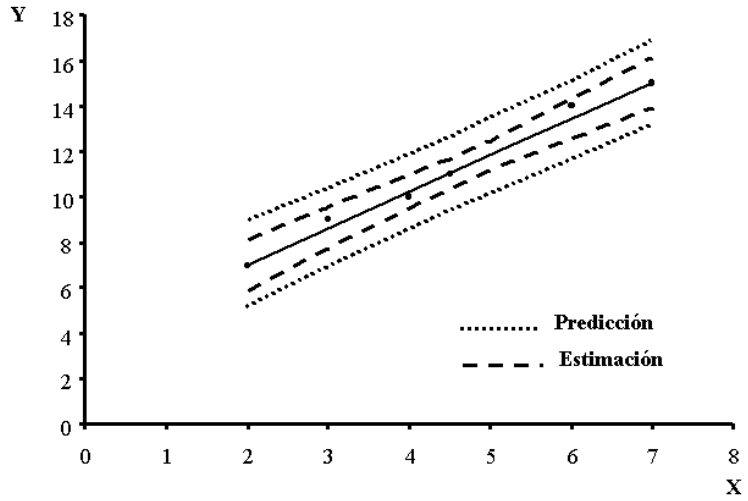


Figura 8.25. Recta de regresión con sus respectivas bandas de límites de predicción y de estimación.

Prueba de significación de β

a) Hipótesis: $H_0 : \beta = 0$
$H_1 : \beta \neq 0$

b) Nivel de significación: $\alpha = 0,05$

c) Estadístico de prueba: Como se desconoce σ_β y $n < 30$ el estadístico de prueba es:

$$t = \frac{\hat{\beta} - \beta}{S_\beta}$$

d) Zona de aceptación: como se trata de una prueba de dos colas, la zona de aceptación será:

$$ZA : \left\{ T \,/\, -t_{(1-\alpha/2;n-2)} < T < + \, t_{(1-\alpha/2;n-2)} \right\}$$

e) Cálculos:

$$S_\beta = \sqrt{S^2_{y/x} \Big/ SC_x} = \sqrt{\left(SC_y - \beta^2 SC_x \right) \Big/ n-2} \Big/ \sqrt{SC_x}$$

$$S_\beta = \sqrt{\frac{[46 - (1,6)^2 (17,5)]}{4}} \Big/ \sqrt{17,5} = 0,5477 / 4,1833 = 0,1309$$

$$t = \frac{\hat{\beta} - \beta}{S_\beta} = \frac{1,6 - 0}{0,1309} = 12,22$$

$$ZA: \left\{ T / -t_{(0,975;4)} < T < +t_{(0,975;4)} \right\} = \left\{ T / -2,78 \ < \ T \ < \ +2,78 \right\}$$

f) Decisión: Como el valor de $t = 12,22$ se encuentra fuera de los límites críticos de la zona de aceptación se rechaza H_0, se concluye que los datos proporcionan suficiente evidencia para aceptar con un 95% de confianza que existe regresión lineal entre las variables X e Y.

Bondad del ajuste

$$r^2 = \frac{SCR}{SCT} = \frac{\hat{\beta}^2 SC_x}{SC_y} = \frac{(1,,6)^2 (17,5)}{46,0} = \frac{44,8}{46,0} = 0,9739$$

Se puede concluir que existe un buen ajuste por regresión entre las variables X e Y, puesto que el 97,39 % de la variación de Y depende de X.

Intervalo de predicción

Los valores del intervalo de predicción se obtienen aplicando la ecuación siguiente:

$$IP = \left[\hat{y} \pm t_{[1-\alpha/2;n-2]} S_{y/x} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2}} \right]$$

Cálculos y datos necesarios:

$$S_{y/x} = \sqrt{\frac{SC_y - \beta^2 SC_x}{n-2}} = \sqrt{\frac{46 - (1,6)^2 17.5}{4}} = \sqrt{\frac{1,2}{4}} = 0,5477$$

$$t_{(1-\alpha/2;n-2)} = t_{(0,975;4)} = 2,78$$

$$\sum_{i=1}^{n} (x_i - \bar{X})^2 = SC_x = 17,5$$

Cuando $x = 2$, se tiene:

$$\hat{y} = 3,8 + 1,6\ x = 3,8 + 1,6(2) = 7,0$$

$$(x_i - \bar{X})^2 = (2 - 4,5)^2 = 6,25$$

$$IP = \left[\hat{y} \pm t_{[1-\alpha/2;n-2]} S_{y/x} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2}} \right] = \left[7 \pm 2{,}78(0{,}5477) \sqrt{1 + \frac{1}{6} + \frac{6{,}25}{17{,}5}} \right] = [7 \pm 1{,}88]$$

El intervalo buscado es $IP = [5{,}12; 8{,}88]$. Se tiene un 95% de confianza que el valor probable de *y* este incluido en dicho intervalo.

Cuando $x = 4{,}5$, se tiene:

$$\hat{y} = 3{,}8 + 1{,}6\ x = 3{,}8 + 1{,}6(4{,}5) = 11{,}0$$

$$(x_i - \bar{X})^2 = (4{,}5 - 4{,}5)^2 = 0$$

$$IP = \left[\hat{y} \pm t_{[1-\alpha/2;n-2]} S_{y/x} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2}} \right] = \left[11 \pm 2{,}78(0{,}5477) \sqrt{1 + \frac{1}{6}} \right] = [11 \pm 1{,}645]$$

El intervalo buscado es $IP = [9{,}36; 12{,}65]$. Se concluye que se tiene un 95% de confianza que el valor probable de *y* este incluido en dicho intervalo.

Cuando $x = 7{,}0$, se tiene:

$$\hat{y} = 3{,}8 + 1{,}6\ x = 3{,}8 + 1{,}6(7) = 15{,}0$$

$$(x_i - \bar{X})^2 = (7{,}0 - 4{,}5)^2 = 8{,}25$$

$$IP = \left[\hat{y} \pm t_{[1-\alpha/2;n-2]} S_{y/x} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2}} \right] = \left[15 \pm 2{,}78(0{,}5477) \sqrt{1 + \frac{1}{6} + \frac{8{,}25}{17{,}5}} \right] = [15 \pm 1{,}88]$$

El intervalo buscado es $[13{,}12; 16{,}88]$. Se concluye que se tiene un 95% de confianza que el valor probable de *y* este incluido en dicho intervalo.

En la Figura 8.25 se muestra la recta de regresión $\hat{y} = 3{,}8 + 1{,}6\ x$ y las líneas que definen los intervalos de predicción de *y* para cada valor de *x*.

8.5.7.2 Estimación de $\mu_{y/x}$ para un dado valor x_i

Puesto que se sabe que $\hat{y}$ se distribuye como *t* con *n-2* grados de libertad, se puede construir el intervalo de confianza a partir del estadístico:

$$t = \frac{\hat{y} - \mu_{y/x}}{S_{(\hat{y}-\mu_{y/x})}}$$

donde:

$\mu_{y/x}$ = valor esperado de Y para un dado valor de X.
$\hat{y}$ = estimación puntual de $\mu_{y/x}$ para un dado x_i .

$$S_{(\hat{y}-\mu_{y/x})} = S_{y/x}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}} = \textit{desviación estándar de } \mu_{y/x}$$

$$S_{y/x} = \sqrt{\frac{SC_y - \beta^2 SC_x}{n-2}} = \textit{desviación del error}$$

El intervalo de confianza será el siguiente:

$$IP = \left[\hat{y} \pm t_{[1-\alpha/2;n-2]} S_{y/x}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}}\right]$$

Ejemplo 8.14.
Construya un intervalo de confianza para $\mu_{y/x}$ a partir de los valores x = 2; x = 4,5 y x = 7, usando los datos del ejemplo 8.13.

Cuando $x = 2{,}0$, se tiene: $\hat{y} = 3{,}8 + 1.6\ x = 3{,}8 + 1{,}6(2) = 7{,}0$

$(x_i - \bar{X})^2 = (2 - 4{,}5)^2 = 6{,}25$

$$IC = \left[\hat{y} \pm t_{[1-\alpha/2;n-2]} S_{y/x}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}}\right] = \left[7 \pm 2{,}78(0{,}5477)\sqrt{\frac{1}{6} + \frac{6{,}25}{17{,}5}}\right] = [7 \pm 1{,}102]$$

El intervalo buscado es $IC = [5,898; 8,102]$. Se tiene un 95% de confianza que este intervalo contenga el valor de $\mu_{y/x}$.

Cuando $x = 4,5$, se tiene: $\hat{y} = 3,8 + 1,6\ x = 3,8 + 1,6(4,5) = 11,0$

$(x_i - \bar{X})^2 = (4,5 - 4,5)^2 = 0$

$$IC = \left[\hat{y} \pm t_{[1-\alpha/2;n-2]} S_{y/x} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2}} \right] = \left[11 \pm 2,78(0,5477) \sqrt{\frac{1}{6}} \right] = [11 \pm 0,6216]$$

El intervalo buscado es $IC = [10,38; 11,62]$. Se tiene un 95% de confianza que este intervalo contenga el valor de $\mu_{y/x}$.

Cuando $x = 7,0$, se tiene: $\hat{y} = 3.8 + 1,6\ x = 3,8 + 1,6(7) = 15,0$

$(x_i - \bar{X})^2 = (7,0 - 4,5)^2 = 6,25$

$$IC = \left[\hat{y} \pm t_{[1-\alpha/2;n-2]} S_{y/x} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2}} \right] = \left[15 \pm 2,78(0,5477) \sqrt{\frac{1}{6} + \frac{6,25}{17,5}} \right] = [15 \pm 1,102]$$

El intervalo buscado es $[13,89; 16,102]$. Se tiene un 95% de confianza que este intervalo contenga el valor de $\mu_{y/x}$.

En la Figura 8.25 se muestra la recta de regresión $\hat{y} = 3.8 + 1,6\ x$ y las líneas que definen los intervalos de confianza de cada $\mu_{y/x}$. Nótese que la banda de valores de los límites de confianza es más estrecha que la banda de valores de los límites de predicción.

8.5.8 Comparando dos rectas de regresión.

Con mucha frecuencia se quiere saber si dos rectas tienen el mismo coeficiente de regresión, es decir, si tienen las mismas pendientes. Puesto que sabemos que el estadístico $\hat{\beta}$ se distribuye normalmente, se puede determinar mediante una prueba la hipótesis si los estadísticos $\hat{\beta}_1$ y $\hat{\beta}_2$ son estimaciones de un mismo parámetro β.

Ejemplo 8.15.
En un estudio sobre la diabetes se determinó la concentración de glucosa en la sangre a dos grupos de varones adultos aparentemente sanos. El grupo A estuvo formado por individuos cuyo peso corporal fue menor a los 80 kg y el grupo B por los individuos con pesos mayores a éste valor.

Grupo A: Peso menor a 80 kg.

Peso (kg)	59	62	64	68	71	73	75	77	78	79
Glucosa (mg/100ml)	89	85	92	99	104	109	105	110	105	115

Grupo B: Peso mayor a 80 kg.

Peso (kg)	82	85	88	89	92	94	99	101	105	115
Glucosa (mg/100ml)	95	97	95	97	99	101	105	101	105	110

Calcular los dos coeficientes de regresión y probar la hipótesis nula de que ambos estiman un mismo coeficiente β.

En la tabla siguiente se muestran los datos necesarios para calcular los valores de los intercepto $(\hat{\alpha}_A;\hat{\alpha}_B)$, los coeficientes de regresión $\left(\hat{\beta}_A;\hat{\beta}_B\right)$ y las varianzas $\left(S^2_{\hat{\beta}_A};S^2_{\hat{\beta}_B}\right)$.

	Grupo A	Grupo B
Número de observaciones	10	10
Media de los pesos $(\bar{X})$	70,6	95,0
Media de la Glucosa $(\bar{Y})$	101,3	101,6
Suma de cuadrados del Peso (X)	450,40	916,00
Suma de cuadrados de Glucosa (Y)	866,10	1000,40
Suma de productos (peso x glucosa)	584,20	905,00

Valores de las pendientes:

$$\hat{\beta}_A = \frac{SP_{xy}}{SC_x} = \frac{584,2}{450,4} = 1,297 \qquad \hat{\beta}_B = \frac{SP_{pxg}}{SC_p} = \frac{905}{916} = 0,988$$

Valores de los intercepto:

$$\hat{\alpha}_A = \bar{Y}_A - \hat{\beta}_A \bar{X}_A = 101,3 - 1,297(70,6) = 9,73$$

$$\hat{\alpha}_B = \bar{Y}_B - \hat{\beta}_B \bar{X}_B = 101,6 - 0,988(95) = 7,74$$

Ecuaciones: $\hat{y}_{i_A} = 9{,}73 + 1{,}297\, x_{i_A}$

$$\hat{y}_{i_B} = 7{,}74 + 0{,}988\, x_{i_A}$$

Hipótesis: $H_0 : \beta_1 = \beta_2$

$H_1 : \beta_1 \neq \beta_2$

Estadístico de prueba: $$t = \frac{\left(\hat{\beta}_A - \hat{\beta}_B\right) - \left(\beta_A - \beta_B\right)}{\sqrt{\frac{S_p^2}{SC_{x_A}} + \frac{S_p^2}{SC_{x_B}}}}$$

Zona de aceptación:

$$ZA = \left\{T \,/\, -t_{(1-\alpha/2;n-2)} < T < +t_{(1-\alpha/2;n-2)}\right\} = \left\{T \,/\, -t_{(0.975;8)} < T < +t_{(0{,}975;8)}\right\}$$

Cálculos necesarios:

$$S_p^2 = \frac{\left\{SC_{y_A} - \left[\left(SP_{xy_{(A)}}\right)^2 \Big/ SC_{x_A}\right]\right\} + \left\{SC_{y_B} - \left[\left(SP_{xy_{(B)}}\right)^2 \Big/ SC_{x_B}\right]\right\}}{n_1 - 2 + n_2 - 2} =$$

$$= \frac{\left\{866{,}1 - \left[(584{,}2)^2 \Big/ 450{,}4\right]\right\} + \left\{1000{,}4 - \left[(905)^2 \Big/ 916\right]\right\}}{16} =$$

$$= \frac{108{,}35 + 106{,}26}{16} = \frac{214{,}62}{16} = 13{,}41$$

Sustituyendo S_p^2 en t se tiene:

$$t = \frac{\left(\hat{\beta}_A - \hat{\beta}_B\right) - \left(\beta_A - \beta_B\right)}{\sqrt{\frac{S_p^2}{SC_{x_A}} + \frac{S_p^2}{SC_{x_B}}}} = \frac{(1.297 - 0.988)}{\sqrt{\frac{13.41}{450.4} + \frac{13.41}{916}}} = \frac{0.309}{0.2107} = 1.47$$

Al comparar el valor de t con los límites de la Zona de Aceptación de H_0:

$$ZA = \left\{T \,/\, -t_{(0{,}975;8)} < T < +t_{(0{,}975;8)}\right\} = \left\{T \,/\, 2{,}31 < T < +2{,}31\right\}$$

Se tiene que $t = 1,47$ pertenece a la zona de aceptación, por lo tanto se concluye que los dos coeficientes de regresión o pendientes de las rectas son iguales.

8.5.9 Advertencias para el uso de la ecuación de regresión.

1. No se debe usar la ecuación de regresión para hacer predicciones o cualesquiera otras inferencias si el valor de β no difiere significativamente de cero.
2. No hacer inferencias fuera del intervalo de valores de la variable independiente. En muchos casos la relación lineal sólo se mantiene dentro de ciertos límites, por lo que es peligroso hacer extrapolaciones fuera del ámbito de valores usados para calcular la ecuación de regresión.
3. Las ecuaciones de regresión deben actualizarse frecuentemente. Muchas veces ecuaciones calculadas algún tiempo atrás o bajo condiciones diferentes no tienen validez actual debido a cambios en las relaciones de las variables.
4. Las inferencias hechas a partir de una ecuación de regresión sólo tienen validez para la población de datos de donde se extrajo la muestra.

8.6 EJERCICIOS

1. Establezca diferencias entre el análisis de correlación y el análisis de regresión.

2. El propietario de una finca citrícola desea saber si las variables número de frutos y producción están relacionadas. Con este propósito seleccionó una muestra de 9 plantas obteniéndose la información siguiente

N° frutos	1000	1532	1175	1370	1710	1010	1080	1200	1600
Peso (kg)	200	260	215	245	280	185	179	238	260

   a. Use los datos anteriores y haga el análisis estadístico más adecuado para responder la inquietud del productor.
   b. ¿Se pueden usar los datos de la pregunta anterior para hacer un análisis de regresión?

3. Para un total de 39 plantas de maíz se calcula que el coeficiente de correlación entre la altura y el diámetro del tallo es igual 0,60. ¿Existe correlación entre las variables estudiadas?

4. Un investigador seleccionó a cinco hombres que en su primera salida de caza mataron un venado con un rifle. Le midió la altura a cada cazador y le preguntó el peso del venado. Los resultados se muestran en la tabla siguiente:

Altura cazador (cm)	165	188	178	182	171
Peso del venado (kg)	45	110	81	87	65

a. Encuentre el valor del coeficiente de correlación y determine si se diferencia significativamente de cero.
b. ¿Se puede afirmar que a mayor altura del cazador mayor será el tamaño de la presa?
c. ¿La existencia de correlación entre dos variables implica causalidad?

5. Se ha efectuado un seguimiento a 20 estudiantes comparando sus calificaciones en el examen de admisión a la universidad y su rendimiento posterior en la carrera, obteniéndose los resultados siguientes:

					Grupo 1					
Examen de admisión	50	35	35	40	55	65	35	60	90	35
Promedio de carrera	10,6	8,2	10,8	11,2	13,6	11,6	9	14	15,8	10,2
					Grupo 2					
Examen de admisión	91	81	62	60	40	59	55	50	65	50
Promedio de carrera	18,2	14,2	15,4	12,0	7,0	15,2	8,0	12,0	13,0	10,0

a. Calcule el coeficiente de correlación entre las notas de admisión y de la carrera para ambos grupos de estudiantes.
b. Ponga a prueba la hipótesis que los coeficientes de correlación de ambos grupos son iguales.
c. Ponga a prueba la hipótesis que el grupo 1 tiene un coeficiente de correlación mayor al de los estudiantes de otra carrera cuyo $r = 0,74$ para las mismas variables.
d. Ponga a prueba la hipótesis que el grupo 2 tiene un coeficiente de correlación diferente al de los estudiantes de otra carrera cuyo $r = 0,68$ para las mismas variables.

6. En un ensayo de maíz se analiza el efecto del número de plantas sobre la producción, obteniéndose la información siguiente:

Número de plantas	20	10	13	18	19	16	12	18	9
Producción (kg/m^2)	1,6	1,2	1,5	2,0	2,0	2,9	2,7	2,2	1,2

a. Determine si existe correlación entre las dos variables.

7. Construya un gráfico de dispersión de puntos con el conjunto de datos siguiente:

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

a. Trace al ojo la mejor recta que se ajusta al conjunto de puntos.
b. Obtenga de la recta anterior los términos necesarios para establecer su ecuación.

8. Encuentre la fórmula de la recta que se ajusta a los datos de la pregunta anterior usando el método de los mínimos cuadrados y compare esta ecuación con la encontrada en forma gráfica.

   a. Use las dos ecuaciones anteriores para predecir el valor de la variable Y cuando $X = 2,57$ y $X = 5,2$
   b. Diga si son válidos las dos extrapolaciones efectuadas en el punto anterior.

9. Si en la pregunta anterior en lugar de solicitarle ajustar una recta por el método de los mínimos cuadrados, se le hubiera pedido encontrar la recta de regresión entre X e Y, cambiaría la ecuación de la recta.

10. ¿Se pueden interpretar las dos rectas anteriores (mínimos cuadrados y regresión) de la misma manera?

11. En un experimento sobre el crecimiento de gramíneas, se le agregó a la tierra dentro del envase donde crecía cada planta cantidades conocidas de nitrógeno asimilable (grs). Después de cierto tiempo se midió la cantidad de nitrógeno presentes en cada planta

N_2 agregado	0,00	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	1,00
N_2 en la planta	0,12	0,18	0,32	0,25	0,36	0,52	0,53	0,63	0,63	0,69	0,73

   a. Encuentre los estimadores de los coeficientes de la recta de regresión.
   b. Estime la cantidad de N_2 en la planta si el contenido inicial del elemento es de 0,45
   c. Construya un intervalo de confianza del 95% para el nitrógeno en la planta cuando el contenido inicial de nitrógeno es de 0.50 grs.
   d. Construya un intervalo de predicción del 95% para el nitrógeno en la planta cuando el contenido inicial de nitrógeno es de 0.50 grs.

12. En un estudio sobre la lechoza (Carica papaya), se quiere determinar si existe relación entre el número de flores y el diámetro del tallo. Para lo cuál a 8 plantas seleccionadas aleatoriamente se les determinó el diámetro del tallo y el número de flores, obteniéndose la información siguiente:

Diámetro (cm)	7,62	3,69	0,6	4,65	3,63	1,75	1,63	0,67
Número de flores	26	22	28	24	27	20	25	23

   a. Pruebe la hipótesis que las dos variables no están correlacionadas con un nivel de significación del 5%.
   b. ¿Se puede utilizar la información anterior para estimar el N° de flores para una planta que tiene un diámetro de 3.0 cm? Explique su respuesta.

13. A cierta cantidad de semen de Toro, se le eliminó el plasma seminal y se resuspendió en un medio buffer, con diferentes concentraciones de fructuosa (mg/100ml). La suspensión fue incubada y se midió la utilización de la fructuosa en la primera hora de incubación, obteniéndose los resultados siguientes:

Fructuosa inicial	100	120	140	160	180	200	220	240	260	280	300
Fructuosa consumida	13,8	57,5	63,3	59,2	66,7	84,9	89,8	121,4	123,4	129,1	134,6

a. Estime la cantidad promedio de fructuosa a ser utilizada, si inicialmente la concentración de fructuosa fue de 190 mg/100ml. La probabilidad de cometer el error tipo I no debe exceder al 1%.

14. Nueve grupos de gorgojos (Tribolium sp.) se sometieron a 6 días de inanición bajo nueve valores de humedad relativa y se midió la pérdida de peso de los individuos para cada valor de humedad relativa. Los resultados fueron los siguientes:

Humedad relativa (%)	0	12	30	43	53	62	76	85	93
Peso perdido (mg)	8,9	8,2	6,7	6,1	5,9	5,8	4,7	4,2	3,7

a. ¿Existe dependencia entre las variables estudiadas?
b. ¿Es realmente menor la pérdida de peso a mayores valores de HR.
c. ¿Que porcentaje de la variación en la pérdida de peso es explicado por la regresión?

15. Un biólogo estudia la relación entre la duración del crecimiento bacteriano y el volumen del medio del cultivo utilizado. Con esa finalidad tomó una muestra formada por 10 determinaciones y obtuvo la información siguiente:

Duración (días)	1,0	3,5	4,0	1,9	4,1	2,5	2,1	1,3	5,3	3,0
Volumen (ml)	100	400	500	200	500	300	200	100	600	300

a. ¿Cuál es la variable dependiente?
b. ¿Se justifica establecer una ecuación de regresión?
c. ¿Que proporción de la variación del tiempo de duración del crecimiento es explicada por los cambios en el volumen del medio?
d. Para un volumen del medio de cultivo de 950 ml, que duración del crecimiento bacterial estimaría Ud. ¿Es correcto hacer ésta estimación? Explique.

16. Bajo la presunción de que hay relación causa efecto entre las variables número de plantas y rendimiento en el cultivo de maíz, se tomó la información siguiente:

Número de plantas	20	35	31	62	54	22	43	43	36
Rendimiento (kg)	10	16	12	48	25	12	28	27	20

a. ¿Se justifica establecer una ecuación de regresión?
b. Grafique la recta de regresión incluyendo los puntos observados
c. Si en las condiciones de la experiencia se siembra una planta por cada metro cuadrado de superficie, calcule el rendimiento que se debe esperar para una superficie de 50 m².
d. Utilizando el gráfico encuentre e indique el rendimiento estimado para una superficie de 50 m². Si encuentra diferencias explique las causas.

17. Cierto compuesto químico se disolvió en 100 ml de dos disolventes a diferentes temperaturas. Para cada caso se registró los gramos disueltos del compuesto hasta que la solución se saturó. Los resultados obtenidos fueron los siguientes:

	Gramos disueltos	
Temperatura (°C)	Disolvente 1	Disolvente 2
0	5	6
15	18	14
30	22	25
45	27	38
60	31	45
75	38	58

a. Establezca una ecuación de regresión para cada disolvente.
b. Construya los límites de confianza para las dos rectas de regresión y y grafique la recta de regresión y las bandas de confianza.
c. Someta a prueba la hipótesis que los coeficientes de regresión (β_1 y β_2) son iguales.

18. Al medir las temperatura de un cuerpo que se enfría desde una temperatura inicial de 100,3°C, en intervalos de medio minuto, se obtuvo:

Tiempo (min)	0,0	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0
Temperatura (°C)	100	90,1	80,2	71,6	64,2	57,5	51,3	46,1	41,2

a. Ajustar una función de regresión que indique la temperatura en función del tiempo.
b. Estimar la temperatura que tendrá el cuerpo a los tres minutos de iniciado el proceso.
c. Cuánto tardará en reducirse a la mitad la temperatura inicial?