AJUSTE DE CURVAS

En múltiples ocasiones se encuentran situaciones en las que se requiere analizar la relación entre dos variables cuantitativas. Los dos objetivos fundamentales de este análisis serán:

- Determinar si dichas variables están asociadas y en qué sentido se da dicha asociación (es decir, si los valores de una de las variables tienden a aumentar –o disminuir- al aumentar los valores de la otra);
- Estudiar si los valores de una variable pueden ser utilizados para predecir el valor de la otra.

La forma correcta de abordar el primer problema es recurriendo a coeficientes de correlación. Sin embargo, el estudio de la correlación es insuficiente para obtener una respuesta a la segunda cuestión: se limita a indicar la fuerza de la asociación mediante un único número, tratando las variables de modo simétrico, mientras que lo que interesa es modelizar dicha relación y usar una de las variables para explicar la otra.

Para tal propósito se recurrirá a la técnica de regresión. Aquí se analizará el caso más sencillo en el que se considera únicamente la relación entre dos variables (x e y). Así mismo, se limita al caso en el que la relación que se pretende modelizar es de tipo lineal. En este caso, la media de la distribución de las y sobre x está dada por $\alpha + \beta.x$.

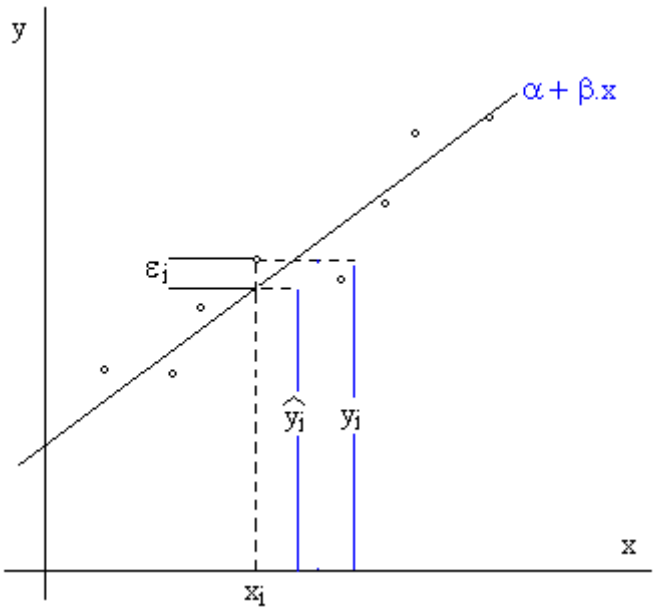
LA RECTA DE REGRESIÓN

Considérese una variable aleatoria respuesta (o dependiente) y, que se supone relacionada con otra variable (no necesariamente aleatoria) que se llamará explicativa, predictora o independiente y que se denotará por x.

A partir de una muestra de n individuos para los que se dispone de los valores de ambas variables, $\{(x_i,y_i),i = 1,...n\}$, se puede visualizar gráficamente la relación existente entre ambas mediante un gráfico de dispersión, en el que los valores de la variable x se disponen en el eje horizontal y los de y en el vertical. El problema que subyace a la metodología de la regresión lineal simple es el de encontrar una recta que ajuste a la nube de puntos del diagrama así dibujado, y que pueda ser utilizada para predecir los valores de y a partir de los de x. La ecuación general de la recta de regresión será entonces de la forma: $\alpha + \beta.x$.

El problema radica en encontrar aquella recta que mejor ajuste a los datos. Tradicionalmente se ha recurrido para ello al método de mínimos cuadrados, que elige como recta de regresión a aquella que minimiza las distancias verticales de las observaciones a la recta.

Cualquier observación i-ésima $\mathbf{y_i}$ diferirá verticalmente de esa recta (por ahora desconocida) en un valor ε_i. Luego ε es el valor de una variable aleatoria.



El valor de ε para cualquier observación determinada dependerá de un posible error de medición y de los valores de otras variables distintas de x que podrían influir sobre y.

Habrá que calcular los valores de α y β de la línea de regresión, es decir la ecuación de la recta que de alguna manera da el mejor ajuste. En referencia al gráfico anterior, es relativamente fácil trazarla a simple vista con un poco de sentido común. Sin embargo, lo habitual es recurrir a un método menos subjetivo.

Para plantear este problema de manera formal, considérese n parejas de observaciones (x_i,y_i) en las cuales es razonable suponer que la regresión de y sobre x es lineal, y se desea determinar la recta del mejor ajuste. Si se predice y por medio de la ecuación:

$$\hat{y} = a + b \cdot x$$

sea e_i el error de predecir el valor de y correspondiente a la x_i es:

$$e_i = y_i - \hat{y}_i$$

Se quiere determinar a y b de modo que estos errores sean, en cierto modo, lo más pequeños posibles. Ya que no se pueden minimizar cada uno de los e_i por separado, esto sugiere intentar

$$\sum_{i=1}^{n} e_i$$

tan cercano a cero como sea posible.

Esto no es aconsejable puesto que errores positivos y negativos se compensarán dando líneas inadecuadas como respuesta. Por lo tanto, se minimizará la suma de los cuadrados de e_i. Es decir, se elegirán a y b de modo que:

$$\sum_{i=1}^{n} \left[y_i - \left(a + b \cdot x_i \right) \right]^2 \qquad \text{sea mínimo}$$

Esto equivale a minimizar la suma de los cuadrados de las distancias verticales a partir de los puntos respecto de la línea. Este método (llamado de los Mínimos Cuadrados) da valores de a y b (estimaciones de α y β) que tienen muchas propiedades convenientes.

Una condición necesaria para que exista un mínimo relativo es la anulación de las derivadas parciales con respecto a a y b:

$$2\cdot\sum_{i=1}^{n}\left[y_i-\left(a+b\cdot x_i\right)\right]\cdot(-1)=0 \qquad \text{derivada respecto de a}$$

$$2\cdot\sum_{i=1}^{n}\left[y_i-\left(a+b\cdot x_i\right)\right]\cdot\left(-x_i\right)=0 \qquad \text{derivada respecto de b}$$

lo que se puede reescribir como:

$$\sum_{i=1}^{n} y_i = a\cdot n + b\cdot\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i\cdot x_i = a\cdot\sum_{i=1}^{n} x_i + b\cdot\sum_{i=1}^{n}\left(x_i\right)^2$$

esto es un conjunto de ecuaciones lineales con incógnitas a y b, denominadas Ecuaciones Normales.

Resolviendo por determinantes:

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

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

Ejemplo: Los siguientes datos son las mediciones de la Tensión Arterial en 14 pacientes de distintas edades:

Edades	17	18	20	22	25	29	33	35	38	40	42	45	47	50
Tensión	114	134	116	130	125	130	144	148	150	153	124	135	156	142

ajustar una línea recta a estos datos por el método de mínimos cuadrados y utilizarla para estimar la tensión arterial para una persona de 36 años.

$$\sum_{i=1}^{n} x_i = 461 \qquad \sum_{i=1}^{n} y_i \cdot x_i = 63892$$

$$\sum_{i=1}^{n} y_i = 1901 \qquad \sum_{i=1}^{n} x_i^{\,2} = 16819$$

de aquí el sistema de ecuaciones queda:

$$1901 = a \cdot 14 + b \cdot 461$$
$$63892 = a \cdot 461 + b \cdot 16819$$

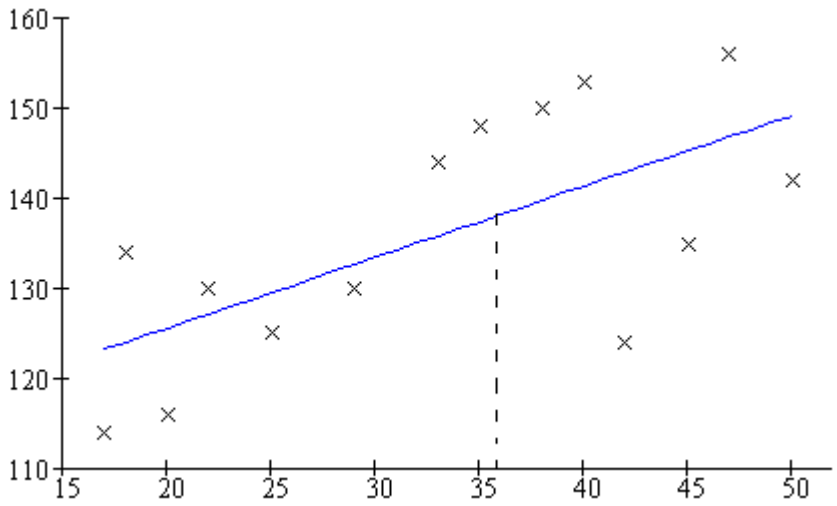
con la soluciones:

$$a = 109.7715 \qquad b = 0.79$$

Para una persona de 36 años de edad:

$$y = 0.79.(36)+109.7715 = 138.2122$$

En el siguiente gráfico se puede apreciar el Diagrama de Dispersión y la recta del mejor ajuste (desde el punto de vista de los mínimos cuadrados) y la estimación para una persona de 36 años de edad:



La siguiente función Matlab permite obtener los resultados vistos del proceso:

```
function recta
% Ajuste lineal de un conjunto de datos por Minimos Cuadrados
% con datos presentes en el archivo ascii regre.txt
```

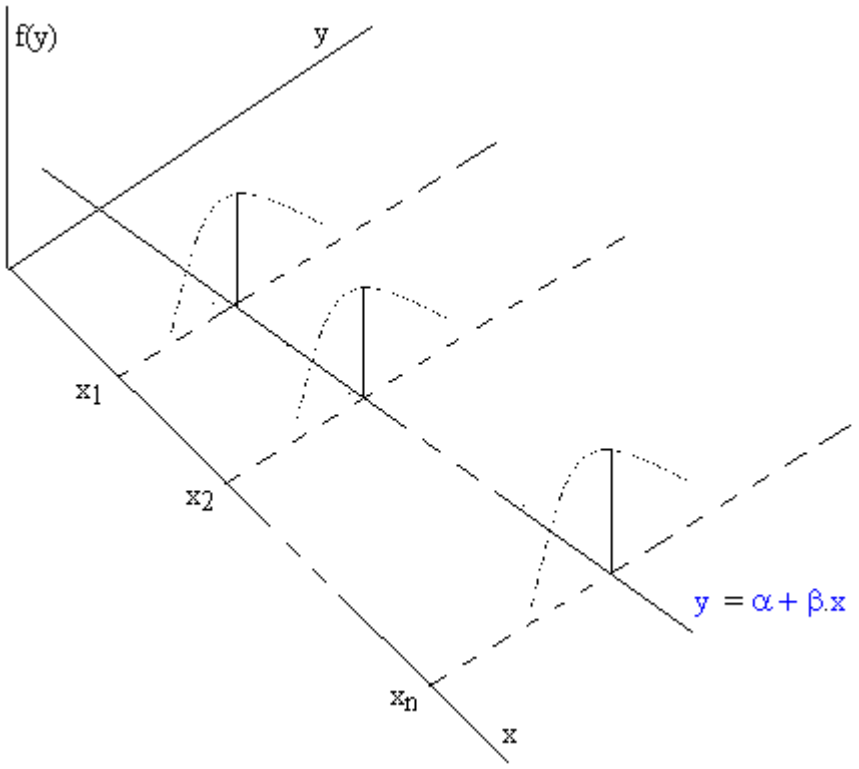
```
% Entradas: u, vector, obtenido del archivo ascii "regre.txt"
% Salida: a, real, Ordenada al origen
%        b, real, pendiente de la recta

load regre.txt;u=regre;n=size(u,1);
sy=0; for i=1:n, sy=sy+u(i,2);end
sx=0; for i=1:n, sx=sx+u(i,1);end
sx2=0; for i=1:n, sx2=sx2+u(i,1)^2;end
sxy=0; for i=1:n, sxy=sxy+u(i,1)*u(i,2);end
A(1,1)=n;A(1,2)=sx;A(2,1)=sx;A(2,2)=sx2;B(1,1)=sy;B(2,1)=sxy;
C=inv(A)*B;a=C(1,1);b=C(2,1);
i=1:n;plot(u(i,1),b*u(i,1)+a,u(i,1),u(i,2),'*')
a
b
```

El Teorema de Gauss-Markov establece: Entre los estimadores insesgados de α y β que son lineales en los y_i, los estimadores de mínimos cuadrados tienen la varianza más pequeña.

INFERENCIAS BASADAS EN ESTIMADORES DE MÍNIMOS CUADRADOS

En lo que sigue se supondrá que la regresión es lineal y, más aún, que las n variables aleatorias que tienen valores y_i (i=1, 2, …, n) son independientes y que están distribuidos normalmente con las medias $\alpha + \beta.x_i$ y la varianza común σ^2.

Si se escribe: $y_i = \alpha + \beta.x_i + \varepsilon_i$ se deriva que los ε_i son valores de variables aleatorias independientes, distribuidas normalmente, y que tienen medias 0 y varianza común σ^2. Gráficamente:



En las suposiciones hechas hasta aquí, como se ilustra, se pueden advertir las distribuciones de los y_i para varios valores de las x_i.

Antes de establecer un teorema relativo a la distribución de los estimadores de mínimos cuadrados de α y β, es conveniente introducir una notación especial:

$$Sxx = n\cdot\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2 = n\cdot sx^2\cdot(n-1)$$

$$Syy = n\cdot\sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i\right)^2 = n\cdot sy^2\cdot(n-1)$$

$$Sxy = n\cdot\sum_{i=1}^{n} x_i\cdot y_i - \sum_{i=1}^{n} x_i\cdot\sum_{i=1}^{n} y_i = n\cdot sxy\cdot(n-1)$$

en base a esto, las ecuaciones normales, resueltas por determinantes, quedan:

$$b = \frac{Sxy}{Sxx} \qquad a = \bar{y} - b\cdot\bar{x}$$

donde $\bar{x}$ e $\bar{y}$ son, respectivamente las medias de las x y de las y. Debe notarse también la estrecha relación entre las Sxx y Syy con las varianzas muestrales respectivas de las x y las y (sx y sy).

La varianza común σ^2 puede estimarse en término de las desviaciones verticales de los puntos muestrales a partir de la línea de mínimos cuadrados. La i-ésima de tales desviaciones es:

$$y_i - (a + b\cdot x_i)$$

De aquí, la estimación, s_e^2, es:

$$s_e^{\,2} = \frac{1}{n-2}\cdot\sum_{i=1}^{n}\left[y_i - (a + b\cdot x_i)\right]^2$$

donde s_e se denomina Error Estándar de Estimación, también la suma de cuadrados, dada por $s_e^2.(n\text{-}2)$, recibe el nombre de *Suma de Cuadrados Residual* o *Suma de Cuadrados de Error*. El número n - 2 en el denominador de (2.23) se llama *grados de libertad* (df). Es igual al número de observaciones menos el número de coeficientes de regresión estimados.

Una fórmula equivalente de esa estimación de σ^2 es:

$$s_e^{\,2} = \frac{Sxx\cdot Syy - Sxy^2}{n\cdot(n-2)\cdot Sxx}$$

el divisor n-2 se emplea para que el estimador resultante de σ^2 sea insesgado.

En base a las suposiciones efectuadas relativas a la distribución de las y, se pueden probar los siguientes teoremas:

Teorema 1: Con las suposiciones dadas, los estadísticos:

con valores de variables aleatorias que tienen la distribución t-Student con n-2 grados de libertad.

Si se requieren intervalos de confianza para los coeficientes de regresión α y β, se sustituye el término medio de $-t_{\alpha/2} < t < t_{\alpha/2}$ por el estadístico t adecuado del teorema anterior. Luego, por medio de cálculos simples, se determinan los correspondientes intervalos de confianza:

$$a - t_{\frac{\alpha}{2}} \cdot s_e \cdot \sqrt{\frac{Sxx + (n \cdot \bar{x})^2}{n \cdot Sxx}} < \alpha < a + t_{\frac{\alpha}{2}} \cdot s_e \cdot \sqrt{\frac{Sxx + (n \cdot \bar{x})^2}{n \cdot Sxx}}$$

$$b - t_{\frac{\alpha}{2}} \cdot s_e \cdot \sqrt{\frac{n}{Sxx}} < \beta < b + t_{\frac{\alpha}{2}} \cdot s_e \cdot \sqrt{\frac{n}{Sxx}}$$

Problema: Los siguientes datos son las mediciones de la velocidad del aire y del coeficiente de evaporación de las gotitas de combustible en una turbina de propulsión:

Velocidad del aire (cm/s)	20	60	100	140	180	220	260	300	340	380
Coeficiente de Evaporación (mm^2/seg)	.18	.37	.35	.78	.56	.75	1.18	1.36	1.17	1.65

Construir un intervalo de confianza del 95% para el coeficiente de regresión α.

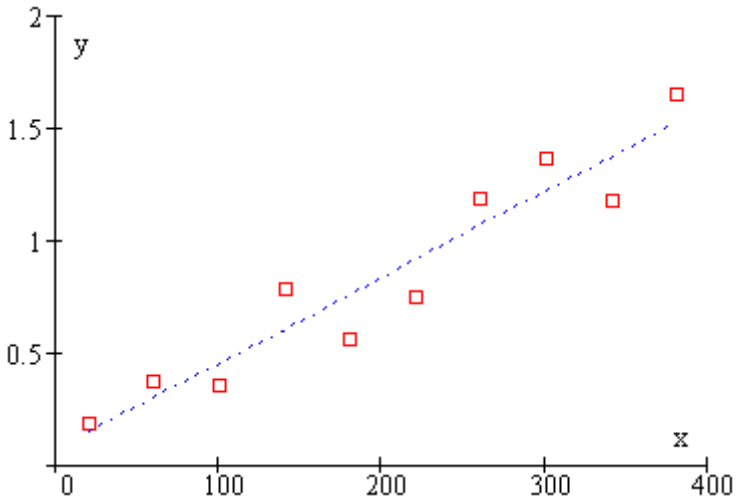
$$Sxx = n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 = 1.32 \times 10^6$$

$$Syy = n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 = 21.375$$

$$Sxy = n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i = 5.054 \times 10^3$$

$$b = \frac{Sxy}{Sxx} = 3.829 \times 10^{-3} \qquad a = \bar{y} - b \cdot \bar{x} = 0.835 - 3.829 \times 10^{-3} \cdot 200 = 0.069$$

Gráficamente:



$1-\alpha = 0.05$; $\alpha/2 = 0.025$; $t_{\alpha/2} = 2.306$ con $\nu = n - 2 = 8$ g.d.l.

los límites de confianza del 95%, para α, se calculan entonces:

$$s_e = \sqrt{\frac{1.32 \times 10^6 \cdot 21.375 - \left(5.054 \times 10^3\right)^2}{10 \cdot (10-2) \cdot \left(1.32 \times 10^6\right)}} = 0.159$$

$$2.306 \cdot 0.159 \cdot \sqrt{\frac{1.32 \times 10^6 + (2000)^2}{10 \cdot \left|1.32 \times 10^6\right|}} = 0.233$$

luego, el intervalo es:

$a - 0.233 = -0.164$ $\qquad$ $a + 0.233 = 0.302$

$-0.164 < \alpha < 0.302$

En las pruebas de hipótesis relativas a los coeficientes de regresión α y β, las que se refieren a β son muy importantes ya que β es la pendiente de la línea de regresión. Esto es, β es el "cambio promedio" de las y correspondiente a un incremento unitario de x. Si $\beta=0$ la línea de regresión es horizontal y la media de las y no "depende linealmente" de x.

Ejemplo: En base al problema anterior, probar la Hipótesis Nula de que $\beta=0$ contra la Hipótesis Alterna que $\beta<>0$, con un nivel de significación de 0.05.

1. Hipótesis nula: $\beta=0$
   Hipótesis alterna: $\beta<>0$
2. Nivel de significación 0.05
3. Criterio: Se rechaza Ho si $t > 2.306$ o $t < -2.306$, con $\nu = n - 2 = 8$ g.d.l.
4. Cálculos:

$$t = \frac{3.829 \times 10^{-3} - 0}{0.159} \cdot \sqrt{\frac{1.32 \times 10^{6}}{10}} = 8.749$$

5. Decisión: Ya que 8.749 > 2.306 Se Rechaza la Hipótesis Nula. Luego, existe relación entre la velocidad del aire y el coeficiente de evaporación promedio (la relación es lineal por las suposiciones que fundamentan la prueba).

Otro problema es estimar $\alpha + \beta.x$, es decir la media de la distribución de las y, para un valor dado de x. Si x se hace igual a un valor fijo x_0 se desea estimar $\alpha + \beta.x_0$ y sería razonable emplear $a + b.x_0$ (con a y b obtenidos por el método de los mínimos cuadrados). Puede verificarse que este estimador es insesgado, y que tiene la varianza:

$$\sigma^2 \cdot \left[\frac{1}{n} + \frac{n \cdot (x_0 - \bar{x})^2}{Sxx} \right]$$

y que los límites de confianza del (1-α).100% para $\alpha + \beta. x_0$ están dados por:

$$(a + b \cdot x_0) \pm t_{\frac{\alpha}{2}} \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{n \cdot (x_0 - \bar{x})^2}{Sxx}} \qquad \text{con } \nu = n-2 \text{ g.d.l}$$

Problema: En relación al ejemplo anterior, construir un intervalo de confianza del 95% para el coeficiente de evaporación medio cuando la velocidad del aire es de 190 cm/seg.

$$2.306 \cdot 0.159 \cdot \sqrt{\frac{1}{10} + \frac{10 \cdot (190 - 200)^2}{1.32 \times 10^{6}}} = 0.116$$

$0.797 - 0.116 = 0.681$ $0.797 + 0.116 = 0.913$

$0.681 < \alpha + \beta. 190 < 0.913$ intervalo de confianza

De mayor importancia aún que la estimación de $\alpha + \beta. x_0$ es la "predicción" de un valor futuro de y cuando $x = x_0$ donde x_0 está dentro del rango de experimentación (se agrega "dentro del rango de experimentación" dado que, la extrapolación es aventurada y se observa que una relación no siempre es válida fuera de tal rango).

Para el primer problema se verifica que para una velocidad de 190 cm/seg (valor situado bien adentro del rango de experimentación) el coeficiente de evaporación es de 0.797 mm^2/seg.

Se describirá un método para construir un intervalo en el cual puede esperarse que una futura observación y se halle con una probabilidad determinada (o confianza) cuando $x = x_0$. Si se conocieran α y β se podría usar el hecho de que y es un valor de una variable aleatoria que tiene distribución normal con la media $\alpha + \beta. x_0$ y varianza σ^2 (o que $y - \alpha - \beta. x_0$ es un valor de una variable aleatoria con distribución normal de media cero y varianza σ^2).

Sin embargo α y β se desconocen, debiéndose considerar la cantidad $y - a - b.x_0$ (donde y, a, b son todas variables aleatorias y la teoría resultante origina los siguientes límites de predicción para y cuando $x = x_0$.

$$\left(a + b \cdot x_0\right) \pm t_{\frac{\alpha}{2}} \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{n \cdot \left(x_0 - \bar{x}\right)^2}{Sxx}} \qquad \text{con } \nu = n\text{-}2 \text{ g.d.l}$$

Problema: Conforme al ejemplo anterior, encontrar los límites de predicción del 95% para una observación del coeficiente de evaporación cuando la velocidad del aire es de 190 cm/seg.

$$2.306 \cdot 0.159 \cdot \sqrt{1 + \frac{1}{10} + \frac{10 \cdot (190 - 200)^2}{1.32 \times 10^6}} = 0.385$$

luego, los límites de predicción son:

$$0.797 - 0.385 = 0.412 \qquad 0.797 + 0.385 = 1.182$$

Comparando con el problema anterior, se ve que si bien la media de la distribución de las y cuando x=190 puede estimarse con bastante precisión, el valor de una simple estimación futura no puede predecirse con mucha precisión.

El ancho del intervalo de predicción depende fundamentalmente de s_e que mide la variabilidad inherente de los datos. Se nota que si se desea extrapolar, el intervalo de predicción (y también el intervalo de confianza para $\alpha + \beta . x_0$) incrementa su ancho.

Problema: Conforme al ejemplo anterior, suponer que la relación de linealidad se cumple más allá del rango de experimentación y calcular los límites de predicción del 95% para una observación del coeficiente de evaporación cuando la velocidad del aire es de 450 cm/seg.

$$2.306 \cdot 0.159 \cdot \sqrt{1 + \frac{1}{10} + \frac{10 \cdot (450 - 200)^2}{1.32 \times 10^6}} = 0.46$$

$$0.797 - 0.46 = 0.337 \qquad 0.797 + 0.46 = 1.257$$

el ancho es de 2*0.46 = 0.92, contra los 2*0.385 = 0.77 del problema anterior.

REGRESIÓN CURVILÍNEA

Se considerará primero el caso en que la graficación en una escala adecuada puede ser lineal. Por ejemplo, si un conjunto de parejas de datos que conste de n puntos (x_i,y_i) "se enderezan" cuando son graficados sobre ejes escalados adecuadamente. E este caso, al ser representados sobre papel semilogarítmico, indican que la curva de regresión de y sobre x es exponencial, es decir para cualquier x considerada, la media de

la distribución está dada por la siguiente ecuación predictora $y = \alpha \cdot \beta^{x}$, tomando logaritmos en ambos miembros:

$$\log(y) = \log(\alpha) + x \cdot \log(\beta)$$

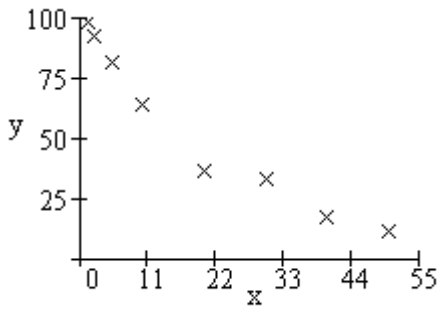
y se puede estimar ahora $\log(\alpha)$ y $\log(\beta)$, y de ahí obtener α y β, aplicando los métodos anteriores a los n pares de valores $[x_i, \log(y_i)]$.

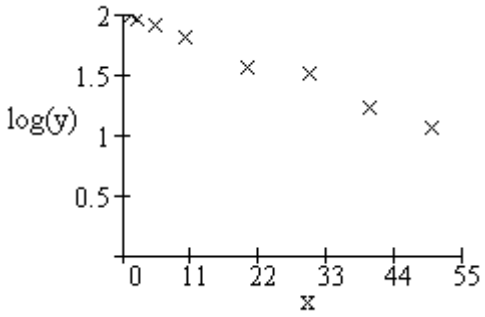
Problema: Las cifras siguientes son datos sobre el porcentaje de llantas radiales producidas por cierto fabricante que aún pueden usarse después de recorrer cierto número de millas:

Miles de Millas recorridas (x)	1	2	5	10	20	30	40	50
Porcentaje útil (y)	98.2	91.7	81.3	64.0	36.4	32.6	17.1	11.3
Log(y)	1.9921	1.9624	1.9101	1.8062	1.5611	1.5132	1.2330	1.0531

a) Graficar los datos proporcionados en escala semilogaritmica para advertir si es razonable que la relación es exponencial.
b) Ajustar una curva exponencial aplicando el método de mínimos cuadrados a las parejas de puntos $[x_i, \log(y_i)]$.
c) Emplear los resultados de la parte b) para estimar qué porcentaje de las llantas radiales del fabricante durarán al menos 25000 millas.

a)





El patrón global (del segundo gráfico) es lineal y esto justifica el ajuste mediante una curva exponencial.

b) Para formar las ecuaciones normales:

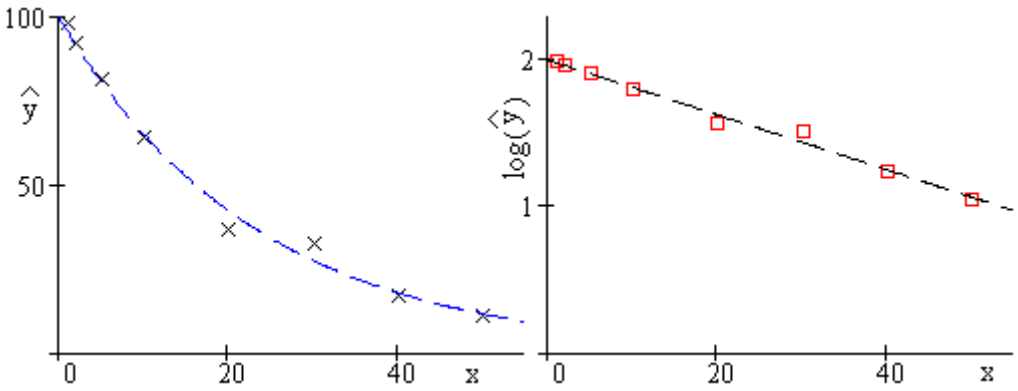
$\Sigma x = 158 \quad \Sigma x^2 = 5530 \quad \Sigma x.\log(y) = 212.1224 \quad \Sigma \log(y) = 13.0312$

$13.0312 = 8 \log(a) + 158 \log(b)$
$212.1224 = 158 \log(a) + 5530 \log(b)$

$\log(a) = 1.9997 \rightarrow a = 99.9408 \qquad \log(b) = -0.0188 \rightarrow b = 0.9577$

Luego, la ecuación de la recta de regresión estimada será:

$$\log(\hat{y}) = 1.9997 - 0.0188 \cdot x \qquad \hat{y} = 99.9408\,(0.9577)^{x}$$



c) Utilizando la última expresión:

$99.9408.(0.9577)^{25} = 33.9$

Vale decir, el 33.9% durarán al menos 25000 millas.

La siguiente función de Matlab, produce los resultados vistos.

```
function logar(x)
% Regresion curvilinea de un conjunto de datos exponencial
% con datos presentes en el archivo ascii expo.txt
% Entradas: u, vector, obtenido del archivo ascii "expo.txt"
%          x, real, valor para el que se quiere hallar la estimacion
% Salida: a, real, Ordenada al origen del ajuste lineal
%         b, real, pendiente de la recta del ajuste linea

load expon.txt;u=expon';n=size(u,1);
slogy=0; for i=1:n, slogy=slogy+log10(u(i,2));end
sx=0; for i=1:n, sx=sx+u(i,1);end
sx2=0; for i=1:n, sx2=sx2+u(i,1)^2;end
slogxy=0; for i=1:n, slogxy=slogxy+u(i,1)*log10(u(i,2));end
A(1,1)=n;A(1,2)=sx;A(2,1)=sx;A(2,2)=sx2;B(1,1)=slogy;B(2,1)=slogxy;
C=inv(A)*B;a=C(1,1);b=C(2,1);
i=1:n;plot(u(i,1),b*u(i,1)+a,u(i,1),log10(u(i,2)),'*')
estima=(10^a)*(10^b)^x
```

de modo que ejecutando:

```
>> logar(25)
estima =
  33.9088
```

Hay dos relaciones más, muy aplicadas: La función potencial $y = \alpha\ x^{\beta}$ y la función recíproca $y = 1/(\alpha + \beta.x)$.

Para el primer caso, al ser representado el conjunto de datos sobre papel doble logarítmico toma la forma de recta, esto significa que los valores siguen una ley potencial.

Si la ecuación predictora está dada por:

$y = \alpha\ .\ x\ \beta$

tomando logaritmos en ambos miembros, queda:

$$\log(y) = \log(\alpha) + \beta \cdot \log(x)$$

En este caso habrá que considerar tanto los logaritmos de los elementos de y como los de x.

Problema: Sea el siguiente conjunto de valores, las lecturas de un experimento donde x es la variable independiente (controlada, medida con poco error) e y la variable resultante.

X	1	2	3	4	5	6	7
Y	6.5	40	90	140	250	500	700

$\Sigma \log(x) = 3.7024$ $\Sigma [\log(x)]^2 = 2.4890$ $\Sigma \log(x).\log(y) = 8.8875$
$\Sigma [\log(y)] = 14.4574$

con lo que las ecuaciones normales quedan:

$14.4574 = 7 \log(a) + 3.7024$ (b)
$8.8875 = 3.7024 \log(a) + 2.4890$ (b)

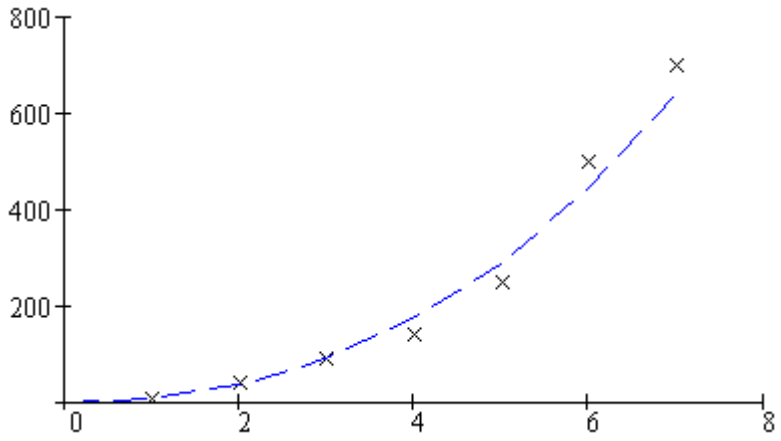
$\log(a) = 0.8289$ ➔ $a = 6.7437$ $b = 2.338$

Luego, la ecuación de la recta de regresión estimada será:

$y' = 0.8229 + 2.338 \cdot x'$

y la función predictora: $y = 6.7437 \cdot x^{2.338}$

Los resultados gráficos son:



La siguiente función de Matlab, produce los resultados vistos.

```
function potencia(x)
% Regresion curvilinea de un conjunto de datos potencial
```

```
% con datos presentes en el archivo ascii poten.txt
% Entradas: u, vector, obtenido del archivo ascii "poten.txt"
%          x, real, valor para el que se quiere hallar la estimacion
% Salida: a, real, Ordenada al origen del ajuste lineal
%         b, real, pendiente de la recta del ajuste linea
%         estima, real, estimacion correspondiente a x
load poten.txt
load poten.txt;u=poten';n=size(u,1);
slogy=0; for i=1:n, slogy=slogy+log10(u(i,2));end
slogx=0; for i=1:n, slogx=slogx+log10(u(i,1));end
slogx2=0; for i=1:n, slogx2=slogx2+log10(u(i,1))^2;end
slogxy=0; for i=1:n, slogxy=slogxy+log10(u(i,1))*log10(u(i,2));end
A(1,1)=n;A(1,2)=slogx;A(2,1)=slogx;A(2,2)=slogx2;B(1,1)=slogy;B(2,1)=slogxy;
C=inv(A)*B;a=C(1,1);b=C(2,1);
a=10^a;
i=1:n;plot(u(i,1),a*u(i,1).^b,u(i,1),u(i,2),'*');
a
b
estima=a*x^b
```

de modo que ejecutando:

```
>> potencia(2)
a =
   6.7431
b =
   2.3377
estima =
  34.0872
```

Para el caso de la <u>función recíproca</u> $y = 1/(\alpha + \beta.x)$, se obtienen α y β, aplicando los métodos anteriores a los n pares de valores $[x_i, 1/y_i]$.

Problema: Sea el siguiente conjunto de valores, las lecturas de un experimento donde x es la variable independiente (controlada, medida con poco error) e y la variable resultante.

X	1	2	3	4	5	6	7
Y	1.5	1	0.8	0.85	0.6	0.5	0.55

$\Sigma (x) = 28 \quad \Sigma x^2 = 140 \quad \Sigma x.1/y = 44.183 \quad \Sigma 1/y = 9.578$

con lo que las ecuaciones normales quedan:

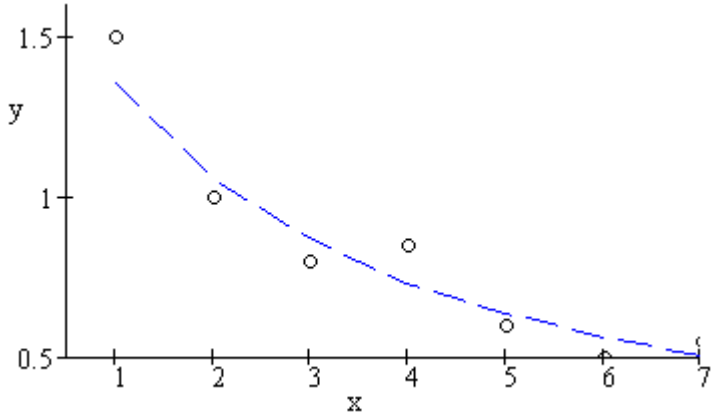
$9.578 = 7\ a + 28\ b$
$44.183 = 28\ a + 140\ b \quad \rightarrow a = 0.53 \qquad b = 0.21$

Luego, la ecuación de la recta de regresión estimada será:

$y' = 0.53 + 0.21\ x'$

y la función predictora: $y = 1/(0.53 + 0.21\ x)$

Los resultados gráficos son:



La siguiente función de Matlab, produce los resultados vistos.

```
function reciproca(x)
% Regresion curvilinea de un conjunto de datos reciprocos
% con datos presentes en el archivo ascii reci.txt
% Entradas: u, vector, obtenido del archivo ascii "reci.txt"
%          x, real, valor para el que se quiere hallar la estimacion
% Salida: a, real, Ordenada al origen del ajuste lineal
%        b, real, pendiente de la recta del ajuste linea
%        estima, real, estimacion correspondiente a x
load reci.txt;u=reci';n=size(u,1);
sy=0; for i=1:n, sy=sy+1/(u(i,2));end
sx=0; for i=1:n, sx=sx+u(i,1);end
sx2=0; for i=1:n, sx2=sx2+u(i,1)^2;end
sxy=0; for i=1:n, sxy=sxy+u(i,1)*1/u(i,2);end
A(1,1)=n;A(1,2)=sx;A(2,1)=sx;A(2,2)=sx2;B(1,1)=sy;B(2,1)=sxy;
C=inv(A)*B;a=C(1,1);b=C(2,1);
i=1:n;plot(u(i,1),1./(a+b*u(i,1)),u(i,1),u(i,2),'*');end
a
b
estima=1/(a+b*x)
```

de modo que ejecutando:

```
>> reciproca(1)
a =
   0.5295
b =
   0.2097
estima =
   1.3528
```

Si no hay ninguna indicación acerca de la forma funcional de la regresión de y sobre x, se supone a menudo que la relación fundamental al menos “se comporta bien” al grado que admita un desarrollo en Serie de Taylor y que los primeros términos constituyen una aproximación bastante buena.

Vale decir, los datos se ajustan a un polinomio o ecuación predictora de la forma

$$y = \beta_0 + \beta_1.x + \beta_2.x^2 + \ldots + \beta_p.x^p$$

donde el grado se determina por observación de los datos o por un método más riguroso como el siguiente: dado un conjunto de datos que consta de n puntos (x_i,y_i) se estiman los coeficientes β_0 , β_1 , β_2,… β_p del polinomio de p-ésimo grado, minimizando:

$$\sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + . + \beta_p \cdot x^p \right) \right]^2$$

diferenciado parcialmente con respecto a β_0 , β_1 , β_2,… β_p , igualando estas derivadas parciales a cero, reacomodando términos e indicando con b_i las estimaciones de β_i, se obtienen las p+1ecuaciones normales:

$$\sum_{i=1}^{n} y_i = b_0 \cdot n + b_1 \cdot \sum_{i=1}^{n} x_i + . + b_p \cdot \sum_{i=1}^{n} \left(x_i \right)^p$$

$$\sum_{i=1}^{n} x_i \cdot y_i = b_0 \cdot \sum_{i=1}^{n} x_i + b_1 \cdot \sum_{i=1}^{n} \left(x_i \right)^2 + . + b_p \cdot \sum_{i=1}^{n} \left(x_i \right)^{p+1}$$

.....

$$\sum_{i=1}^{n} \left(x_i \right)^p \cdot y_i = b_0 \cdot \sum_{i=1}^{n} \left(x_i \right)^p + b_1 \cdot \sum_{i=1}^{n} \left(x_i \right)^{p+1} + . + b_p \cdot \sum_{i=1}^{n} \left(x_i \right)^{2 \cdot p}$$

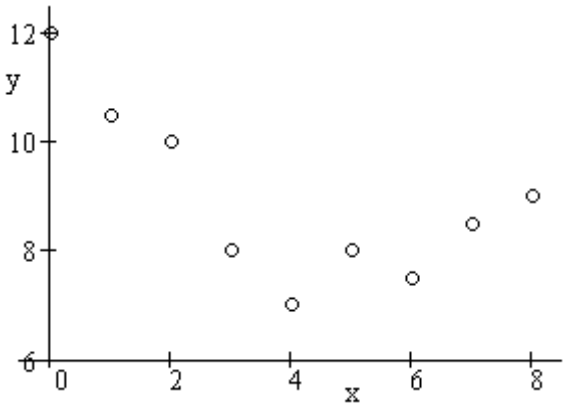
siendo b_0, b_1, …, b_p las p+1 incógnitas.

Problema: Los datos siguientes corresponden al tiempo de secado de cierto barniz y a la cantidad de un aditivo con que se intenta reducir el tiempo de secado:

Cantidad de aditivo (en gr.)	0	1	2	3	4	5	6	7	8
Tiempo de secado (en seg.)	12.0	10.5	10.0	8.0	7.0	8.0	7.5	8.5	9.0

a) Dibujar el diagrama de dispersión de modo que permita advertir si es razonable una relación parabólica.
b) Ajustar un polinomio de segundo grado por el método de mínimos cuadrados.
c) Emplear el resultado de b) para predecir el valor del tiempo de secado cuando se han utilizado 6.5 gr. del aditivo.

a)



b) Cálculos:

$$\sum_{i=1}^{n} x_i = 36 \qquad \sum_{i=1}^{n} x_i^{\,2} = 204 \qquad \sum_{i=1}^{n} x_i^{\,3} = 1.296 \times 10^3$$

$$\sum_{i=1}^{n} x_i^{\,4} = 8.772 \times 10^3 \qquad \sum_{i=1}^{n} y_i^{\,2} = 740.75$$

$$\sum_{i=1}^{n} y_i = 80.5 \qquad \sum_{i=1}^{n} x_i^{\,2} \cdot y_i = 1.697 \times 10^3 \qquad \sum_{i=1}^{n} x_i \cdot y_i = 299$$

con lo que las ecuaciones normales quedan:

$$80.5 = 9 \cdot b_0 + 36 \cdot b_1 + 204 \cdot b_2$$

$$299 = 36 \cdot b_0 + 204 \cdot b_1 + 1296 \cdot b_2$$

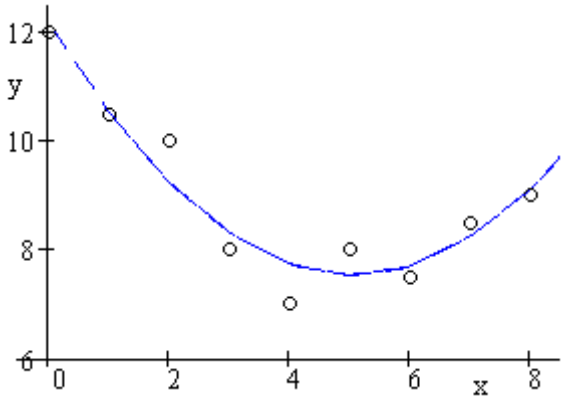
$$1697 = 204 \cdot b_0 + 1296 \cdot b_1 + 8772 \cdot b_2$$

$b_0 = 12.2$ $b_1 = -1.85$ $b_2 = 0.183$

La ecuación del polinomio será:

$$\hat{y} = 12.2 - 1.85 \cdot x + 0.183 \cdot x^2$$

gráficamente:



d) sustituyendo x = 6.5, da:

$$\hat{y} = 12.2 - 1.85 \cdot 6.5 + 0.183 \cdot 6.5^2 = 7.907$$

La siguiente función de Matlab, produce los resultados vistos.

```
function parabola(x)
% Regresion curvilinea de un conjunto de datos de origen cuadratico
% con datos presentes en el archivo ascii parabo.txt
% Entradas: u, vector, obtenido del archivo ascii "parabo.txt"
%           x, real, valor para el que se quiere hallar la estimacion
% Salida: b0, b1, b2, reale, coeficientes del polinomio de ajuste
%         estima, real, estimacion correspondiente a x
load parabo.txt;u=parabo';n=size(u,1);
sy=0; for i=1:n, sy=sy+u(i,2);end
sx=0; for i=1:n, sx=sx+u(i,1);end
sxy=0; for i=1:n, sxy=sxy+u(i,1)*u(i,2);end
sx2=0; for i=1:n, sx2=sx2+u(i,1)^2;end
sx3=0; for i=1:n, sx3=sx3+u(i,1)^3;end
sx4=0; for i=1:n, sx4=sx4+u(i,1)^4;end
sx2y=0; for i=1:n, sx2y=sx2y+(u(i,1)^2)*u(i,2);end
sy2=0; for i=1:n, sy2=sy2+u(i,2)^2;end
A(1,1)=n;A(1,2)=sx;A(1,3)=sx2;
A(2,1)=sx;A(2,2)=sx2;A(2,3)=sx3;
A(3,1)=sx2;A(3,2)=sx3;A(3,3)=sx4;
B(1,1)=sy;B(2,1)=sxy;B(3,1)=sx2y;
C=inv(A)*B;b0=C(1,1);b1=C(2,1);b2=C(3,1);
C
i=1:n;plot(u(i,1),b0+b1.*u(i,1)+b2.*u(i,1).^2,u(i,1),u(i,2),'*');end
estima=b0+b1*x+b2*x^2
```

de modo que ejecutando:

```
>> parabola(6.5)
C =
   12.1848
   -1.8465
    0.1829
estima =
    7.9099
```

En la práctica, puede ser difícil determinar el grado del polinomio que se ajusta a un conjunto de parejas de datos. Como siempre, es posible hallar un polinomio de grado

n-1 que pase a través de los n puntos correspondientes a n valores distintos de x. Debe ser claro el objetivo de encontrar un polinomio de grado mínimo que "describa adecuadamente" a los datos. A menudo es posible determinar el grado con la simple observación de los datos.

Existe también un método más estricto para determinar el grado de un polinomio que se ajuste a un conjunto de datos. En esencia, consiste en ajustar inicialmente a una línea recta, así como a un polinomio de segundo grado y probar la Hipótesis Nula $\beta_2=0$. Es decir, nada se gana incluyendo el término cuadrático.

Si esta Hipótesis Nula puede rechazarse, entonces se ajusta con un polinomio de tercer grado y se prueba la Ho $\beta_3=0$. Es Es decir, nada se gana incluyendo el término cúbico.

Este procedimiento se continua hasta que la Ho $\beta_t=0$ no pueda ser rechazada en dos etapas sucesivas, no existe pues ventaja en utilizar términos adicionales. Para aplicar estas pruebas, se requieren las suposiciones de normalidad, independencia y varianzas iguales introducidas al principio.

AJUSTE POLINOMIAL MEDIANTE LA VARIAZA RESIDUAL

Como se ha dicho más arriba, cuando se ajusta un polinomio a un conjunto de parejas de datos, se suele empezar ajustando una línea recta y se prueba la Ho $\beta_1=0$. Entonces se ajusta un polinomio de segundo grado y se prueba si vale la pena conservar el término cuadrático comparando $\hat{\sigma}_1^2$, la varianza residual después de ajustar la línea recta, con $\hat{\sigma}_2^2$, la varianza residual después de ajustar el polinomio de segundo grado.

Cada una de estas varianzas residuales está dada por:

$$\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\text{grados de libertad}}$$

con $\hat{y}$ determinada, respectivamente, de la ecuación de la recta y de la ecuación de segundo grado. Los grados de libertad se determinan restando el número de puntos considerados y los coeficientes estimados.

Este proceso se reitera, hacia grados superiores, hasta que la varianza residual produzca "el salto decreciente más significativo".

Problema: Dado el siguiente conjunto de datos:

x	.5	1.5	2.5	5.5	6.5	9.5	10.5	12.5	14.5	15.5
y	3	7	12.5	14.5	16	14.5	16	16	21	23

encontrar el polinomio de mejor ajuste.

Se intenta en primer lugar, el ajuste lineal. Para ello se determinan y resuelven las ecuaciones normales:

$\Sigma x = 79 \quad \Sigma x^2 = 888.5 \quad \Sigma x.y = 1394 \quad \Sigma y = 143.5$

$143.5 = 10\, b_0 + 79\, b_1$

$1394 = 79\ b_0 + 888.5\ b_1$ ➔ $b_0 = 6.578$ $b_1 = 0.984$

luego, se calcula la varianza residual:

$$res_1 = \frac{\sum_{i=1}^{10} \left(y_i - b_0 - b_1 \cdot x_i \right)^2}{10 - 2} = 7.207$$

Se sigue con un ajuste cuadrático:

$\Sigma\ x^2\ y = 16720$ $\Sigma\ x^3 = 11200$ $\Sigma\ x^4 = 149400$

$143.5 = 10\ b_0 + 79\ b_1 + 888.5\ b_2$
$1394 = 79\ b_0 + 888.5\ b_1 + 11200\ b_2$
$16720 = 888.5\ b_0 + 11200\ b_1 + 149400\ b_2$ ➔ $b_0 = 5.399$ $b_1 = 1.5$ $b_1 = -0.033$

$$res_2 = \frac{\sum_{i=1}^{10} \left[y_i - b_0 - b_1 \cdot x_i - b_2 \cdot \left(x_i \right)^2 \right]^2}{10 - 3} = 7.58$$

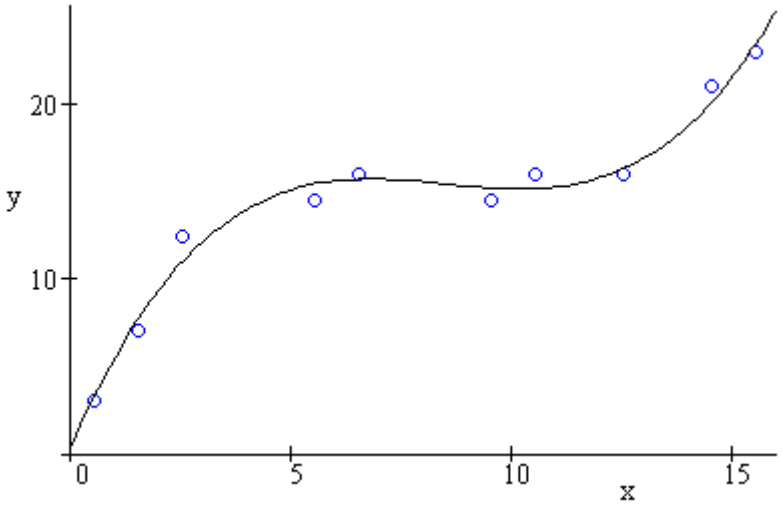
y así hasta llegar a un polinomio de grado 9, que pasará por todos los puntos (varianza residual nula).

Una tabla con las varianzas residuales para cada ajuste es la siguiente (hasta el orden cuártico):

Lineal	Cuadrático	Cúbico	Cuartico
7.207	7.58	1.021	0.992

Se ve que el salto mayor se produce entre el ajuste cuadrático y el cúbico, por lo tanto el mejor estimador lo constituye el ajuste cúbico.

Gráficamente:



La siguiente función de Matlab, permite encontrar las varianzas residuales para ajuste lineal y de mayor orden, lo que permite decidir el ajuste más conveniente a realizar:

```
function residual
% Ajuste de un conjunto de datos por medio de un polinomio
% con datos presentes en el archivo ascii resi.txt
% por el metodo de la varianza residual
% Entradas: u, vector, obtenido del archivo ascii "resi.txt"
% Salida: re, real, varianza residual para cada ajuste (desde lineal)
load resi.txt;u=resi';n=size(u,1);
A(1,1)=n;
B(1,1)=0; for i=1:n, B(1,1)=B(1,1)+u(i,2);end
B(2,1)=0; for i=1:n, B(2,1)=B(2,1)+u(i,2)*u(i,1);end
A(1,2)=0; for i=1:n, A(1,2)=A(1,2)+u(i,1);end
A(2,1)=A(1,2);
A(2,2)=0; for i=1:n, A(2,2)=A(2,2)+u(i,1)^2;end
C=inv(A)*B;re=0; for i=1:n, re=re+(u(i,2)-C(1,1)-C(2,1)*u(i,1))^2/(n-2);end
z=3;
re
while z<7,
  for j=1:z
     A(j,z)=0; for i=1:n, A(j,z)=A(j,z)+u(i,1)^(z+j-2);end
     if j<z, A(z,j)=A(z-1,j+1);end
  end
  B(z,1)=0; for i=1:n, B(z,1)=B(z,1)+u(i,2)*u(i,1)^(z-1);end
  C=inv(A)*B;
  re=0;
  for i=1:n,
       aju=0;
         for j=1:z,
            aju=aju+(C(j,1)*u(i,1)^(j-1));
         end
         dif(i)=aju;
         re=re+(u(i,2)-dif(i))^2;
     end
     re=re/(n-z); z=z+1;
     re
  end
```

Ejecutando:

```
>> residual
re =
   7.2069
re =
   7.5798
re =
   1.0206
re =
   0.9917
re =
   1.2390
```

REGRESIÓN MÚLTIPLE

Es necesario señalar que las curvas obtenidas (y las superficies a obtener) no sólo se utilizan para hacer predicciones . A menudo también se emplean para fines de

optimización, es decir, para determinar los valores de la variable independiente (o variables) de tal manera que la variable dependiente sea un máximo o un mínimo. En el caso del barniz del problema de ajuste cuadrático, el tiempo de secado tiene un mínimo cuando la cantidad de aditivo es de 5.1 gramos.

Esto se obtiene derivando $\hat{y} = 12.2 - 1.85 \cdot x^2 + 0.183 \cdot x^2$ e igualando a cero.

Los métodos estadísticos de predicción y optimización suelen ser incluidos bajo el título general de Análisis de las Superficies de Respuesta.

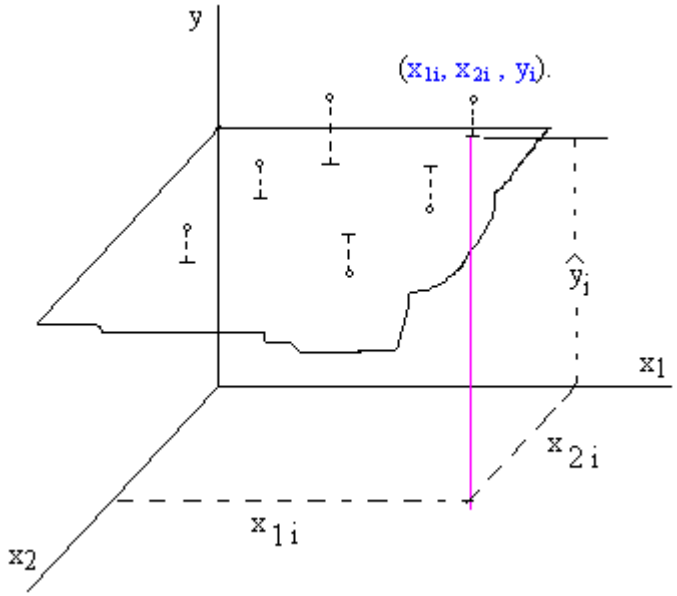
En la regresión múltiple, se manejan datos que constan de n (r+1) coordenadas (x_{1i}, x_{2i} , …, x_{ri}, y_i) donde otra vez se supone que las x_i se conocen sin error, mientras que las y son valores de variables aleatorias. Datos de esta clase aparecen en:

- Estudios diseñados para determinar el efecto que ejercen en la resistencia mecánica de un metal la corrosión bajo varias condiciones climáticas.
- El efecto que la temperatura de horneado, humedad y contenido de hierro tienen en la resistencia mecánica de un revestimiento cerámico.
- El efecto de la producción industrial, nivel de consumo y existencias almacenadas producen en el precio de un producto.

Como en el caso de una sola variable, en primer término se aborda el problema en que la ecuación de regresión es lineal, es decir cuando para cualquier conjunto determinado de valores x_1, x_2 , …, x_r la media de la distribución es:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x_r$$

En el caso de dos variables independientes, el problema es ajustar un plano a un conjunto de n puntos con coordenadas (x_{1i}, x_{2i} , y_i). Gráficamente:



Aplicando el método de los mínimos cuadrados para obtener estimaciones de β_0, β_1 y β_2, se minimiza la suma de los cuadrados de las distancias verticales de los puntos del plano, es decir minimizar:

$$\sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} \right) \right]^2$$

Las ecuaciones normales resultantes de la aplicación de derivadas parciales son:

$$\sum_{i=1}^{n} y_i = b_0 \cdot n + b_1 \cdot \sum_{i=1}^{n} x_{1i} + b_2 \cdot \sum_{i=1}^{n} x_{2i}$$

$$\sum_{i=1}^{n} y_i \cdot x_{1i} = b_0 \cdot \sum_{i=1}^{n} x_{1i} + b_1 \cdot \sum_{i=1}^{n} \left(x_{1i} \right)^2 + b_2 \cdot \sum_{i=1}^{n} x_{1i} \cdot x_{2i}$$

$$\sum_{i=1}^{n} y_i \cdot x_{2i} = b_0 \cdot \sum_{i=1}^{n} x_{2i} + b_1 \cdot \sum_{i=1}^{n} x_{1i} \cdot x_{2i} + b_2 \cdot \sum_{i=1}^{n} \left(x_{2i} \right)^2$$

Estas son las ecuaciones normales para regresión múltiple con r=2. Donde b_0, b_1 y b_2 son estimadores de mínimos cuadrados para β_0, β_1 y β_2.

Problema: Los datos siguientes provienen del número de torsiones necesarios para romper una barra hecha con cierto tipo de aleación y los porcentajes de metales que la integran:

Nro, de Torsiones (x)	38	40	85	59	40	60	68	53	31	35	42	59	18	34	29	42
Porc. Del elemento A (x_1)	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Porc. Del elemento B (x_2)	5	5	5	5	10	10	10	10	15	15	15	15	20	20	20	20

Ajustar un plano de regresión por mínimos cuadrados y emplear su ecuación para estimar el número de torsiones requeridas para romper una de las barras cuando x_1= 2.5 y x_2 = 12.

Sustituyendo en las ecuaciones normales anteriores:

$\Sigma x_1 = 40$ $\Sigma x_2 = 200$ $\Sigma x_1^2 = 120$ $\Sigma x_1 x_2 = 500$ $\Sigma x_2^2 = 3000$
$\Sigma y = 733$ $\Sigma x_1 y = 1989$ $\Sigma x_2 y = 8285$

$733 = 16\, b_0 + 40\, b_1 + 200\, b_2$
$1989 = 40\, b_0 + 200\, b_1 + 500\, b_2$
$8285 = 200\, b_0 + 500\, b_1 + 3000\, b_2$ ➔ $b_0 = 48.2$ $b_1 = 7.83$ $b_1 = -1.76$

el plano de regresión tiene entonces la ecuación:

$$\hat{y} = 48.2 + 7.83 \cdot x_1 - 1.76 \cdot x_2$$

sustituyendo por x_1= 2.5 y x_2 = 12:

$$\hat{y} = 48.2 + 7.83 \cdot 2.5 - 1.76 \cdot 12 = 46.7$$

La siguiente función de Matlab, permite encontrar los coeficientes del plano de regresión y la estimación para una par de valores x_1 y x_2.

```
function multiple(x1,x2)
% Regresion multiple de un conjunto de datos
% con datos presentes en el archivo ascii multip.txt
% Entradas: u, matriz, obtenida del archivo ascii "multip.txt"
%          x1,x2, reales, valores para los que se quiere hallar la estimacion
% Salida: b0, b1, b2, reales, coeficientes del polinomio de ajuste
%         estima, real, estimacion correspondiente a x
load multip.txt;u=multip';n=size(u,1);
sy=0; for i=1:n, sy=sy+u(i,1);end
sx1=0; for i=1:n, sx1=sx1+u(i,2);end
sx2=0; for i=1:n, sx2=sx2+u(i,3);end
sx12=0; for i=1:n, sx12=sx12+u(i,2)^2;end
sx22=0; for i=1:n, sx22=sx22+u(i,3)^2;end
sx1x2=0; for i=1:n, sx1x2=sx1x2+u(i,2)*u(i,3);end
sx1y=0; for i=1:n, sx1y=sx1y+u(i,2)*u(i,1);end
sx2y=0; for i=1:n, sx2y=sx2y+u(i,3)*u(i,1);end
A(1,1)=n;A(1,2)=sx1;A(1,3)=sx2;
A(2,1)=sx1;A(2,2)=sx12;A(2,3)=sx1x2;
A(3,1)=sx2;A(3,2)=sx1x2;A(3,3)=sx22;
B(1,1)=sy;B(2,1)=sx1y;B(3,1)=sx2y;
C=inv(A)*B;b0=C(1,1);b1=C(2,1);b2=C(3,1);
estima=b0+b1*x1+b2*x2
```

Ejecutando:

```
>> multiple(2.5,12)
C =
  48.1875
   7.8250
  -1.7550
estima =
  46.6900
```

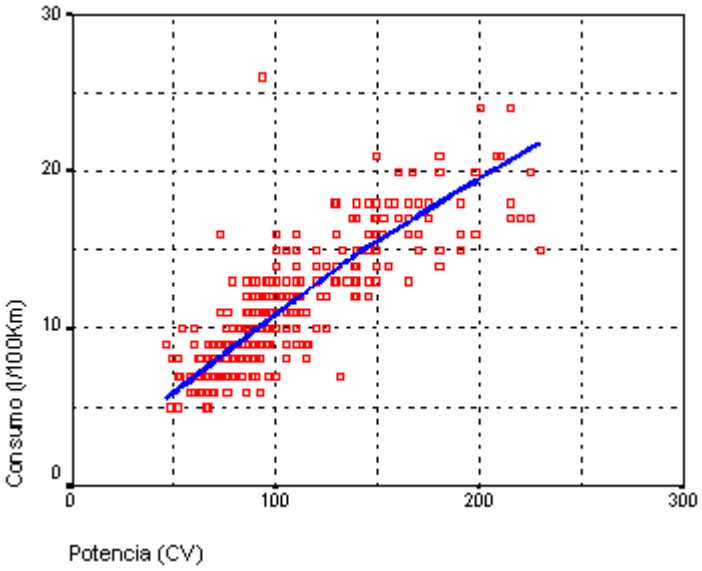
COEFICIENTE DE CORRELACIÓN SIMPLE DE PEARSON (MODELO RECTILÍNEO)

El coeficiente de correlación es una medida de asociación entre dos variables y se simboliza con la literal r. Los valores de la correlación van de + 1 a – 1, pasando por el cero, el cual corresponde a ausencia de correlación. Los primeros dan a entender que existe una correlación directamente proporcional e inversamente proporcional. El coeficiente de correlación permite predecir si entre dos variables existe o no una relación o dependencia matemática.

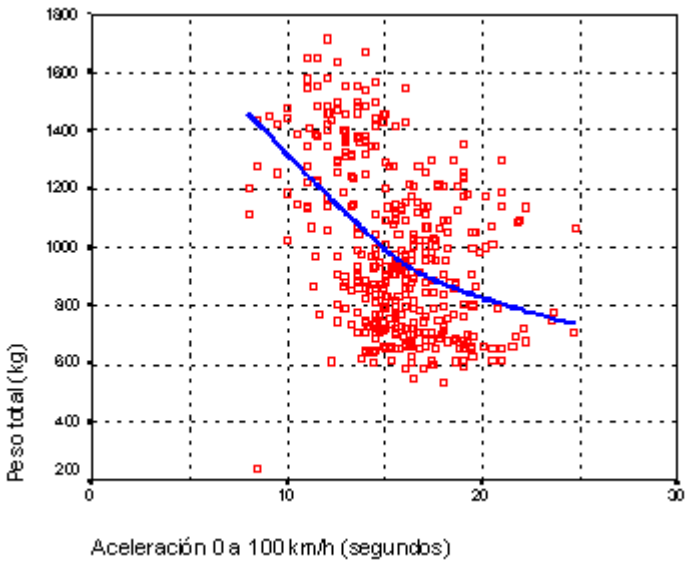
Supóngase que se quiere estudiar la correlación existente entre peso y altura de un grupo de personas tomadas al azar. Se someten los datos recogidos de peso y altura al análisis de correlación y se encuentra el coeficiente de correlación entre ambas, resultando r = 0.78. Esto significa que a mayor altura correspondería mayor peso.

Se puede representar la correlación entre las dos variables a través de una gráfica de dos ejes (abscisas y ordenadas) cartesianos.

En el siguiente gráfico se observa la correlación entre potencia de motor de un automóvil y consumo en Litros por cada 100 Km. El r = 0.87 (correlación positiva). (SPSS). Evidentemente a mayor potencia se observa mayor consumo de combustible. Esto quiere decir que la correlación entre potencia y consumo no es aleatoria.



En el siguiente gráfico se encuentra la relación existente entre peso del automóvil en kg. y aceleración 0 a 100 Km. / hora en segundos. Para este caso, r=- 0.56. Esto significa que existe una correlación negativa significativa, entre peso del auto y respuesta de la aceleración.



Para interpretar el coeficiente de correlación, **Colton** ha dado los siguientes lineamientos generales:

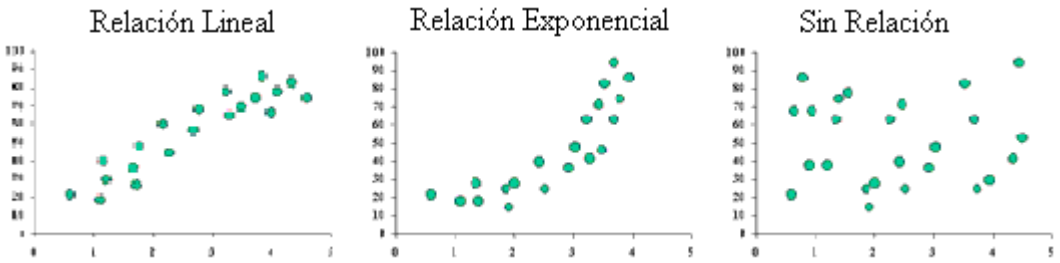
- Valor de r de 0 a 0.25 implica que no existe correlación entre ambas variables.
- Valor de r de 0.25 a 0.50 implica una correlación baja a moderada.
- Valor de r de 0.50 a 0.75 implica correlación moderada a buena.
- Valor de r de 0.75 o mayor, implica una muy buena a excelente correlación.

- Estos rangos de valores se pueden extrapolar a correlaciones negativas también.

Se debe tener cuidado al analizar la correlación entre dos variables, de que ambas varíen juntas permanentemente. Esto parece redundante, pero es importante. Por ejemplo, si se correlaciona edad y altura. La altura irá aumentando con la edad hasta un determinado punto en donde ya no aumentará más.

Puede que exista una relación que no sea lineal, sino exponencial, parabólica, etc. En estos casos, el coeficiente de correlación lineal mediría mal la intensidad de la relación de las variables, por lo que convendría utilizar otro tipo de coeficiente más apropiado.

Para ver, por tanto, si se puede utilizar el coeficiente de correlación lineal, lo mejor es representar los pares de valores en un gráfico y ver que forma describen.



Para calcular el **coeficiente de correlación lineal** se considera el caso en que las dos variables son aleatorias. Dada una muestra:

$$(x_1,y_1)\ (x_2,y_2)\ \ \ldots\ldots\ \ (x_n,y_n)$$

de tamaño n proveniente de una población bidimensional (X,Y), se pueden determinar:

$$\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i \qquad s_x^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$$

del mismo modo:

$$\bar{y} = \frac{1}{n}\cdot\sum_{i=1}^{n} y_i \qquad s_y^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$$

Se define como Covarianza:

$$s_{xy} = \frac{1}{n-1}\cdot\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)\right]$$

de modo que el coeficiente de correlación de la muestra es:

$$r = \frac{s_{xy}}{s_x\cdot s_y}$$

ya que $s_x > 0$ y $s_y > 0$, el producto $s_x\ s_y > 0$ y r puede ser positivo, negativo o nulo según lo sea s_{xy}.

En base a lo expresado, la suma de cuadrados de error en un ajuste lineal es:

$$q = \sum_{i=1}^{n} \left(y_i - a - b \cdot x_i \right)^2$$

además:

$$s_{xy} = r \cdot s_x \cdot s_y \qquad b = \frac{s_{xy}}{s_x^2} \qquad \bar{y} = a + b \cdot \bar{x}$$

queda:

$$q = \sum_{i=1}^{n} \left(y_i - \bar{y} + \frac{s_{xy}}{s_x^2} \cdot \bar{x} - \frac{s_{xy}}{s_x^2} \cdot x_i \right)^2 = \sum_{i=1}^{n} \left[\left(y_i - \bar{y} \right) - \frac{s_{xy}}{s_x^2} \cdot \left(x_i - \bar{x} \right) \right]^2$$

$$q = \sum_{i=1}^{n} \left[\left(y_i - \bar{y} \right)^2 - 2 \cdot \left(y_i - \bar{y} \right) \cdot \frac{s_{xy}}{s_x^2} \cdot \left(x_i - \bar{x} \right) + \frac{s_{xy}^2}{s_x^4} \cdot \left(x_i - \bar{x} \right)^2 \right]$$

$$q = (n-1) \cdot s_y^2 - 2 \cdot (n-1) \cdot \frac{s_{xy}^2}{s_x^2} + (n-1) \cdot s_x^2 \cdot \frac{s_{xy}^2}{s_x^4}$$

$$q = (n-1) \cdot s_y^2 - (n-1) \cdot \frac{s_{xy}^2}{s_x^2} = (n-1) \cdot \left(s_y^2 - r^2 \cdot s_y^2 \right)$$

$$q = (n-1) \cdot s_y^2 \cdot \left(1 - r^2 \right)$$

puesto que q es una suma de cuadrados, debe ser mayor que cero. Por lo tanto, conforme a la última expresión, los tres factores deben ser positivos. Luego:

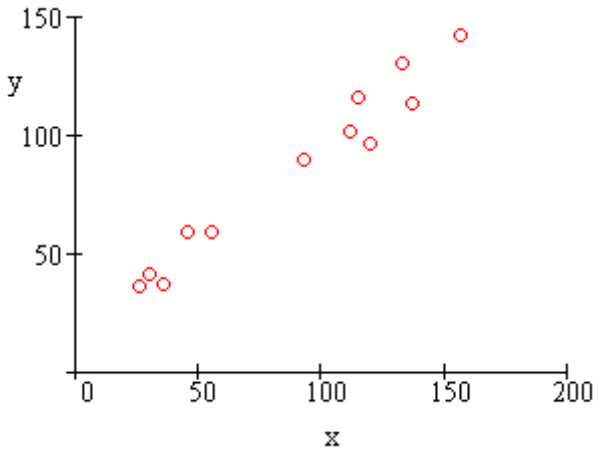
$1 - r^2 \geq 0$ o bien $r^2 \leq 1$ por lo tanto: **$-1 \leq r \leq 1$**

Los valores de la muestra (x_1,y_1) (x_2,y_2) …… (x_n,y_n) se localizan sobre una recta si y sólo si el coeficiente de correlación tiene los valores +1 ó –1.

Problema: Sea la siguiente muestra:

X	26	45	111	92	119	114	136	156	132	55	30	35
Y	36	59	102	90	97	116	114	143	131	59	41	37

El diagrama de dispersión correspondiente es:



y para calcular el coeficiente de correlación, se procede del siguiente modo:

$s_x = 51.639 \qquad s_y = 41.808 \qquad s_{xy} = 2118$

$r = s_{xy} / (s_x s_y) = 2118/(51.639\ 41.808) = 0.981$

COEFICIENTE DE CORRELACIÓN DE LA POBLACIÓN

Hasta aquí se usó una muestra de n parejas (x_1,y_1) (x_2,y_2) …… (x_n,y_n) que se tomaron de una población XY. A los promedios $\bar{x}$ e $\bar{y}$ les corresponden los valores medios μ_X de X y μ_Y de Y, respectivamente:

$\mu_X = E[X]$ y $\mu_Y = E[Y]$

a las varianzas s_X^2 y s_Y^2 les corresponden las varianzas:

$\sigma_X^2 = E[(X - \mu_X)^2]$ y $\sigma_Y^2 = E[(Y - \mu_Y)^2]$

la cantidad

$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$

se llama covarianza de las variables aleatorias X e Y. El cociente:

$\rho = \sigma_{XY} / (\sigma_X \sigma_Y)$

se llama coeficiente de correlación de X e Y.

Si $\rho = 0$, se dice que X e Y son no correlacionadas, también si X e Y son independientes $\sigma_{XY} = 0$ y $\rho = 0$.

Teorema: Si las variables aleatorias X e Y son independientes entonces son no correlacionadas. Lo recíproco no es cierto.

Ejemplo: Suponer que X es una variable aleatoria que toma los valores –1, 0 y +1 con probabilidad p=1/3. Luego, $\mu_X = 0$. Sea $Y = X^2$, entonces:

$\sigma_{XY} = E[XY] - E[X]\,E[Y] = E[X^3] - 0.\ E[Y]$

$$\sigma_{XY} = (-1)^3\ 1/3 + (0)^3\ 1/3 + (1)^3\ 1/3 = 0$$

luego, por ser $\sigma_{XY} = 0$, $\rho = 0$, las variables aleatorias X e Y no están correlacionadas, pero no son independientes entre sí, ya que están ligadas por una relación funcional.

En el ejemplo se ve que ρ no es una medida de la dependencia general, pero si se verá que es una medida de la dependencia lineal.

VARIACIÓN EXPLICADA Y NO EXPLICADA

La variación total de una variable aleatoria Y se define como:

$$\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2$$

es decir, la suma de los cuadrados de las desviaciones de los valores de la variable aleatoria Y respecto de su media $\bar{y}$.

Partiendo de la siguiente igualdad:

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

elevando al cuadrado en ambos miembros y tomando sumatoria

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

desarrollando

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

dado que: $\hat{y}_i = a + b.x_i$

$$2.\sum_{i=1}^{n} (y_i - \hat{y}_i).(\hat{y}_i - \bar{y}) = 2.\sum_{i=1}^{n} (y_i - a - b.x_i).(\hat{y}_i - \bar{y})$$

y por las ecuaciones normales $y_i - a - b.x_i = 0$

La misma se puede reescribir como:

$$\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 + \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2$$

el primer término del segundo miembro se llama variación no explicada, mientras que el segundo se llama variación explicada, y esto es así porque las desviaciones $\hat{y}_i - \bar{y}$ tienen un patrón definido, mientras que las desviaciones $y_i - \hat{y}_i$ se comportan en forma aleatoria o no previsible. Resultados análogos se obtienen para la variable X.

La razón de la variación explicada a la variación total se llama Coeficiente de Determinación. Si la variación explicada es cero, es decir, la variación total es toda no explicada, esta razón es cero. Si la variación no explicada es cero, es decir, la variación total es toda explicada, la razón es uno. En los demás casos la razón se encuentra entre 0 y 1.

Puesto que la razón es siempre no negativa, se denota por r^2. La cantidad r es lo que conocemos como coeficiente de correlación y otra forma de definirlo es como:

$$r = \pm \sqrt{\frac{\text{variación explicada}}{\text{variación total}}} = \pm \sqrt{\frac{\sum_i \left(\hat{y}_i - \bar{y}\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2}}$$

los signos ± se utilizan para la correlación lineal positiva y negativa, respectivamente. Nótese que r es una cantidad sin dimensiones, es decir no depende de las unidades empleadas.

Problema: Los datos siguientes corresponden al número de minutos x que tardan 10 mecánicos en ensamblar cierta pieza de una maquinaria en la mañana, e y representa el tiempo que ocupan en la tarde.

x	11.1	10.3	12	15.1	13.7	18.5	17.3	14.2	14.8	15.3
y	10.9	14.2	13.8	21.5	13.2	21.1	16.4	19.3	17.4	19

calcular r.

Para un ajuste lineal, se forma el sistema de ecuaciones normales.

$\Sigma x = 142.3 \quad \Sigma x^2 = 2085 \quad \Sigma x.y = 2435 \quad \Sigma y = 166.8$

$166.8 = 10\,a + 142.3\,b$
$2435 = 142.3\,a + 2085\,b \quad \rightarrow a = 2.274 \quad b = 1.012$

var_explicada = 61.88 var_total = 115.576

de aquí, el coeficiente de correlación al cuadrado (r^2) vale:

$r^2 = 61.88 / 115.576 = 0.535$

Esto implica que r^2.100% (es decir 53.5%) de la variación entre los tiempos de la tarde responden a las diferencias correspondientes entre los tiempos de la mañana.

Siempre que un valor de r se fundamenta en una muestra aleatoria de una población normal bivariada, se puede practicar una prueba de significación (como $\rho = \rho_0$) o construir un intervalo de confianza para ρ, en base a la transformación:

$$Z = \frac{1}{2}\cdot\ln\left(\frac{1+r}{1-r}\right)$$

Z es un estadístico con distribución normal con media y varianza dadas por las siguientes expresiones:

$$\mu_Z = \frac{1}{2}\cdot\ln\left(\frac{1+\rho}{1-\rho}\right) \qquad \sigma_Z^2 = \frac{1}{n-3}$$

Luego las inferencias respecto de ρ serán:

$$z = \frac{Z-\mu_Z}{\frac{1}{\sqrt{n-3}}} = \sqrt{n-3}\cdot\ln\left[\frac{(1+r)\cdot(1-\rho)}{(1-r)\cdot(1+\rho)}\right]$$

donde z es una variable aleatoria con distribución normal estándar.

En particular, se puede probar la Hipótesis Nula de que no hay correlación ($\rho = 0$) con el estadístico:

$$z = \sqrt{n-3}\cdot Z = \frac{\sqrt{n-3}}{2}\cdot\ln\left(\frac{1+r}{1-r}\right)$$

Ejemplo: En relación con el problema anterior (donde n=10 y r=0.73) probar la Hipótesis Nula que $\rho = 0$ contra la alternativa $\rho <> 0$ con un nivel de significancia de 0.05.

1 – Hipótesis Nula $\rho = 0$
Hipótesis Alternativa $\rho <> 0$ (bilateral)
2 - Nivel de significancia: $\alpha = 0.05$. $z_{\alpha/2} = 1.96$
3- Criterio: se rechaza Ho si $z < -1.96$ ó $z > 1.96$, donde z vale:
4-

$$z = \sqrt{n-3}\cdot Z = \frac{\sqrt{n-3}}{2}\cdot\ln\left(\frac{1+r}{1-r}\right)$$

4 – Cálculos:

$$z = \frac{\sqrt{10-3}}{2}\cdot\ln\left(\frac{1+0.73}{1-0.73}\right) = 2.457$$

5- Dado que 2.457 > 1.96 se Rechaza la Hipótesis Nula, por lo tanto se acepta la Hipótesis Alternativa, esto es existe relación entre el tiempo que ocupa en la mañana y en la tarde, un mecánico para ensamblar un determinado tipo de maquinaria.

Si se quiere construir un intervalo de confianza para ρ se debe empezar por construir uno para μ_Z:

$$-z_{\frac{\alpha}{2}} < \frac{Z-\mu_Z}{\frac{1}{\sqrt{n-3}}} < z_{\frac{\alpha}{2}} \qquad \frac{-z_{\frac{\alpha}{2}}}{\sqrt{n-3}} < Z-\mu_Z < \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}$$

$$\frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}} + Z > \mu_Z > \frac{-z_{\frac{\alpha}{2}}}{\sqrt{n-3}} + Z \qquad Z - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}} < \mu_Z < Z + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}$$

Ejemplo: Si $r=0.70$ para las calificaciones en Física y Matemática de 30 estudiantes, construir un intervalo de confianza con un nivel de confianza del 95% para el coeficiente de correlación de la población.

$$Z = \frac{1}{2}\cdot\ln\left(\frac{1+0.7}{1-0.7}\right) = 0.867 \qquad z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

luego:

$$0.867 - \frac{1.96}{\sqrt{27}} < \mu_Z < 0.867 + \frac{1.96}{\sqrt{27}} \qquad 0.49 < \mu_Z < 1.244$$

y como:

$$\mu_Z = \frac{1}{2}\cdot\ln\left(\frac{1+\rho}{1-\rho}\right) \qquad e^{2\cdot\mu_Z} = \frac{1+\rho}{1-\rho}$$

$$e^{2\cdot\mu_Z}\cdot(1-\rho) = 1+\rho \qquad \rho = \frac{e^{2\cdot\mu_Z}-1}{e^{2\cdot\mu_Z}+1}$$

el intervalo para ρ queda:

$$\frac{e^{2\cdot 0.49}-1}{e^{2\cdot 0.49}+1} < \rho < \frac{e^{2\cdot 1.244}-1}{e^{2\cdot 1.244}+1} \qquad 0.454 < \rho < 0.847$$

Ejemplo: Si $r=0.20$ para una muestra aleatoria de $n=40$ parejas de datos, construir un intervalo de confianza del 95% para ρ.

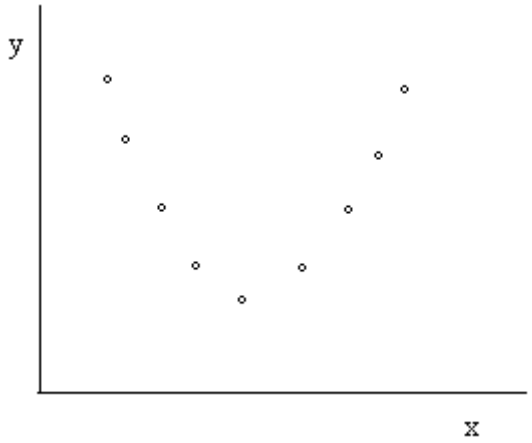
$$Z = \frac{1}{2}\cdot\ln\left(\frac{1+0.2}{1-0.2}\right) = 0.203 \qquad z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

$$0.203 - \frac{1.96}{\sqrt{37}} < \mu_Z < 0.203 + \frac{1.96}{\sqrt{37}} \qquad -0.119 < \mu_Z < 0.525$$

$$\frac{e^{2\cdot(-0.119)}-1}{e^{2\cdot(-0.119)}+1} < \rho < \frac{e^{2\cdot 0.525}-1}{e^{2\cdot 0.525}+1} \qquad -0.118 < \rho < 0.482$$

En ambos ejemplo los intervalos de confianza son grandes para ρ. Esto ilustra el hecho de que los coeficientes de correlación basados en muestras relativamente chicas suelen ser poco confiables.

Existen varias trampas peligrosas en la interpretación de ρ. r es una estimación de la fuerza de la relación entre los valores de dos variables aleatorias. En la siguiente figura r puede ser muy cercana a cero, aún cuando hay una fuerte relación funcional (parabólica, no lineal).



En segundo lugar, una correlación significativa no necesariamente implica una relación causal entre las dos variables aleatorias.