CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE

De la determinación de la función matemática que explica las fluctuaciones de una variable en función de la otra se encarga la denominada **Teoría de la Regresión**. Mientras que el estudio del grado de dependencia que pueda existir entre las variables es propio de la denominada **Teoría de la Correlación.**

Tipos de relación: en algunos casos, dos variables están relacionadas de una forma:

Determinista: son del tipo $y = f(x)$

No determinista: Son del tipo $y = f(x) + u$

DIAGRAMAS DE DISPERSIÓN

Con el fin de analizar visualmente el tipo de relación existente entre dos variables consideradas X e Y, una de las representaciones gráficas más usuales de una distribución de frecuencias conjunta es la nube de puntos o **diagrama de dispersión**, que consiste en trazar un plano de coordenadas sobre cuyo eje de abscisas representamos los valores de la variable X, reservando el eje de ordenadas para los valores de Y. Estos diagramas nos permiten apreciar si existe relación entre las variables, si esta relación es lineal o no lineal, si es directa o inversa y la intensidad de la misma.

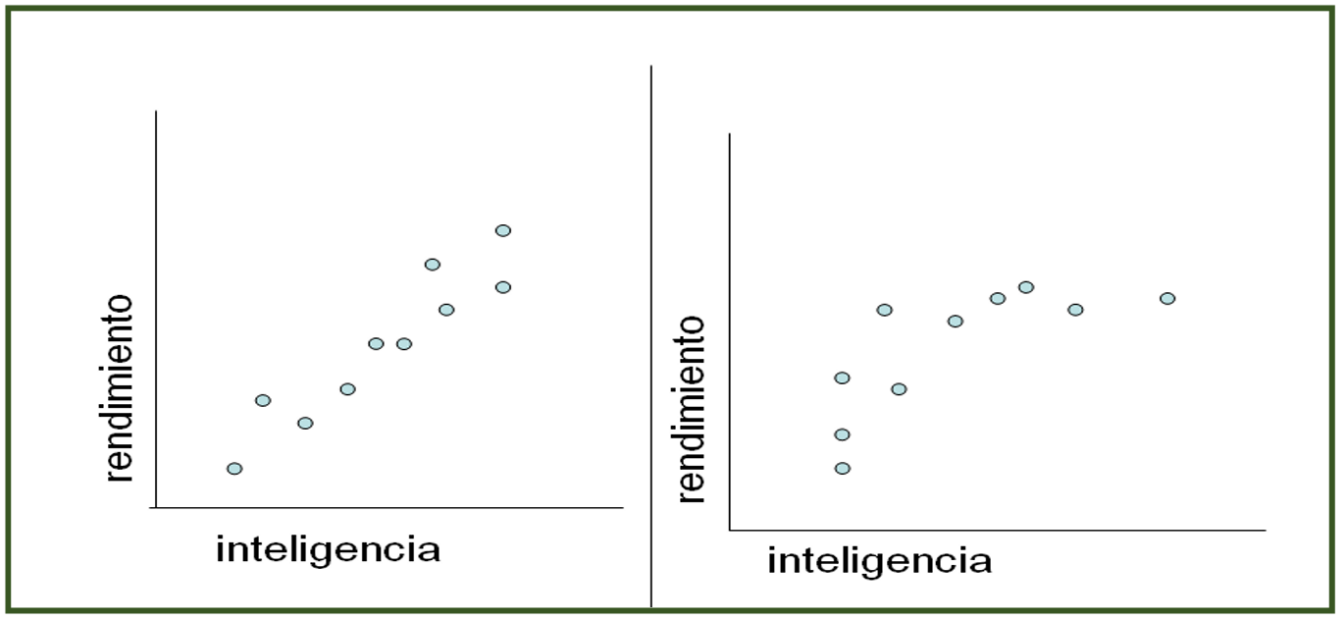


Figura 1: Relación lineal vs. Relación no lineal (ambas son directas)

Medidas de dependencia lineal

La relación entre dos variables también puede expresarse en forma numérica. La covarianza es una medida de la asociación lineal entre dos variables que resume la información existente en un gráfico de dispersión.

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

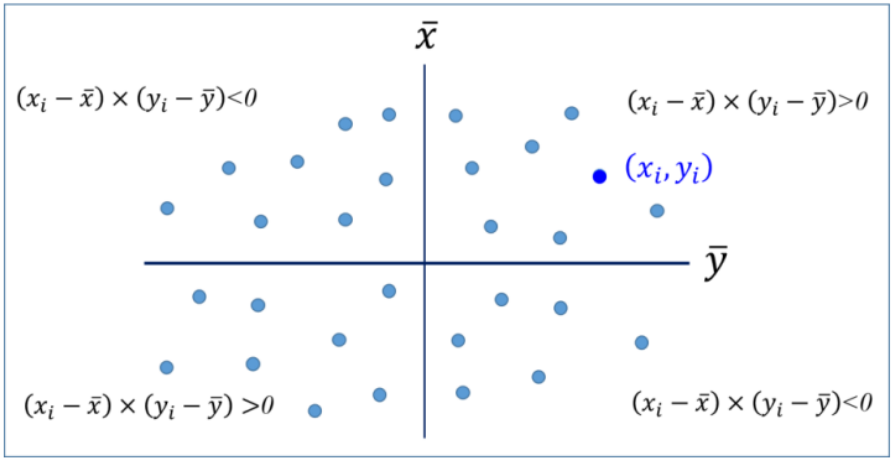
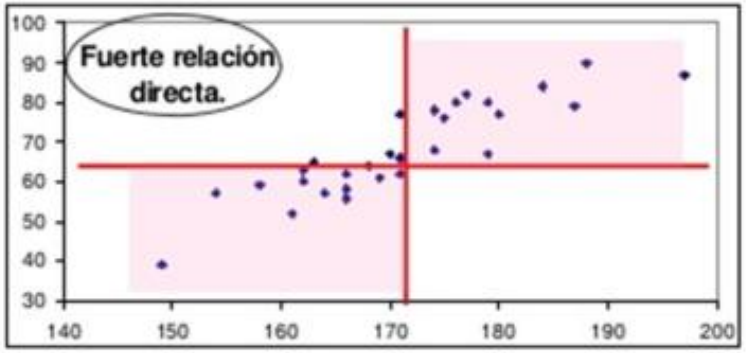
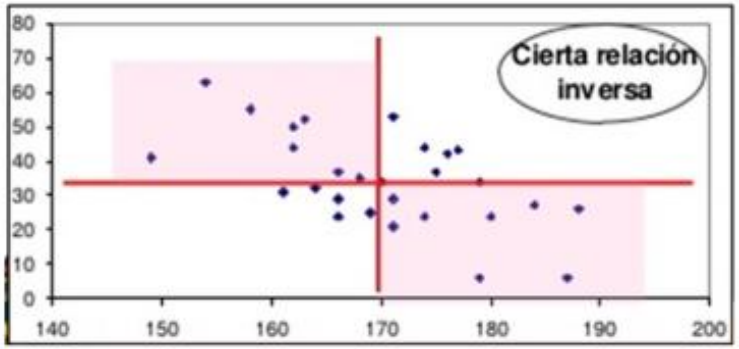


Figura 2: Signo de los términos de la covarianza

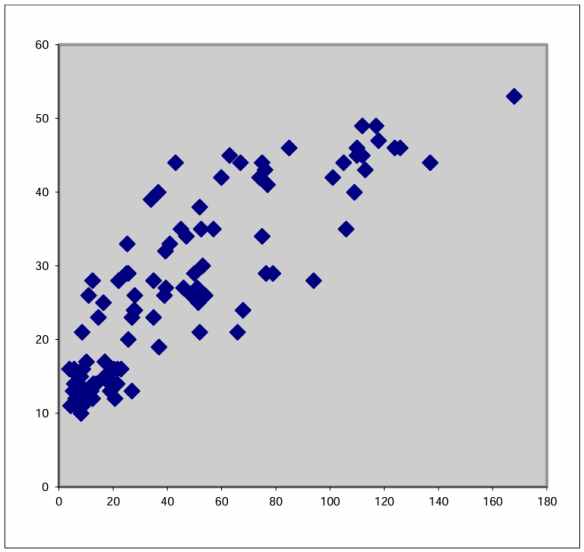


Para valores mayores de x que la media le corresponden valores de y mayores.
Para los valores de x menores que la media le corresponden valores de y menores.
Esto se llama relación directa

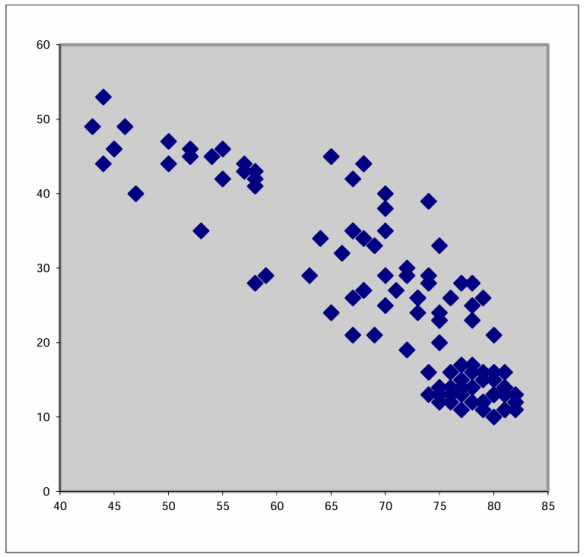


Para valores de x mayores que la media le corresponden valores de y menores.
Para los valores de x menores que la media le corresponden valores de y mayores.
Esto se llama relación inversa.

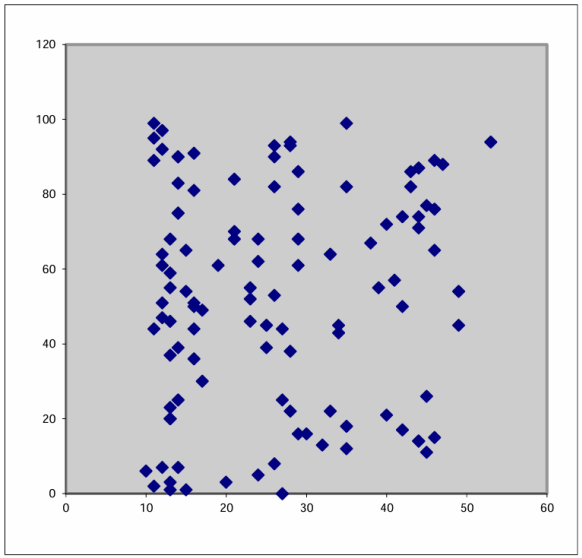
Figura 2: Signo de los términos de la covarianza



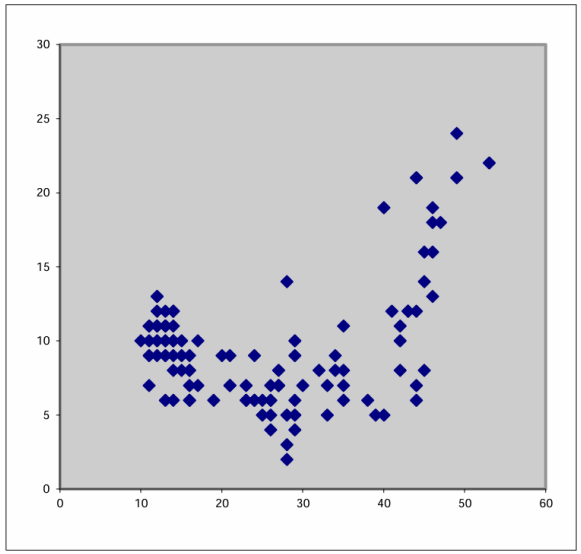
a)



b)



c)



d)

Figura 3: Diagramas de dispersión correspondientes a variables (X,Y) con $S_{XY} > 0$ (a), $S_{XY} < 0$ (b), $S_{XY} \sim 0$ (c) y (d).

- Si hay relación lineal positiva, la covarianza será positiva y grande.

- Si hay relación lineal negativa, la covarianza será negativa y grande en valor absoluto.

- Si no hay relación entre las variables o la relación es marcadamente no lineal, la covarianza será próxima a cero.

- La covarianza S_{XY} es invariante ante cambios de origen, pero no ante cambios de escala.

La covarianza depende de las unidades de medida de las variables.

El coeficiente de correlación lineal

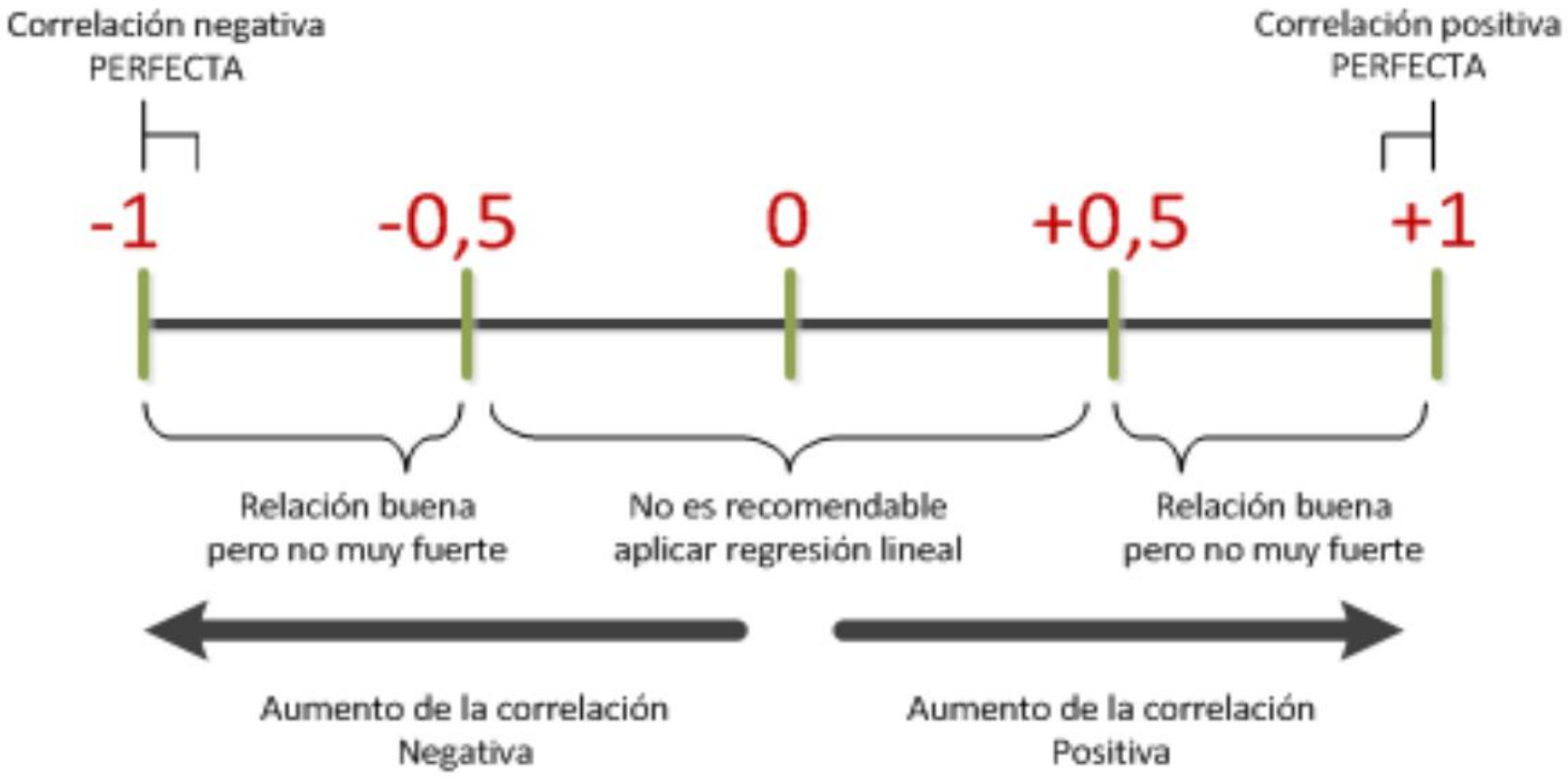
Una medida de la dependencia lineal que no depende de las unidades de medida es el coeficiente de correlación lineal:

$$r = \frac{S_{XY}}{S_X S_Y}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \quad \text{y} \quad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}$$

- $-1 \leq r \leq 1$

- Es adimensional
- Está relacionado con la medición de la bondad de ajuste de la recta de regresión.
- Mide la fuerza de una relación lineal. No está diseñada para medir la fuerza de una relación que no sea lineal


Correlación negativa
PERFECTA
Correlación positiva
PERFECTA
-1
-0,5
0
+0,5
+1
Relación buena
pero no muy fuerte
No es recomendable
aplicar regresión lineal
Relación buena
pero no muy fuerte
Aumento de la correlación
Negativa
Aumento de la correlación
Positiva

El problema de la **Regresión** consiste en la obtención de la ecuación de una curva que pase “cerca” de los puntos representados en el diagrama de dispersión, y que se adapte lo mejor posible al conjunto de los mismos, cumpliendo determinadas condiciones. Por lo tanto, cuando se pretende hacer un ajuste nos encontramos con dos problemas:

(a) Elegir el tipo de curva que mejor se adapte a los datos disponibles, es decir, que mejor represente la relación entre las variables endógena y exógena. En esta fase suele ser de gran utilidad la representación gráfica como orientación para la elección.

(b) Fijado el tipo de curva a través de su ecuación en forma explícita con un cierto número de parámetros, determinar éstos mediante las condiciones que se impongan según el procedimiento de ajuste empleado.

El modelo de regresión lineal simple

El modelo de regresión lineal simple supone que

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y_i representa el valor de la variable respuesta para la observación i-ésima
x_i representa el valor de la variable explicativa para la observación i-ésima.
ε_i representa el error para la observación i-ésima que se asume normal,

$$\varepsilon_i \sim N(0, \sigma)$$

β_0 y β_1 son los coeficientes de regresión:

- β_0 : intercepto
- β_1 : pendiente

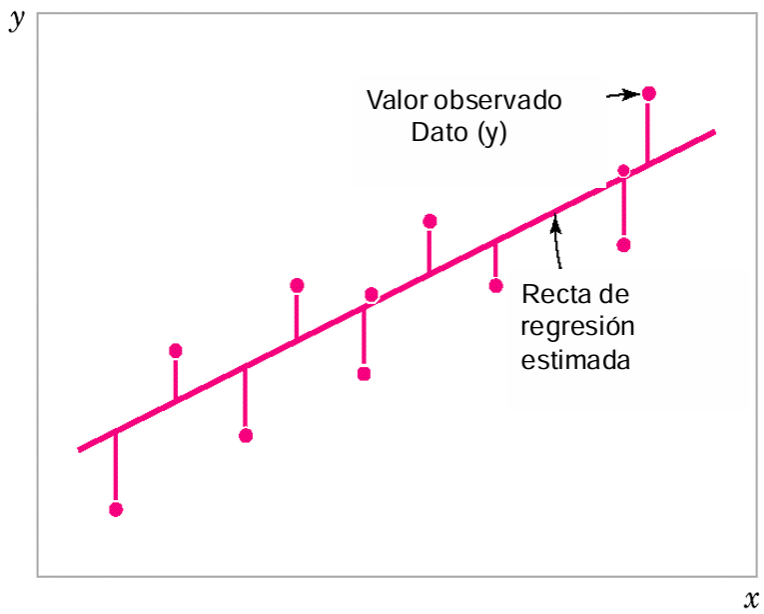
Los parámetros que hay que estimar son: β_0, β_1 y σ.

El objetivo es obtener estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ de β_0 y β_1 para calcular la **recta de regresión**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

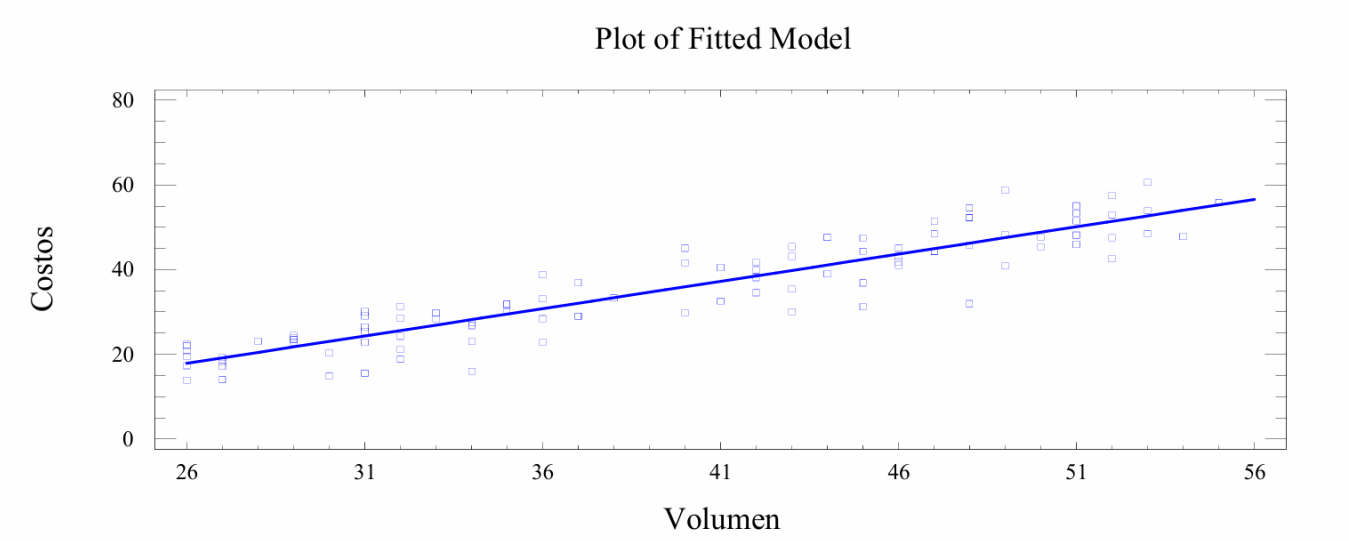
que se ajuste lo mejor posible a los datos.
La diferencia entre cada valor y_i de la variable respuesta y su estimación $\hat{y}_i$ se llama residuo:

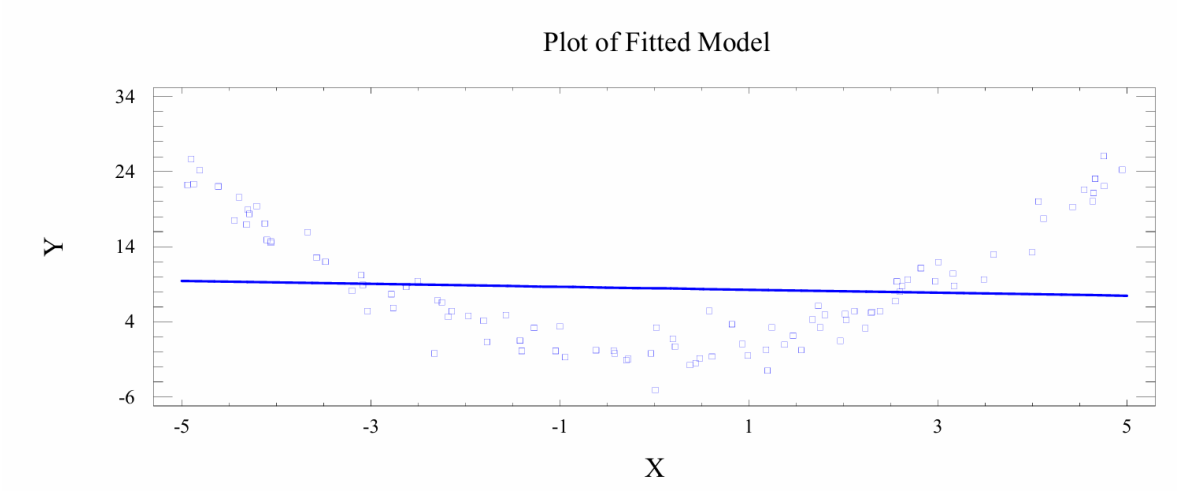
$$e_i = y_i - \hat{y}_i$$



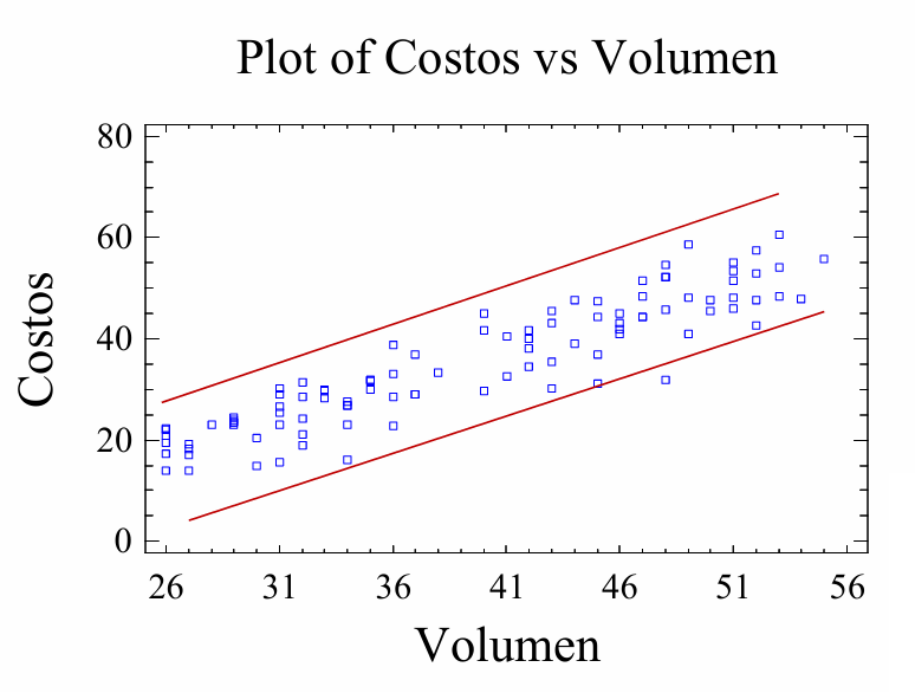
HIPÓTESIS DEL MODELO DE REGRESIÓN LINEAL SIMPLE

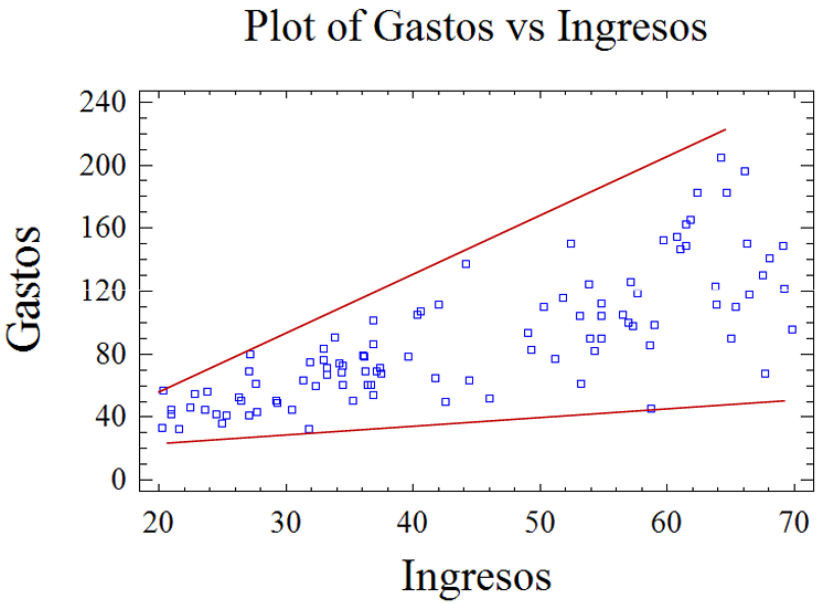
Linealidad: los datos deben ser razonablemente rectos





Homocedasticidad: la dispersión de los datos debe ser constante para que los datos sean homocedásticos.



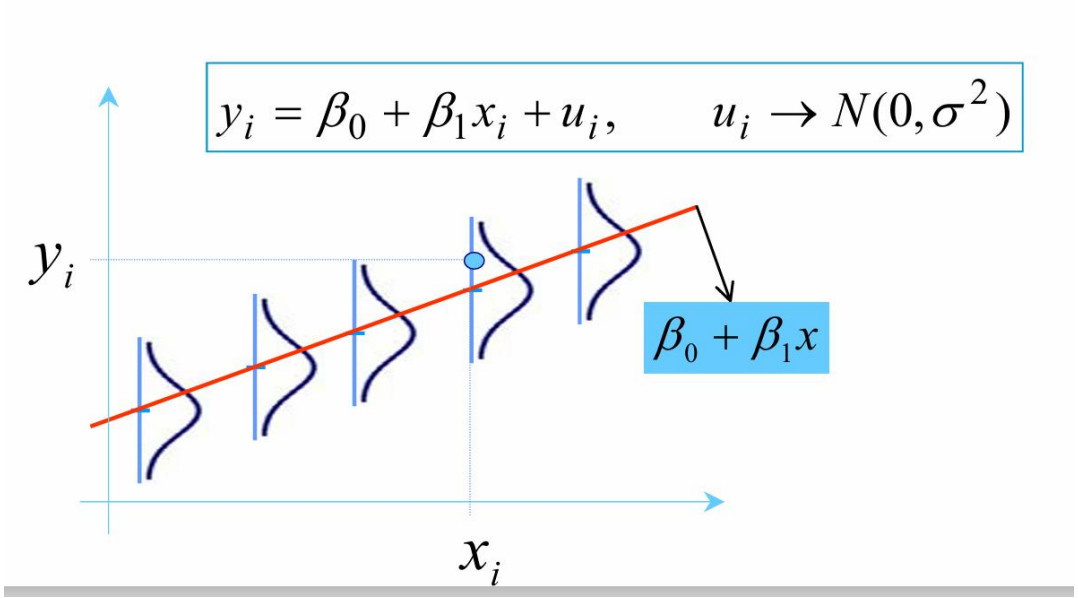


Independencia

- Los datos deben ser independientes.
- Una observación no debe dar información sobre las demás. Habitualmente, se sabe por el tipo de datos si son adecuados o no para el análisis. En general, las series temporales no cumplen la hipótesis de independencia.

Normalidad

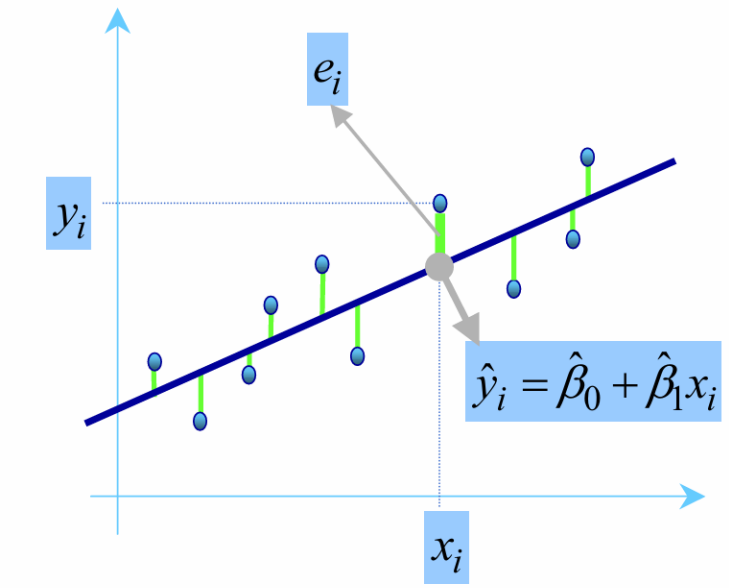
Se asume que los datos son normales a priori.



Estimadores de mínimos cuadrados

Usaremos el método de mínimos cuadrados para obtener los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ que mejor se ajustan a los datos. El método consiste en minimizar la suma de los cuadrados de las distancias verticales entre los datos y las estimaciones, es decir, minimizar la suma de los residuos al cuadrado,

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)^2$$



$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)^2$$

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \qquad \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i = 0$$

$$\frac{\partial SSE}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0 \qquad \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \hat{\beta}_0 x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

considerando:

$$\sum_{i=1}^{n} \hat{\beta}_0 = n\hat{\beta}_0$$

Tenemos:

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \quad (1)$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \quad (2)$$

Despejando $\hat{\beta}_0$ de la ecuación (1):

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\Longrightarrow \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Sustituyendo $\hat{\beta}_0$ en la ecuación (2):

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \left(\overline{y} - \hat{\beta}_1 \overline{x}\right) \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i$$

$$\Longrightarrow \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i}$$

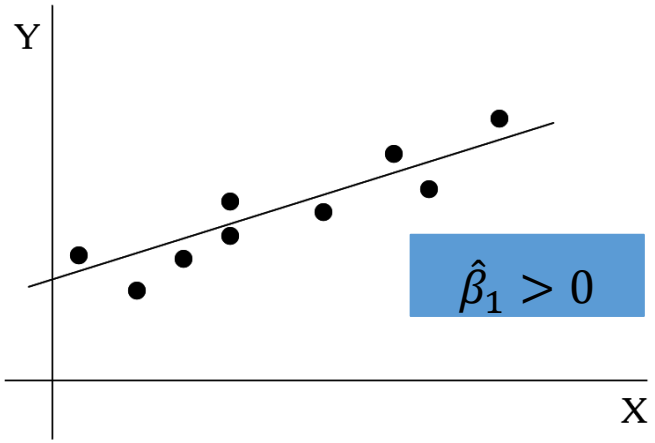
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{y}\,\overline{x}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

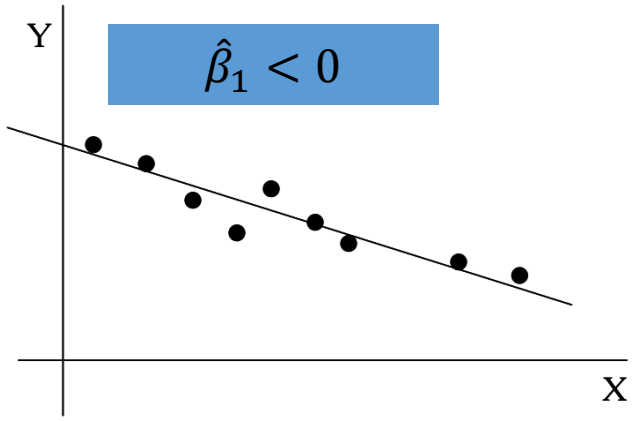
El resultado que se obtiene es:

$$\hat{\beta}_1 = \frac{cov(x,y)}{S_X^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Signo de la dependencia lineal



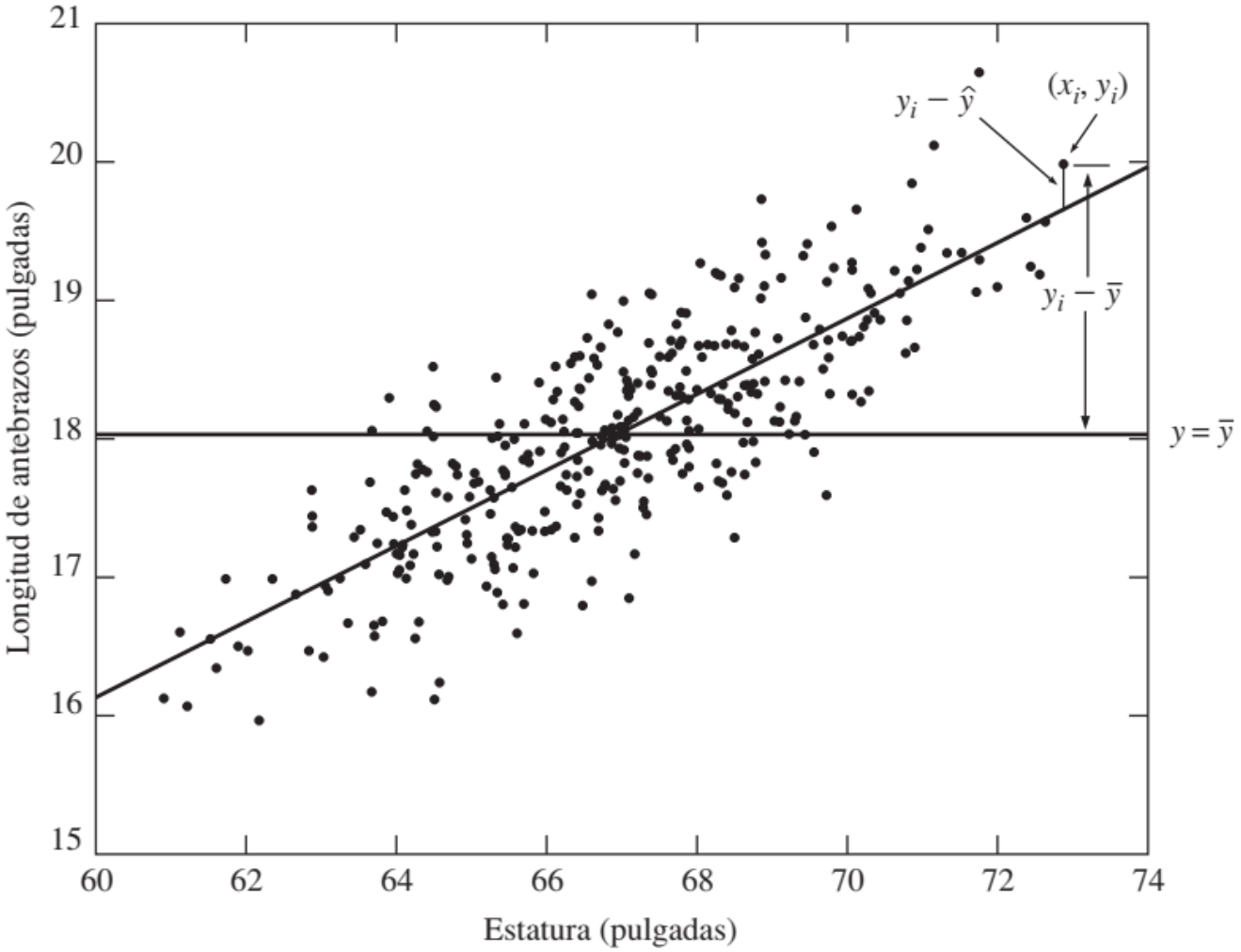


BONDAD DE AJUSTE Y CORRELACIÓN

El objetivo fundamental de la Teoría de la **Regresión** es poder expresar una variable en función de otra y hacer uso de dicha función para realizar predicciones, de manera que, si se quiere predecir un valor de Y, y el ajuste realizado ha sido lineal, se deberá considerar la recta de regresión de Y/X, mientras que si la predicción que se quiere efectuar es de X se hará uso de la recta de regresión de X/Y. Ahora bien, cualquier predicción debería ir acompañada de un estudio de la fiabilidad que ésta merece. Del estudio de la fiabilidad y del grado de dependencia que pueda existir entre las dos variables se encarga la denominada **Teoría de la Correlación**.

Un estadístico de la bondad de ajuste representa una cantidad que mide qué tan bien un modelo explica un conjunto específico de datos

Un modelo lineal ajusta bien si hay una fuerte relación lineal entre *x* e *y*. Entonces **r** es un estadístico de la bondad del ajuste para el modelo lineal.


21
20
19
18
17
16
15
60
62
64
66
68
70
72
74
Longitud de antebrazos (pulgadas)
Estatura (pulgadas)
$y_i - \hat{y}$
(x_i, y_i)
$y_i - \bar{y}$
$y = \bar{y}$

La fuerza de la relación lineal se mide al calcular la reducción obtenida en la suma de los errores pronosticados al cuadrado con $\hat{y}_i$ en lugar de $\bar{y}$.

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

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

$$\sum_{i=1}^{n}(y_i - \hat{y})^2$$ Suma de los cuadrados de los errores

$$\sum_{i=1}^{n}(y_i - \overline{y})^2$$ Suma total de los cuadrados

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y})^2$$ Suma de los cuadrados de la regresión

Suma total de los cuadrados = Suma de los cuadrados de la regresión + Suma de los cuadrados de los errores

$$r^2 = \frac{Suma\ de\ los\ cuadrados\ de\ la\ regresión}{Suma\ total\ de\ los\ cuadrados} = \frac{Variación\ explicada}{Variación\ total}$$

Análisis de la Varianza para la regresión lineal

En un análisis de regresión, la respuesta $\boldsymbol{y}$ se relaciona con la variable independiente $\boldsymbol{x}$. De aquí que la variación total en la variable de respuesta $\boldsymbol{y}$ dada por

$$SS_{total} = S_{yy} = \sum (y_i - \overline{y})^2$$

Se divide en dos partes

$$SS_{total} = SSR + SSE$$

SSR: suma de cuadrados para la regresión, mide la cantidad de variación explicada al usar la recta de regresión con una variable independiente x.

SSE: suma de cuadrados para el error, mide la variación residual en los datos que no es explicada por la variable independiente x.

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

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}\right)^2 = \sum_{i=1}^{n} \left(\overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i - \bar{y}\right)^2 =$$

$$= \hat{\beta}_1{}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} = \frac{\left(S_{xy}\right)^2}{S_{xx}}$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SS_{total} - SSR = S_{yy} - \frac{\left(S_{xy}\right)^2}{S_{xx}}$$

Incertidumbres en los coeficientes de mínimos cuadrados

Suponga que se tienen n puntos de datos $(x_1, y_1), \ldots, (x_n, y_n)$ y se ajusta la recta de mínimos cuadrados. Con la finalidad de que los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ sean útiles, se necesita estimar qué tan grande son sus incertidumbres. Para realizar este cálculo se necesita que los errores ε_i deben satisfacer cuatro supuestos importantes:

1) Los errores ε_i son aleatorios e independientes.
2) Todos los errores ε_i tienen media 0.
3) Todos los errores ε_i tiene la misma varianza, que se denota por σ^2.
4) Los errores ε_i están distribuidos normalmente.

La dispersión de los puntos alrededor de la recta se puede medir con la suma de los residuos al cuadrado $\sum_{i=1}^{n} e_i^2$. El estimador de la varianza del error σ^2 es la cantidad s^2

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

Bajo los supuestos 1 a 4, las observaciones y_i son también variables aleatorias independientes que siguen la distribución normal.

$$\mu_{y_i} = \beta_0 + \beta_1 xi$$
$$\sigma^2_{y_i} = \sigma^2$$

Tanto $\hat{\beta}_0$ como $\hat{\beta}_1$ se pueden expresar como combinaciones lineales de las y_i ,

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] y_i$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] y_i$$

Tenemos:

$$\mu_{\hat{\beta}_0} = \beta_0 \qquad \mu_{\hat{\beta}_1} = \beta_1$$

$$\sigma_{\hat{\beta}_0} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Las desviaciones estándar de $\hat{\beta}_0$ y $\hat{\beta}_1$ se calculan con:

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

donde $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$

Prueba de la utilidad del modelo de regresión lineal

Es útil la recta de regresión de mínimos cuadrados?
La ecuación de regresión que usa la información proporcionada por x es sustancialmente mejor que el predictor $\overline{y}$ $(que\ no\ depende\ de\ x)$?
Hay una relación lineal entre x e y?

Prueba de hipótesis o intervalo de confianza para la pendiente $\boldsymbol{\beta_1}$ y $\boldsymbol{\beta_0}$

Se basa en la distribución muestral de $\hat{\beta}_1$ el estimador muestral de la pendiente β_1

$DM(\hat{\beta}_1) \sim N\left(E(\hat{\beta}_1), s_{\hat{\beta}_1}\right)$ $\qquad$ $DM(\hat{\beta}_0) \sim N\left(E(\hat{\beta}_0), s_{\hat{\beta}_0}\right)$

$t = \dfrac{\hat{\beta}_1 - \beta}{s_{\hat{\beta}_1}}$, $\qquad$ $t = \dfrac{\hat{\beta}_0 - \beta}{s_{\hat{\beta}_0}}$ $\qquad$ $ambos\ con\ n-2\ grados\ de\ libertad$

Los intervalos con un nivel de confianza de $(1-\alpha)$ para $\hat{\beta}_0$ y $\hat{\beta}_1$ están dados por

$\hat{\beta}_0 \pm t_{n-2,\alpha/2} s_{\widehat{\beta}_0}$ $\hat{\beta}_1 \pm t_{n-2,\alpha/2} s_{\widehat{\beta}_1}$

Contraste sobre la pendiente

1) Hipótesis nula $H_0: \beta_1 = 0$

2) Hipótesis alternativa $H_1: \beta_1 \neq 0$

$\beta_1 > 0$

$\beta_1 < 0$

3) $t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$ Estadístico de prueba con distribución t de Student con (n-2) grados de libertad.

4) Rechazo H_0 cuando:

- $t > t_{\alpha/2}$ ó $t < -t_{\alpha/2}$
- $t > t_{\alpha/2}$
- $t < t_{\alpha/2}$

o cuando $p - valor < \alpha$

El rechazo de H_0 implica que hay una relación lineal significativa.

ANÁLISIS DE CORRELACIÓN

El coeficiente de correlación como una medida de la fuerza de la relación lineal entre dos variables. El coeficiente de correlación, r (coeficiente de correlación muestral momento – producto de Pearson)

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{s_{xy}}{s_{xx}s_{yy}}$$