Regresión lineal simple

José Gabriel Palomo Sánchez
gabriel.palomo@upm.es

E.U.A.T.
U.P.M.

Julio de 2011

Índice I

Índice II

Índice III

El problema general. Dependencia e independencia de variables.

DEFINICIÓN

Dos variables son dependientes cuando el conocimiento del valor de una de ellas en un individuo aporta información sobre el valor de la otra en ese individuo.

DEFINICIÓN

Cuando dos variables no son dependientes se dice que son independientes.

Dependencia determinista I. Ejemplo

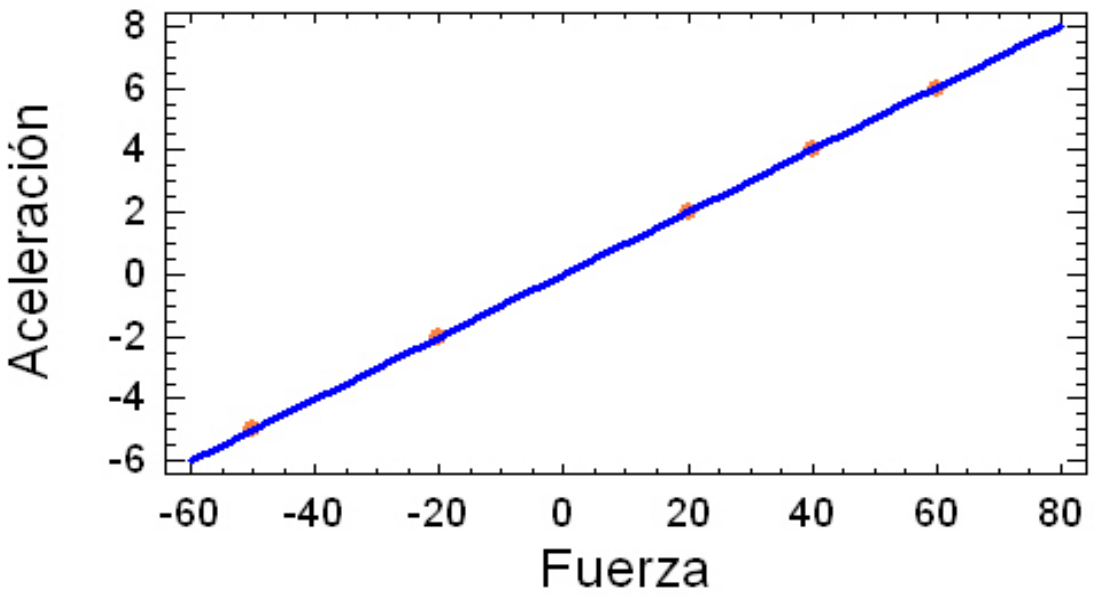
- Si a un cuerpo de masa **m** se le aplica una fuerza **F**, esta fuerza comunica una aceleración al cuerpo, cuyo módulo viene expresado por la ecuación:

$$a = \frac{F}{m}.$$

- Esta ecuación permite calcular con exactitud, el módulo de la aceleración que una fuerza determinada comunicará a un cuerpo de masa conocida.

Dependencia determinista II. Ejemplo

El siguiente gráfico muestra los distintos valores de las aceleraciones provocadas sobre un cuerpo de masa 10 Kg, por distintas fuerzas ejercidas sobre él.



La ecuación $a = \frac{F}{10}$ es el modelo que explica la relación de dependencia entre estas variables.

Dependencia determinista III. Ejemplo

- El espacio recorrido por un cuerpo en caída libre, en el vacío, viene dado por la expresión:

$$e = \frac{1}{2} g t^2,$$

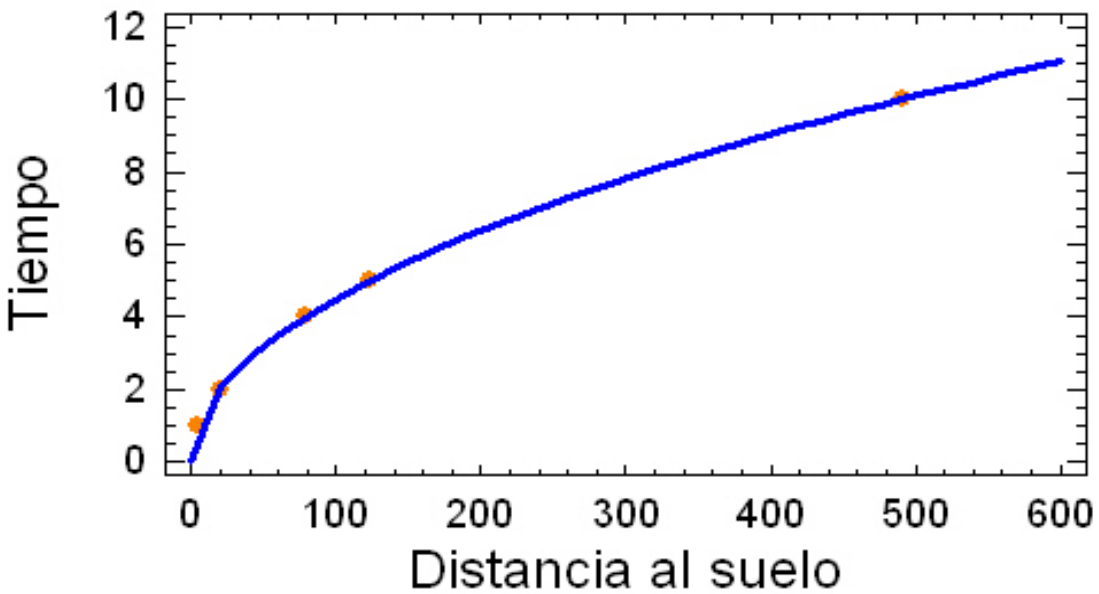
donde g representa el valor de la aceleración de la gravedad, y t es el valor del tiempo transcurrido.

- Despejando:

$$t = \sqrt{\frac{2e}{g}}$$

Dependencia determinista IV. Ejemplo

El siguiente gráfico muestra los distintos valores del tiempo transcurrido hasta que un cuerpo en caída libre alcanza el suelo, en función de la distancia entre éste y el punto en el que inicia la caída.



La ecuación $t = \sqrt{\frac{2e}{g}}$ es el modelo que explica la relación de dependencia entre estas variables.

Dependencia determinista V

- Cuando el conocimiento del valor de una variable permite el cálculo exacto de otra, se dice que entre ellas hay una relación de dependencia **determinista o funcional**.
- La ecuación que posibilita este cálculo determina el **modelo** que explica la relación entre ambas variables.

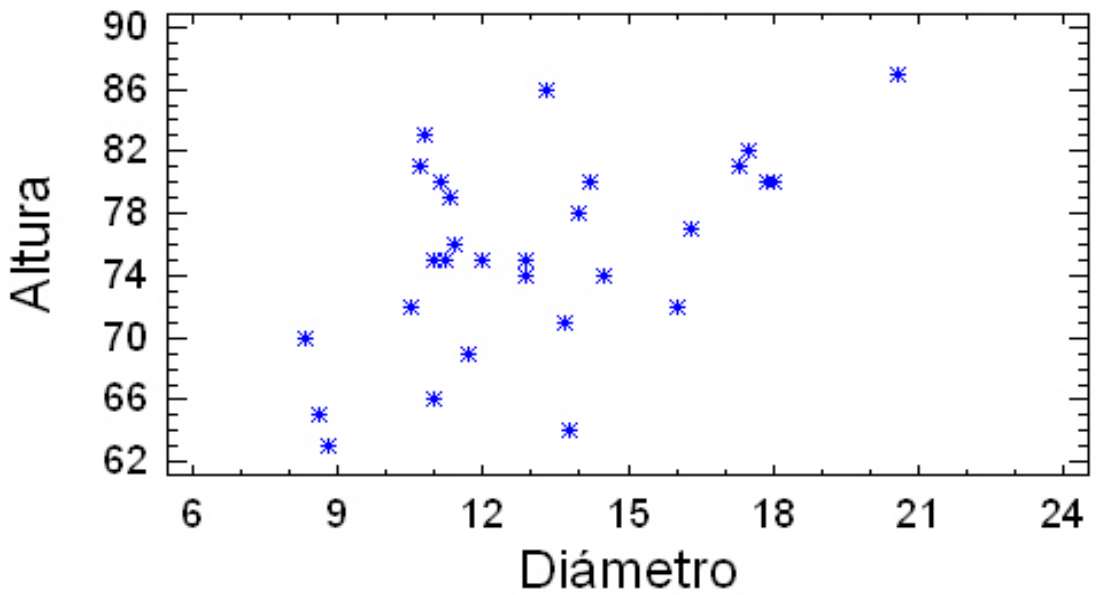
Dependencia estadística I

En ocasiones, cuando dos variables son dependientes, **NO** se puede calcular con exactitud el valor de una variable cuando el de la otra es conocido.

En estos casos se dice que la relación de dependencia entre las variables es **estadística o aleatoria**.

Dependencia estadística II. Ejemplo

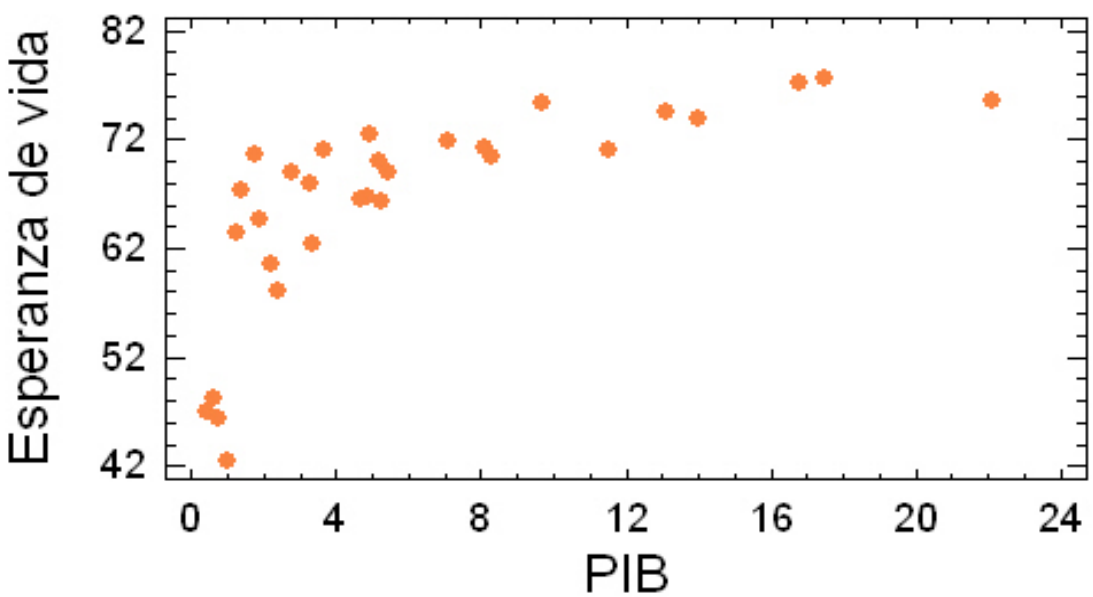
El siguiente gráfico representa los diámetros en la base del tronco, y las alturas, de un conjunto de cerezos.



¿Qué altura le corresponde a un cerezo que tenga un diámetro en la base de 14 unidades?

Dependencia estadística III. Ejemplo

El siguiente gráfico representa la esperanza de vida en un conjunto de paises en función de su producto interior bruto, (en el gráfico las unidades del PIB son miles de millones de dólares).



¿Qué esperanza de vida le corresponde a un país que tenga un PIB de 15 unidades?
¿Y a otro con un PIB de 5 unidades?

Dependencia estadística IV

Problema

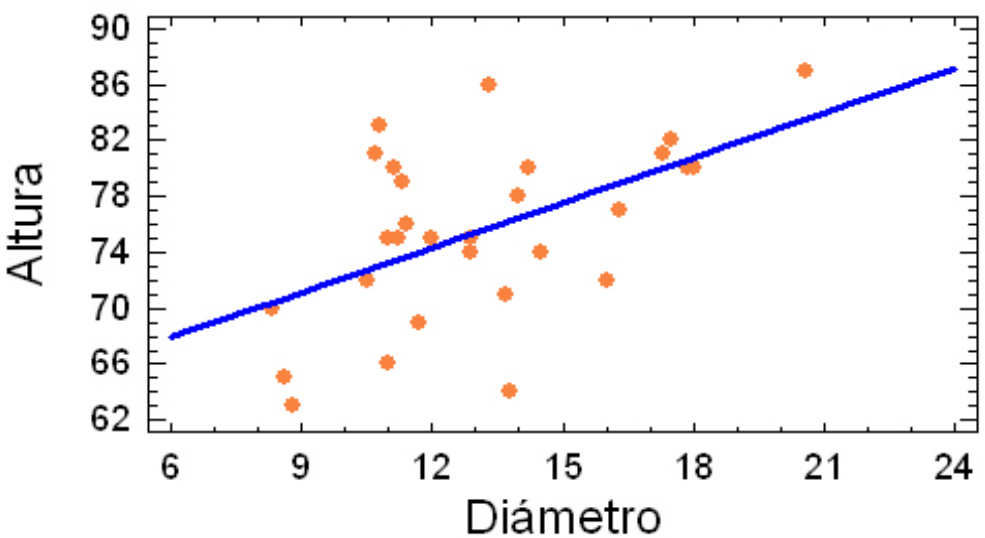
En los casos de dependencia estadística **no existe** un modelo matemático (ecuación) que permita calcular con exactitud el valor de una variable, cuando la otra es conocida.

Solución

En ocasiones se puede establecer un modelo que permita calcular, de manera aproximada, el valor de una variable aleatoria, cuando el de la otra, también aleatoria, es conocida.

Dependencia estadística V. Ejemplo

La recta del gráfico permite el cálculo aproximado de la altura de un cerezo, conocido su diámetro en la base.

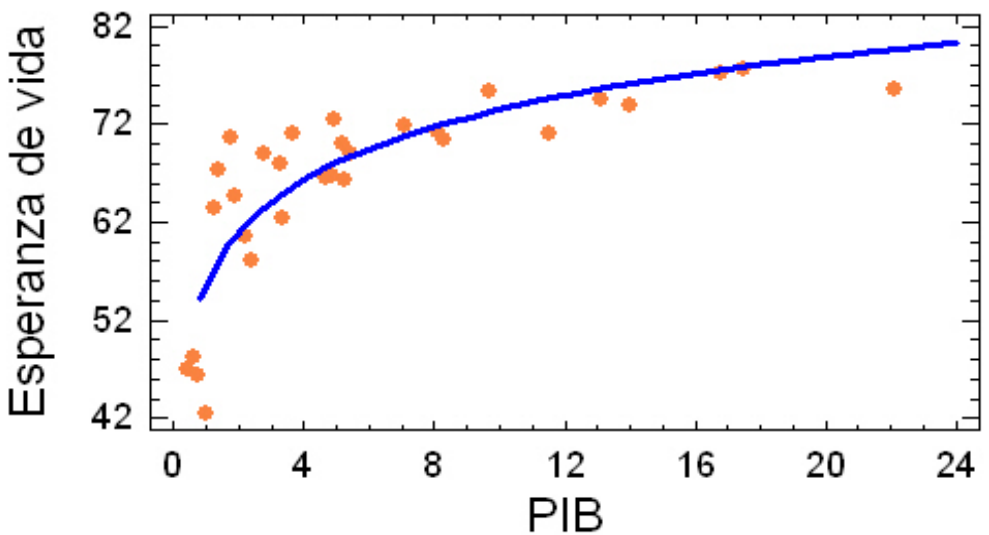


Su ecuación es:
$y = 61'55 + 1'066x$

La altura aproximada de un cerezo, cuyo diámetro en la base sea 14, será:
$y = 61'55 + 1'066 \times 14 = 76'47$

Dependencia estadística VI. Ejemplo

La curva del gráfico permite el cálculo aproximado de la esperanza de vida de un país, conocido el número de miles de millones de su PIB.



Su ecuación es:
$y = 2'03 + 7'76 \times \ln(x)$

La esperanza de vida aproximada en un país de 5000 millones de dólares de PIB es: $y = 2'03 + 7'76 \times \ln(5000) = 68'12$

Dependencia estadística VII. Resumen

1. Cuando dos variables son dependientes, el conocimiento del valor de una de ellas aporta información sobre el valor de la otra.

2. En el caso de dependencia funcional, conocido el valor de una de las variables, la ecuación del modelo, $y = f(x)$, permite el cálculo exacto del valor de la otra.

3. En el caso de dependencia estadística, el conocimiento del valor de una variable aleatoria permite, sólo, el cálculo aproximado del valor de la otra.

Los modelos de regresión. Generalidades I

Definición

Un modelo de regresión es una expresión matemática que permite calcular, **de forma aproximada**, el valor de una variable aleatoria en un individuo, cuando se conoce el valor de una o varias variables en ese mismo individuo (regresores), que también son aleatorias.

Cuando se contempla únicamente un regresor se trata de un modelo de regresión simple. En el caso en que se trate más de un regresor se tratará de un modelo de regresión múltiple. En este capítulo, solo se tratarán modelos de regresión simple.

Los modelos de regresión. Generalidades II

A lo largo de este capítulo se tratará de dar respuesta a las siguientes preguntas:

1. ¿Cuándo es útil un modelo de regresión?
2. ¿Cómo se calcula un modelo de regresión?
3. ¿Cómo se emplea un modelo de regresión?
4. ¿Qué fiabilidad ofrece un modelo de regresión?

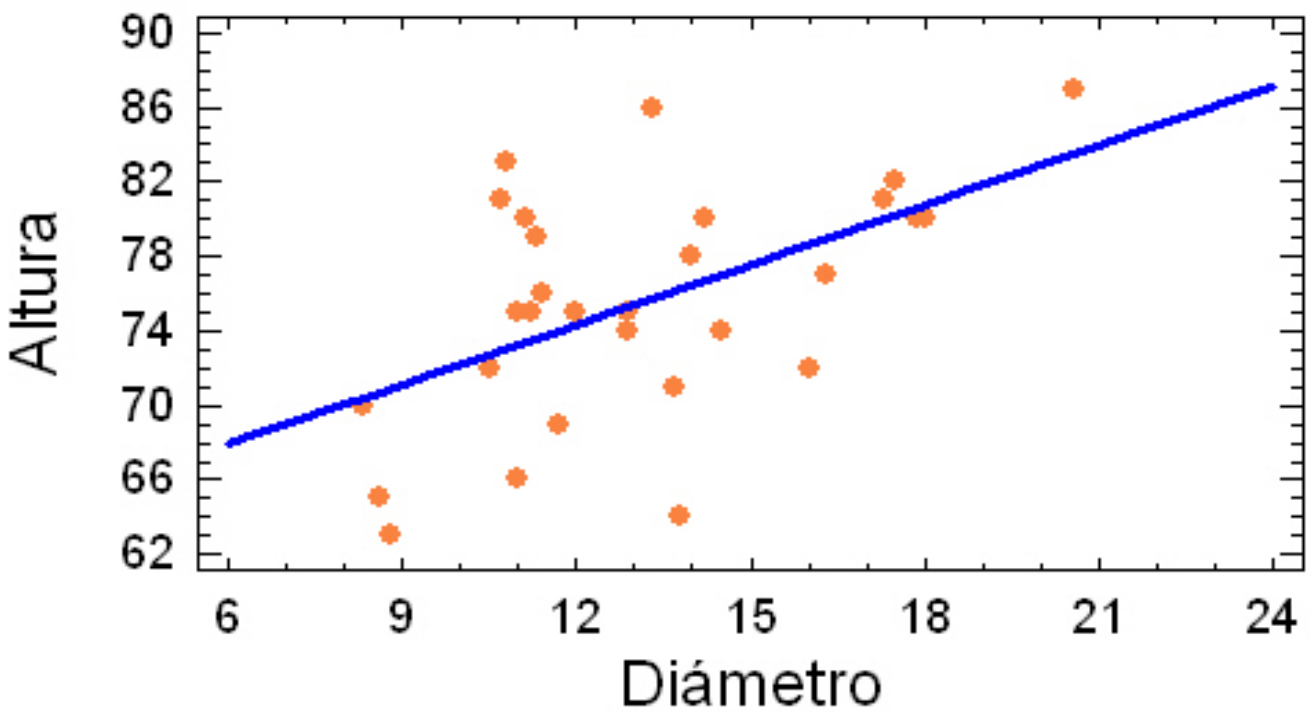
Los modelos de regresión. Generalidades III

Principio básico

Un modelo de regresión es útil cuando describe correctamente la relación de dependencia entre variables.

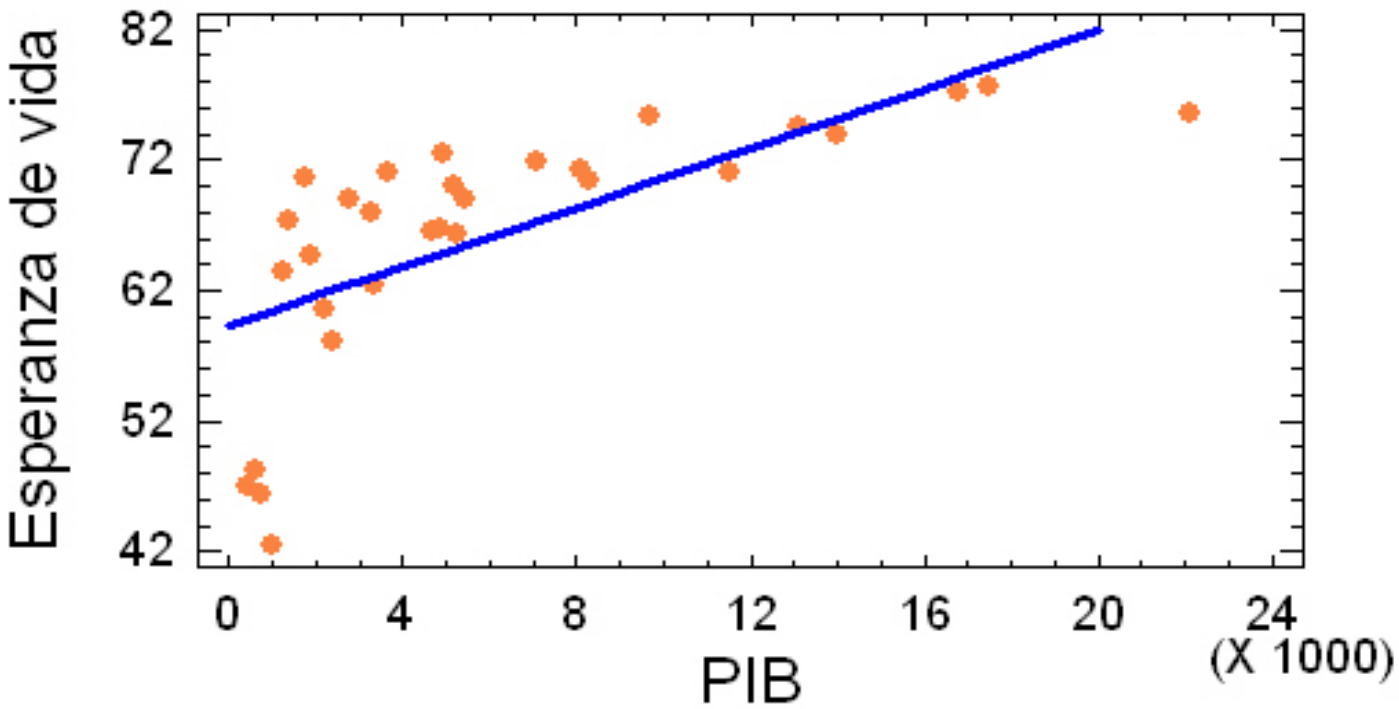
Los modelos de regresión. Generalidades IV. Ejemplo

La recta del siguiente gráfico describe, de forma aproximada, y según la información disponible, la relación entre la altura de los cerezos y su diámetro en la base.



Los modelos de regresión. Generalidades V. Ejemplo

La recta del siguiente gráfico no describe, de forma aproximada, y según la información disponible, la relación entre la esperanza de vida en un país y su producto interior bruto.



Cálculo de un modelo de regresión

Para el cálculo de un modelo de regresión es necesario establecer una metodología que tenga en cuenta:

- La clase de modelo que explique la relación de dependencia entre las variables, (lineal, polinómico, logarítmico,...).
- La estructura matemática de dicho modelo.
- Un criterio de cálculo de los parámetros del modelo.

Conjetura del modelo I

¿Qué modelo es el adecuado?

La conjetura de la conveniencia de un modelo de regresión, para explicar la relación de dependencia entre variables, se realiza, en primer lugar, a través del análisis gráfico de la información disponible.

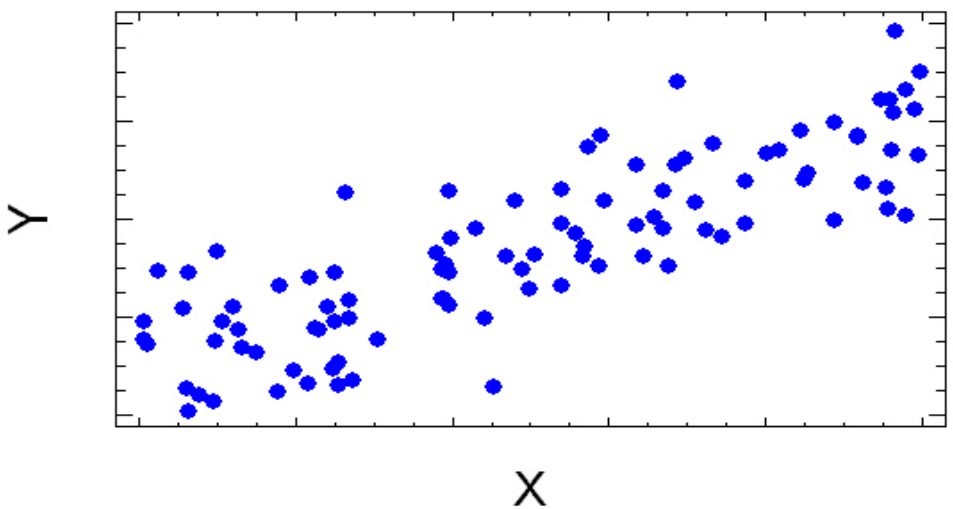
Conjetura del modelo II. Ejemplo

Para analizar la relación de dependencia entre dos variables aleatorias X e Y se toman datos (pareados), según la tabla:

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

Conjetura del modelo III. Ejemplo

Gráficamente,

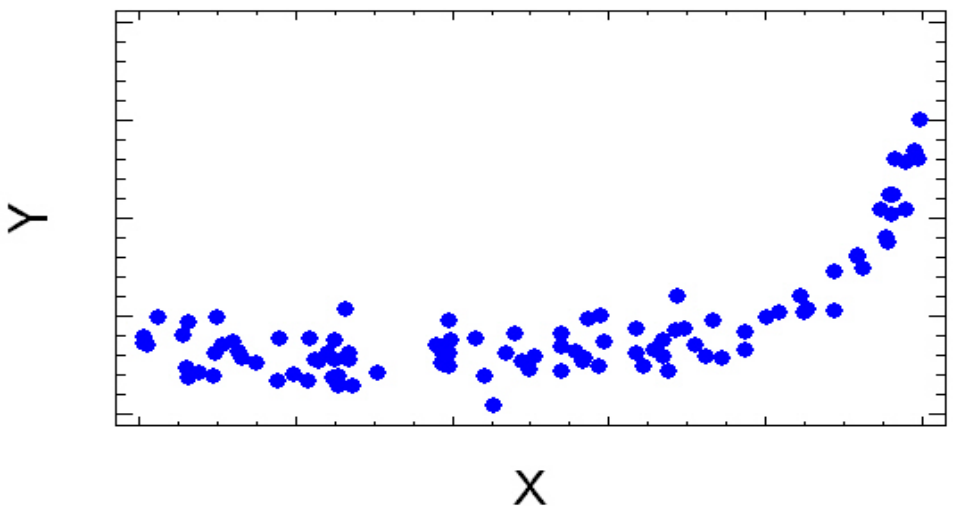


¿Qué tipo de modelo explicaría esta relación de dependencia entre X e Y?

Parece razonable, en este caso, conjeturar una recta como el modelo adecuado.

Conjetura del modelo IV. Ejemplo

El siguiente gráfico resume la información de un conjunto de datos, obtenidos para analizar la relación de dependencia entre las variables aleatorias X e Y.



¿Qué tipo de modelo explicaría esta relación de dependencia entre X e Y?

No parece razonable, en este caso, conjeturar una recta como el modelo adecuado.

Conjetura del modelo V. El caso lineal

En el caso en que la nube de puntos sugiera una relación lineal, con forma de recta, entre las variables, existen dos coeficientes que complementan la información gráfica:

- Covarianza.
- Coeficiente de correlación lineal.

El coeficiente de covarianza I

- El coeficiente de covarianza se construye para medir la intensidad de la dependencia lineal entre dos variables.

- Supóngase que para medir esta relación de dependencia se dispone de una muestra de datos pareados como los expuestos en la siguiente tabla:

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

El coeficiente de covarianza II

DEFINICIÓN

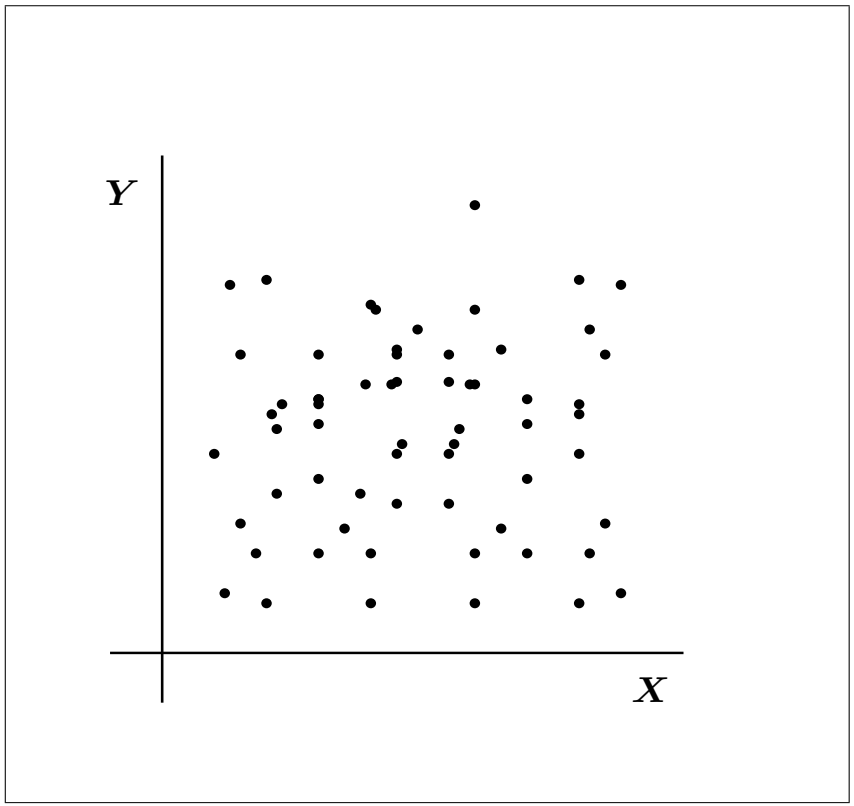
Se define el coeficiente de covarianza entre X e Y como:

$$COV(X, Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Donde $\bar{x}$ e $\bar{y}$ representan las medias muestrales de X e Y, respectivamente.

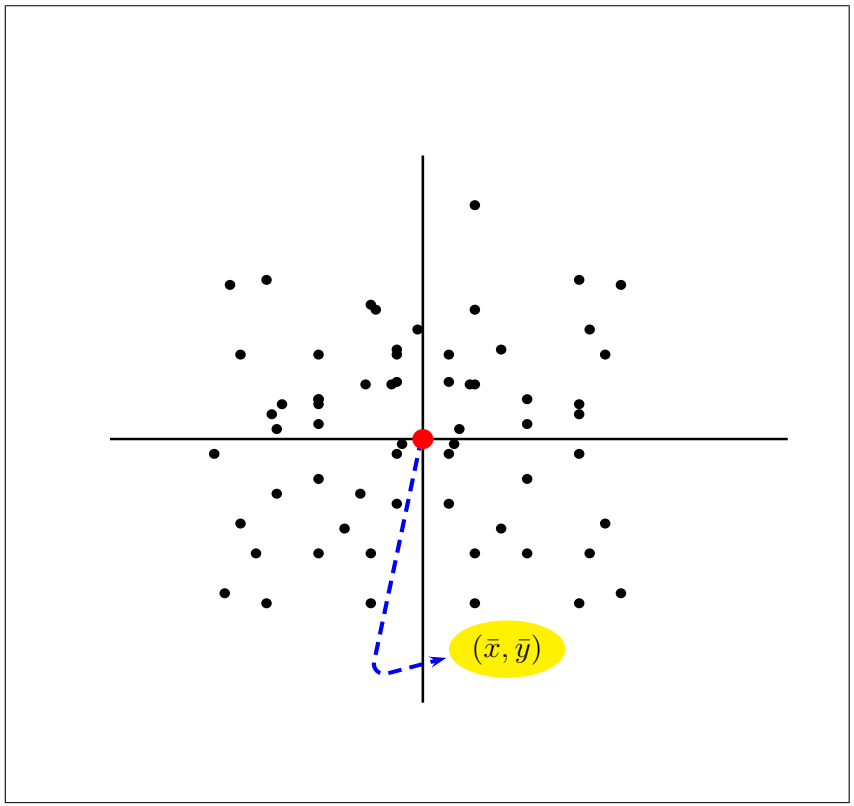
Interpretación de la covarianza I

Para interpretar el significado del coeficiente de covarianza, considérese la representación gráfica de los datos de la tabla.



Interpretación de la covarianza II

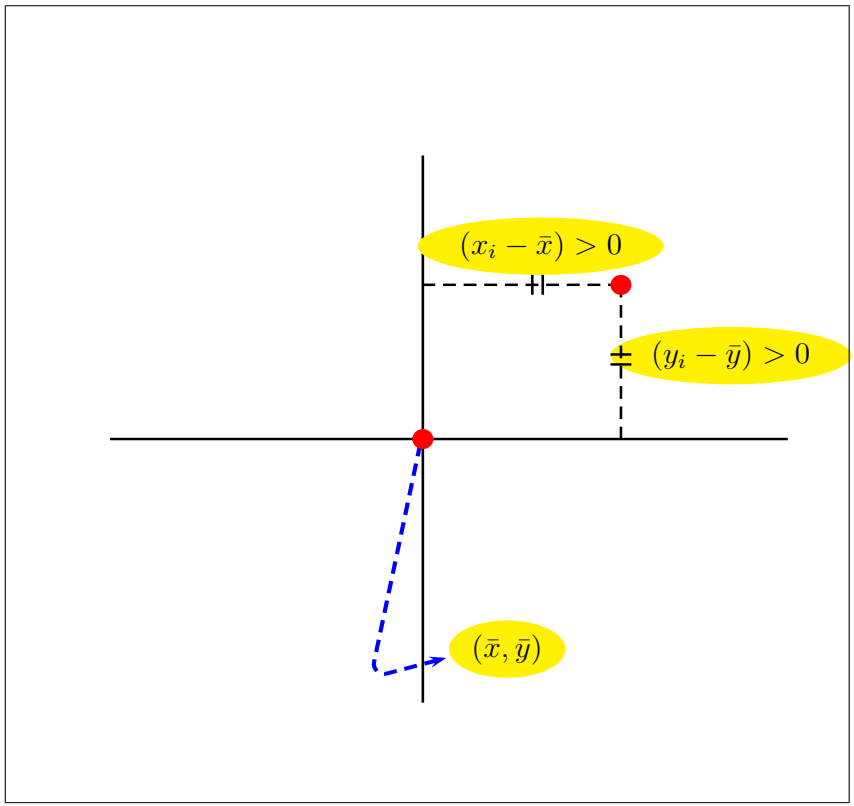
Considéres una traslación de los ejes al punto $(\bar{x}, \bar{y})$:



Interpretación de la covarianza III

Para todo punto del primer cuadrante, se observa que:

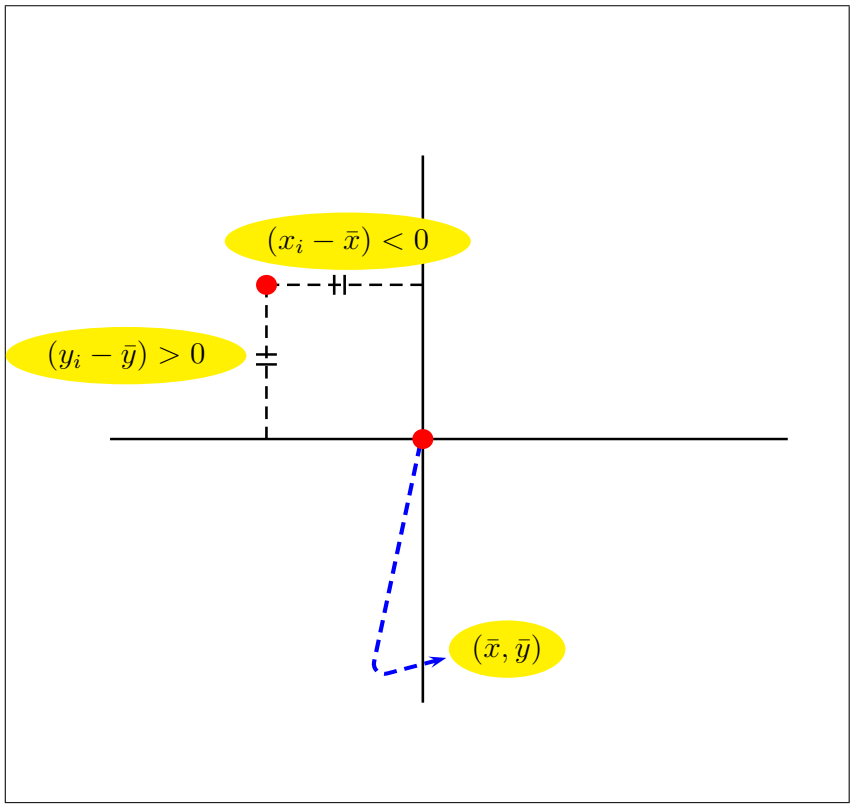
$$(x_i - \bar{x})(y_i - \bar{y}) > 0$$



Interpretación de la covarianza IV

Del mismo modo, para los puntos del segundo cuadrante:

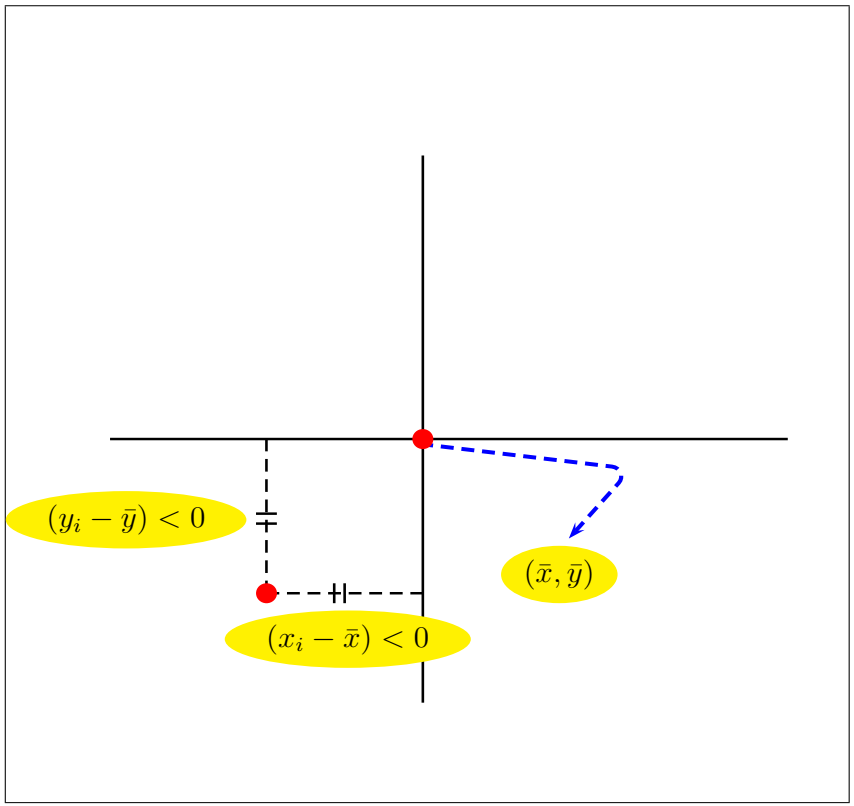
$$(x_i - \bar{x})(y_i - \bar{y}) < 0$$



Interpretación de la covarianza V

De forma similar, en el tercer cuadrante:

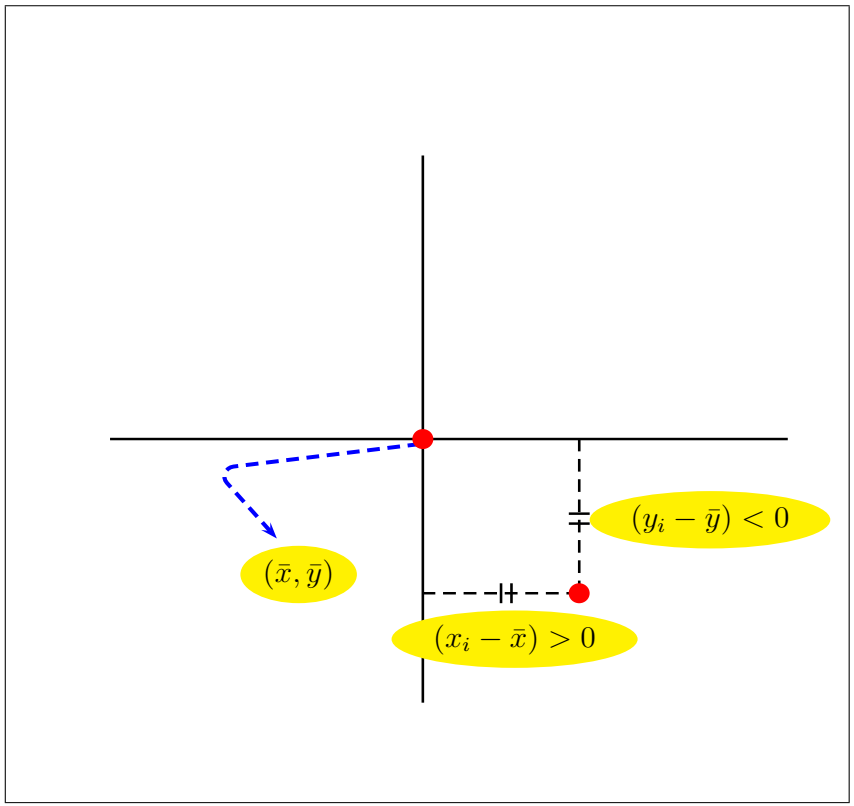
$$(x_i - \bar{x})(y_i - \bar{y}) > 0$$



Interpretación de la covarianza VI

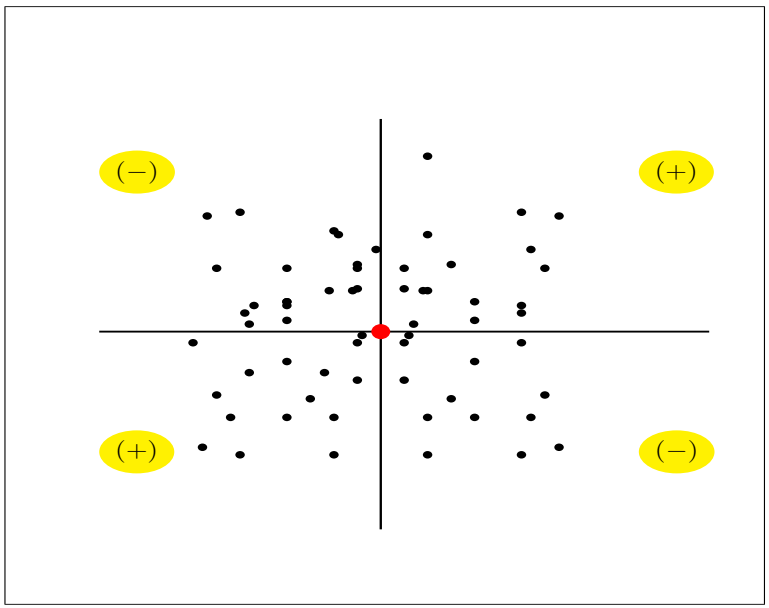
Y en el cuarto cuadrante:

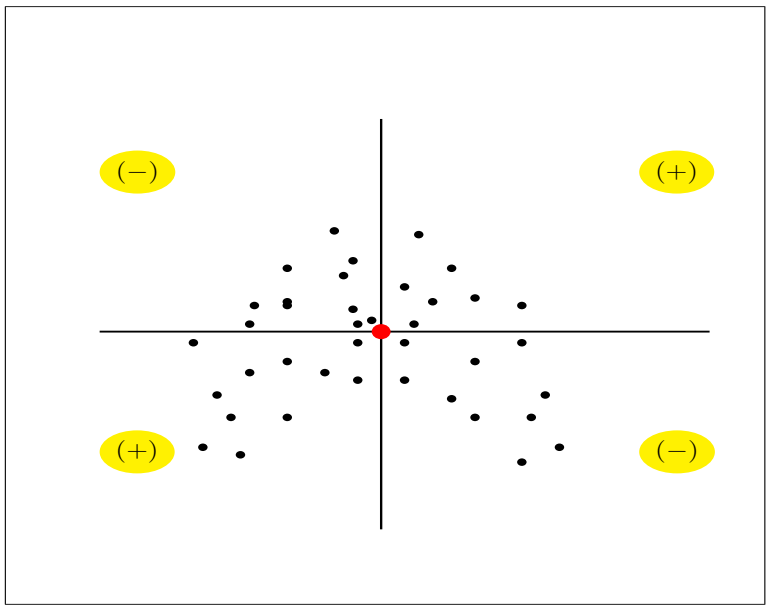
$$(x_i - \bar{x})(y_i - \bar{y}) < 0$$



Interpretación de la covarianza VII

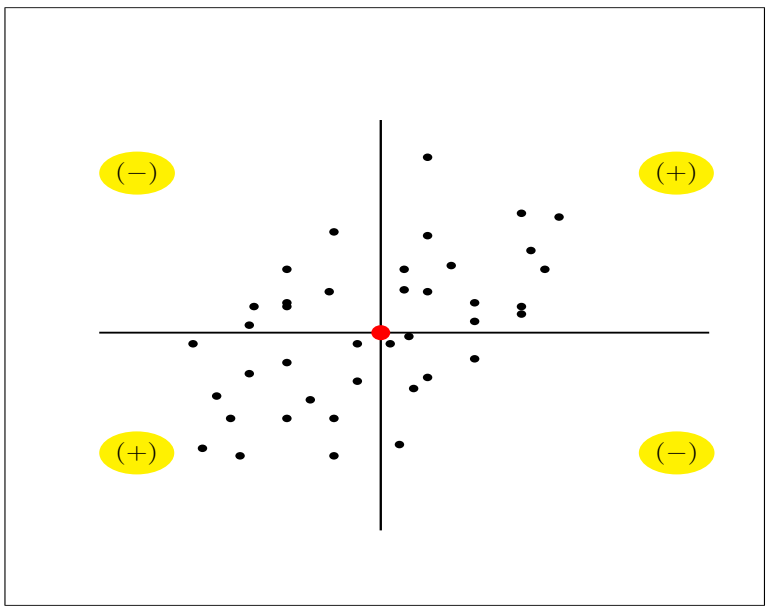
Por lo tanto, en distribuciones de puntos como las de las figuras adjuntas cabe esperar un coeficiente de covarianza próximo a cero.

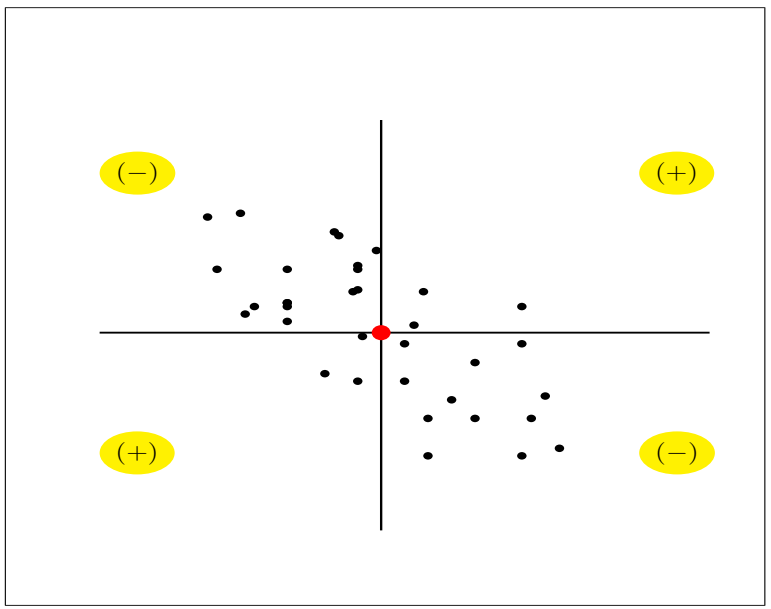




Interpretación de la covarianza VIII

Sin embargo, en distribuciones de puntos como las de las figuras adjuntas cabe esperar un coeficiente de covarianza alto en valor absoluto.





Propiedades de la covarianza

- La covarianza tiene unidades, las de la variable X multiplicadas por las de la variable Y.
- La covarianza no tiene escala y se puede hacer, en valor absoluto, arbitrariamente grande o pequeña con el mismo conjunto de datos.

El coeficiente de correlación lineal

- Para corregir los inconvenientes de la covarianza se define el coeficiente de correlación, que también mide la intensidad de la dependencia lineal entre dos variables.

Definición

El coeficiente de correlación entre dos variables es:

$$\rho = \frac{COV(X,Y)}{s_X s_Y}$$

Donde s_X y s_Y representan las desviaciones típicas de X e Y, respectivamente.

Propiedades del coeficiente de correlación I

El coeficiente de correlación tiene las siguientes propiedades:

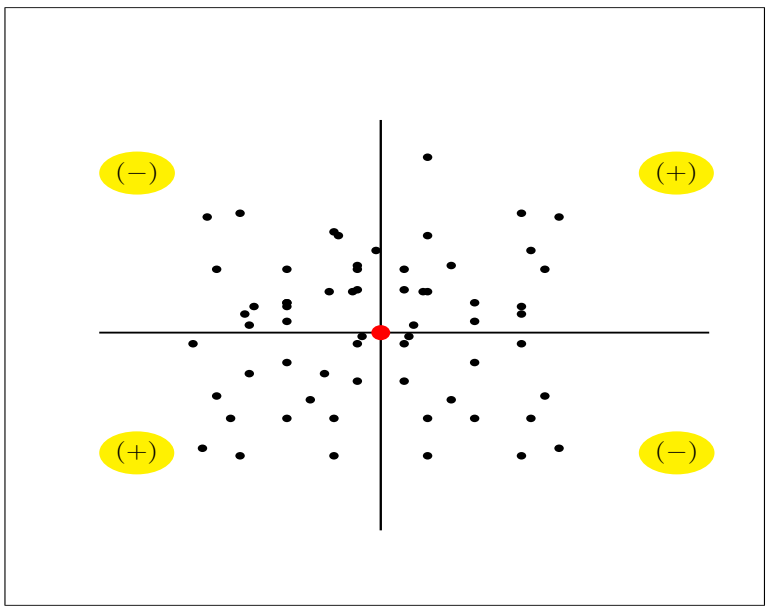
- Es un número adimensional.
- En todo caso:

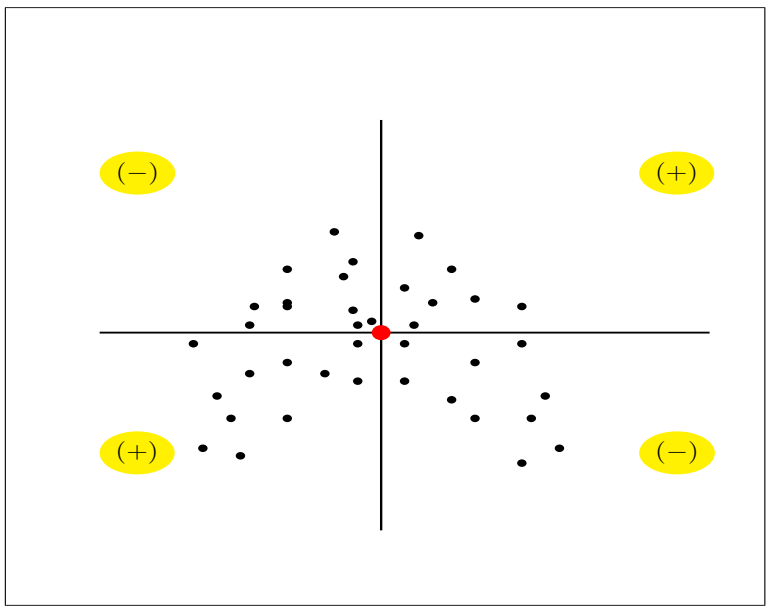
$$-1 \leq \rho \leq 1$$

- $|\rho| = 1$ implica dependencia lineal exacta entre X e Y.
- $\rho = 0$ implica falta de dependencia lineal entre X e Y.

Propiedades del coeficiente de correlación II

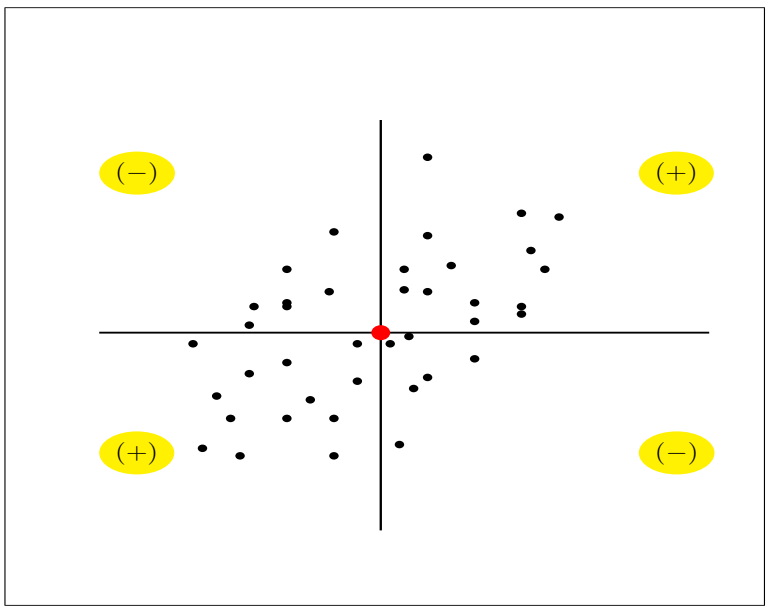
En situaciones como las que muestran los siguinetes gráficos, cabe esperar un coeficiente de correlación próximo a cero.

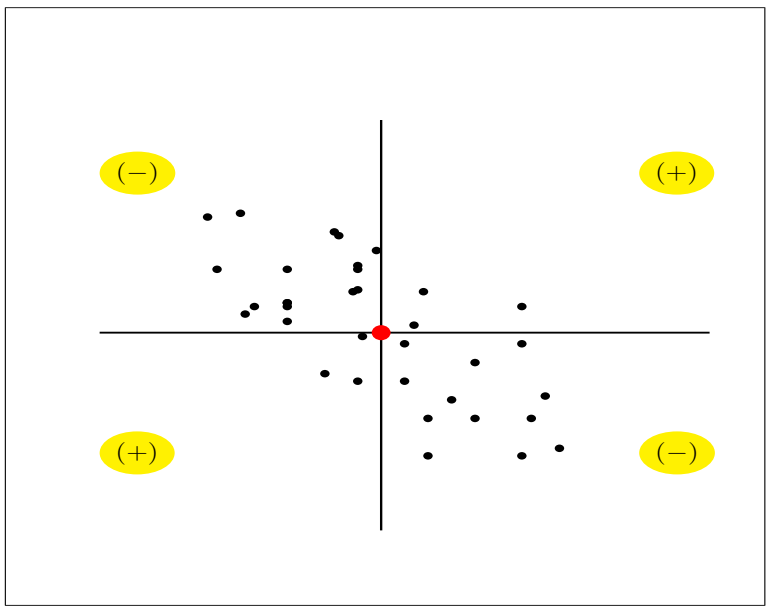




Propiedades del coeficiente de correlación III

Sin embargo, en los casos que resumen los siguientes gráficos cabe esperar un coeficiente de correlación próximo a uno en valor absoluto.

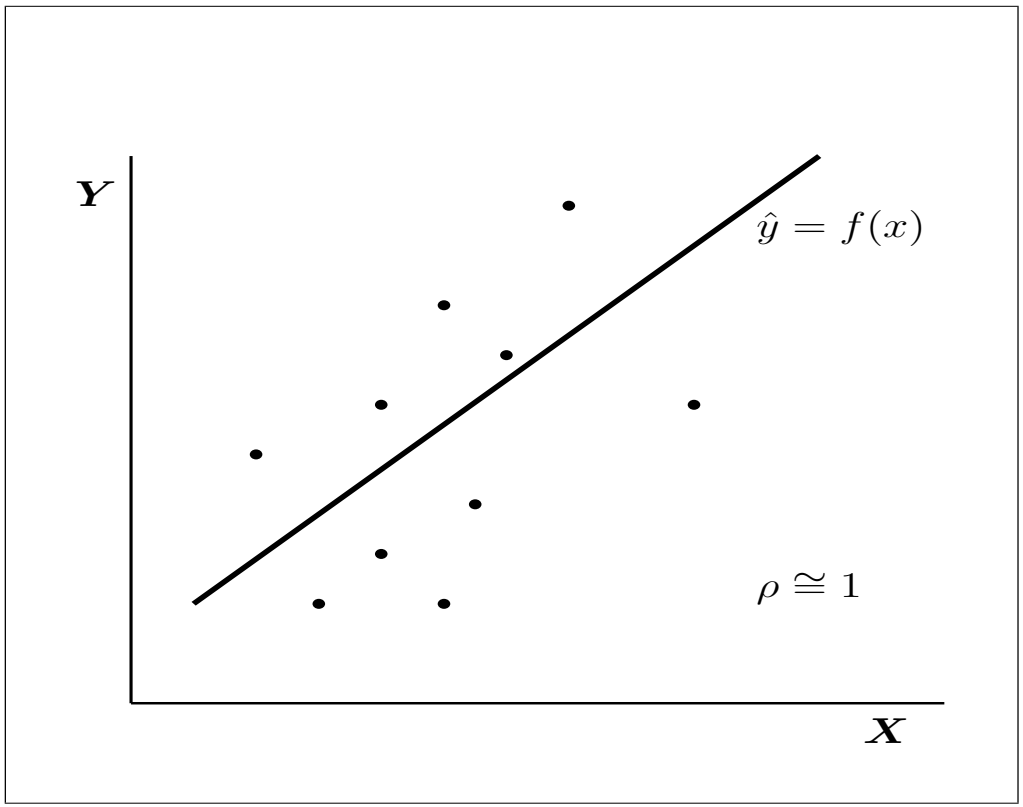




Estructura de un modelo de regresión simple. Partes determinista y aleatoria I

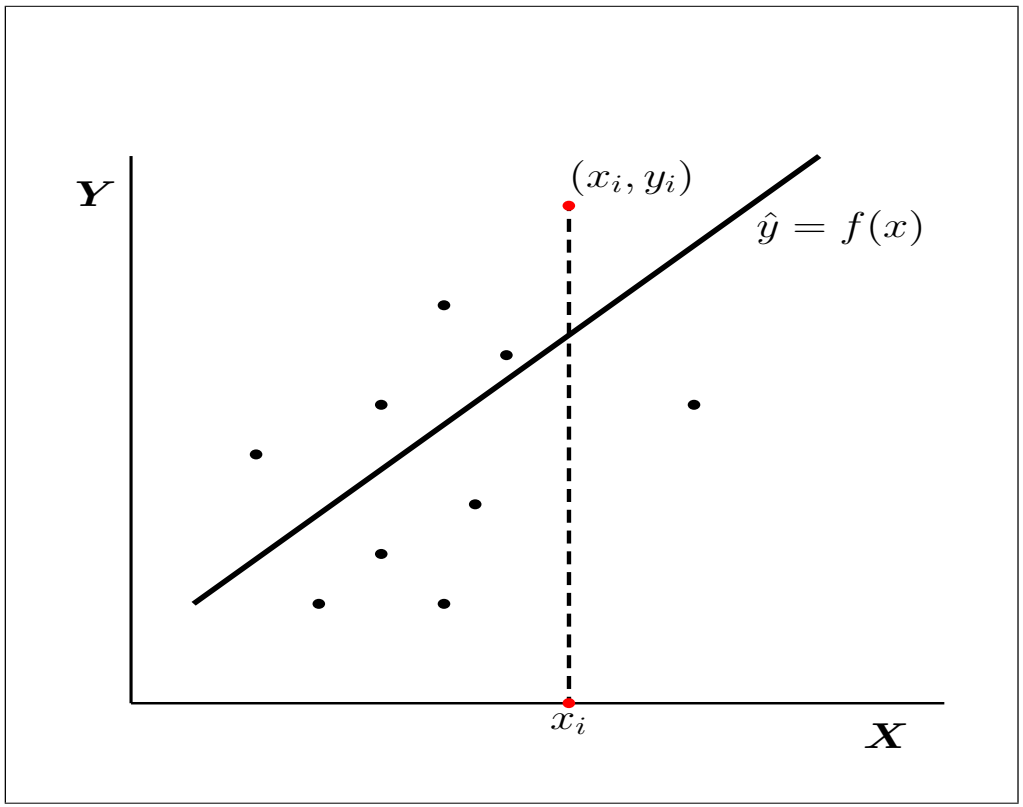
Para analizar la estructura de un modelo de regresión, supóngase que se ha ajustado uno de estos modelos a un conjunto de datos.

Sin pérdida de generalidad, se supondrá que se analiza el caso de dependencia entre dos variables, y que se puede considerar que el modelo adecuado es una recta:



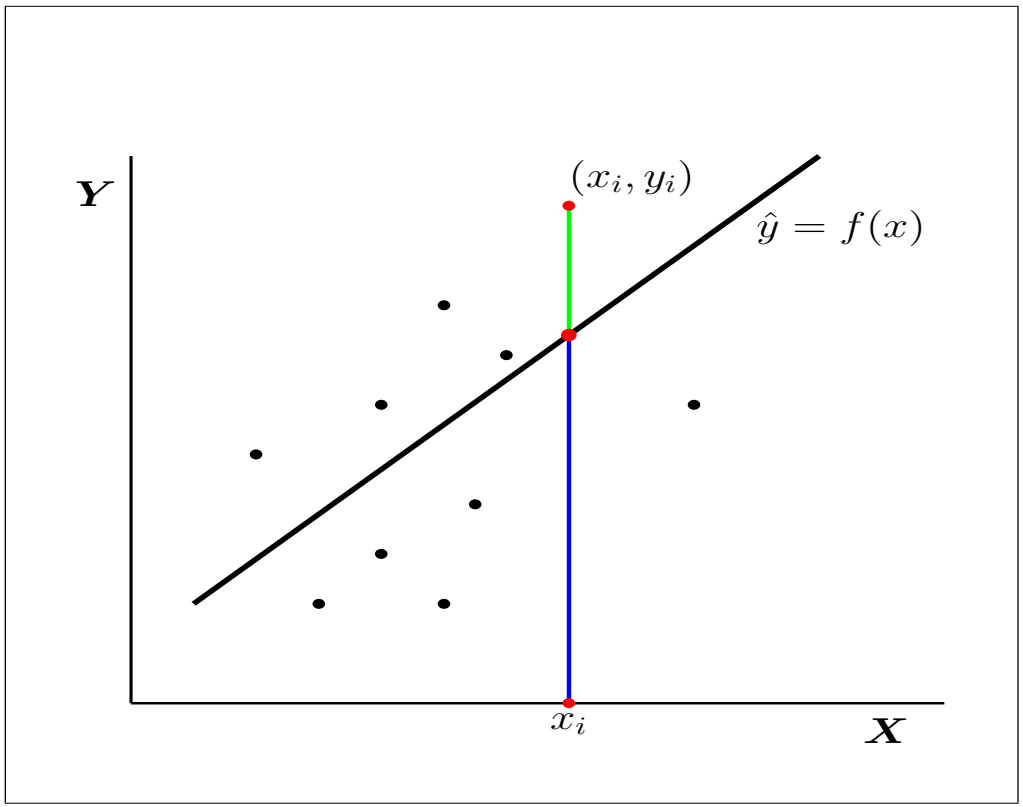
Estructura de un modelo de regresión simple. Partes determinista y aleatoria II

Sea (x_i, y_i) un punto correspondiente a un dato cualquiera del conjunto:



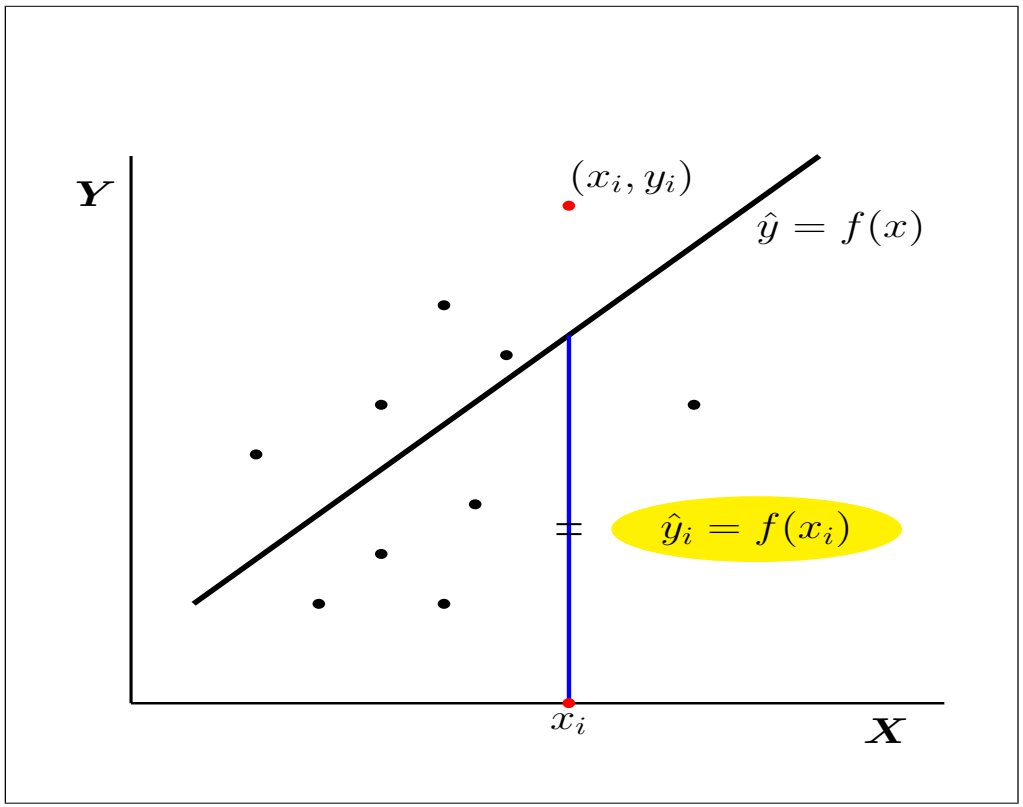
Estructura de un modelo de regresión simple. Partes determinista y aleatoria III

y_i se puede descomponer como se describe en el gráfico:



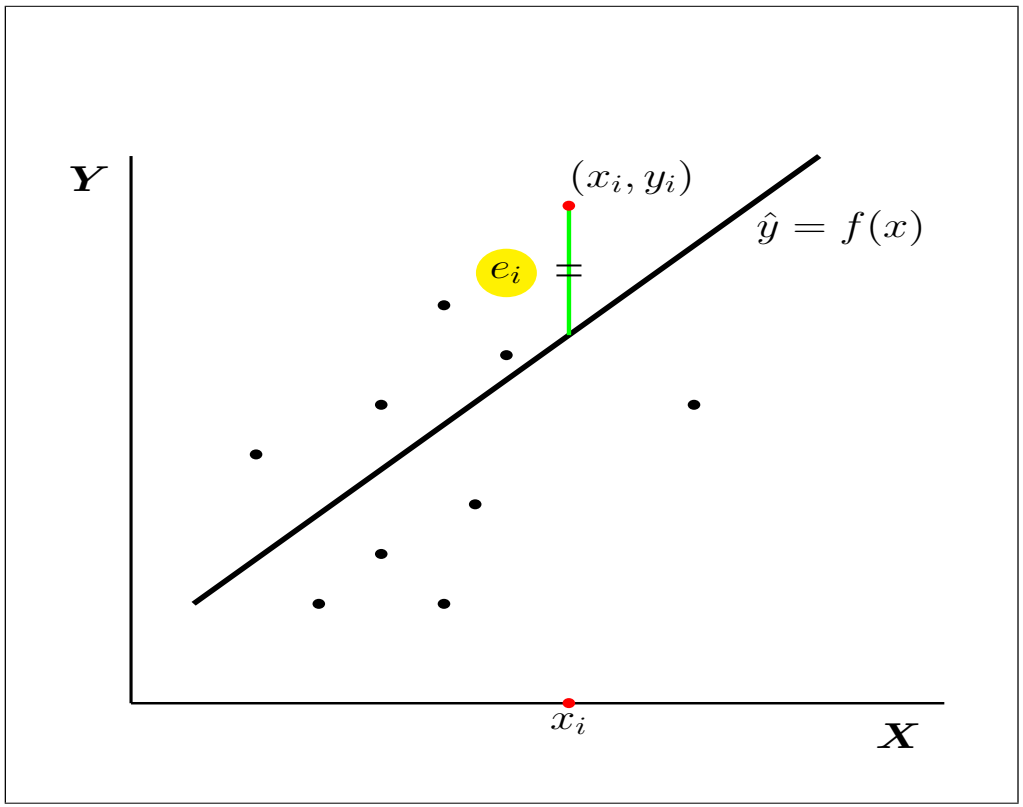
Estructura de un modelo de regresión simple. Partes determinista y aleatoria IV

La parte inferior, $\hat{y}_i = f(x_i)$, representa el valor que el modelo prevé para la variable Y, en un individuo cuyo valor en X es x_i.



Estructura de un modelo de regresión simple. Partes determinista y aleatoria V

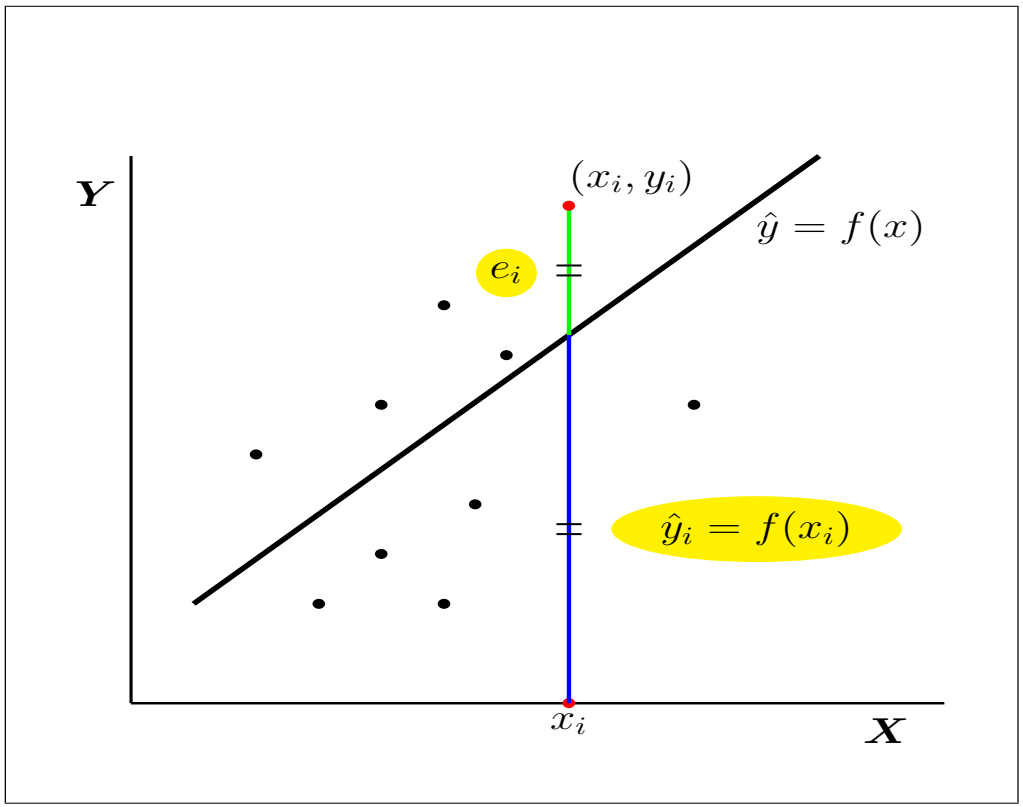
La parte superior, e_i, es la diferencia entre el valor observado de Y en el individuo y_i, y el previsto por el modelo, $\hat{y}_i$, para ese individuo.



Estructura de un modelo de regresión simple.Partes determinista y aleatoria VI

En consecuencia,

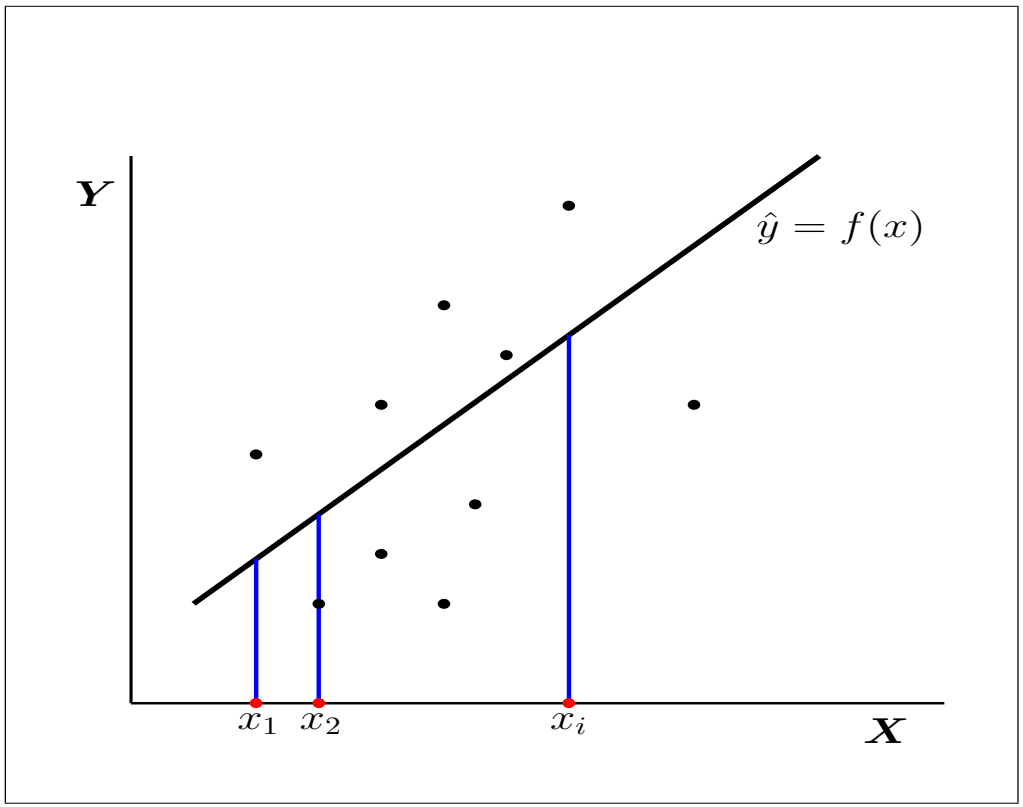
$y_i = \hat{y}_i + e_i$.



Estructura de un modelo de regresión simple. La parte determinista.

Calculado el modelo, el valor de $\hat{y}_i$ queda determinado para cada x_i, $\hat{y}_i = f(x_i)$

$\hat{y}_i = f(x_i)$ es la parte **determinista, o funcional** del modelo.

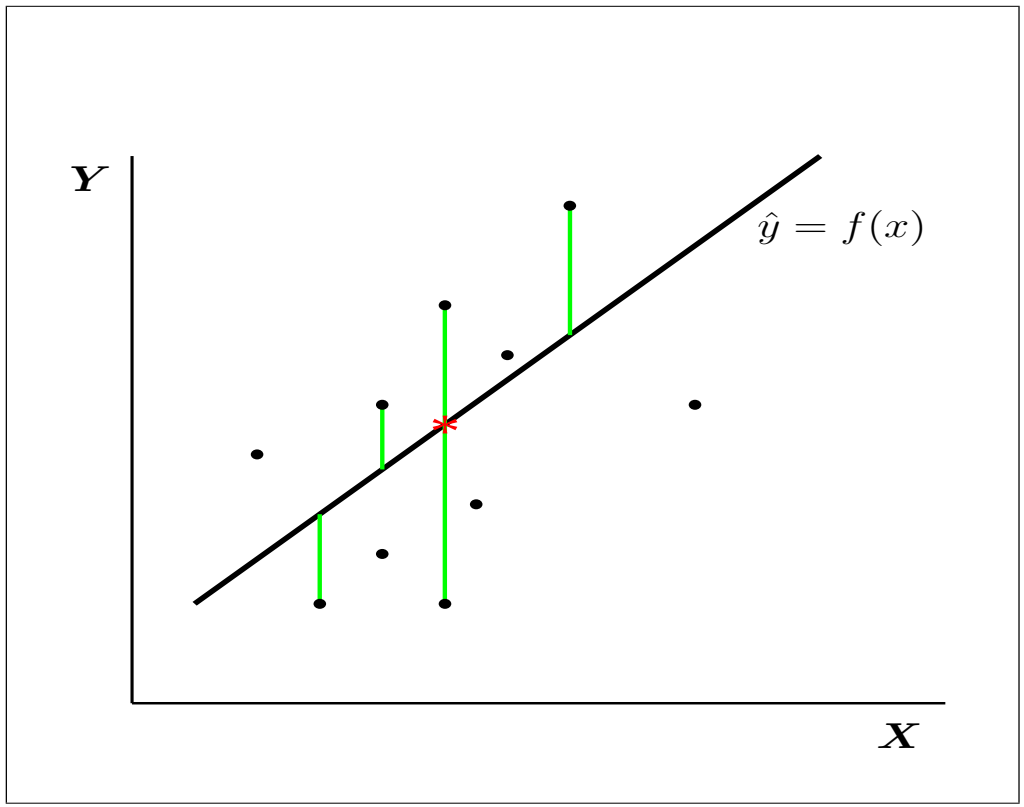


Estructura de un modelo de regresión simple. La parte aleatoria.

Calculado el modelo, el valor de e_i no queda determinado por x_i

Puede haber dos observaciones con el mismo x_i y distinto e_i

$e_i = y_i - \hat{y}_i$ es la parte **aleatoria** del modelo. (Error aleatorio.)



Estructura de un modelo de regresión simple. Resumen

En consecuencia, la estructura de un modelo de regresión simple es:

$$\underbrace{y_i}_{\text{Valor observado}} = \underbrace{f(x_i)}_{\text{Parte determinista, } \hat{y}_i} + \underbrace{e_i}_{\text{Error aleatorio}}$$

De manera resumida:

$$\boxed{y=f(x)+E}$$

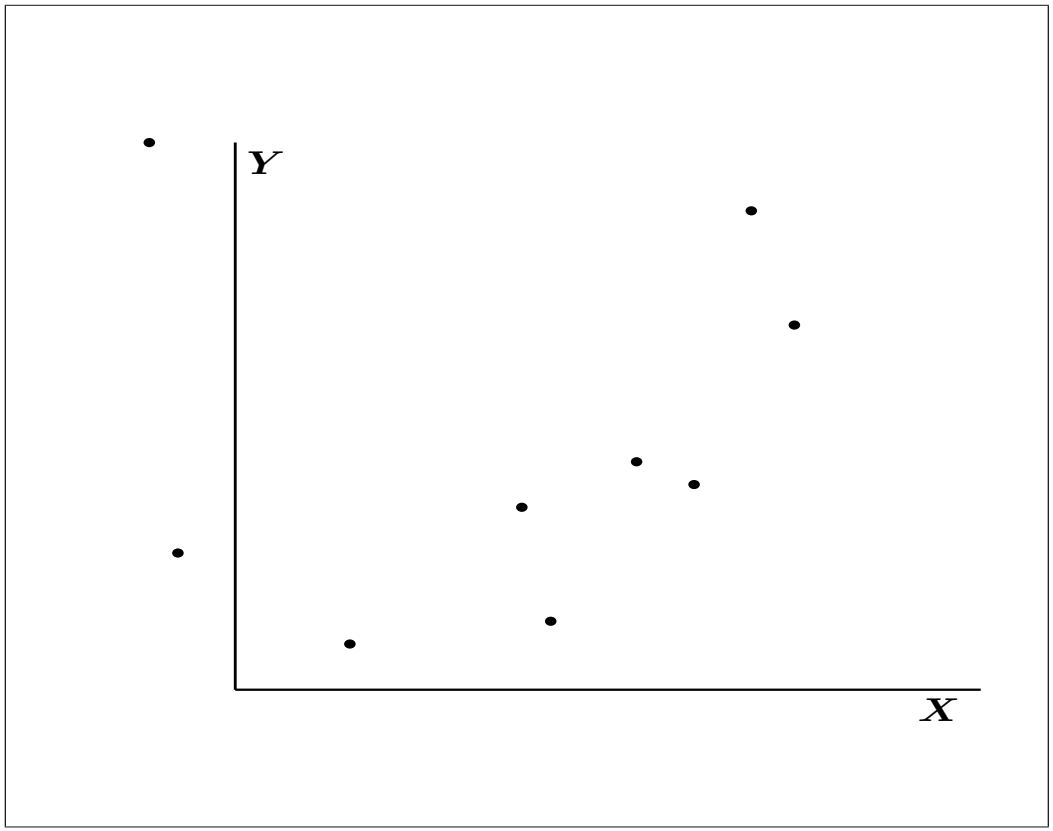
Nomenclatura de un modelo de regresión simple $y = f(x) + E$

- y es la variable explicada, dependiente o respuesta.
- x es la variable explicativa, el regresor o la variable independiente.
- E representa el error aleatorio. Contiene el efecto sobre y de todas las variables distintas de x.

Cálculo de los parámetros del modelo de R.S. Mínimos cuadrados I

Supóngase que un conjunto de datos sugiere que entre dos variables, X e Y, existe una relación de dependencia.

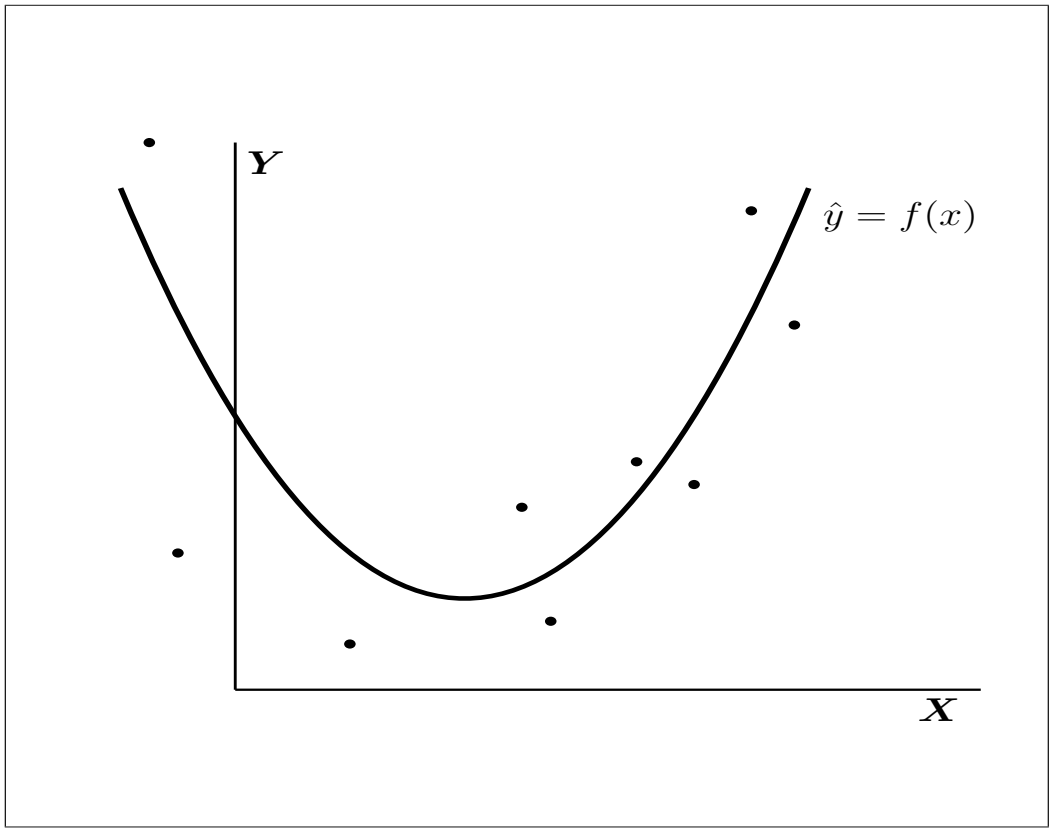
Gráficamente,



Cálculo de los parámetros del modelo de R.S. Mínimos cuadrados II

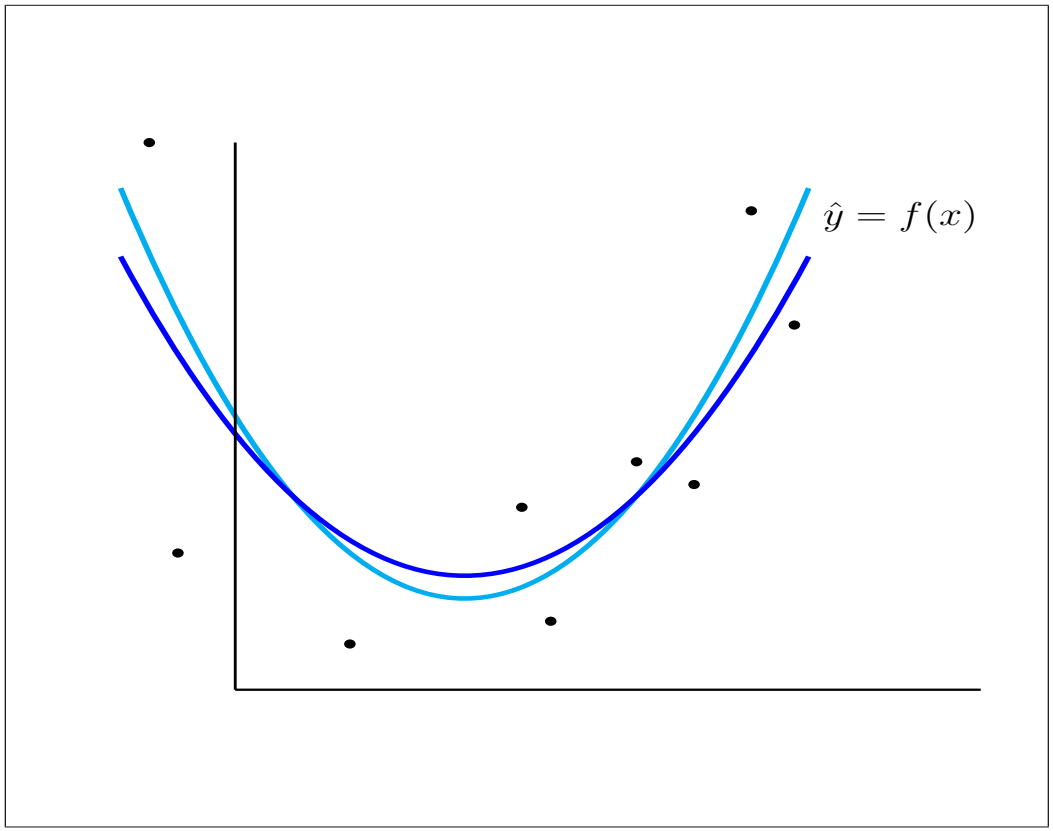
A la vista del gráfico se conjetura como un modelo posible una parábola de la forma:

$\hat{y} = c(x - h)^2 + k$



Cálculo de los parámetros del modelo de R.S. Mínimos cuadrados III

¿Qué valores de k, c y h se deben tomar?
Distintos valores de los parámetros modifican la ecuación del modelo ajustado.

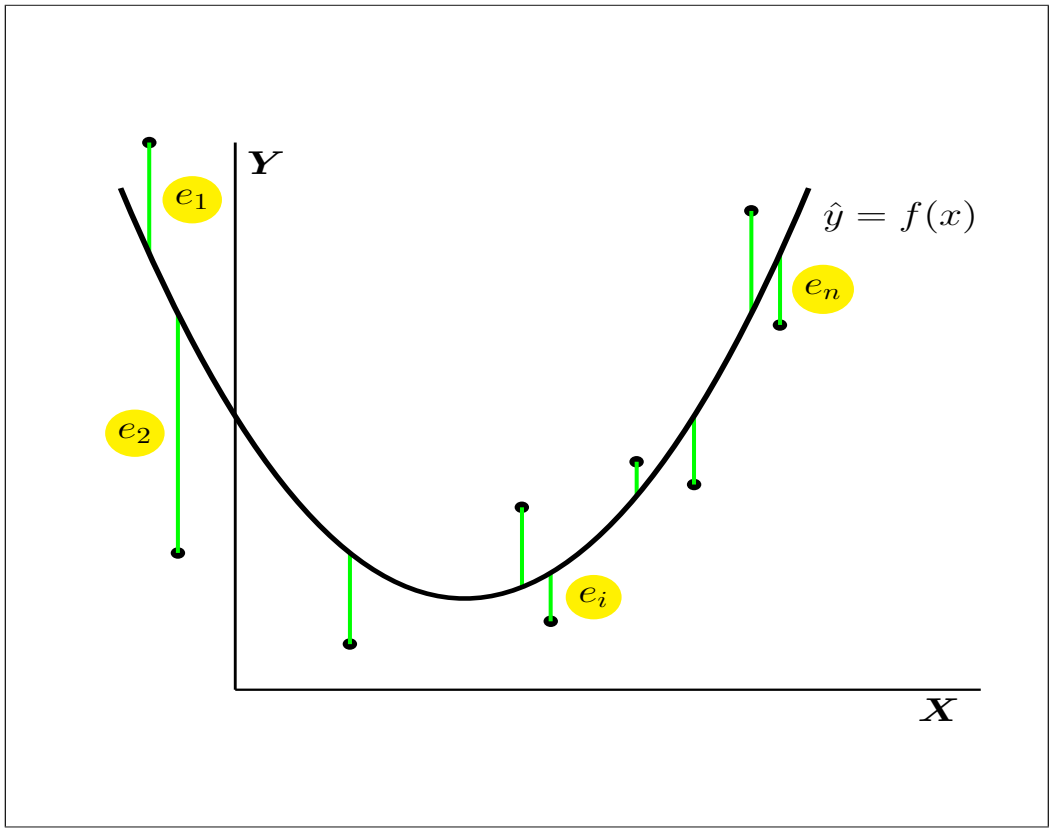


Cálculo de los parámetros del modelo de R.S. Mínimos cuadrados IV

Recuérdese que, para cualquier modelo ajustado, cada valor observado lleva asociado su error aleatorio:

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

Interesaría que, globalmente, el error cometido por el modelo fuera mínimo.



Cálculo de los parámetros del modelo de R.S. Mínimos cuadrados V

¿Cómo se minimiza globalmente el error asociado al modelo?

Cálculo de los parámetros del modelo de R.S. Mínimos cuadrados VI

Criterio de mínimos cuadrados:

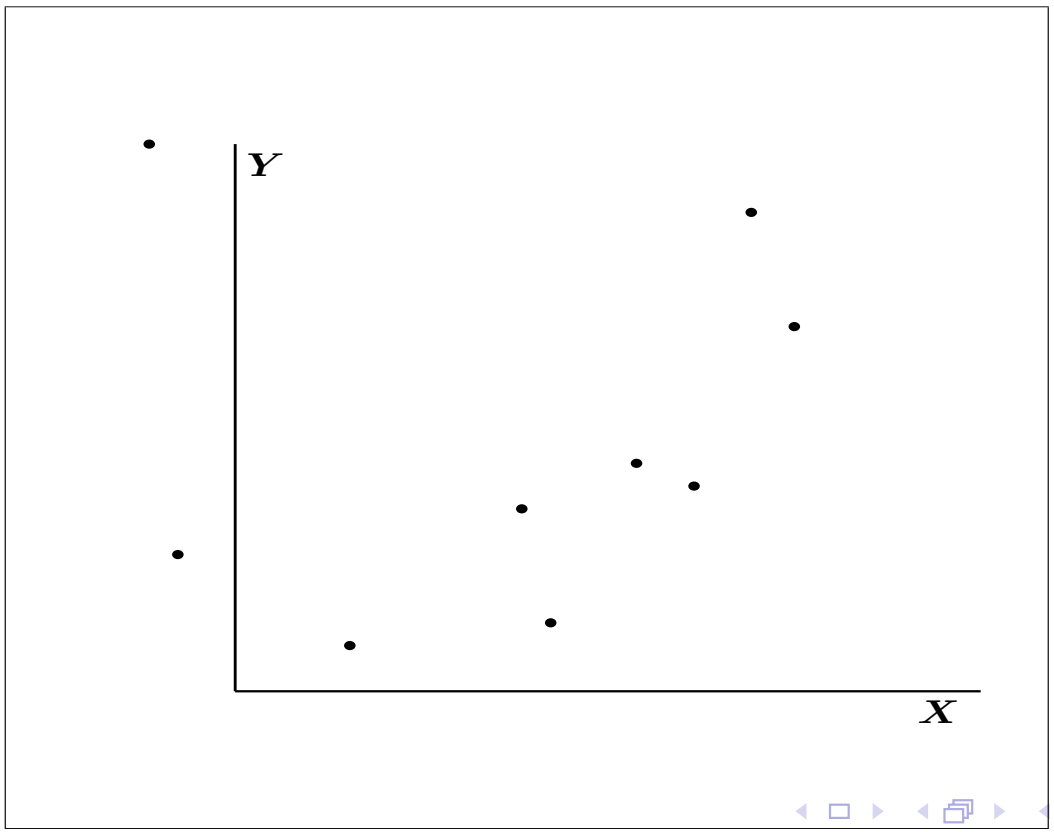
- Sea $e = (e_1, e_2, \ldots, e_n)$ el vector de errores asociado al modelo.
- El módulo de este vector viene dado por la expresión:

$$|e| = \sqrt{e_1^2 + e_2^2 + \cdots + e_n^2}$$

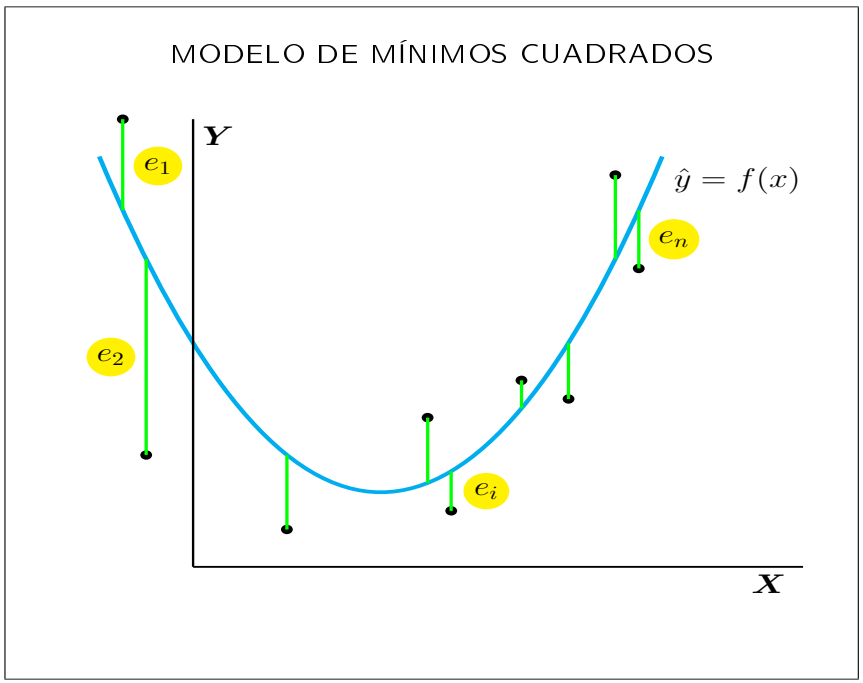
- El criterio de mínimos cuadrados selecciona los valores de los parámetros del modelo que **minimizan** el módulo del vector error, (equivalentemente el $\sum(e_i^2)$.)

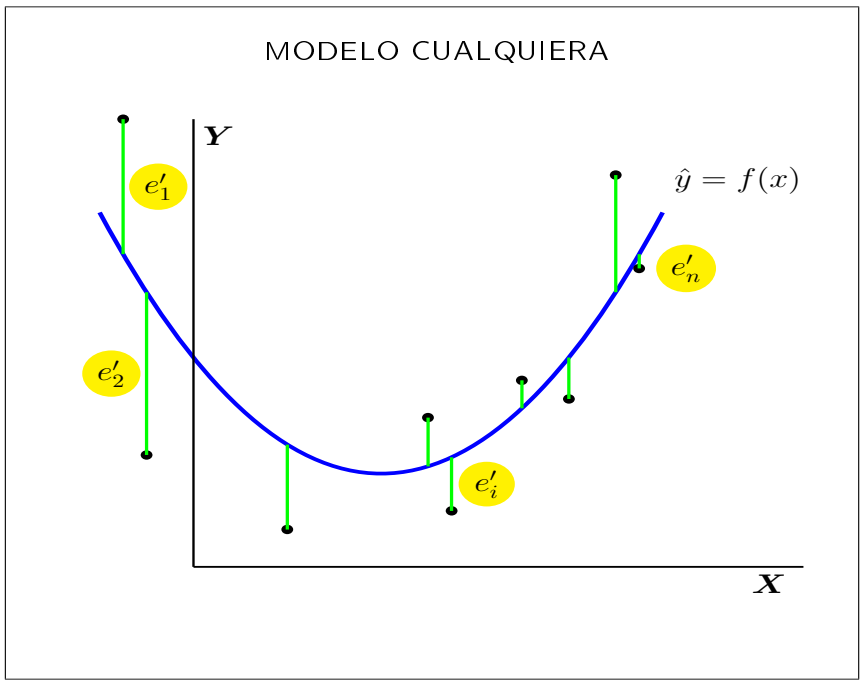
Cálculo de los parámetros del modelo de R.S. Mínimos cuadrados VII. Ejemplo

Se se ajustan dos modelos de regresión a una nube de puntos, y uno de ellos es el de mínimos cuadrados:



Cálculo de los parámetros del modelo de R.S. Mínimos cuadrados VIII. Ejemplo





Necesariamente,

$$\sum e_i^2 < \sum (e'_i)^2$$

Los parámetros del modelo de regresión lineal por mínimos cuadrados I

El modelo de regresión lineal con una variable independiente tiene la forma:

$$\underbrace{y = \beta_0 + \beta_1 x}_{Recta} + E$$

Los parámetros del modelo de regresión lineal por mínimos cuadrados II

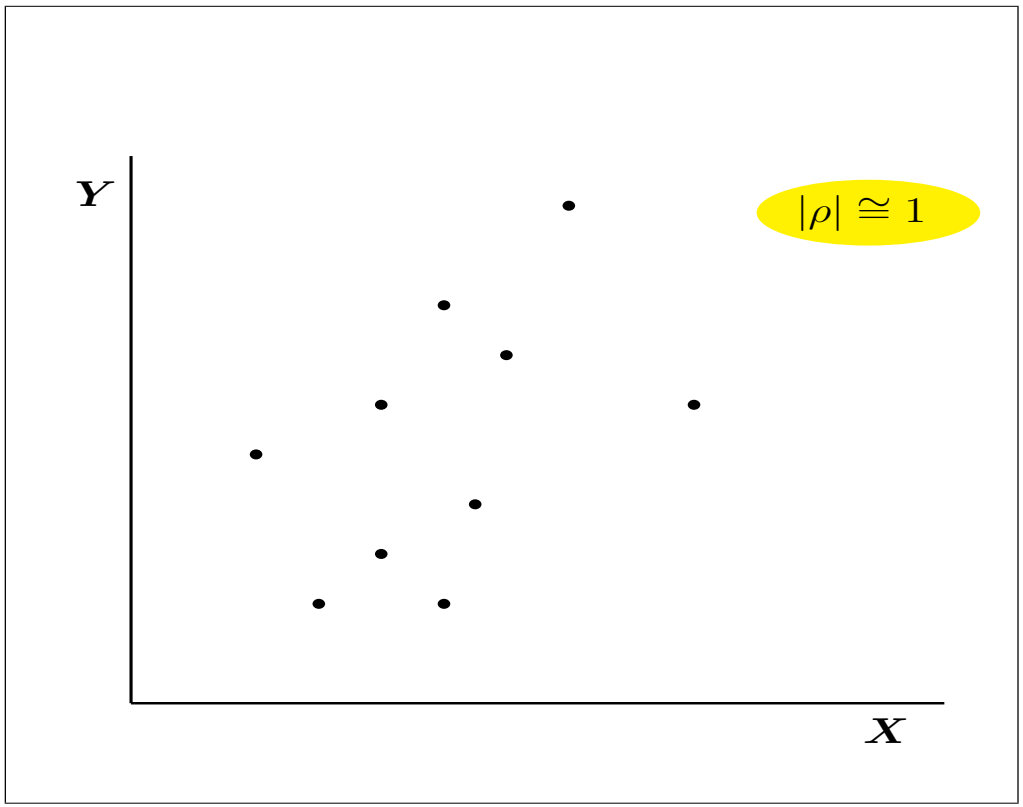
- El modelo de regresión lineal simple es el modelo de regresión más sencillo.

Se utiliza cuando:

1. La nube de puntos se asemeja a una recta.
2. El coeficiente de correlación lineal es alto en valor absoluto.

Los parámetros del modelo de regresión lineal por mínimos cuadrados III

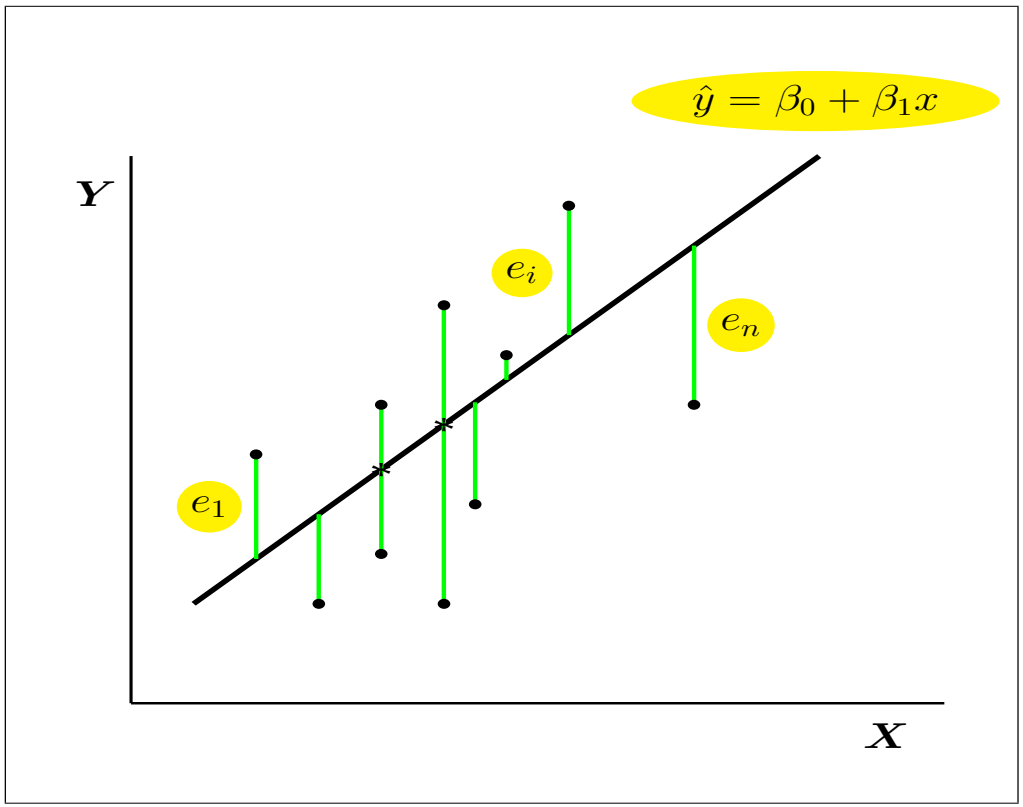
Supóngase que la relación entre dos variables sugiere una alta relación lineal.



Los parámetros del modelo de regresión lineal por mínimos cuadrados IV

Para ajustar una recta por mínimos cuadrados hay que minimizar:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2$$



Los parámetros del modelo de regresión lineal por mínimos cuadrados V

Como S es función de β_0 y de β_1, para que S sea mínimo:

$$\frac{\partial S}{\partial \beta_0} = 0$$

y

$$\frac{\partial S}{\partial \beta_1} = 0$$

Los parámetros del modelo de regresión lineal por mínimos cuadrados VI

Ahora bien, como

$$e_i = y_i - \hat{y}_i, \quad \text{con} \quad \hat{y}_i = \beta_0 + \beta_1 x_i,$$

se tiene que:

$$e_i = y_i - (\beta_0 + \beta_1 x_i)$$

Los parámetros del modelo de regresión lineal por mínimos cuadrados VII

De donde:

$$\frac{\partial S}{\partial \beta_0} = \frac{\partial \left[\sum_{i=1}^{n} e_i^2\right]}{\partial \beta_0} = \frac{\partial \left[\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2\right]}{\partial \beta_0} = 0$$

y

$$\frac{\partial S}{\partial \beta_1} = \frac{\partial \left[\sum_{i=1}^{n} e_i^2\right]}{\partial \beta_1} = \frac{\partial \left[\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2\right]}{\partial \beta_1} = 0$$

Los parámetros del modelo de regresión lineal por mínimos cuadrados VIII

Operando para resolver el sistema anterior se tiene que:

$$\sum_{i=1}^{n} e_i = 0.$$

$$\sum_{i=1}^{n} e_i x_i = 0, \quad \text{e}$$

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

Siendo $\hat{\beta}_0$ y $\hat{\beta}_1$ las soluciones del sistema.

Los parámetros del modelo de regresión lineal por mínimos cuadrados IX

Resolviendo el sistema, se tiene que:

$$\hat{\beta}_1 = \frac{COV(X,Y)}{s_x^2}$$

Por lo que la ecuación de la recta de regresión es:

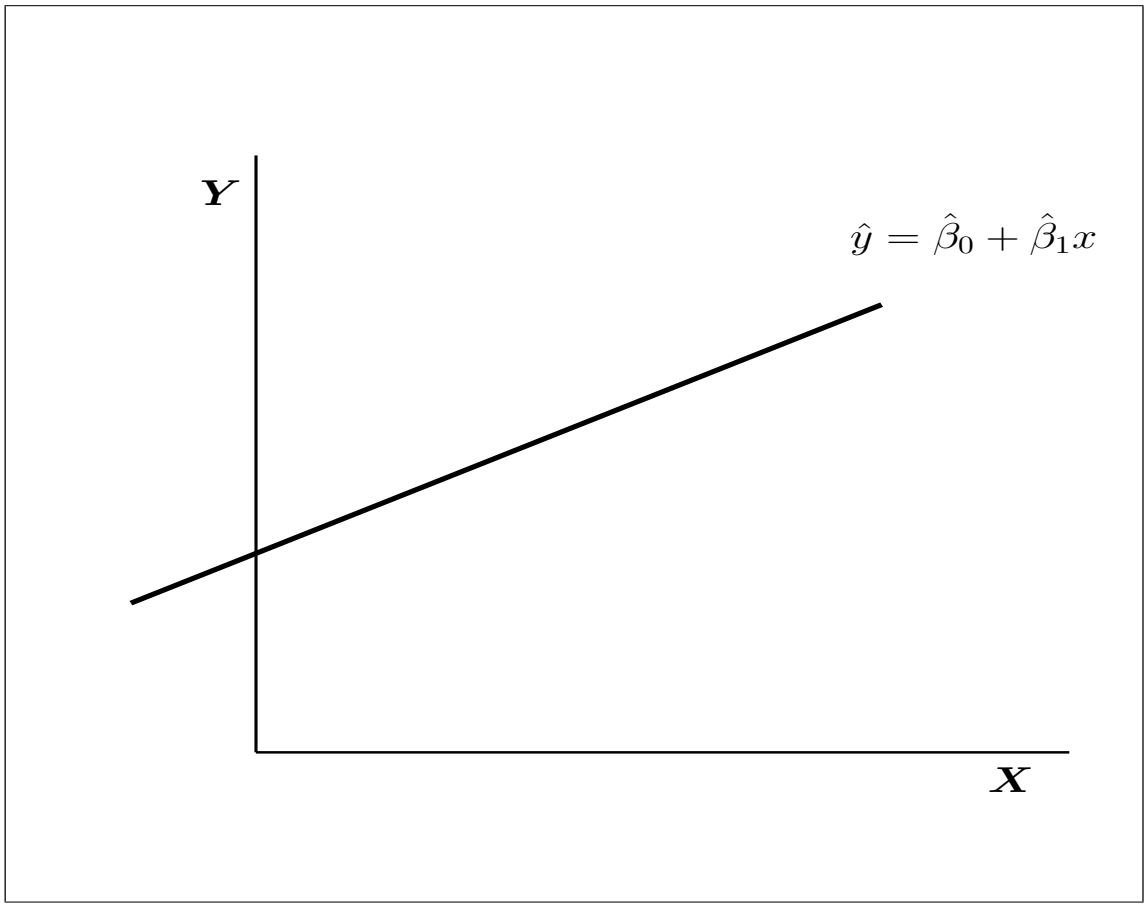
$$\boxed{(y - \bar{y}) = \frac{COV(X,Y)}{s_x^2}(x - \bar{x})}$$

Interpretación de los parámetros de un modelo de regresión lineal simple I

- En el modelo $y = \beta_0 + \beta_1 x$ que relaciona las variables X e Y:
  - β_0 representa el valor medio de la variable $Y|X = 0$, que en muchas ocasiones carece de sentido.
  - β_1 representa la variación de la variable Y, cuando X aumenta o disminuye una unidad.

Interpretación de los parámetros de un modelo de regresión lineal simple II

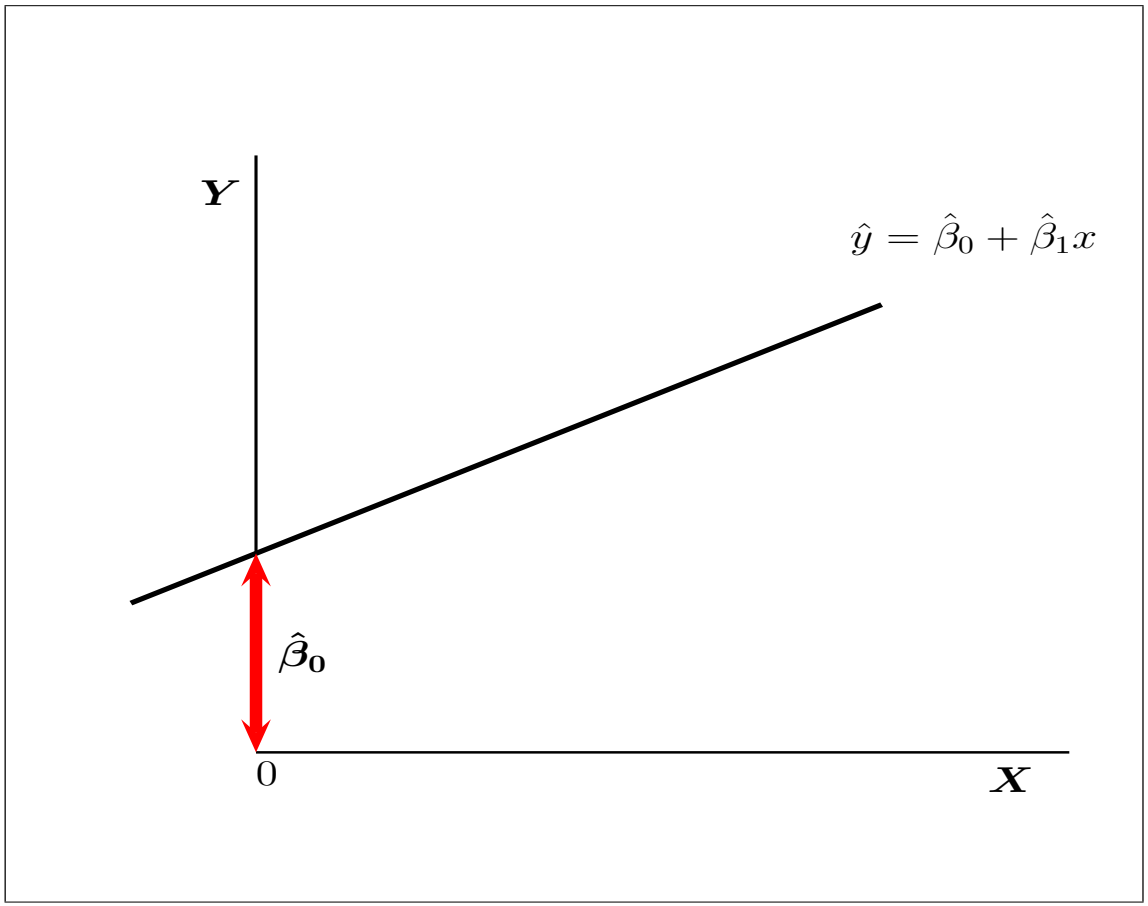
Si $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ es la recta de regresión calculada por mínimos cuadrados, asociada a una muestra,



Interpretación de los parámetros de un modelo de regresión lineal simple III

$\hat{\beta}_0 = \hat{y}(0)$. Gráficamente,

- Obsérvese que $\hat{\beta}_0$ no siempre tiene significado físico.



Interpretación de los parámetros de un modelo de regresión lineal simple IV

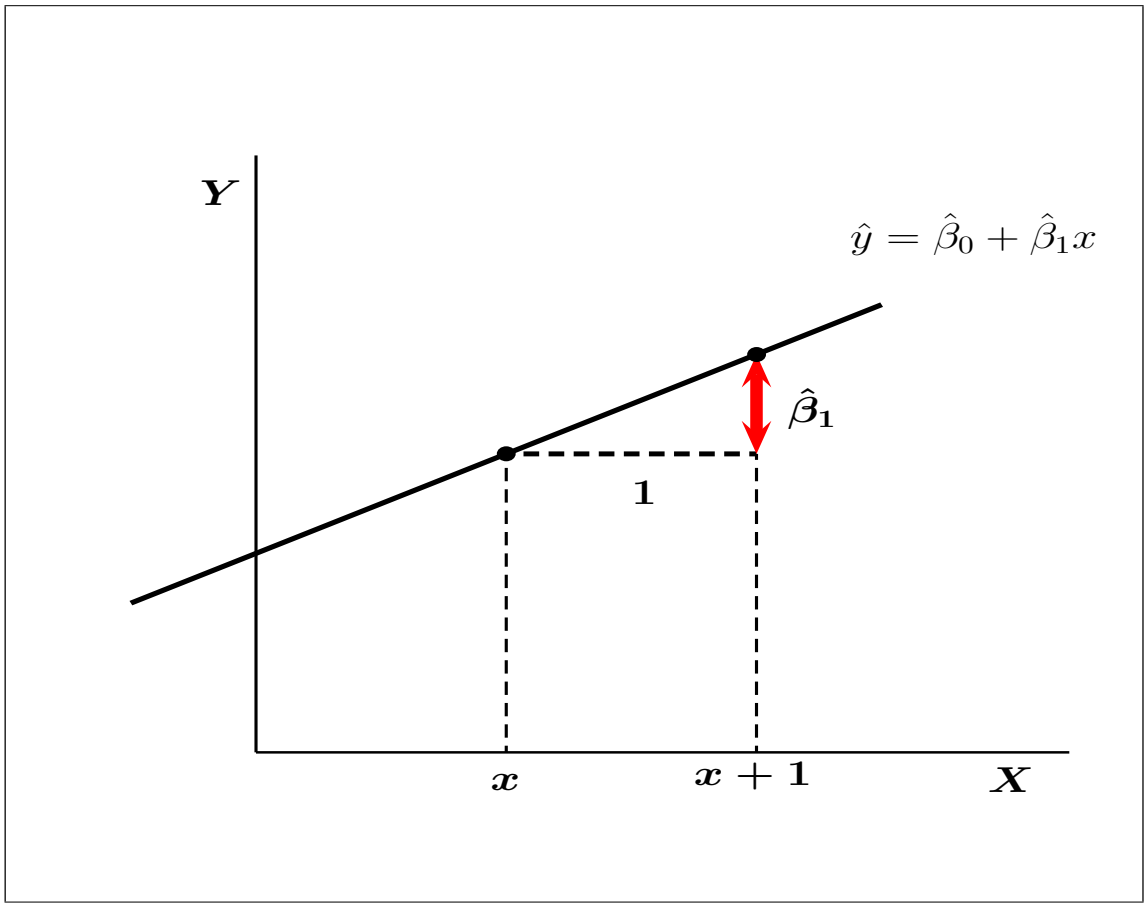
$\hat{\beta}_1$ representa la variación de la variable Y cuando X aumenta o disminuye una unidad. En efecto:

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

$$\hat{y}(x+1) = \hat{\beta}_0 + \hat{\beta}_1 (x+1),$$

De donde,

$$\hat{y}(x+1) - \hat{y}(x) = \hat{\beta}_1.$$



Problemas abiertos

Una vez calculado un modelo de regresión, cabe preguntarse

1. ¿Cómo se emplea un modelo de regresión?
2. ¿Qué fiabilidad ofrecen las previsiones de un modelo de regresión?

Hipótesis del modelo I

Idea clave

Para poder usar correctamente un modelo de regresión y para analizar su fiabilidad es necesario controlar el error.

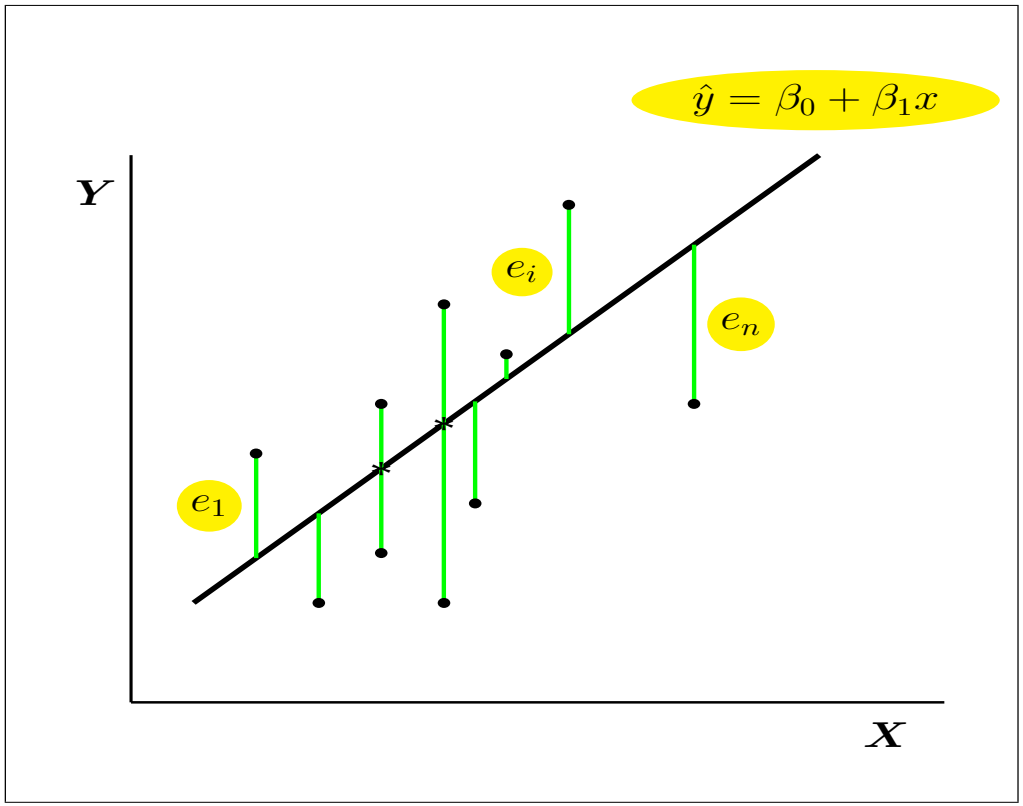
Hipótesis del modelo II

Recordando que para cada observación, (x_i, y_i)

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

Se tiene que

- Cada error, e_i, es una **variable aleatoria**.



Hipótesis del modelo III

Al ajustar un modelo de regresión lineal simple, se supondrá que se verifican las siguientes hipótesis:

1. Para un valor fijo de X, x_i, se tiene que $y_i = \beta_0 + \beta_1 x_i + e_i$ donde β_0 y β_1 son constantes desconocidas.
2. Cada error $e_i \approx N\left(0, \sigma^2\right)$.
   - La hipótesis de normalidad se basa en el teorema central del límite.
   - El hecho de que la varianza sea constante recibe el nombre de homocedasticidad.
3. Cualquier par de errores e_i y e_j son independientes.

Consecuencias de las hipótesis del modelo I

Las hipótesis impuestas al modelo tienen las siguientes consecuencias:

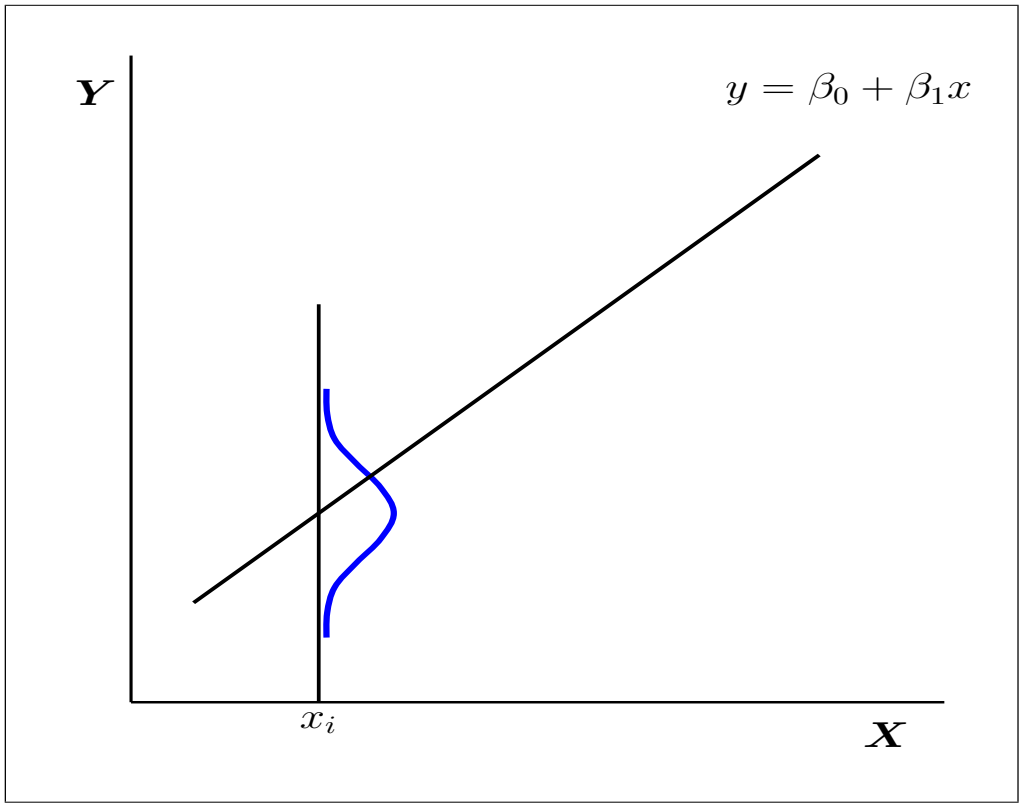
1. Para cada valor, x_i, de X la variable aleatoria $(Y|X = x_i)$ tiene una distribución:

$$(Y|X = x_i) \approx N(\beta_0 + \beta_1 x_i, \sigma^2)$$

2. Las observaciones y_i de la variable Y son independientes.

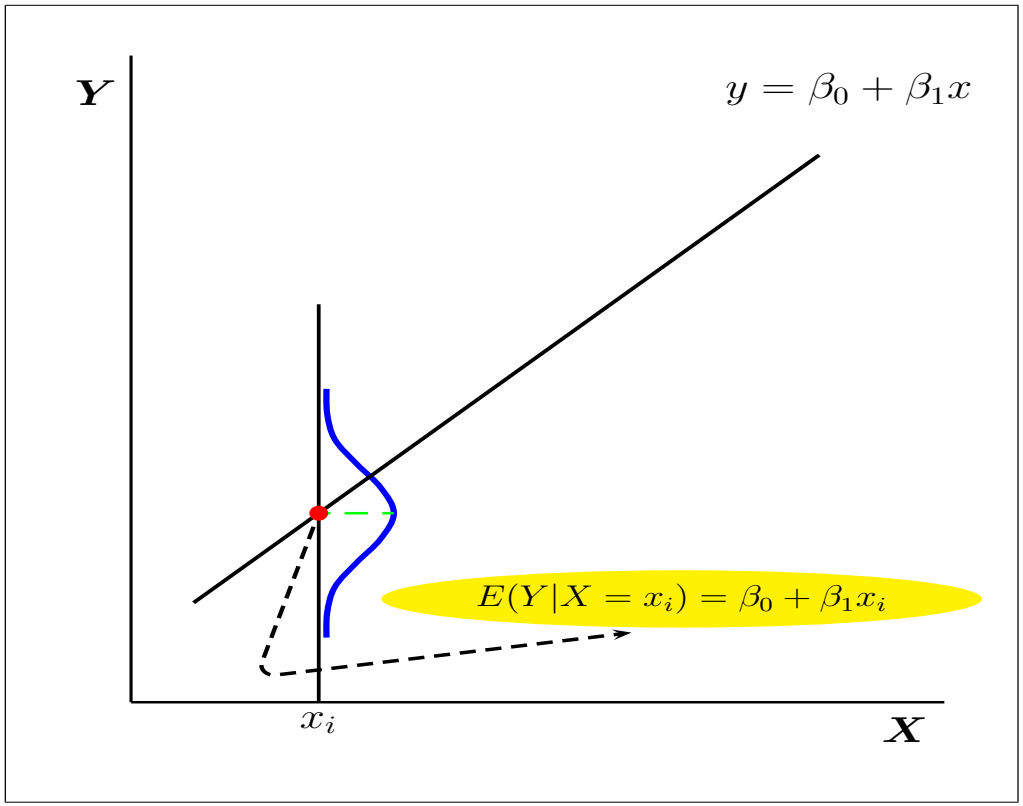
Consecuencias de las hipótesis del modelo II

Gráficamente, si las hipótesis del modelo son ciertas, cuando $X = x_i$, Y es una V.A. normal.



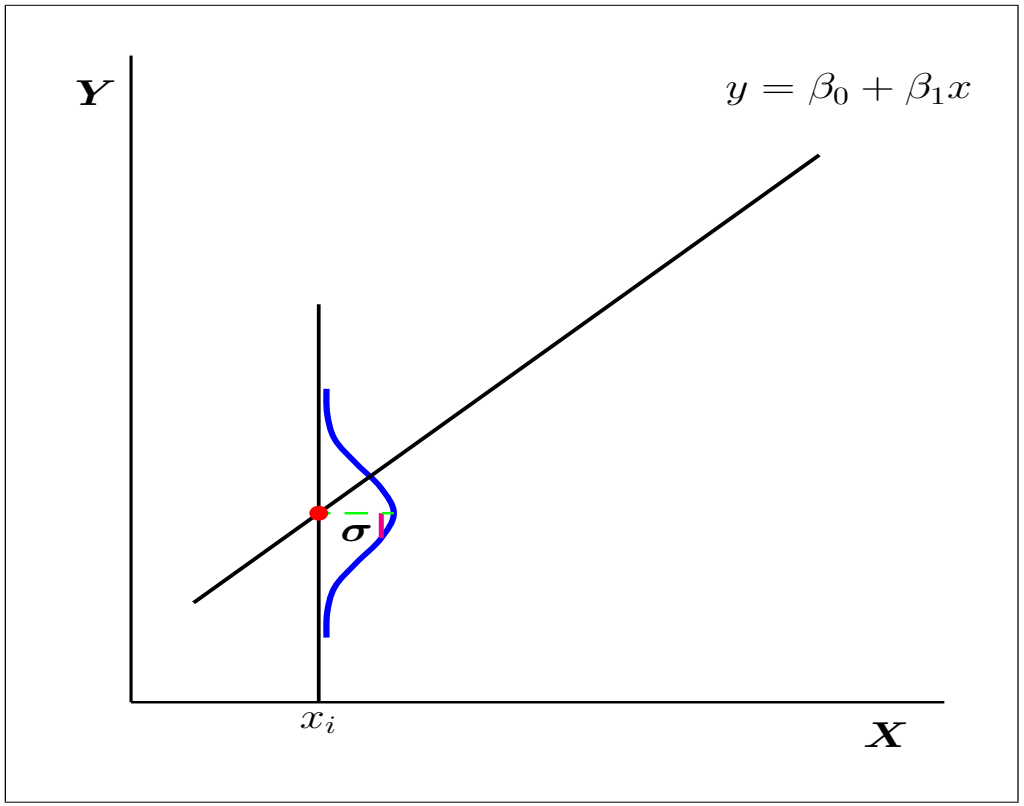
Consecuencias de las hipótesis del modelo III

La esperanza matemática de esta distribución es $\beta_0 + \beta_1 x_i$.



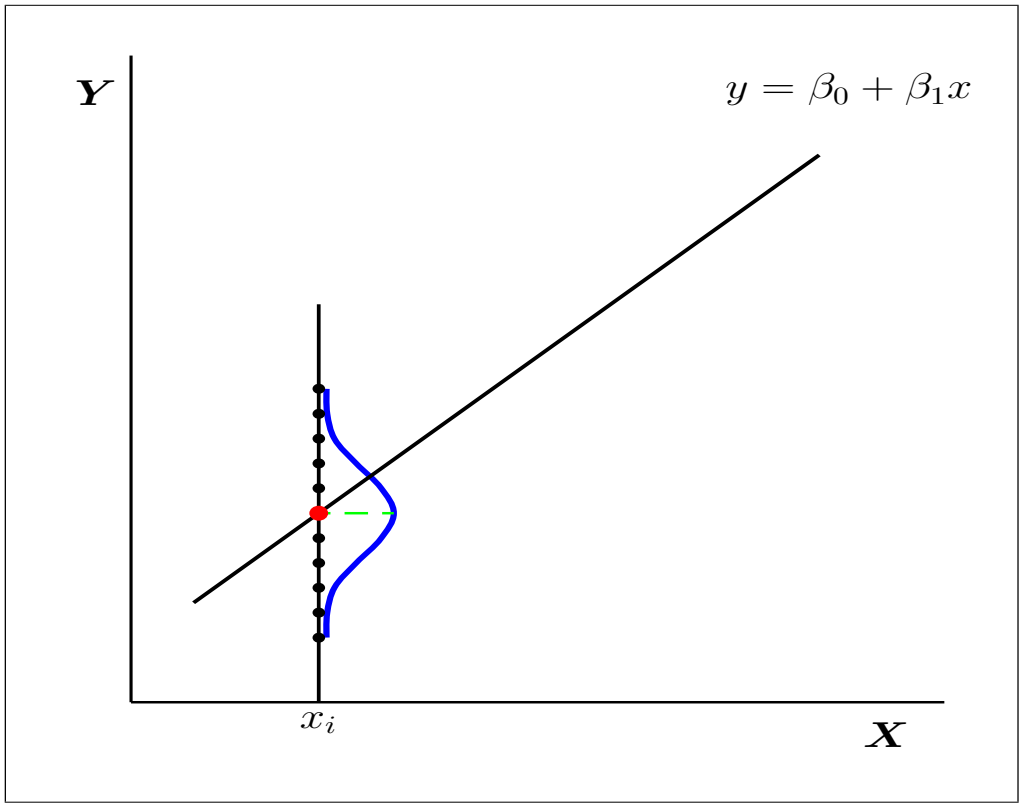
Consecuencias de las hipótesis del modelo IV

La desviación típica de esta distribución coincide con la del error aleatorio, σ.



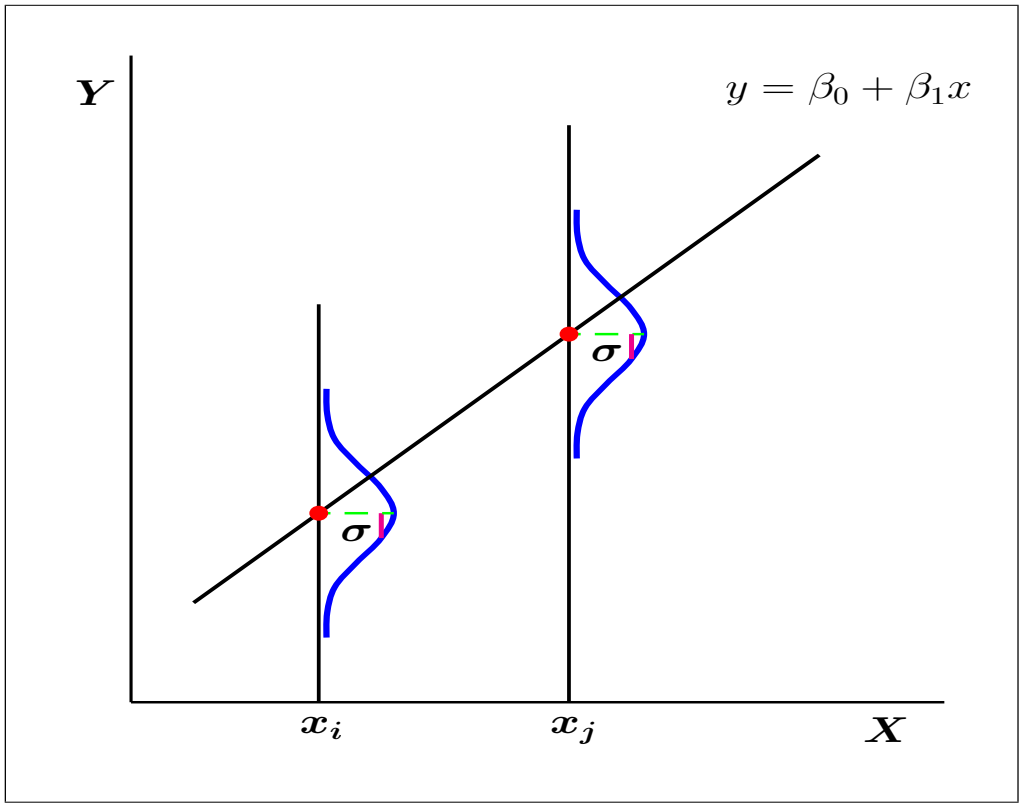
Consecuencias de las hipótesis del modelo V

En general, si el modelo es correcto,los valores de la variable Y, cuando $X = x_i$, se encontrarán en el intervalo $(\beta_0 + \beta_1 x_i) \pm 3\sigma$, con una probabilidad 0'997.



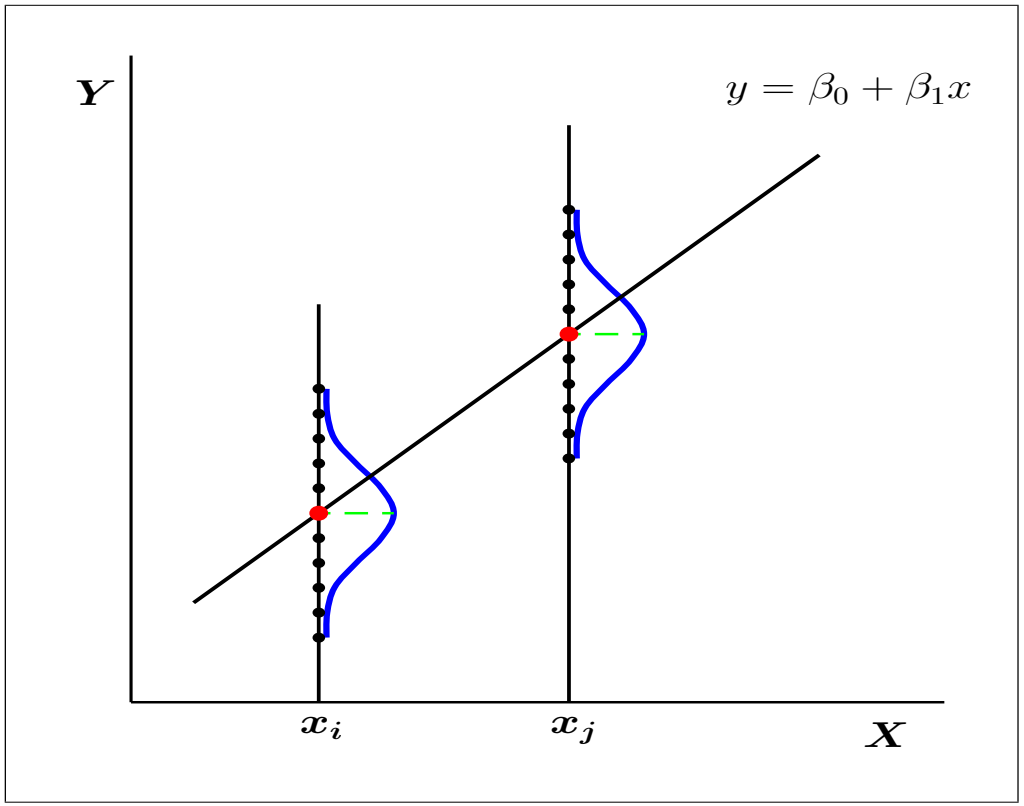
Consecuencias de las hipótesis del modelo VI

Para dos valores distintos de X, $X = x_i$ y $X = x_j$, las distribuciones de Y serán:



Consecuencias de las hipótesis del modelo VII

Y los individuos de $Y|X = x_i$ y de $Y|X = x_i$ se situarán, respectivamente, como muestra la figura:



Consecuencias de las hipótesis del modelo VIII.Resumen

Si las hipótesis del modelo son ciertas:

1. Existe una recta, $y = \beta_0 + \beta_1 x$ que, para cada valor de $X = x_i$, permite obtener el valor de la esperanza de $(Y|X = x_i)$:
$$E(Y|X = x_i) = \beta_0 + \beta_1 x_i$$
2. La varianza de la distribución de $(Y|X = x_i)$, que es normal, no depende de x_i y coincide con la varianza del error, σ^2.

Estimadores de los parámetros de la recta de regresión I

Problema

Si existe una recta, $y = \beta_0 + \beta_1 x$, que pasa por los puntos (x_i, μ_{x_i}), donde μ_{x_i} representa la media de la distribución de Y condicionada por $X = x_i$, ¿coincide con la recta $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ calculada por mínimos cuadrados?

Estimadores de los parámetros de la recta de regresión II

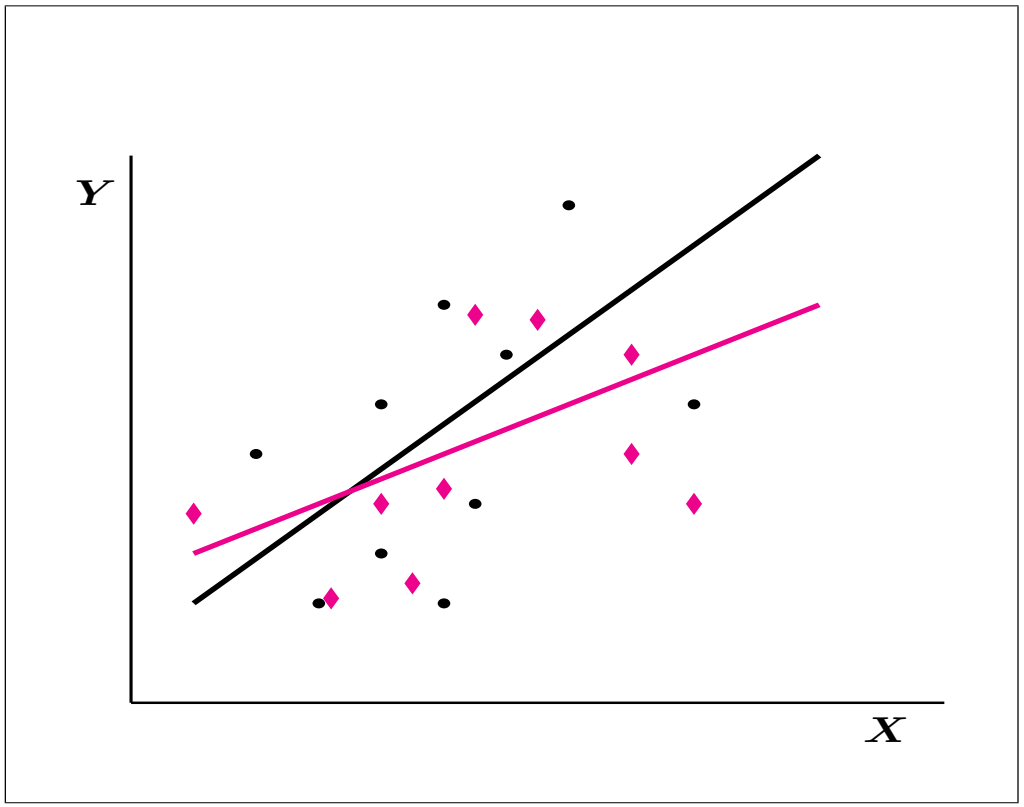
Discusión del problema

1. Si existe una recta, $y = \beta_0 + \beta_1 x$, que pasa por los puntos (x_i, μ_{x_i}), donde μ_{x_i} representa la media de la distribución de Y condicionada por $X = x_i$, ésta debería ser única.

2. La recta $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ calculada por mínimos cuadrados depende de la muestra $(x_1, y_1), \ldots, (x_n, y_n)$

Estimadores de los parámetros de la recta de regresión III

Gráficamente se observa cómo dos muestras distintas darían lugar a rectas distintas.



Estimadores de los parámetros de la recta de regresión IV

Conclusión

- La recta $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ es una aproximación de la recta $y = \beta_0 + \beta_1 x$.
- Los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimaciones de β_0 y β_1, respectivamente.
- $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores de β_0 y β_1.

Propiedades de los estimadores de los parámetros de la recta de regresión I

Recordando que los estimadores de un parámetro siempre son variables aleatorias, se puede demostrar que:

1. $$\hat{\beta}_1 \approx N\left(\beta_1, \frac{\sigma}{s_x\sqrt{n}}\right).$$

2. $$\hat{\beta}_0 \approx N\left(\beta_0, \frac{\sigma}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_x^2}}\right),$$

   donde σ representa la desviación típica del error experimental, y $\bar{x}$ y s_x son la media y la desviación típica de los valores observados de X, respectivamente.

Propiedades de los estimadores de los parámetros de la recta de regresión II

Observaciones

1. Tanto $\hat{\beta}_0$ como $\hat{\beta}_1$ son estimadores centrados de β_0 y de β_1, respectivamente.
2. Las desviaciones típicas de ambos estimadores crecen con el error experimental, σ, y disminuyen cuando aumenta la varianza de los valores observados de X.
3. La realización de un estudio inferencial para β_0 y β_1, requiere el conocimiento de σ.

Estimador de la varianza del error experimental. La varianza residual I

- La estimación por mínimos cuadrados no aporta información sobre la variabilidad del error experimental.
- La información sobre el error experimental se encuentra en los valores de e_i, con $i = 1, \ldots, n$

Estimador de la varianza del error experimental. La varianza residual II

- Los métodos de los momentos y de máxima verosimilitud proponen como estimador de σ^2, la varianza de los residuos:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n}$$

- Este estimador de σ^2 no tiene en cuenta las relaciones de dependencia entre los residuos:

$$\sum e_i = 0 \quad \text{y} \quad \sum e_i x_i = 0,$$

y origina un estimador no centrado de σ^2, es decir:

$$E(\hat{\sigma}^2) \neq \sigma^2.$$

Estimador de la varianza del error experimental. La varianza residual III

Alternativamente, se define la varianza residual en la forma:

$$\hat{s}_R^2 = \frac{\sum e_i^2}{n-2}.$$

$\hat{s}_R^2$ será el estimador habitual de σ^2.

Propiedades de la varianza residual

1. $\hat{s}_R^2$ es un estimador centrado de σ^2, esto es:

$$E(\hat{s}_R^2) = \sigma^2$$

2. Además,

$$\frac{\sum e_i^2}{\sigma^2} = \frac{(n-2)\hat{s}_R^2}{\sigma^2} \longrightarrow \chi^2_{n-2}.$$

   - Esta distribución permite realizar inferencia respecto del valor de σ^2.

Cálculo de intervalos de confianza para el coeficiente de regresión, β_1 I

Como

$$\hat{\beta}_1 \approx N\left(\beta_1, \frac{\sigma}{s_x\sqrt{n}}\right),$$

se deduce que:

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{s}_R}{s_x\sqrt{n}}} \longrightarrow t_{n-2},$$

por lo que, con el $(1-\alpha)\times 100\,\%$ de confianza,

$$\beta_1 \in \left(\hat{\beta}_1 \pm t_{\alpha/2\,;(n-2)} \times \frac{\hat{s}_R}{s_x\sqrt{n}}\right)$$

Cálculo de intervalos de confianza para el coeficiente de regresión, β_1 II. Ejemplo

Al calcular una recta de regresión que describa la relación entre el tamaño de un conjunto de siete guisantes con el de sus descendientes, se obtuvieron los siguientes resultados:

$$\hat{\beta}_1 = 0'21. \quad s_x = 2'00002871. \quad \text{Y} \quad \hat{s}_R = 0'204324741.$$

¿Cuál sería un intervalo de confianza al 95 % para β_1?

Cálculo de intervalos de confianza para el coeficiente de regresión, β_1 III. Ejemplo

Como

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{s}_R}{s_x\sqrt{n}}} \longrightarrow t_{n-2},$$

con el 95 % de probabilidad,

$$-2'57 \leq \frac{0'21 - \beta_1}{\frac{0'204324741}{2'000002871 \times \sqrt{7}}} \leq 2'57.$$

Cálculo de intervalos de confianza para el coeficiente de regresión, β_1. Ejemplo III

Y operando,

$$-2'57 \leq \frac{0'21 - \beta_1}{0'03861} \leq 2'57,$$

de donde se deduce que, con el 95 % de confianza,

$$\beta_1 \in (0'21 - 2'57 \times 0'03861, 0'21 + 2'57 \times 0'03861).$$

Es decir, al 95 %,

$$\boxed{\beta_1 \in (0'11076, 0'30923).}$$

El contraste de regresión I

- Se denomina contraste de regresión al análisis de la hipótesis $H_0 : \beta_1 = 0$, frente a la hipótesis alternativa $H_1 : \beta_1 \neq 0$.

- La realización del contraste se realiza teniendo en cuenta la distribución:

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{s}_R}{s_x\sqrt{n}}} \longrightarrow t_{n-2}.$$

El contraste de regresión II

- Por lo que, si la hipótesis nula, $\beta_1 = 0$, es cierta, debería ser

$$\frac{\hat{\beta}_1}{\frac{\hat{s}_R}{s_x\sqrt{n}}} \longrightarrow t_{n-2},$$

lo que permite discutir el resultado del contraste.

- Si

$$-t_{\alpha/2\,;(n-2)} \leq \frac{\hat{\beta}_1}{\frac{\hat{s}_R}{s_x\sqrt{n}}} \leq t_{\alpha/2\,;(n-2)}$$

se aceptará la hipótesis nula, rechazándose en caso contrario.

El contraste de regresión III. Ejemplo

Al calcular una recta de regresión que describa la relación entre el tamaño de un conjunto siete de guisantes con el de sus descendientes, se obtuvieron los siguientes resultados:

$$\hat{\beta}_1 = 0'21. \quad s_x = 2'00002871. \quad \text{Y} \quad \hat{s}_R = 0'204324741.$$

¿Se aceptaría, con una confianza del 95 %, la hipótesis de que $\beta_1 = 0$?

El contraste de regresión IV. Ejemplo

Si la hipótesis nula, $\beta_1 = 0$, es cierta, debería ser

$$\frac{\hat{\beta}_1}{\frac{\hat{s}_R}{s_x\sqrt{n}}} \longrightarrow t_{n-2},$$

por lo tanto, con el 95 % de confianza, debería cumplirse que:

$$-2'57 \leq \underbrace{\frac{0'21}{\frac{0'204324741}{2'00002871\sqrt{7}}}}_{5'438} \leq 2'57.$$

El contraste de regresión V. Ejemplo

Y como

$$5'438 \notin (-2'57, 2'57)$$

se rechaza la hipótesis nula de que $\beta_1 = 0$, y se acepta que $\beta_1 \neq 0$.

- Naturalmente, se podría haber llegado a la misma conclusión con el análisis del intervalo de confianza para β_1 calculado anteriormente, que no contiene al 0.

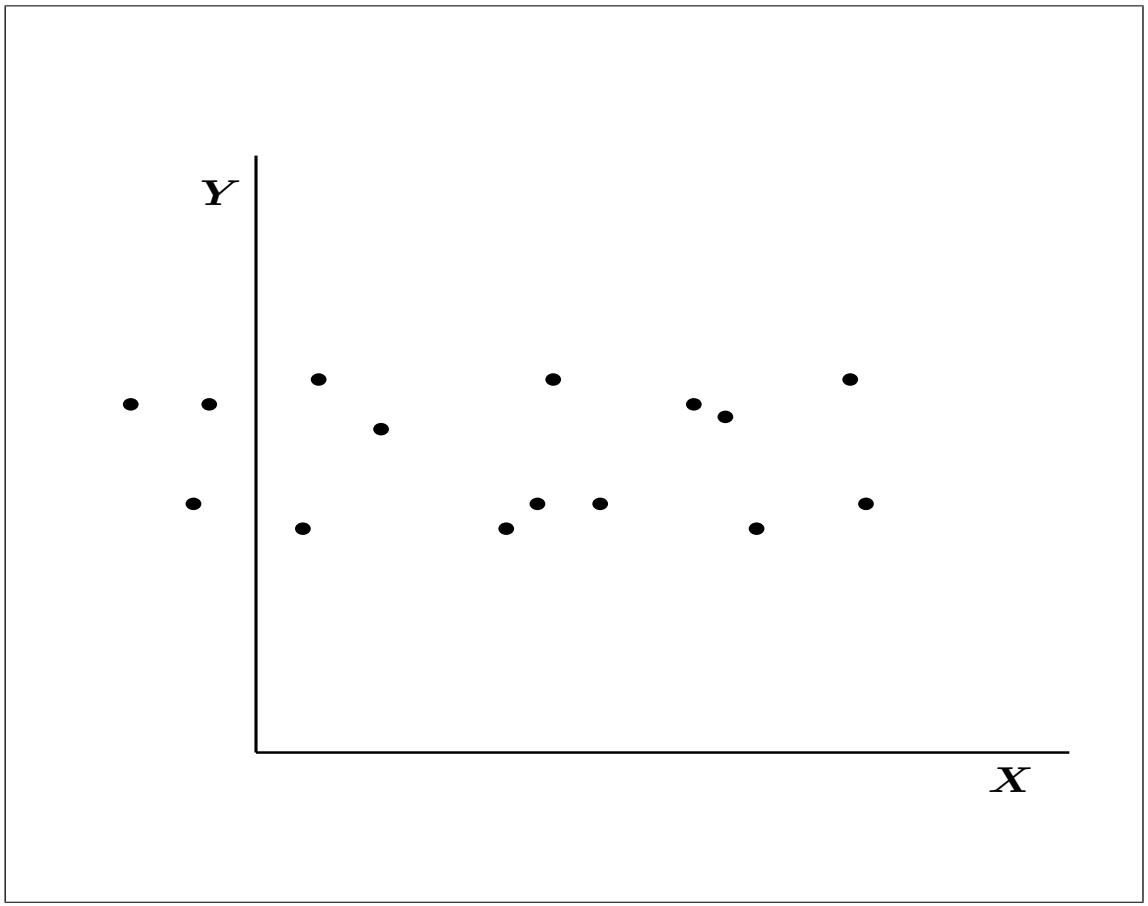
El contraste de regresión VI. Interpretación

Observaciones:

- La aceptación del contraste de regresión, $\beta_1 = 0$, se interpreta como **falta de relación lineal** entre las variables y, por lo tanto, supone la inutilidad del modelo de regresión.
  - Si $\beta_1 = 0$, puede ser debido a que X e Y sean independientes.
  - Si $\beta_1 = 0$, puede ser debido, también, a que entre X e Y haya una relación NO lineal.

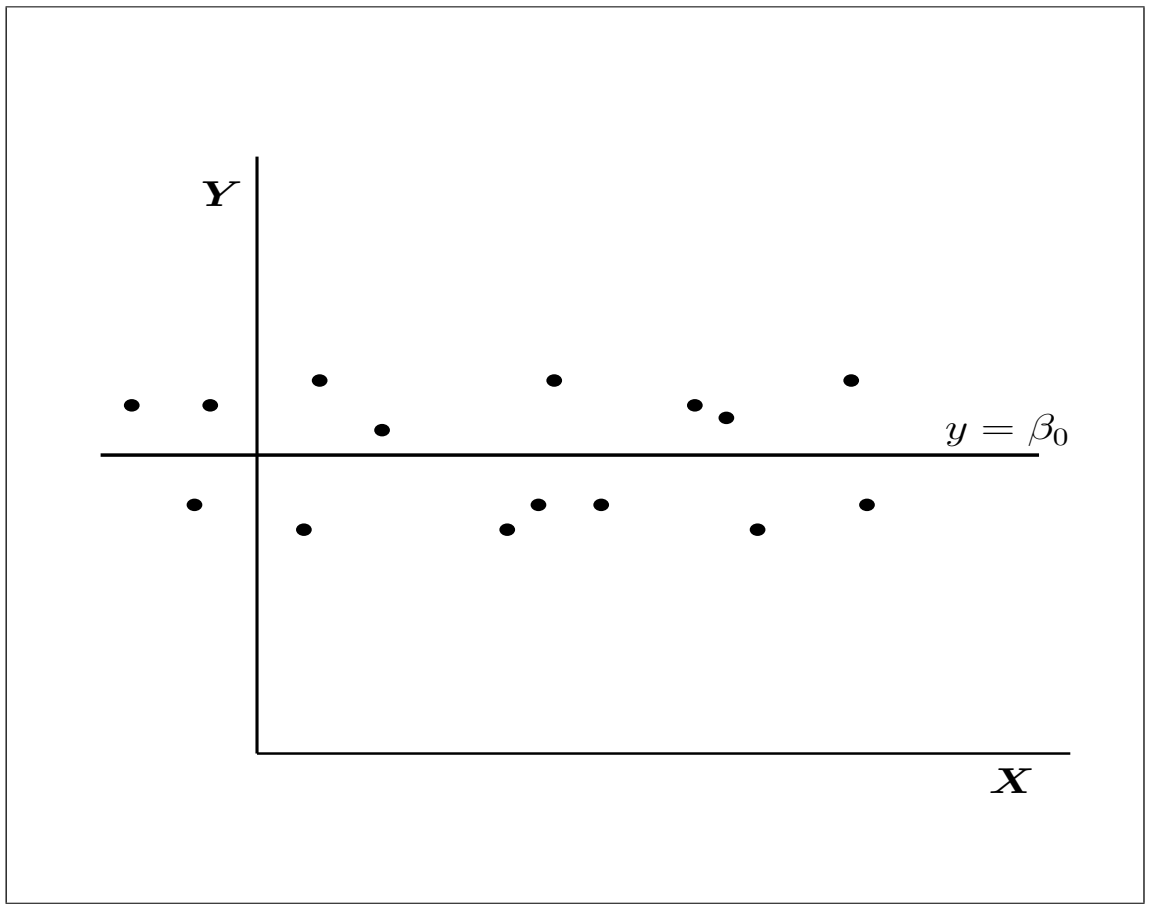
El contraste de regresión VII. Interpretación. Ejemplo

Los puntos del gráfico muestran cómo no existe relación de dependencia entre las variables X e Y.



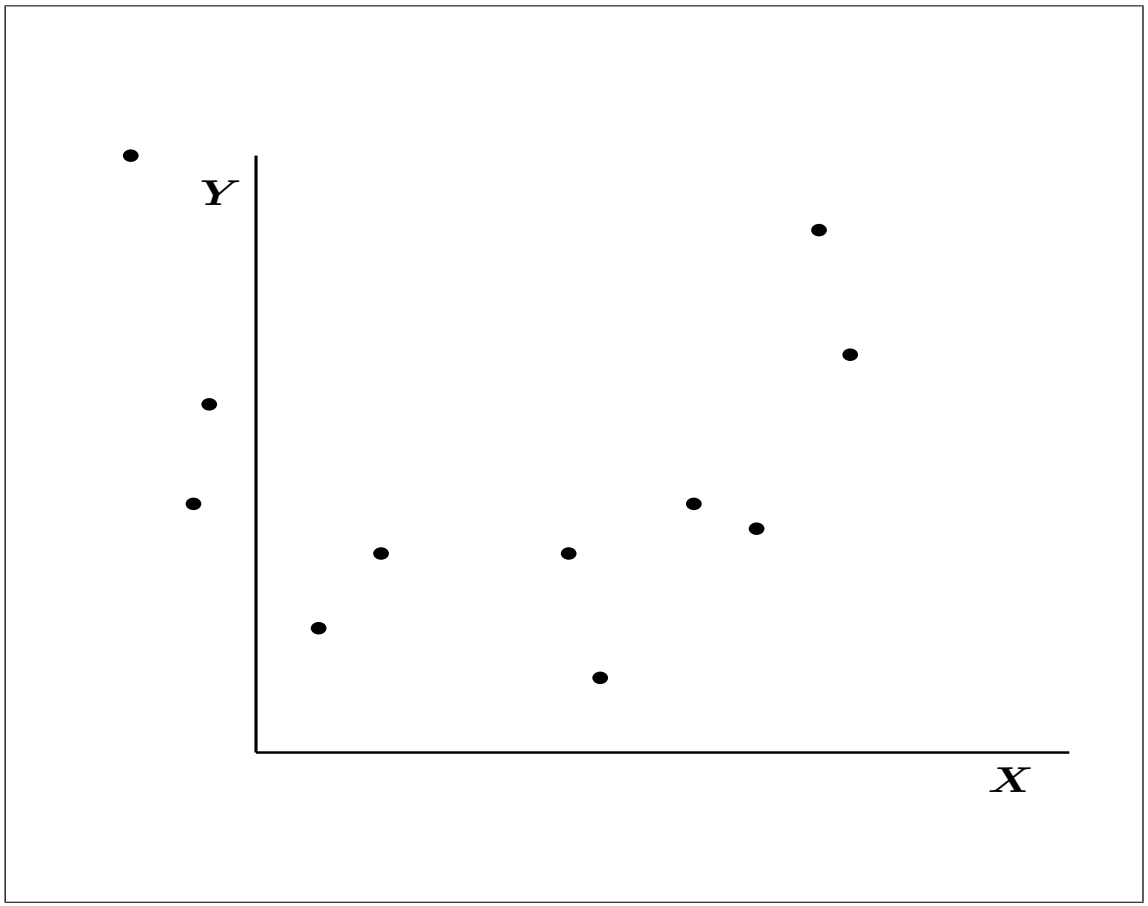
El contraste de regresión VIII. Interpretación. Ejemplo

En este caso se aceptaría la hipótesis nula, $\beta_1 = 0$. Gráficamente,



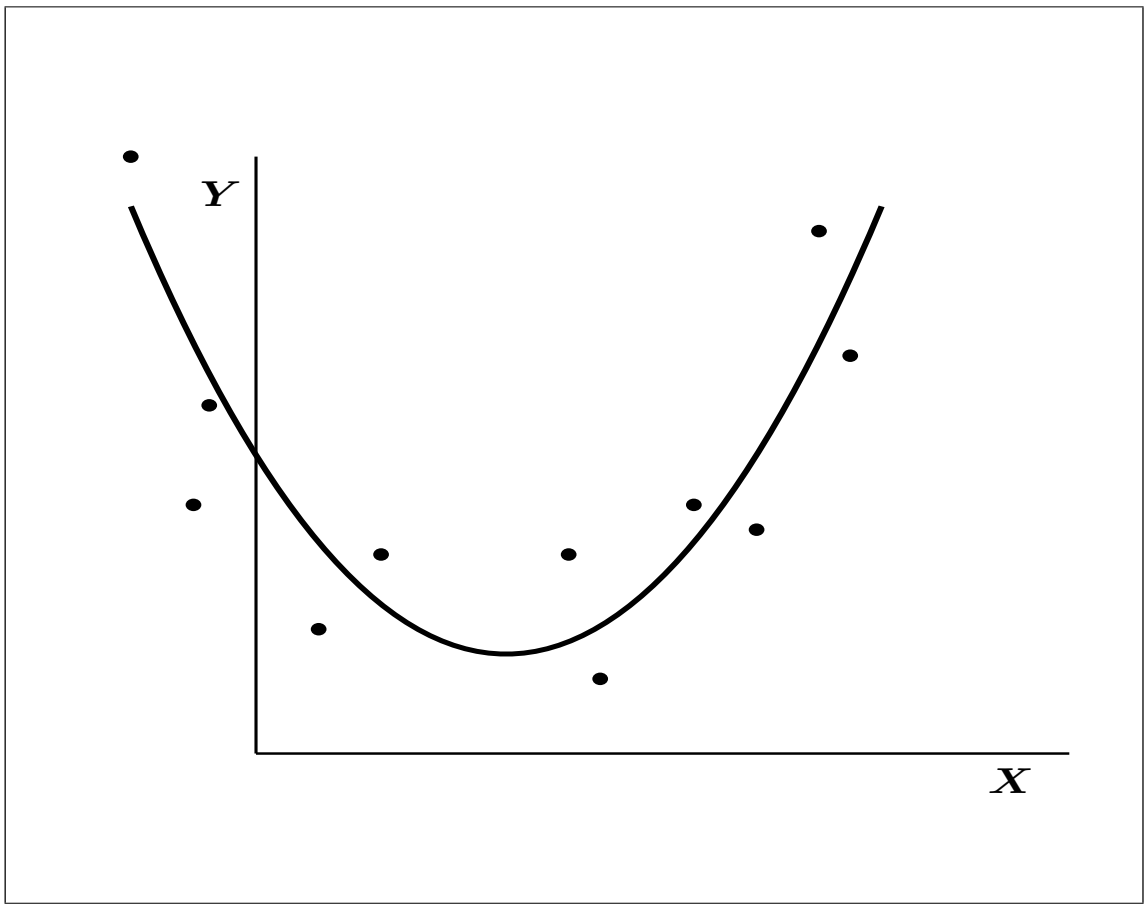
El contraste de regresión IX. Interpretación. Ejemplo

Los puntos del gráfico muestran cómo existe una relación de dependencia no lineal entre las variables X e Y.



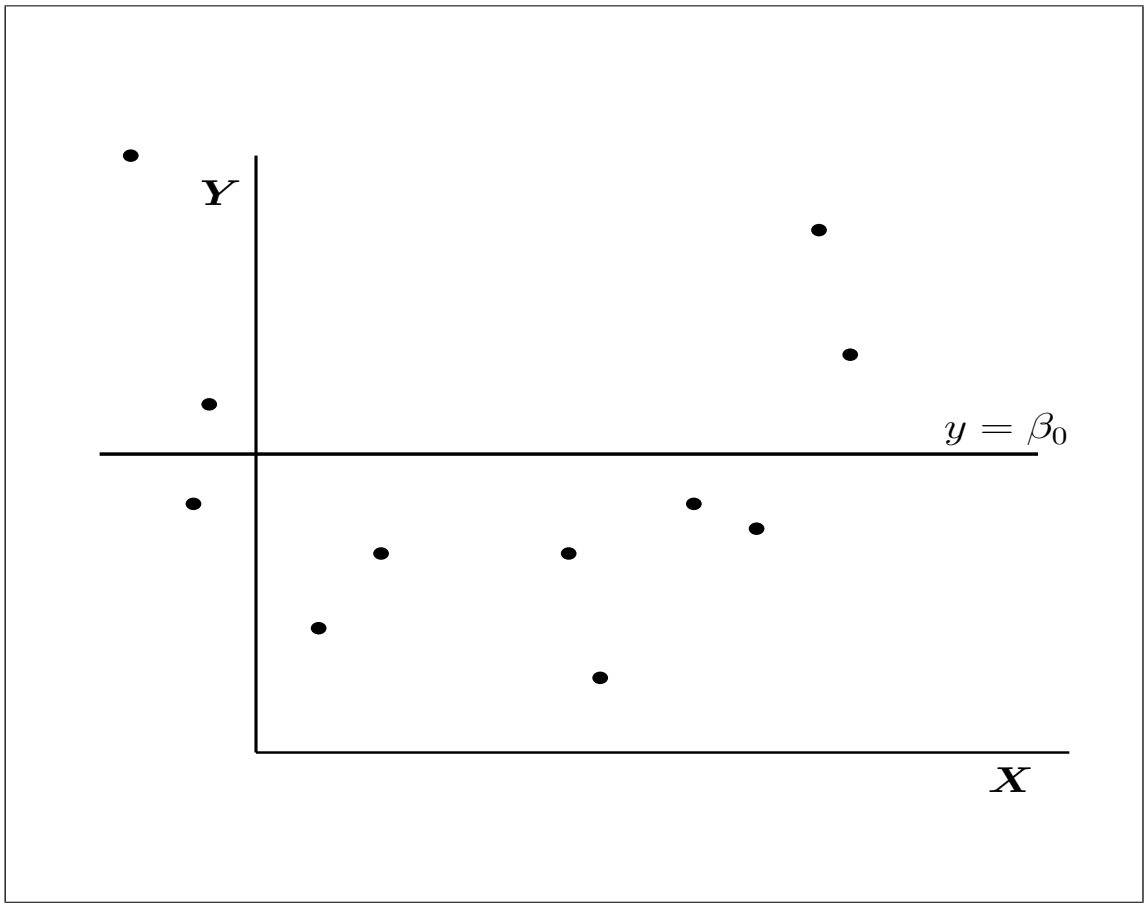
El contraste de regresión X. Interpretación. Ejemplo

Esta relación sería, posiblemente, descriptible por un modelo cuadrático. Gráficamente,



El contraste de regresión XI. Interpretación. Ejemplo

En este caso se aceptaría la hipótesis nula, $\beta_1 = 0$. Gráficamente,



El contraste de regresión XII. Interpretación

Observaciones:

- El rechazo del contraste de regresión, $\beta_1 = 0$, supone la aceptación de la hipótesis alternativa $\beta_1 \neq 0$, y se interpreta como **síntoma** de la existencia de relación lineal entre las variables X e Y, resumida por la recta de regresión.

- La aceptación de que $\beta_1 \neq 0$ no garantiza por sí sola la bondad del modelo de regresión.

Diagnosis y validación del modelo I

- Una vez calculado el modelo de regresión siguiendo los pasos anteriores, antes de emplearlo, es necesario verificar las hipótesis de linealidad y las de normalidad, homocedasticidad e independencia de los errores, impuestas anteriormente.
- Este proceso se conoce como la **validación** o **diagnosis** del modelo.
    - Observación: Debe tenerse en cuenta que para que un modelo de regresión pueda utilizarse, es imprescindible que supere el requisito de su validación.

Diagnosis y validación del modelo II

- La diagnosis del modelo se realiza a través de los gráficos de los **residuos**.
- Cada residuo, e_i, está definido por la diferencia:

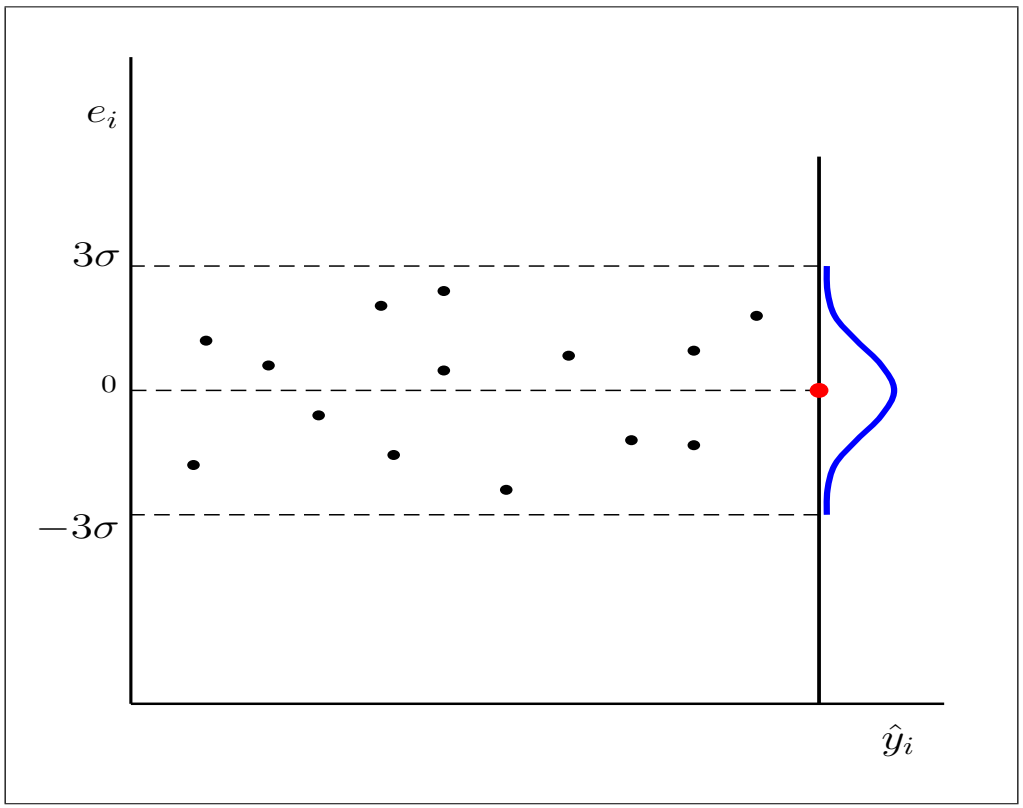
$$e_i = y_i - \hat{y}_i.$$

Diagnosis y validación del modelo III. Gráficos de residuos

- Las hipótesis de linealidad, homocedasticidad e independencia se contrastan a través del gráfico que enfrenta los valores de los residuos con los previstos para cada valor de x_i observado.
- La hipótesis de independencia se contrasta también a través del gráfico que enfrenta los valores de los residuos con el orden de la obtención de datos.

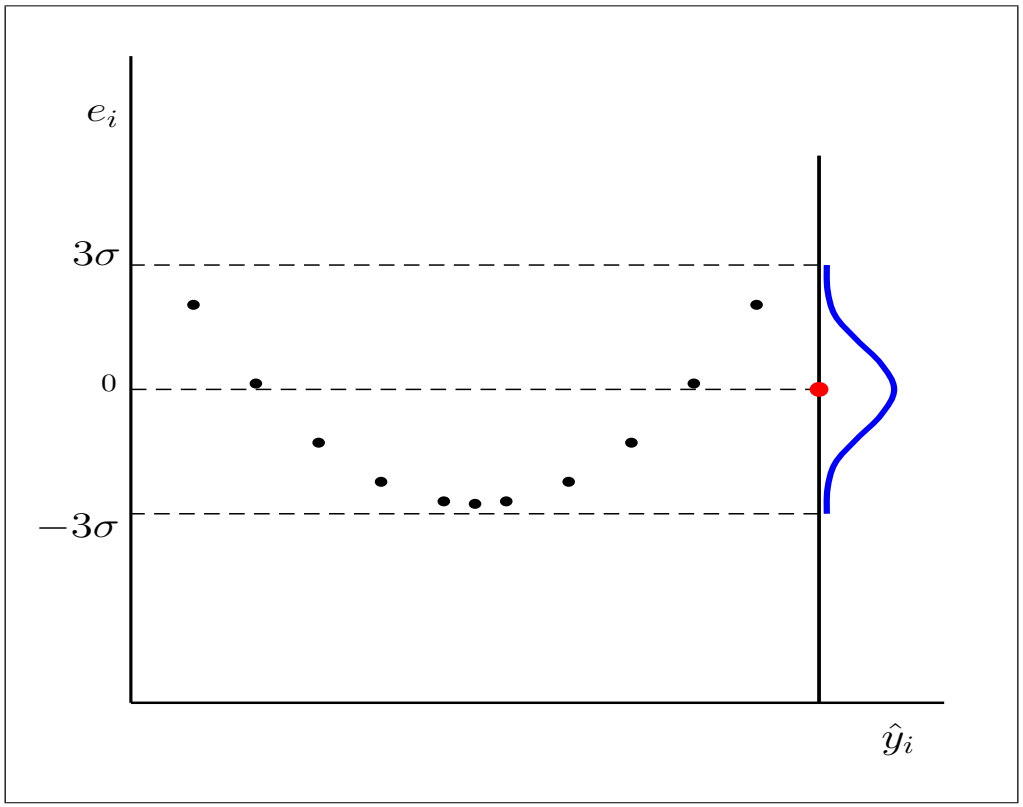
Diagnosis y validación del modelo IV. Gráficos de residuos

Al representarlos gráficamente, los residuos deberían formar una nube de puntos sin estructura, y con, aproximadamente, la misma variabilidad por todas las zonas del gráfico. Gráficamente,



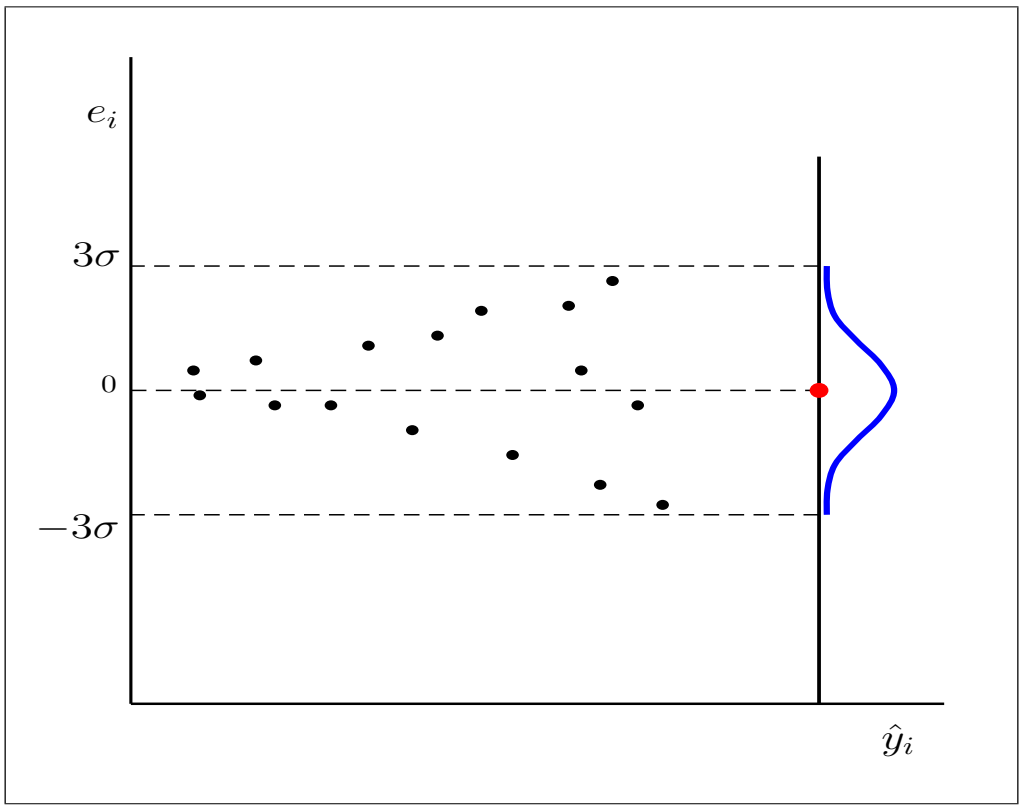
Diagnosis y validación del modelo V. Gráficos de residuos. Ejemplo

Los residuos de la figura muestran una estructura que sugiere una relación no lineal entre las variables:



Diagnosis y validación del modelo VI. Gráficos de residuos. Ejemplo

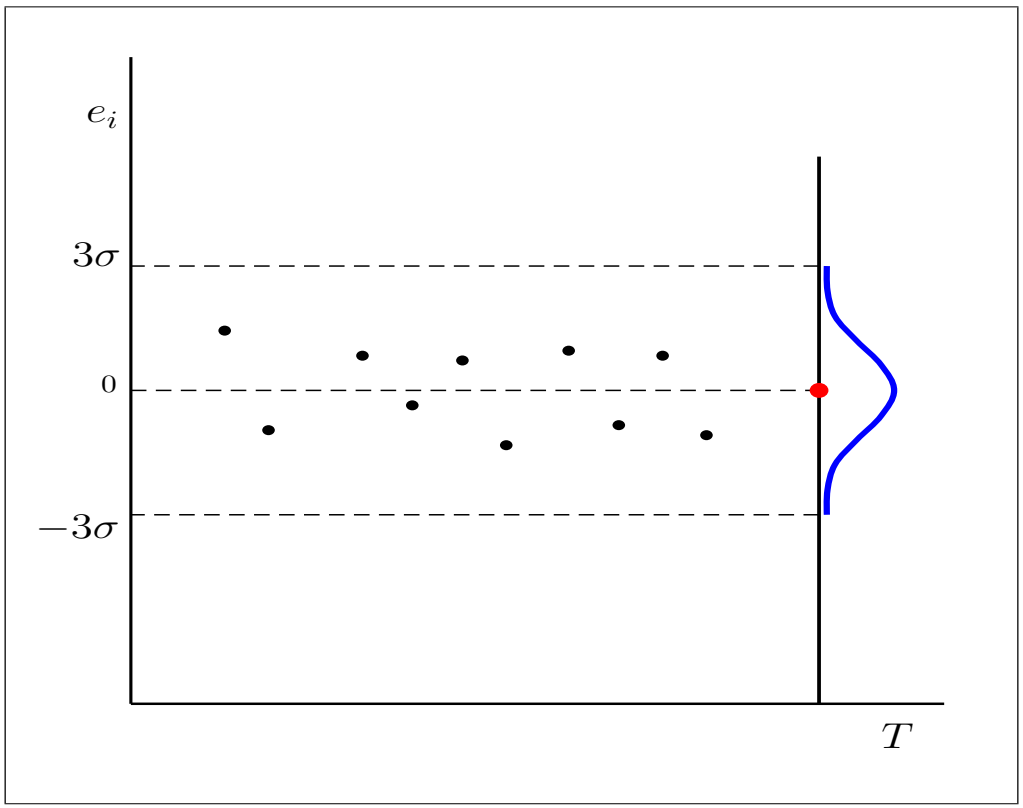
Los residuos de la figura sugieren la asusencia de homocedasticidad (heterocedasticidad).



Diagnosis y validación del modelo VII. Gráficos de residuos. Ejemplo

El gráfico de la figura contiene una representación temporal de los residuos.

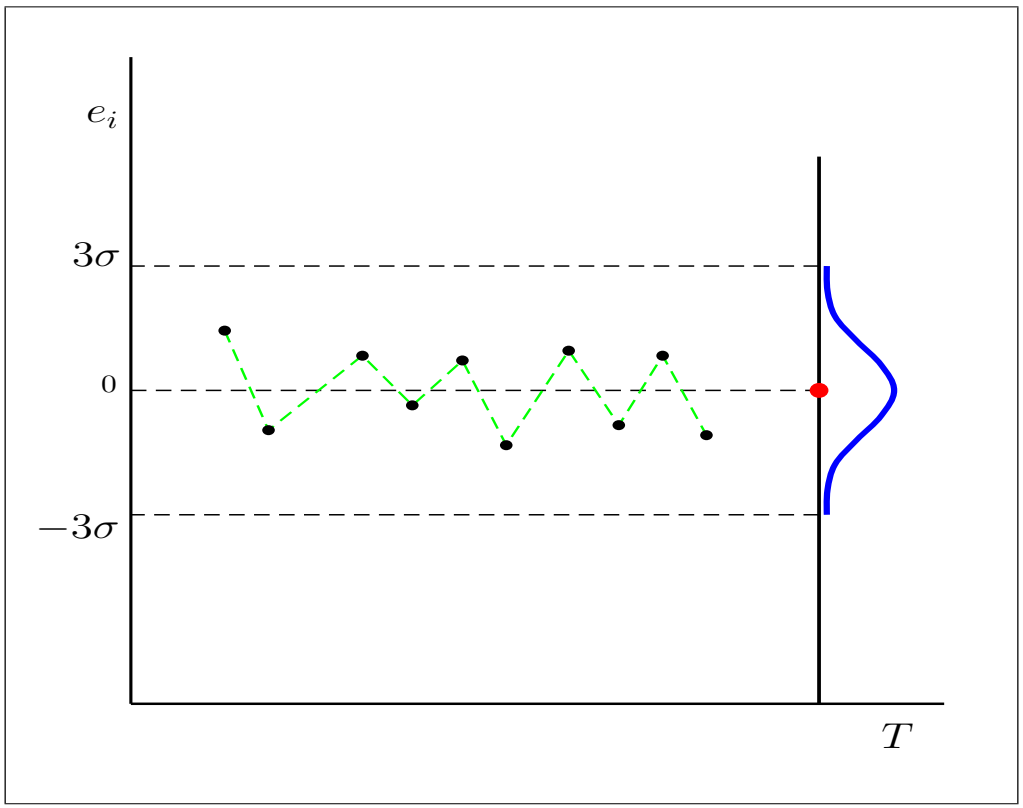
El eje de abscisas indica el orden de obtención de los datos, y la estructura del gráfico sugiere falta de independencia en los mismos:



Diagnosis y validación del modelo VIII. Gráficos de residuos. Ejemplo

La unión de los puntos por medio de una línea ayuda a detectar la falta de independencia en los residuos.

¿Sabría colocar aproximadamente el siguiente residuo en el gráfico?



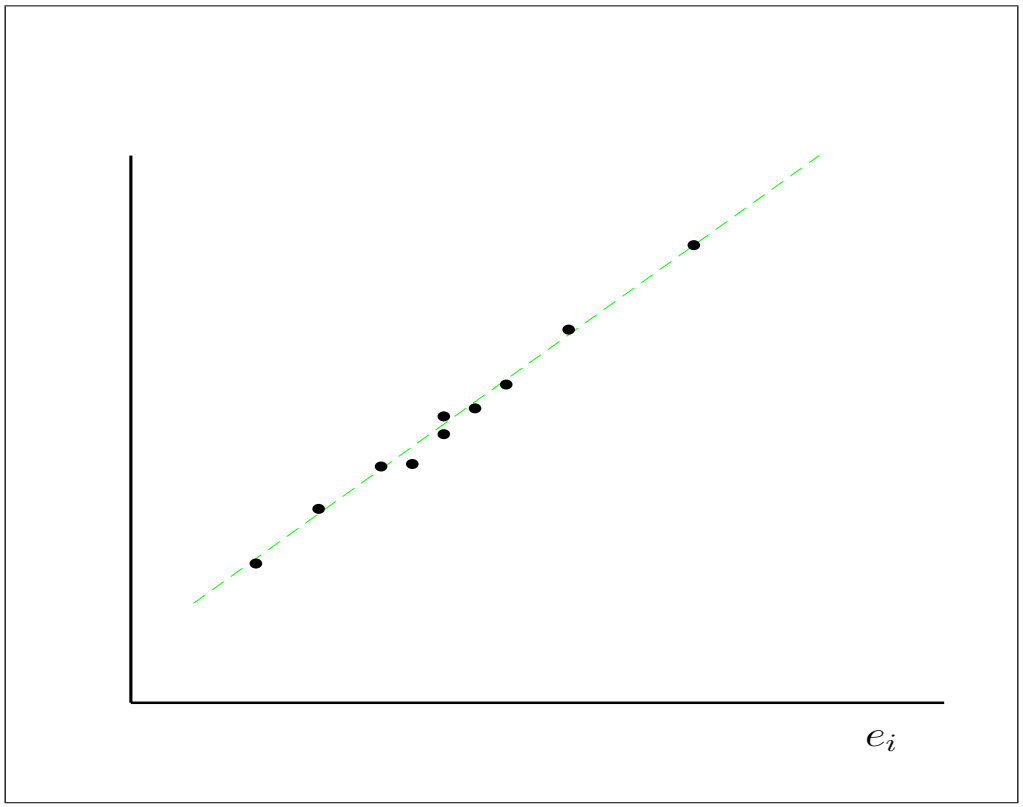
Diagnosis y validación del modelo IX. Gráficos de residuos. Ejemplo

- La representación de los residuos en papel probabilístico normal permite contrastar la hipótesis de normalidad. Esta hipótesis será aceptada cuando los residuos originen, aproximadamente, una línea recta.

- Observación: Esta hipótesis puede, en el caso en el que el número de datos sea grande, contrastarse por medio del test de la chi cuadrado, aunque los residuos no son independientes, ya que existen dos relaciones algebraicas que los relacionan, como se vió anteriormente.

Diagnosis y validación del modelo X. Gráficos de residuos. Ejemplo

El gráfico de la figura representa un conjunto de residuos sobre papel probabilístico normal, que hace razonable la aceptación de la hipótesis de normalidad.



Transformaciones I

- En el caso en el que el análisis de los residuos no permita validar el modelo, bien por
  - Falta de linealidad en la relación entre las variables X e Y.
  - Falta de homocedasticidad.
  - Falta normalidad.

En ocasiones se puede obtener un modelo lineal que sí verifique las hipótesis a través de transformaciones en X, en Y, o en ambas.

Transformaciones II. Algunos Modelos linealizables

Modelo real (desconocido)	Transformación	Modelo lineal
$y = \beta_0 + \beta_1 x^k$	$z = x^k$	$y = \beta_0 + \beta_1 z$
$y = \beta_0 + \beta_1 \ln(x)$	$z = \ln(x)$	$y = \beta_0 + \beta_1 z$
$y = \beta_0 e^{\beta_1 x}$	$v = \ln(y)$	$v = \ln(\beta_0) + \beta_1 x$
$y = K x^{\beta_1}$	$v = \ln(y)$	$v = \beta_0 + \beta_1 \ln x$

Transformaciones III. Interpretación de los parámetros de regresión

Observaciones

- Cuando se realiza una transformación, la interpretación de los parámetros del modelo estimado se modifica.
- Pueden encontrarse las interpretaciones de los parámetros del modelo, cuando se realizan algunas transformaciones de interés, por ejemplo las logarítmicas, en Peña (2002).

Predicción en regresión simple

Una vez calculada la recta de regresión, y validado el modelo, se puede emplear dicha recta para hacer predicciones.

1. Se puede emplear $\hat{y}(x_i)$ para predecir el valor de $E(Y|X = x_i)$, la media de la variable $(Y|X = x_i)$.
2. También se puede emplear $\hat{y}(x_i)$ para predecir el valor de un individuo de la variable $(Y|X = x_i)$.

    - Obsérvese que los dos valores se estiman por el mismo número.

Predicción en regresión simple. Precisión de la estimación de $E(Y|X = x_i)$ I

- Se puede demostrar que si $\mu_{X_i} = E(Y|X = x_i)$, se cumple que:

$$\frac{\hat{y}(x_i) - \mu_{X_i}}{DT(\hat{y}(x_i))} \longrightarrow t_{n-2},$$

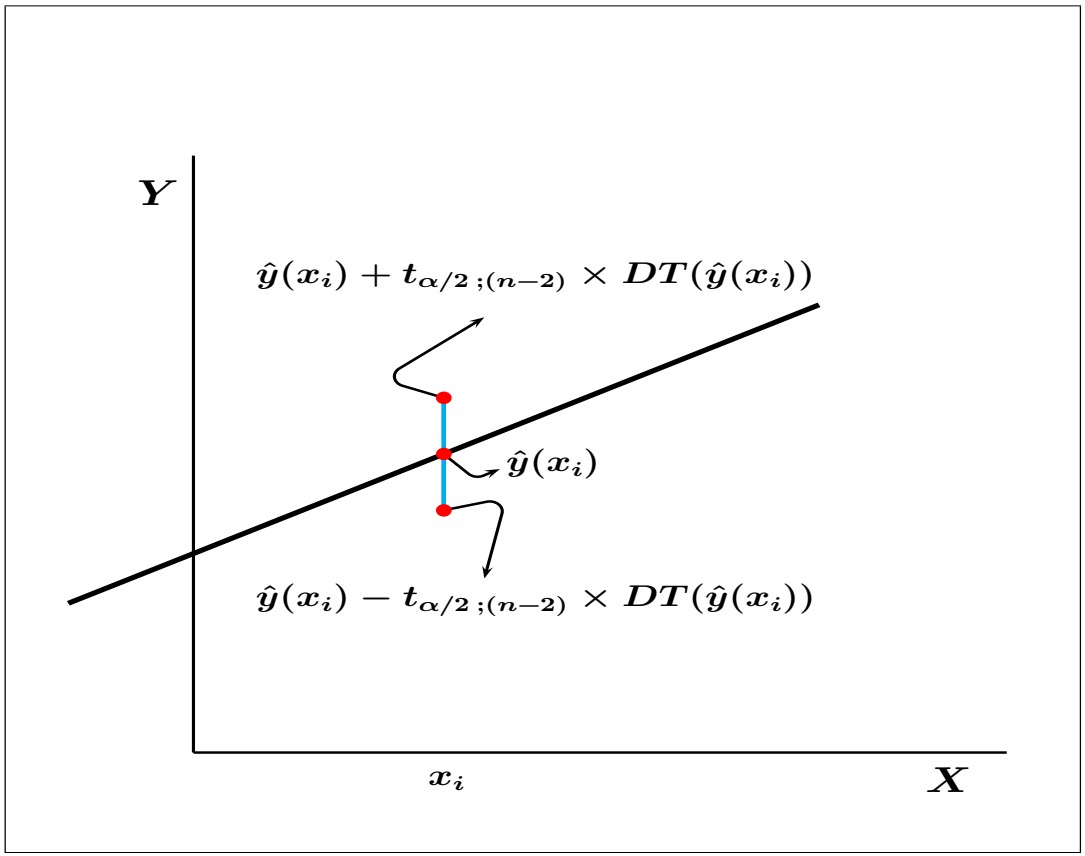
lo que permite calcular un intervalo de confianza para μ_{X_i}, siendo $DT(\hat{y}(x_i))$ la desviación típica de $\hat{y}(x_i)$

- Con el $(1 - \alpha) \times 100\,\%$ de confianza,

$$\mu_{X_i} \in (\hat{y}(x_i) \pm t_{\alpha/2\,;(n-2)} \times DT(\hat{y}(x_i)))$$

Predicción en regresión simple. Precisión de la estimación de $E(Y|X = x_i)$ II

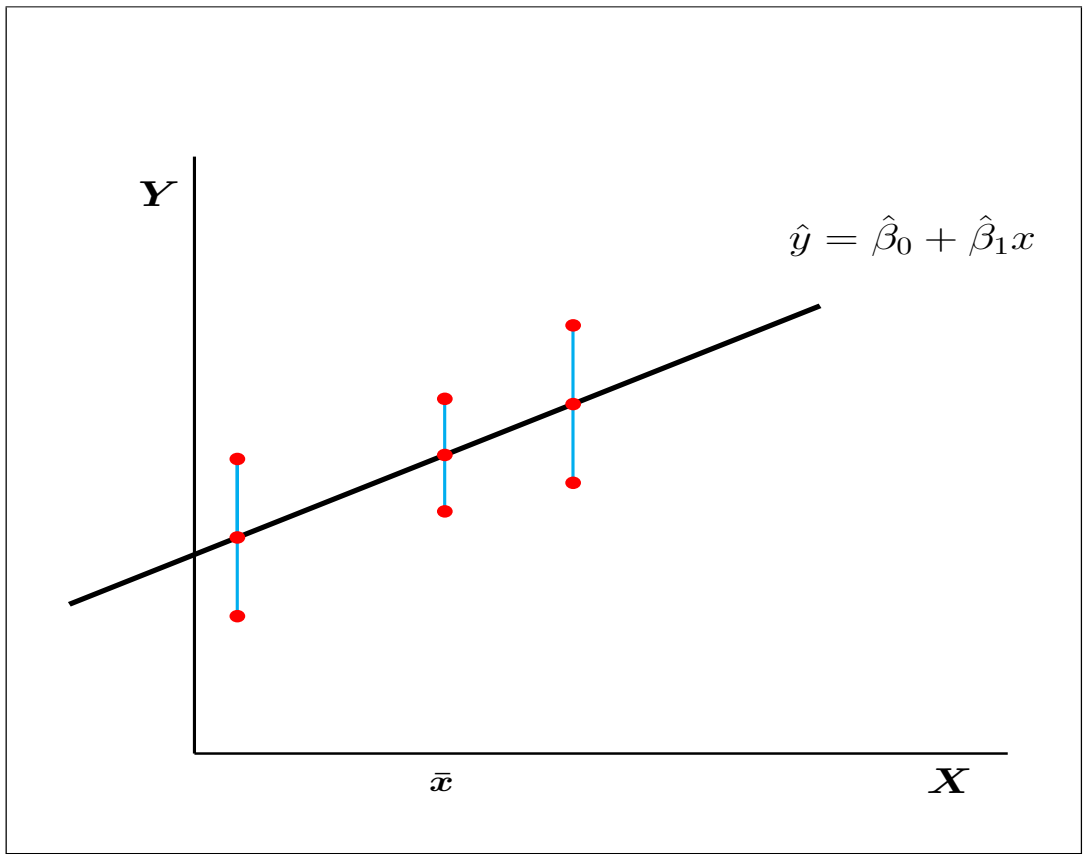
Gráficamente:



Predicción en regresión simple. Precisión de la estimación de $E(Y|X = x_i)$ III

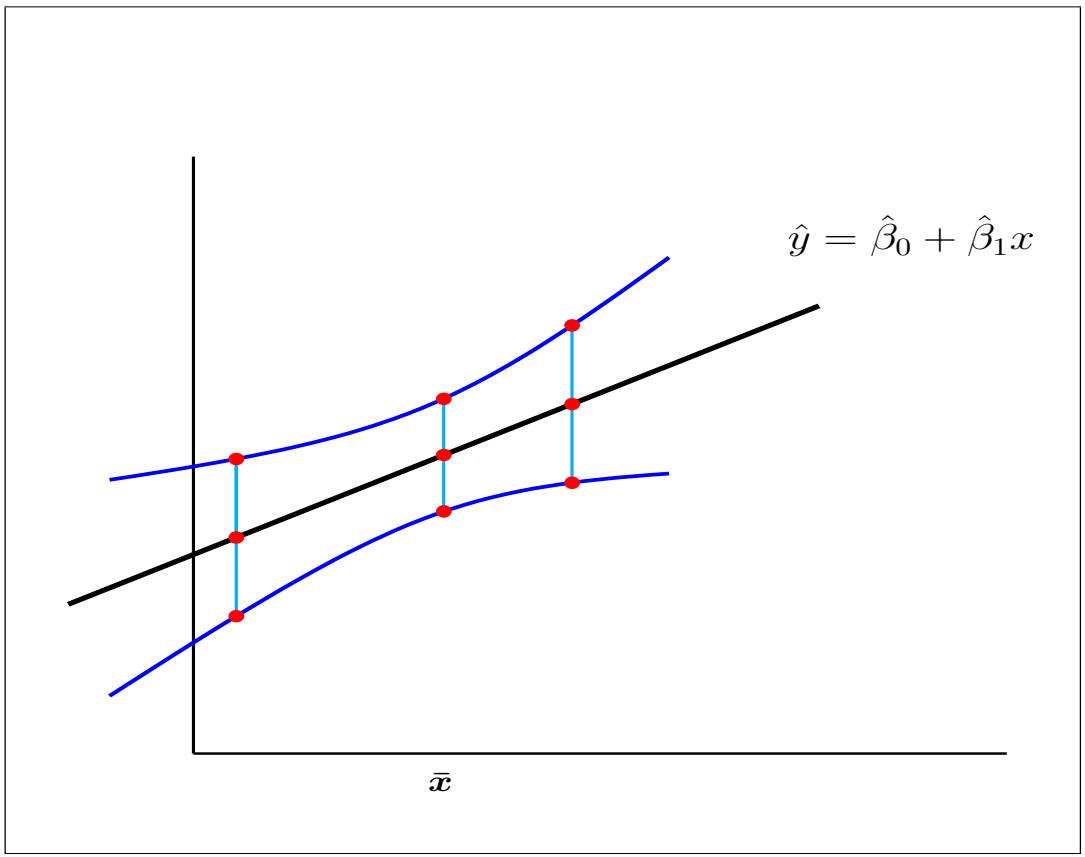
Observaciones:

- El valor exacto de $DT(\hat{y}(x_i))$ puede consultarse en Peña (2002).
- Se puede comprobar que $DT(\hat{y}(x_i))$ aumenta cuando (x_i) se aleja de $\bar{x}$.



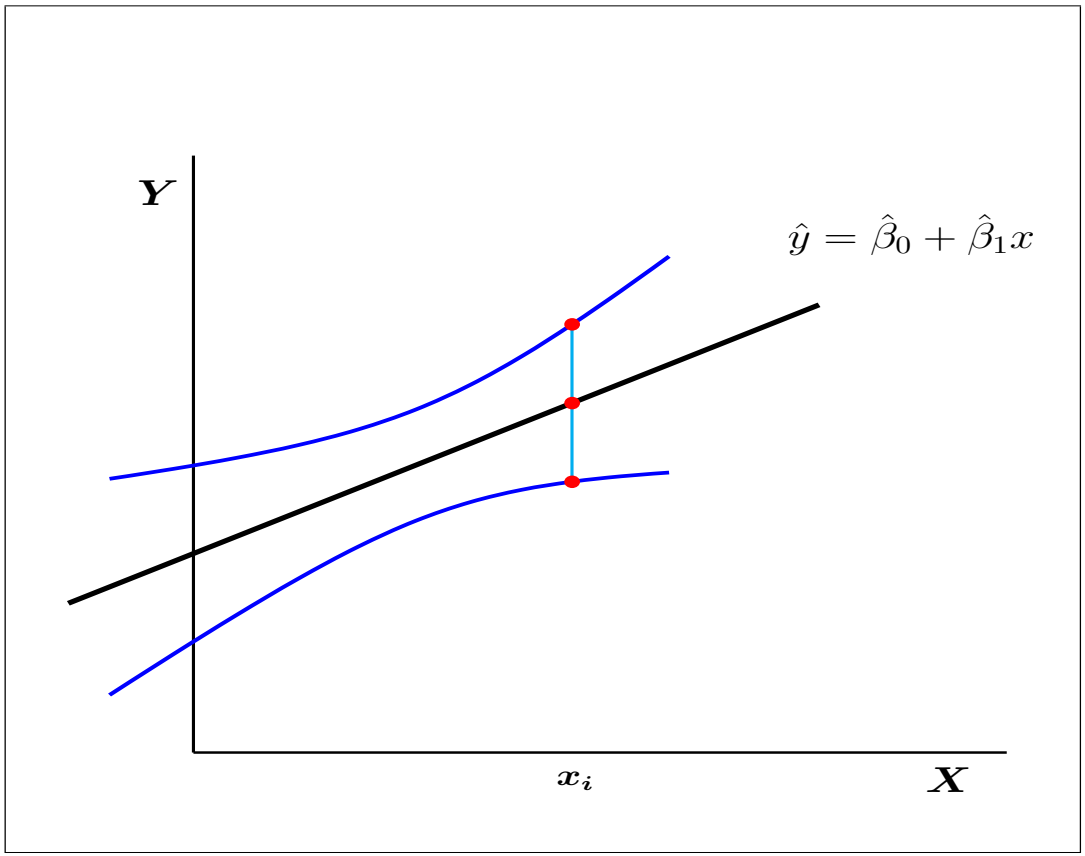
Predicción en regresión simple. Precisión de la estimación de $E(Y|X = x_i)$ IV

Uniendo los extremos de todos los intervalos de confianza de μ_x, para todo x, se observa cómo la precisión de la estimación disminuye cuando x se aleja de $\bar{x}$, originándose la hipérbola que se representa en el gráfico.



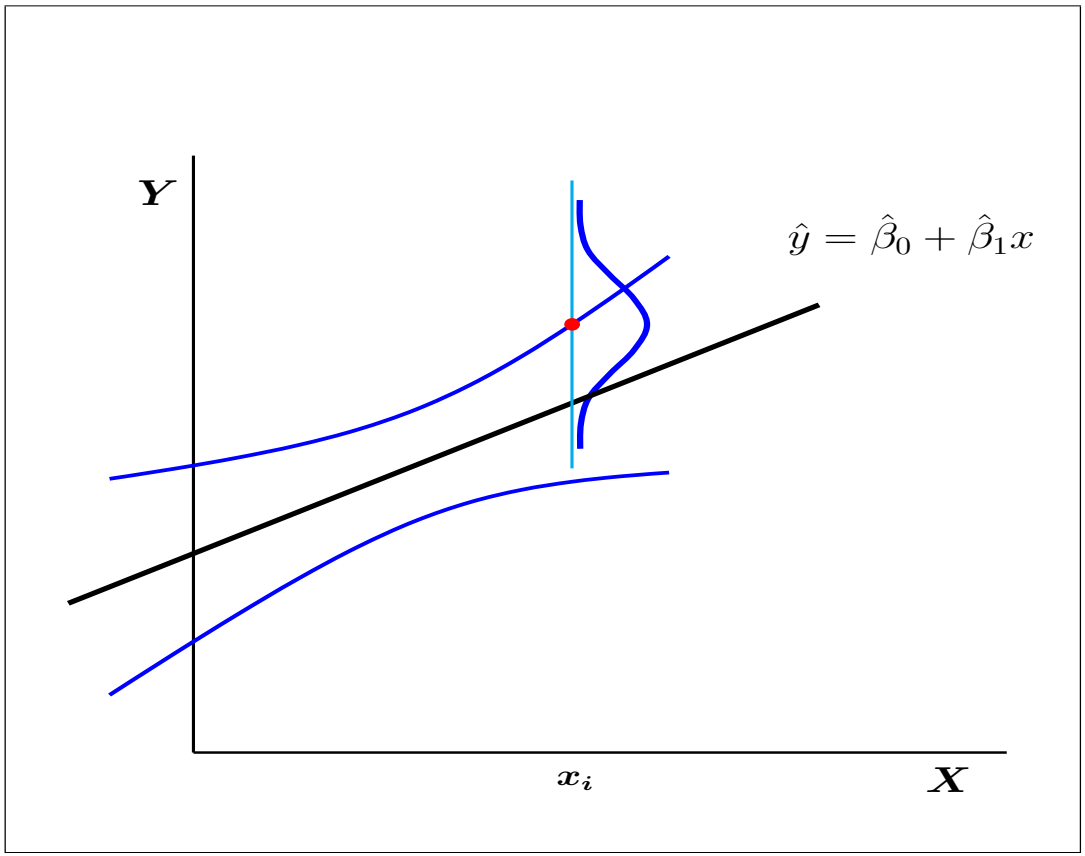
Predicción en regresión simple. Precisión de la estimación de una observación. I

Si se utiliza $\hat{y}(x_i)$ para predecir el valor de un individuo de la población $Y|X = x_i$, teniendo en cuenta el intervalo de confianza para μ_{X_i} calculado anteriormente, cuya representación gráfica es



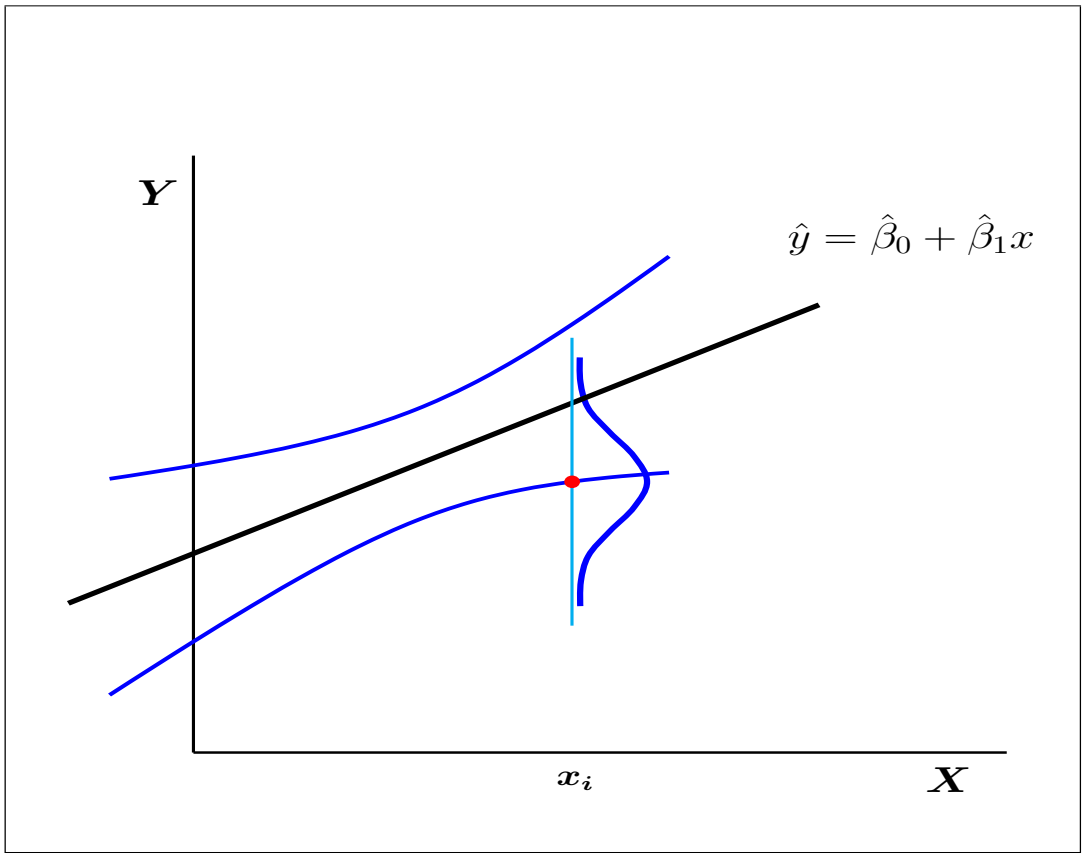
Predicción en regresión simple. Precisión de la estimación de una observación. II

La distribución de $(Y|X = x_i)$, para los posibles valores extremos de μ_{X_i}, sería, gráficamente:



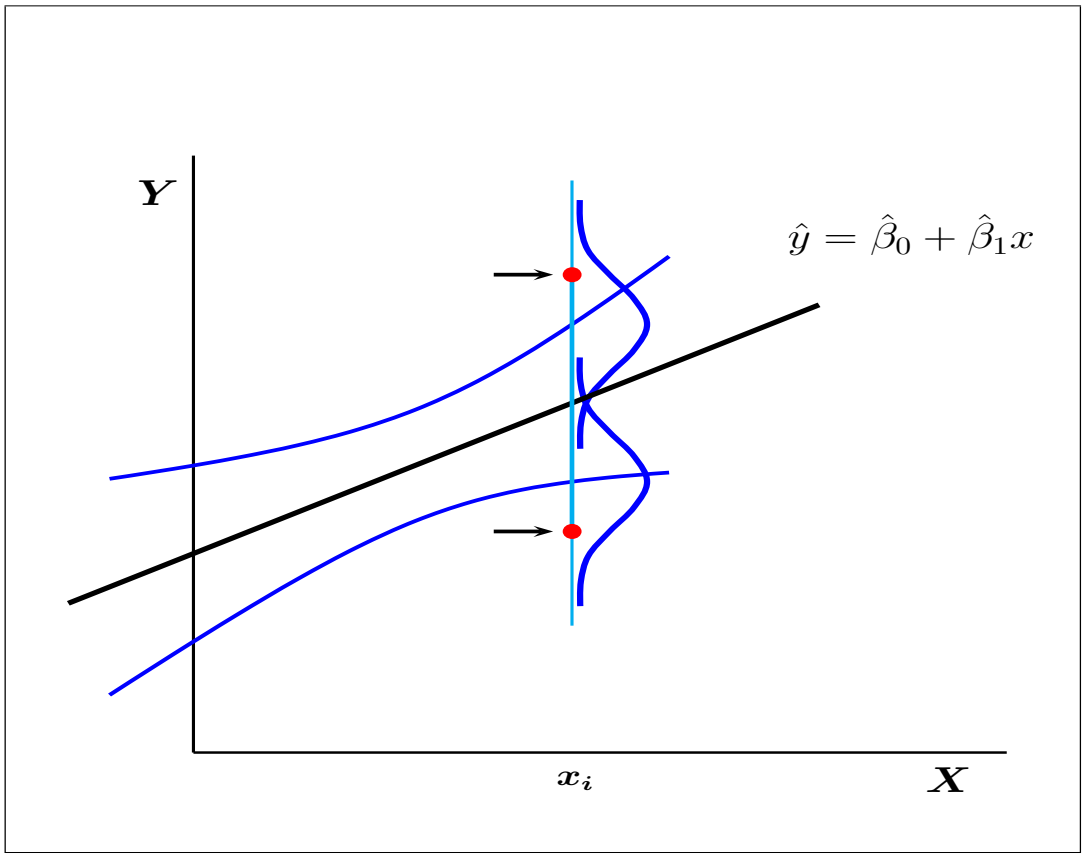
Predicción en regresión simple. Precisión de la estimación de una observación. III

O bien:



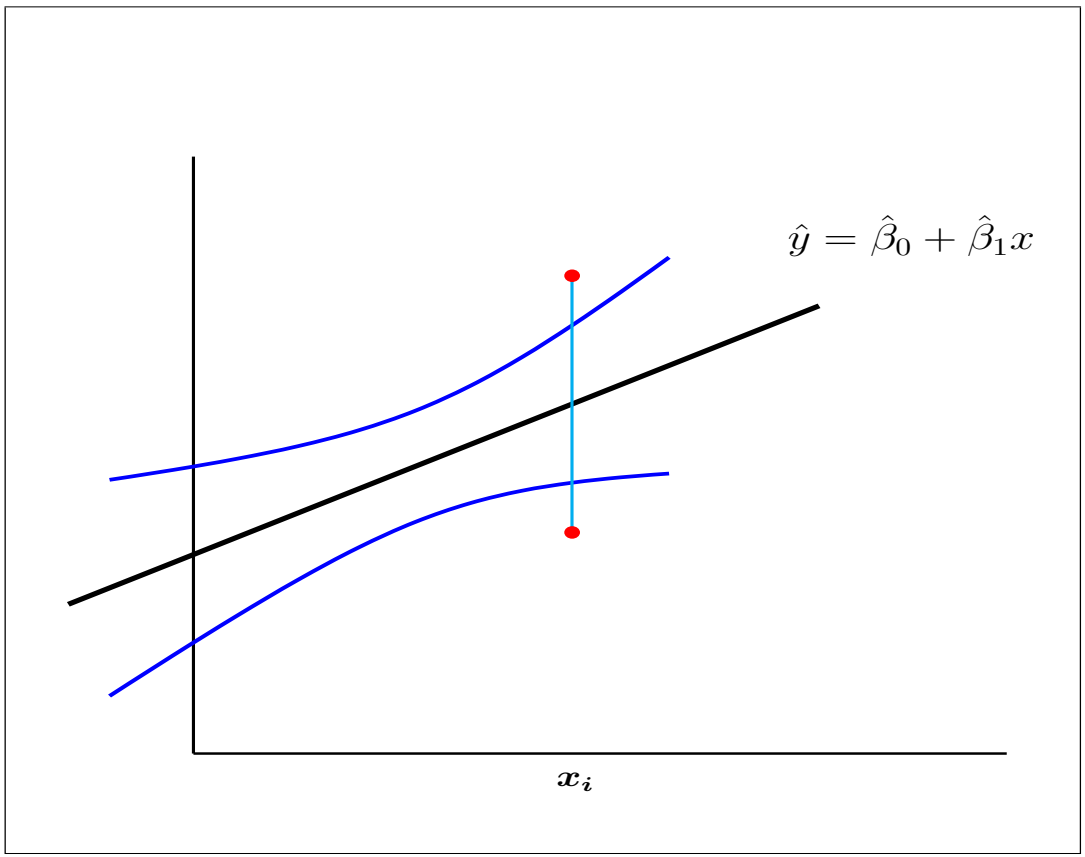
Predicción en regresión simple. Precisión de la estimación de una observación. IV

Por lo tanto, cabría esperar que los individuos de la variable $(Y|X = x_i)$ se encuentren en el intervalo:



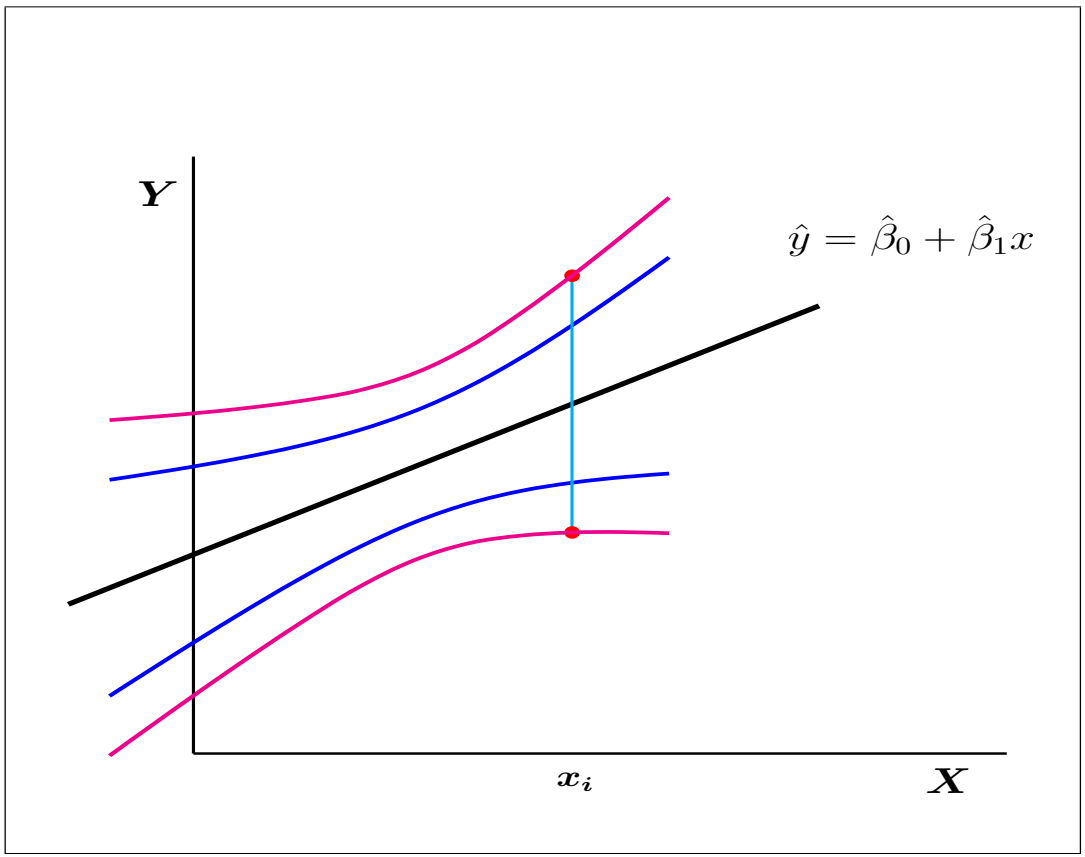
Predicción en regresión simple. Precisión de la estimación de una observación. V

Con el nivel de confianza deseado, una observación de la variable $(Y|X = x_i)$ se encontraría en el intervalo:



Predicción en regresión simple. Precisión de la estimación de una observación. VI

Uniendo los extremos de los intervalos de confianza para una observación de $(Y|X = x)$, para todo x, se observa cómo la precisión de la estimación disminuye cuando x se aleja de $\bar{x}$, originándose la hipérbola que se representa en el gráfico.



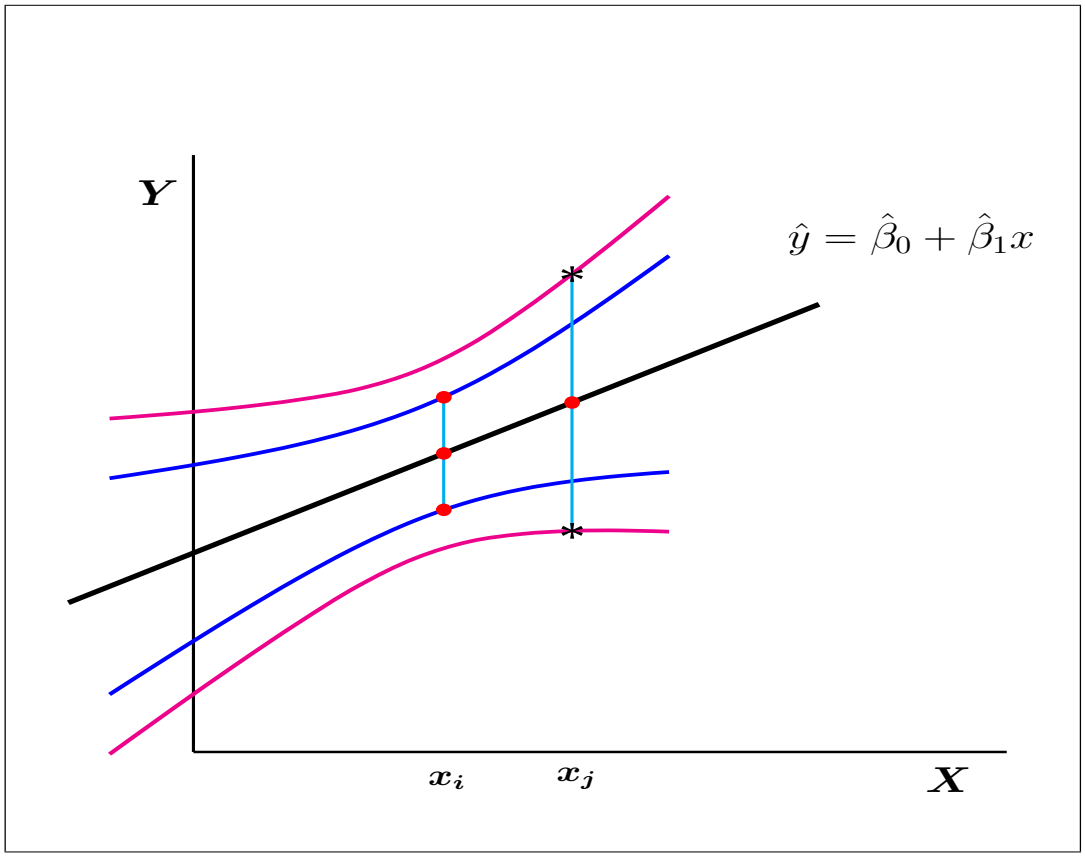
Predicción en regresión simple. Resumen I

- El valor de $\hat{y}(x)$ se puede emplear para estimar tanto μ_x, como una observación de $(Y|X = x)$.
- La precisión de la estimación disminuye al aumentar la distancia de x a $\bar{x}$.
- La precisión de la estimación de μ_x es mayor que la de una observación de $(Y|X = x)$.

Predicción en regresión simple. Resumen II

Gráficamente, la hipérbola interior ofrece intervalos de confianza para el valor de μ_X.

Y la exterior para el valor de un individuo de $(Y|X = x)$.

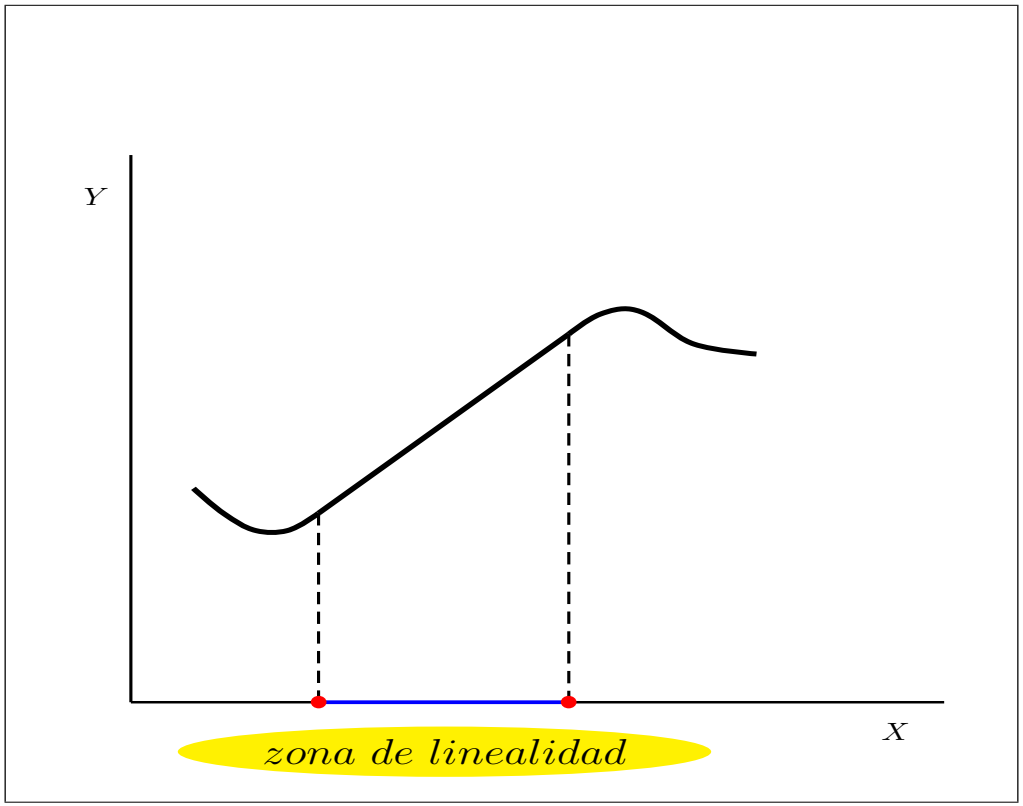


Predicción en regresión simple. Observaciones

- Es importante no emplear la recta para hacer previsiones fuera del rango muestral.
- Fuera de este rango no hay garantía de que la recta de regresión describa correctamente la relación entre las variables.

Predicción en regresión simple. Observaciones

- Puede observarse, como ejemplo, el siguiente gráfico.
- La recta de regresión sólo es útil en la zona de linealidad.
- Esta zona, en general, se descubre experimentalmente.

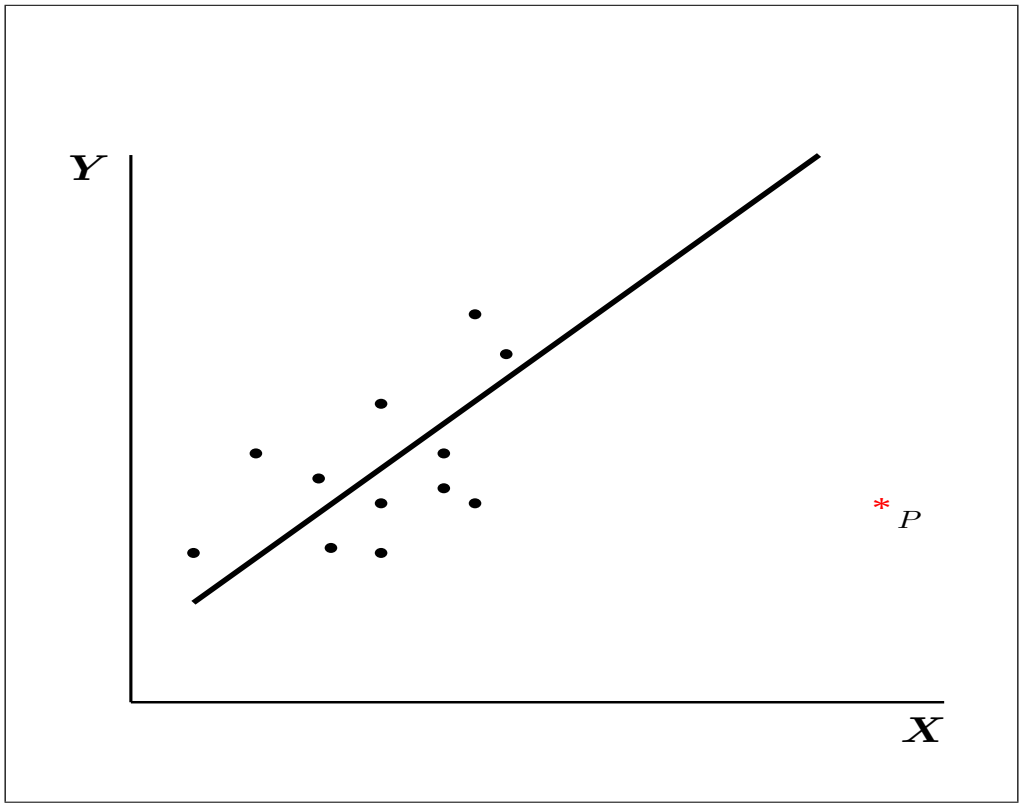


Los valores atípicos en regresión I.

- Un punto atípico, en regresión, es un punto muy separado del resto.
- Un punto atípico es **influyente** si modifica sustancialmente la ecuación de la recta de regresión.
- Los puntos atípicos en la variable X, **puntos palanca**, son los que poseen mayor potencialidad de influencia.
- Los puntos atípicos en Y pueden no afectar a la pendiente de la recta.

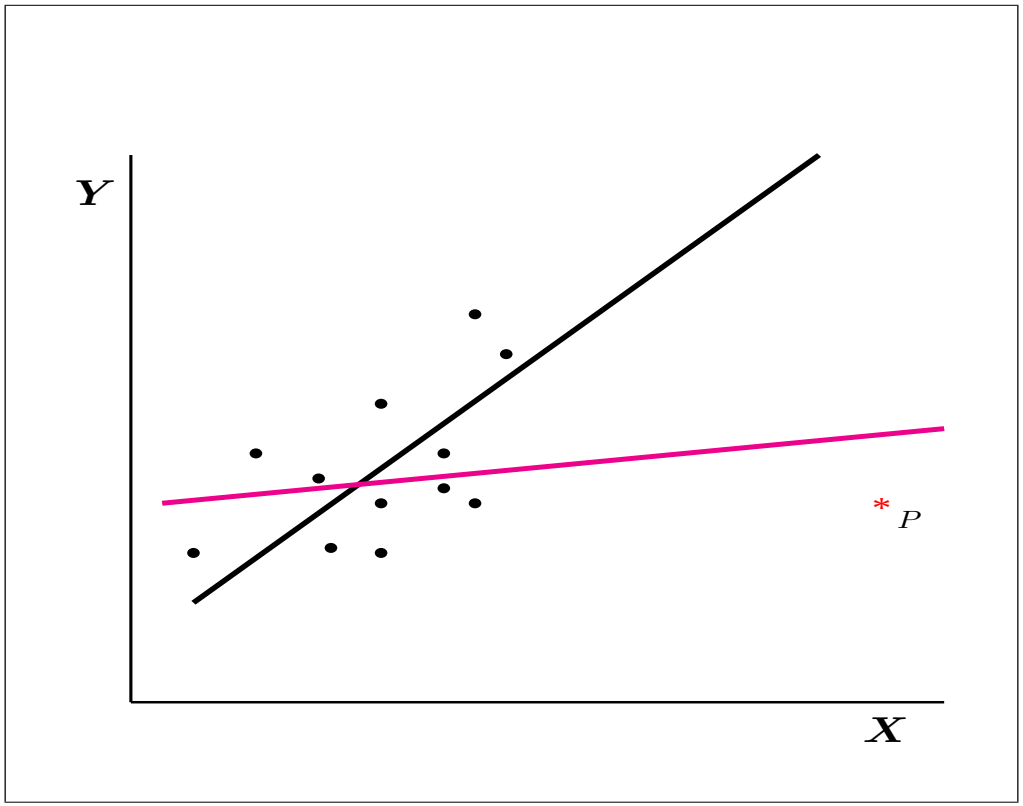
Los valores atípicos en regresión II. Ejemplo

El gráfico de la figura representa la recta de regresión calculada sin considerar el punto P.



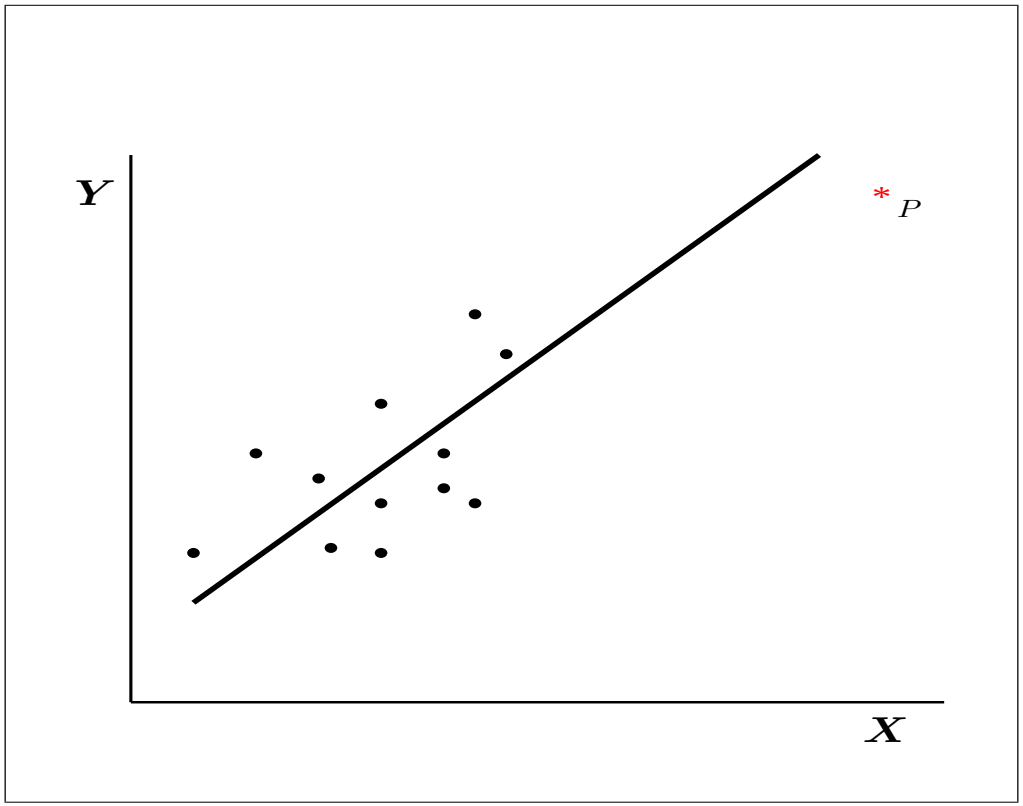
Los valores atípicos en regresión III. Ejemplo

El punto P es influyente, puesto que su inclusión modifica sustancialmente la recta de regresión.



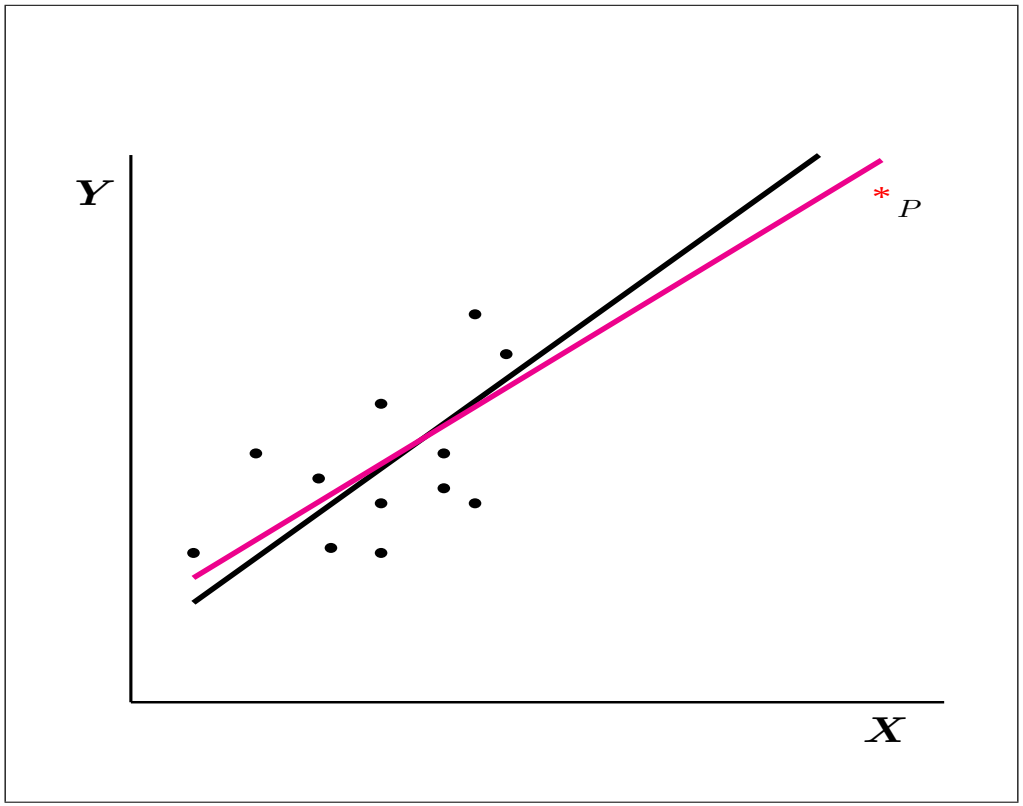
Los valores atípicos en regresión IV. Ejemplo

El gráfico de la figura representa la recta de regresión calculada sin considerar el punto P.



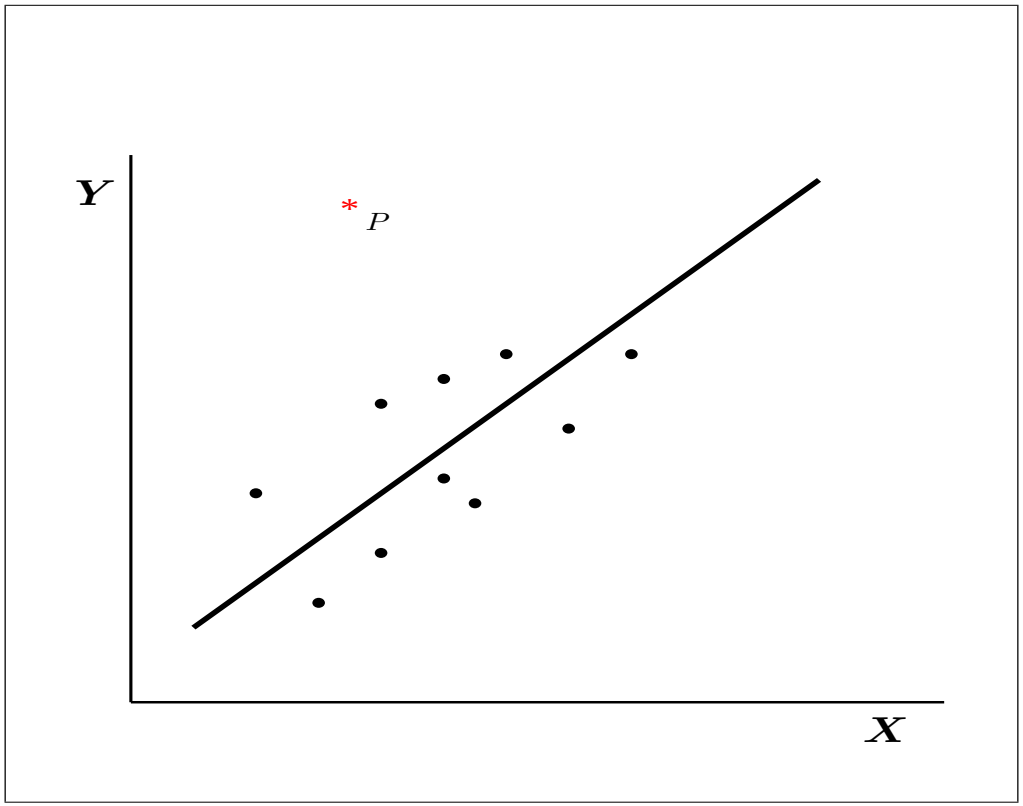
Los valores atípicos en regresión V. Ejemplo

El punto P NO es influyente, puesto que su inclusión NO modifica sustancialmente la recta de regresión.



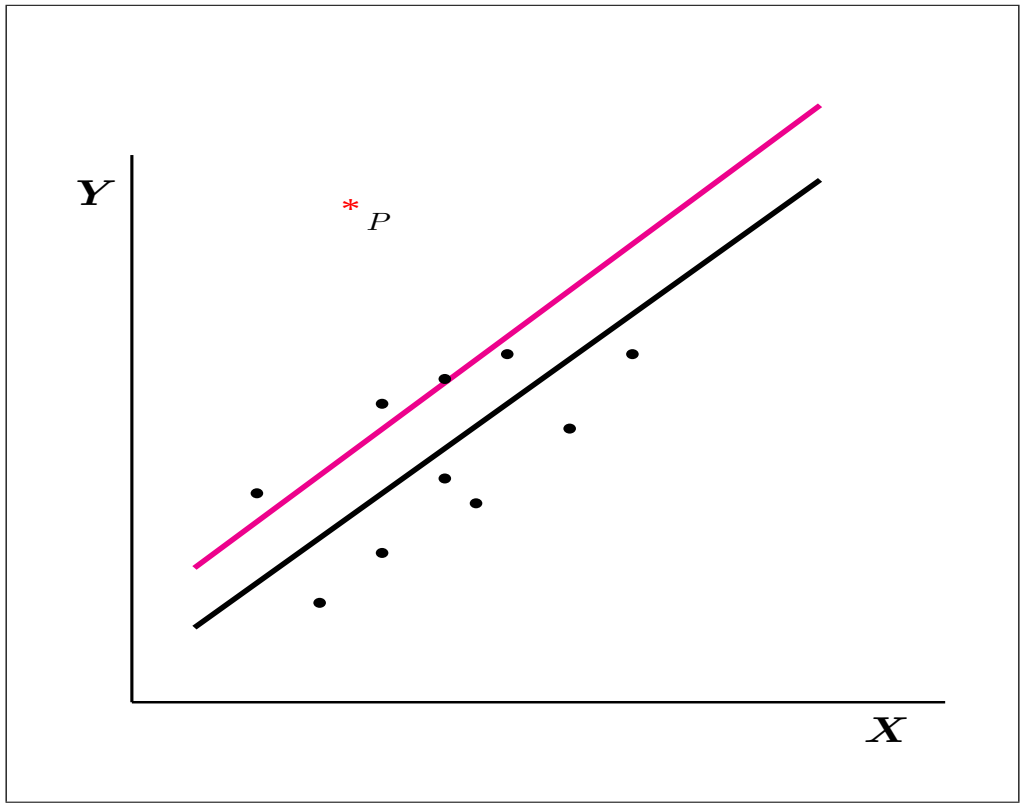
Los valores atípicos en regresión VI. Ejemplo

El gráfico de la figura representa la recta de regresión calculada sin considerar el punto P.



Los valores atípicos en regresión VII. Ejemplo

La inclusión del punto P no supone variación significativa en la pendiente de la recta de regresión estimada.



Estrategia ante los valores atípicos.

Si en un análisis se observan valores atípicos, una estrategia recomendable es la siguiente:

1. Descartar que se trata de un error.
2. Analizar si el punto es influyente.
3. Si el punto es influyente, calcular las rectas de regresión incluyéndole y excluyéndole, eligiendo la que mejor se adapte al conocimiento del problema y a las observaciones futuras.
    - Observación: En caso de duda, se debe utilizar el modelo con precaución. No se debe descartar, en ningún caso, recabar más información.