Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

Abril 2021

10.- Regresión lineal simple. Efectos fijos (I)

Introducción

La regresión lineal simple, múltiple, el análisis de varianza (ANOVA), el de covarianza (ANCOVA) etc. surgen del Modelo Lineal General (GLM) y por reducción de los modelos aditivos generales (GAM).

Modelo Lineal General:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

La regresión se podría definir como: el estudio de la **relación** que existe entre las variables explicativas y las variables respuestas. En el caso de la regresión lineal, se basa en un modelo lineal y si además es simple, sólo existirá una variable explicativa y una variable respuesta, mal llamadas independiente y dependiente respectivamente.

La regresión (relación) y el ANOVA (diferenciación) son caras de la misma moneda. Son las técnicas de análisis mayoritariamente usadas en estadística, sobre todo la regresión y la correlación.

En la Población

En la regresión lineal, las diferentes posibilidades de y son y_i, para un x_i hay diferentes y_i

Las medias de los diferentes y_i para cada x están sobre una recta. $Y = media\ de\ las\ y\ para\ cada\ x$, definida por la ecuación:

$$\boxed{Y = A + BX}$$

Para cada valor de x la media de todos los y_i están sobre una recta, con dos suposiciones principales:

1. Para cada x_i las y_i están distribuidas normalmente

$$Y \cap N(\mu, \sigma^2)$$

$$\mu = media\ de\ las\ y\ para\ cada\ x \Rightarrow \mu = Y$$

$$\boxed{Y \cap N(A + BX,\ \sigma^2)} \tag{1}$$

2. Las varianzas de todos los y_i son constantes

$$\boxed{\sigma^2 = cte} \tag{2}$$

Pero esto mismo es muy importante verlo bajo el "prisma" de los residuos. Otra aproximación al modelo podría ser, que para cada x_i, valor fijo de X, se cumple la ecuación:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + e_i}$$

donde β_0 y β_1 son *constantes* desconocidas.

Las hipótesis básicas del modelo son:

1. **Independencia en los residuos**. Sin correlación (auto-correlación) de los residuos

$$Cor(e_i, e_j) = 0$$

   Cualquier par de errores e_i y e_j son independientes.

2. **Esperanza nula** de los residuos es:

$$E(e_i) = 0$$

3. **Varianza constante** de los residuos (homocedasticidad):

$$Var(e_i) = \sigma^2$$

4. **Normalidad** de los residuos:

$$e_i \sim N(0, \sigma^2)$$

De lo que se deduce:

- Cada valor y_i de la variable aleatoria Y tiene una distribución

$$\boxed{(Y \mid X = x_i) \approx N(\beta_0 + \beta_1 x_i,\ \sigma^2)}$$

- Las observaciones y_i de la variable Y son independientes.

Si las hipótesis del modelo son ciertas, gráficamente se tiene que:

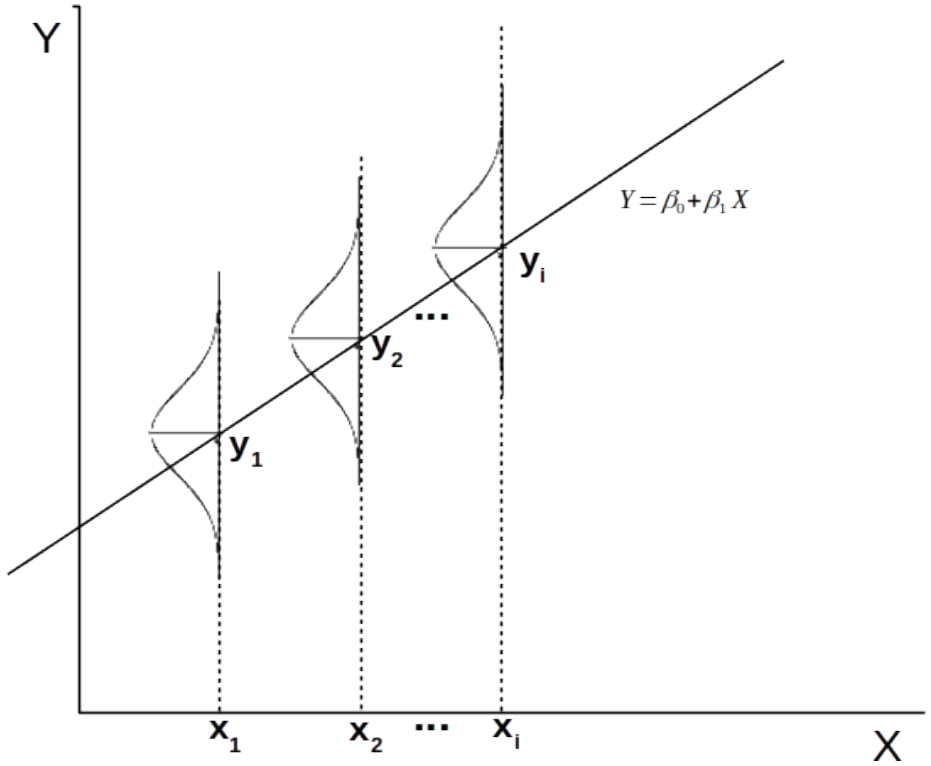


Figura 1: Modelo gráfico de la recta de regresión.

En la muestra

Criterio de selección

Las X **no** tienen criterio probabilístico, se decide arbitrariamente (por el investigador) cuánto vale. Por eso, a veces se le llama "media variable o semi-variable". Por lo tanto, X **NO** es una variable aleatoria, no tiene una probabilidad asociada. Su selección es arbitraria, tiene cierta libertad, pero una vez seleccionada las x_i, las y_i **SÍ** que se seleccionan de forma aleatoria.

El caso más sencillo, para cada x, es que la muestra tenga un tamaño 1. Lo que quiere decir: que para cada x sólo se escoge una y de todas las posibles y_s.

Generalmente, el criterio de selección suele ser que: para cada valor arbitrario de x_i elijo un y_i aleatorio simple (a.s.), con lo cual la muestra total tiene n pares de valores.

A la X se le llamaba variable independiente, pero es un nombre antiguo y lo correcto, en este contexto, es llamarle: variable auxiliar, variable regresora o en general, **variable explicativa**.

En el Muestreo

En el mundo del Muestreo (no confundir con el hecho de muestrear) es fundamental saber qué inferencia se pretende realizar; i.e. responder a: ¿Cuál es el objetivo que se quiere conocer?

En la Población para cada valor de X, la Y (media de y para cada x) está sobre una recta: es una recta. La distribución de Y para cada X es una normal con $\mu = A + BX$ y con la misma varianza σ^2. También lo podemos expresar como $y \cap (Y,\ \sigma^2)$. Es decir:

$$\boxed{Y = A + BX}$$

Para resolver la ecuación anterior, ¿Cuántos parámetros se tienen que calcular?

Se tiene A y B y $\cdots$ σ^2 (común para todos los x) e. . . . $Y_{(}media)$ que también lo desconozco, pero si se conoce A y B ya se puede calcular.

Con lo cual el objetivo que se pretende es: que a partir de la muestra, **estimar** A, B y σ^2 para posteriormente calcular la Y.

Es a partir de este punto cuando ya se pueden hacer pruebas de hipótesis.

Estimación de $A, B,\ \sigma^2$ e Y

Habra que estimar estos tres parámetros: a, b *e* $\hat{Y}$.
Estimadores puntuales de A, B *e* Y respectivamente.
Obviamente, los importantes serán los dos primeros (a y b).

Observación		Pretensión	Estimación
x	y	Y	$\hat{Y}$
x_1	y_1	$A + B\,x_1$	$a + b\,x_1$
x_2	y_2	$A + B\,x_2$	$a + b\,x_2$
⋮	⋮	⋮	⋮
x_i	y_i	$A + B\,x_i$	$a + b\,x_i$
⋮	⋮	⋮	⋮
x_n	y_n	$A + B\,x_n$	$a + b\,x_n$

Figura 2: Tabla de regresión.

Método de mínimos cuadrados

Este método de estimación es **uno** de los más habitualmente utilizados. Concebido para el cálculo de la estimación puntual de a y b, basado en reducir (minimizar) la distancia al cuadrado entre el y de la muestra menos el estimador ($\hat{Y}$) de y:

$$\sum_{i=1}^{n}(y_i - \hat{Y}_i)^2 = S^2$$

esa diferencia de valores se llama *residuo* y por tanto S^2 es la suma de los cuadrados de los residuos. Que también se puede expresar como:

$$S^2 = \sum[y_i - (a + bx_i)]^2$$

Para calcular a y b, simplemente se calculan las derivadas parciales de la suma de cuadrados de los residuos con respecto a cada estimador para que estas derivadas sean igual a cero, ya que lo que se busca es minimizar esa función para cada parámetro.

Este método sencillo de ajuste se utiliza en muchas ocasiones en estadística, ya que los estimadores que genera son:

- No sesgados
- Eficientes(Varianza pequeña en el Muestreo)
- Consistentes. Si sesgados, $\Uparrow n$ tiende a ser insesgado

En resumen, para la estimación de los parámetros lo que se busca es minimizar la suma de cuadrados de los residuos $(y_i - \hat{Y}_i)^2$ en el ajuste. También se puede expresar como: $S = S^2_{residual}$.

Otra forma de verlo es: que lo que se busca es reducir la diferencia entre lo que se observa (y_i) y lo que se estima $(\hat{Y})$ o espera obtener a partir del ajuste.

Nota: Cuidado con la notación usada para las sumas de cuadrados porque a veces es confusa y depende de los autores y/o de su ubicación formal.

Cálculo de los estimadores por mínimos cuadrados

Para el cálculo de los parámetros de la regresión lo primero que se necesita es recordar cómo se calcula la suma de cuadrados para x e y y la suma de cuadrados del producto $x \cdot y$ (S_{xy}):

$$S_{xx} = \sum (x_i)^2 - \bar{x}(\sum x_i)$$
$$S_{yy} = \sum (y_i)^2 - \bar{y}(\sum y_i)$$
$$S_{xy} = \sum x_i \cdot y_i - \bar{x}(\sum y_i)$$
$$o$$
$$S_{xy} = \sum x_i \cdot y_i - \bar{y}(\sum x_i)$$

Una vez obtenidos las suma de cuadrados de las variables y obviamente sus medias aritméticas, ya se pueden calcular todas las estimaciones puntuales necesarias sabiendo que:

1. Estimador de **b**

$$\boxed{b = \frac{S_{xy}}{S_{xx}}}$$

2. Estimador de **a**

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

3. Estimador de $\hat{\mathbf{Y}}$

$$\boxed{\hat{Y} = a + b\ X}$$

4. Estimador para $\sigma^2 \rightarrow s^2$

$$\boxed{s^2 = \frac{S^2_{residual}}{n-2}}$$

Donde la $S^2_{residual}$ es:

$$S^2_{residual} = S_{yy} - \frac{S^2_{xy}}{S_{xx}}$$

$$o$$

$$S^2_{residual} = S_{yy} - b^2 S_{xx}$$

Nota: Cuidado con la notación, ya que puede aparecer tanto $S_{xx/yy/xy}$ *como* $S^2_{x/y/xy}$

Teoría de Muestreo (TM) sobre los parámetros

Los estimadores de nuestros parámetros son:

Parámetros	A	B	σ^2	Y
Estimadores	a	b	s^2	$\hat{Y}$

Y la TM, ¿Qué es lo que dice sobre cada uno de estos estimadores?

1. Estimador de a

- Esperanza

$$E[a] = A \qquad \rightarrow no\ sesgado$$

- Varianza

$$V[a] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \qquad \rightarrow \sigma^2\ conocida$$

$$var[a] = s^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \qquad \rightarrow \sigma^2\ desconocida$$

- Error

$$Error[a] = \sqrt{V[a]} \qquad \rightarrow \sigma^2\ conocida$$

$$error[a] = \sqrt{var[a]} \qquad \rightarrow \sigma^2\ desconocida$$

- Comportamiento (Distribución)

$$a \sim N(E, V) \rightarrow Si\ \sigma^2\ desconocida: \ a \sim t_\nu$$

2. Estimador de b

- Esperanza

$$E[b] = B \qquad \rightarrow no\ sesgado$$

- Varianza

$$V[b] = \frac{\sigma^2}{S_{xx}} \qquad \rightarrow \sigma^2\ conocida$$

$$var[b] = \frac{s^2}{S_{xx}} \qquad \rightarrow \sigma^2\ desconocida$$

- Error

$$Error[b] = \sqrt{V[b]} \qquad \rightarrow \sigma^2\ conocida$$

$$error[b] = \sqrt{var[b]} \qquad \rightarrow \sigma^2\ desconocida$$

- Comportamiento

$$b \sim N(E, V) \rightarrow \qquad Si\ \sigma^2\ desconocida: \ b \sim t_\nu$$

3. Estimador de $\hat{Y}$

- Esperanza

$$E[\hat{Y}] = Y = A + B\ X \qquad \rightarrow no\ sesgado$$

- Varianza

$$V[\hat{Y}] = \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right] \rightarrow \sigma^2\ conocida$$

$$var[\hat{Y}] = s^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right] \rightarrow \sigma^2\ desconocida$$

- Error

$$Error[\hat{Y}] = \sqrt{V[\hat{Y}]} \qquad \rightarrow \sigma^2\ conocida$$

$$error[\hat{Y}] = \sqrt{var[\hat{Y}]} \qquad \rightarrow \sigma^2\ desconocida$$

- Comportamiento

$$\hat{Y} \sim N(E, V)$$

4. Estimador s^2

- Esperanza

$$E[s^2] = \sigma^2$$

- Varianza/Error
  (ahora no es relevante)

- Comportamiento

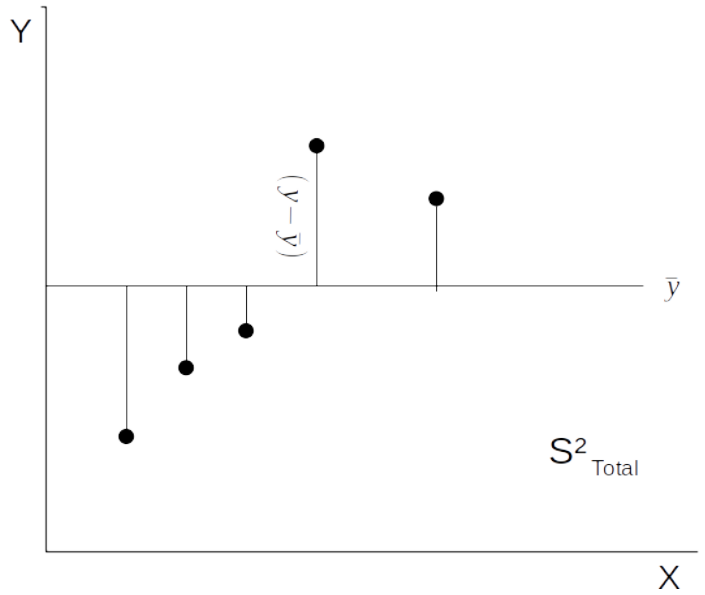
$$\frac{S^2_{residual}}{\sigma^2} \sim \chi^2_{(n-2)}$$

Nota: Recordad la ecuación para la varianza residual.

La importancia de lo Residual

La suma de cuadrados de la "variación" total, no es más que la suma de las diferencias (distancias o dispersiones) cuadráticas de cada valor de la variable respuesta (y) y su media aritmética ($\bar{y}$). Se puede expresar, a partir de las **diferencias**, como:

$$(y - \bar{y}) : S_{yy} = S^2_{Total}$$



Suma de cuadrados Total.

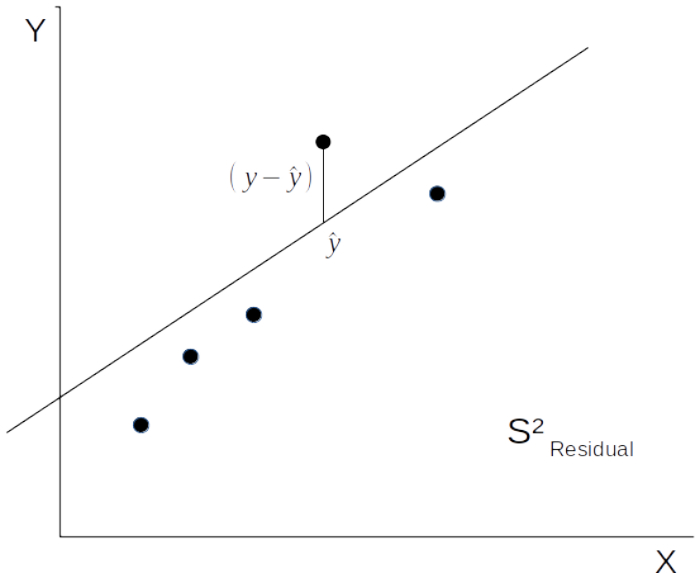
Para cada valor de x_i hay un punto y sobre la recta horizontal de y, que no es más que la media aritmética de y: $\bar{y}$. Por lo tanto, cada segmento representa la diferencia de $(y_i - \bar{y})$.

La S^2_{Total} da en general una medida básica y directa de la dispersión de los puntos. En el caso de la regresión, de la dispersión de los puntos de la variable respuesta.

Precisamente, esto es lo que se quiere conocer y lo que da sentido al modelo de regresión. Saber cómo es la variación de la variable respuesta en relación (a la "influencia") de otra variable que por eso se llama variable explicativa. De ahí que se denomine **Suma de cuadrados total**.

Integrando ahora la recta de regresión, se observan otros tipos de diferencias, como son las diferencias de cada valor de y con respecto a la estimación de $\hat{y}$ derivada del modelo de regresión: $(y - \hat{y})$, a esta diferencia se le llama *residuo*. Por lo tanto, la **Suma de cuadrados residual**, es una medida de la variación residual, también llamada *variación no explicada por la recta de regresión*:

$$(y - \hat{y}) : S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{Residual}^2$$



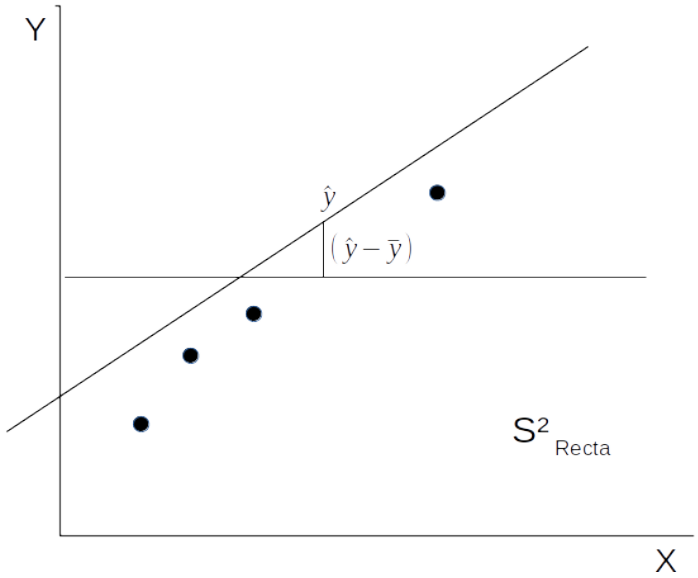
Suma de cuadrados Residual.

Dicho de otra manera, la variación residual es la variación no explicada por la recta (es residual), es lo que resta del valor observado de la recta después de que se haya ajustado un modelo de regresión.

Resultando ser una herramienta muy poderosa a tener muy en cuenta porque da, entre otras cosas, una medida del ajuste.

La diferencia entre la variación total y la residual será la variación debida a la recta. Aquella parte de la variación que es *explicada* por la recta de regresión. Por lo tanto, la variación de la recta se mide a través de la **Suma de cuadrados de la recta**, que será:

$$(\hat{y} - \bar{y}) : \frac{S_{xy}^2}{S_{xx}} = S_{Recta}^2$$



Suma de cuadrados de la Recta.

En resumen y bajo la perspectiva de las desviaciones, tenemos que:

$$(y - \bar{y}) \equiv (y - \hat{y}) + (\hat{y} - \bar{y})$$

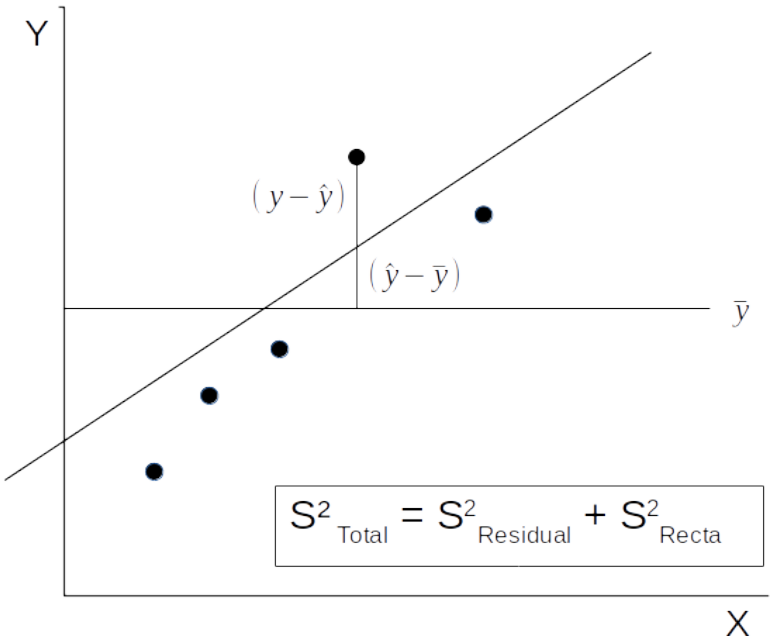
Las desviaciones totales para cada puntuación, son idénticas a la suma de las desviaciones hasta la recta y las desviaciones de la recta hasta la media.

Si se eleva al cuadrado y se suma cada desviación, se obtiene que la variación total es igual a la suma de la variación residual más la variación explicada por la recta.

$$\boxed{S^2_{Total} = S^2_{Residual} + S^2_{Recta}}$$

Esta es la descomposición fundamental de la variación para el modelo de regresión. También se expresa, como: la variación total es la suma de la variación no explicada (residual) y la variación explicada por la recta.

Evidentemente, cuánto mayor es la S^2_{Recta} menor será la $S^2_{Residual}$ lo que implica que será mejor el ajuste a la recta.



Suma de cuadrados Total de la recta de regresión.

Si la relación anterior se divide por la S^2_{Total} queda:

$$\boxed{1 = \frac{S^2_{Recta}}{S^2_{Total}} + \frac{S^2_{Residual}}{S^2_{Total}}}$$

Con lo cual se obtiene en porcentaje (o tanto por uno), sobre cual es la importancia relativa que explica o no, la variación total de la recta de regresión.

La relación $\frac{S^2_{Recta}}{S^2_{Total}}$ es el **coeficiente de determinación.**

Nota: Realmente llamarlo coeficiente de determinación es un mal nombre. Sería más apropiado llamarlo: *coeficiente de explicación*.

El coeficiente de determinación (*c.d.*) es una medida de la bondad del ajuste, ya que indica el porcentaje de variación total que es explicado por la recta de regresión. Muchas veces se expresa como r^2 y es una notación muy equívoca.

Esta denominación se suele, o se puede, relacionar erróneamente con el *coeficiente de correlación*.

La correlación considera que la variable explicativa es aleatoria (modelo de tipo II, o de efectos variables). Por lo tanto, asume aleatoriedad en ambas variables, lo que **no** ocurre con el modelo de regresión simple(modelo de tipo I, o de efectos fijos).

Nota: Este es un error conceptual muy extendido, donde la variable explicativa **no** es aleatoria.

Con respecto a sus valores se puede interpretar que:

- Si $cd|r^2 = 0$ La recta no explica nada
- Si $cd|r^2 = 1$ La recta explica toda la variación
- $0 \leq cd|r^2 \leq 1$ La recta explica parte de la variación

Como resumen, se puede decir que el *coeficiente de determinación*, sirve para:

- **Medida de la bondad de ajuste.**
  Cuan cercanos son los puntos a la recta.
- **Medida del grado de linealidad.**
  "Acercamiento" residual.
- **Medida directa de la mejora en el modelo.**
  Mejora al introducir información al modelo: más datos, mejores datos, una u otra variable, mejores variables etc.

Análisis visual de los residuos

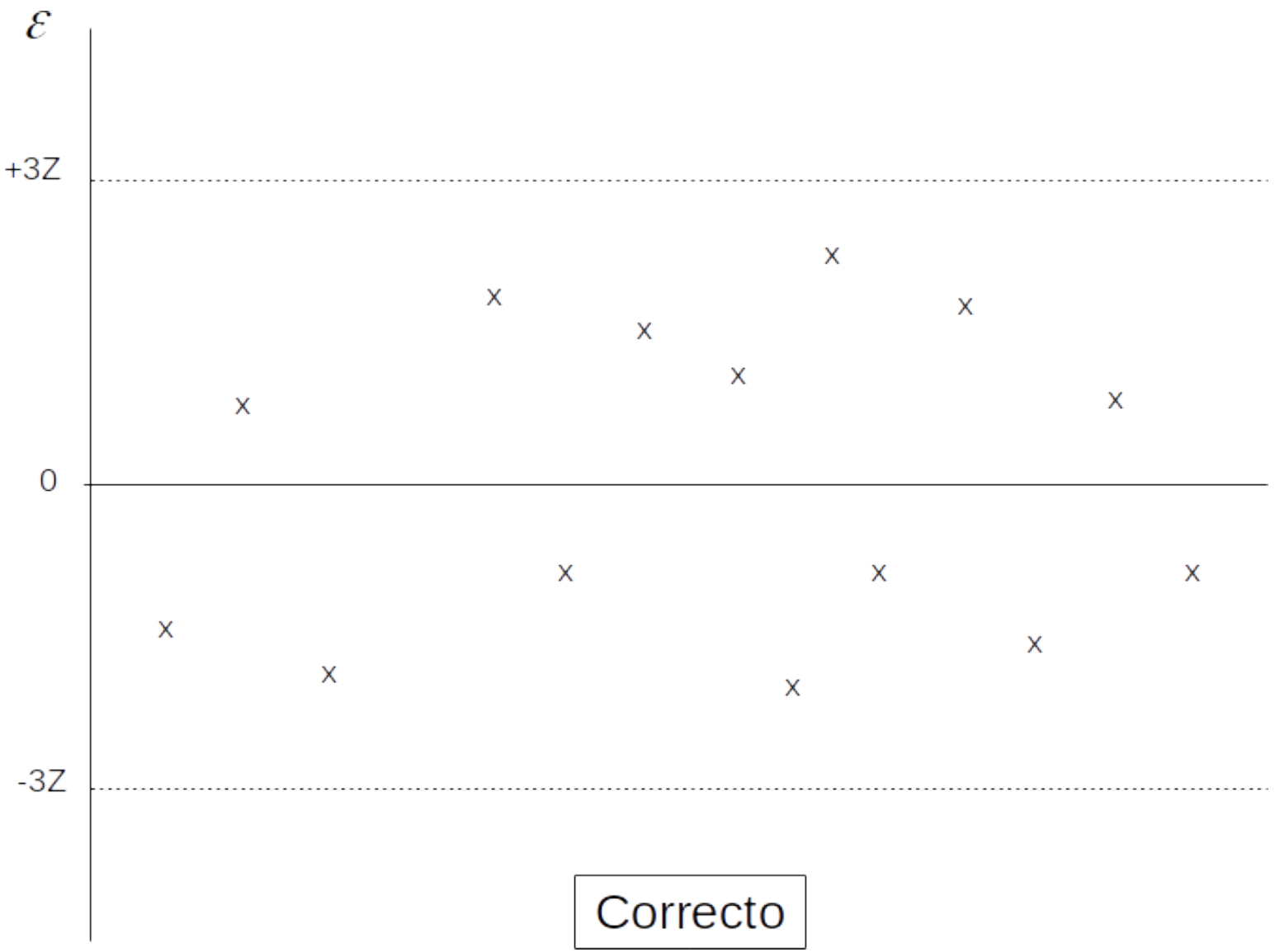
El residuo en el modelo de regresión lineal simple se había definido como:

$$Residuo \;=\; Error \;= \hat{y}_i - y_i = \mathcal{E}_{ij} = \beta_0 + \beta_1 x_1 - y_i$$

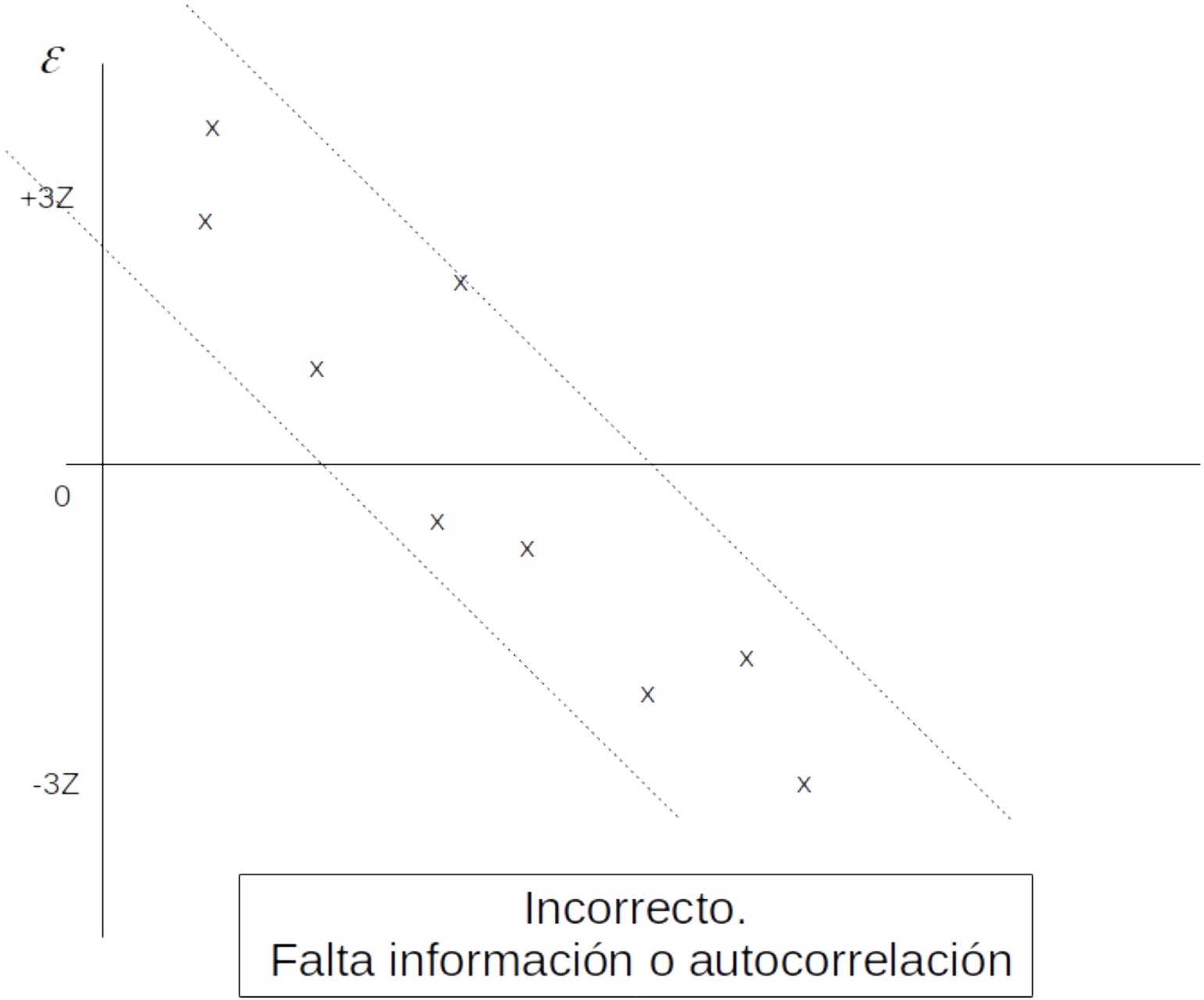
Lo primero que hay que hacer es ver si se cumple:

1. $E[\mathcal{E}_i] = 0$ Aleatoriamente y con Esperanza nula
2. $cov(\mathcal{E}_i, \mathcal{E}_j) = 0; \quad \forall\, i \neq j \Rightarrow \nexists\, rachas$ Sin auto-covariación
3. $Var[\mathcal{E}_i] = \sigma^2 = cte$ Homocedasticidad
4. $\mathcal{E}_i \cap N(0, \sigma^2)$ Distribución normal de los residuos

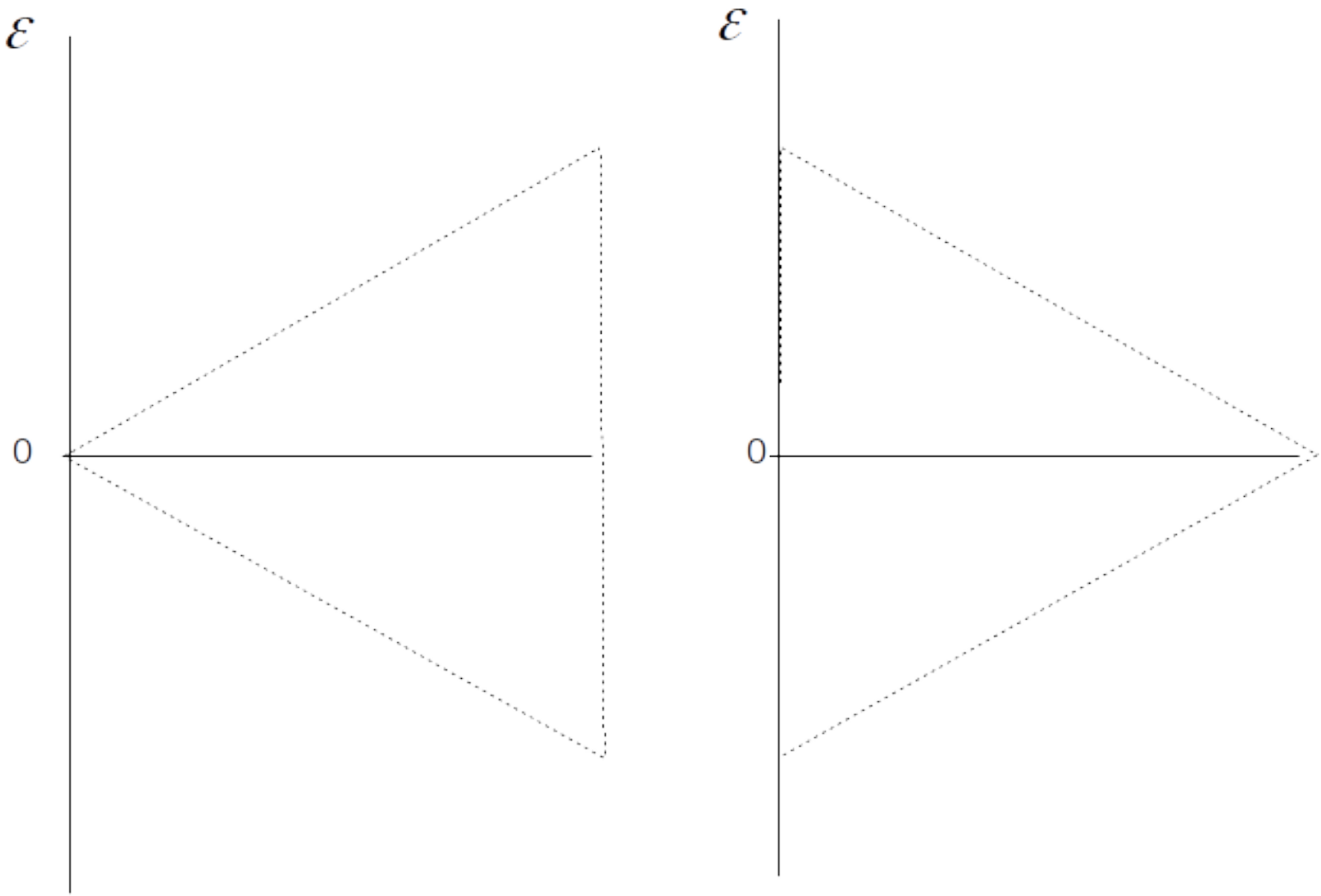
Patrones de representación visual sobre los residuos



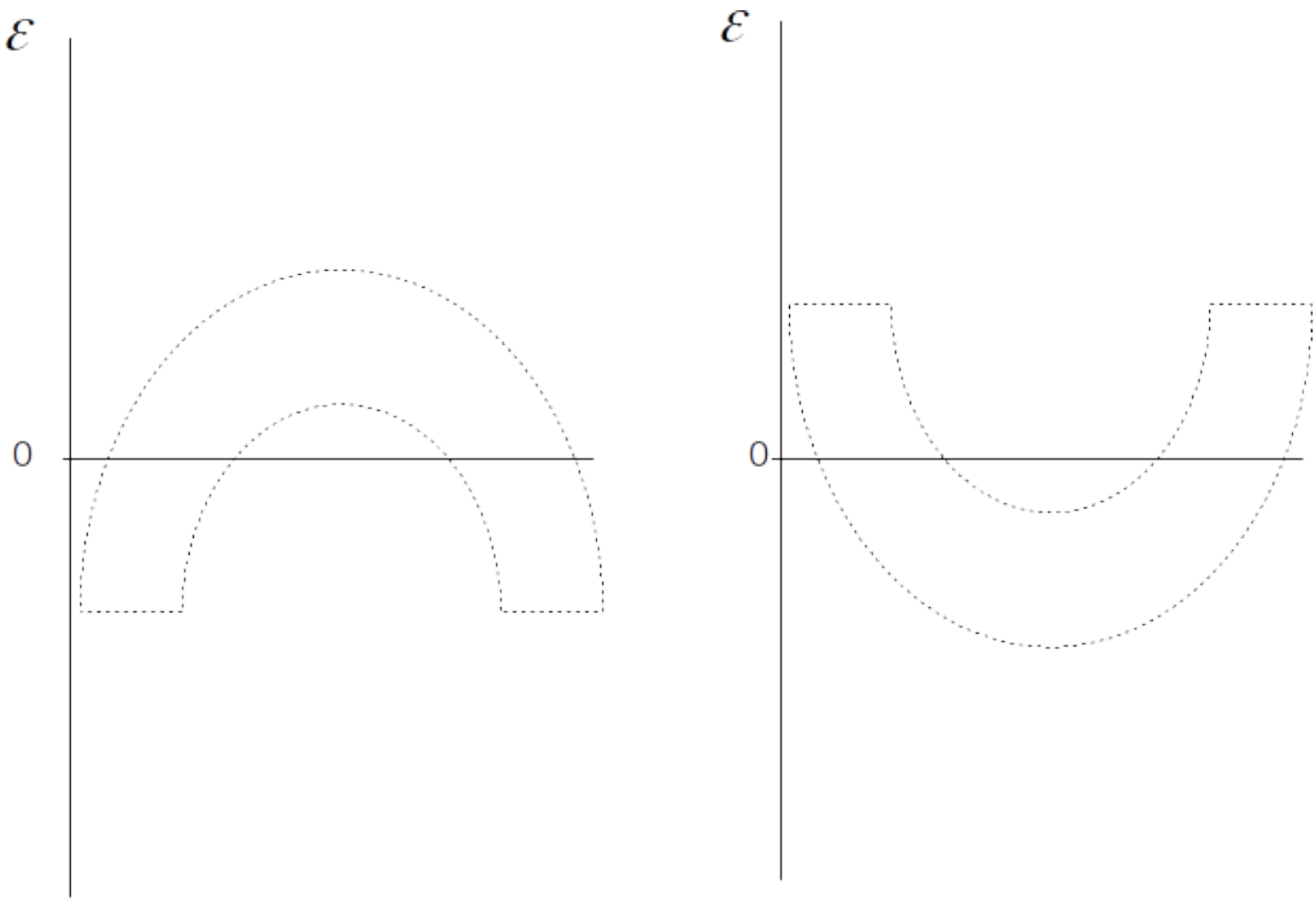
Representación correcta/esperable de los Residuos en un modelo de regresión.



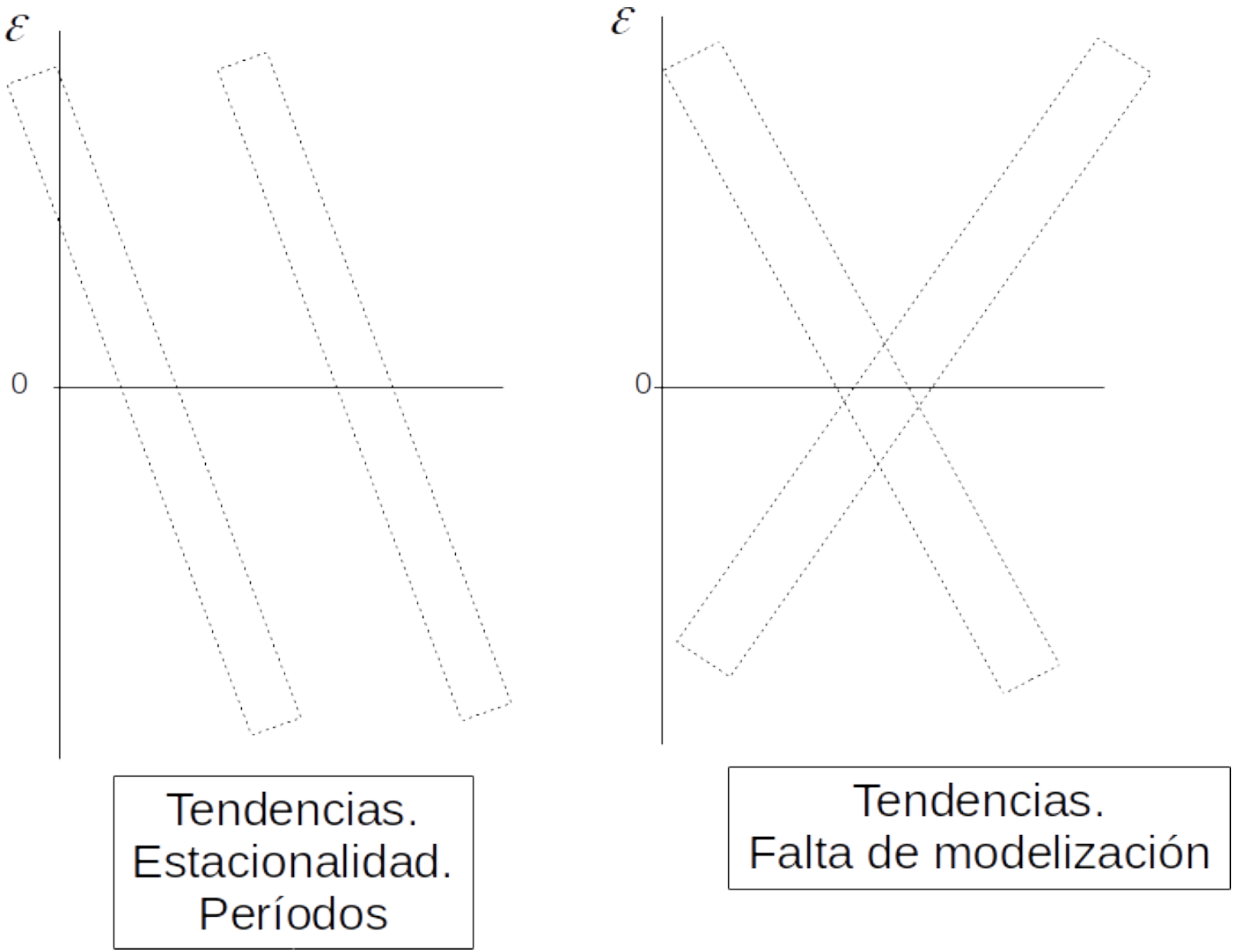
Representación incorrecta de los Residuos en un modelo de regresión.



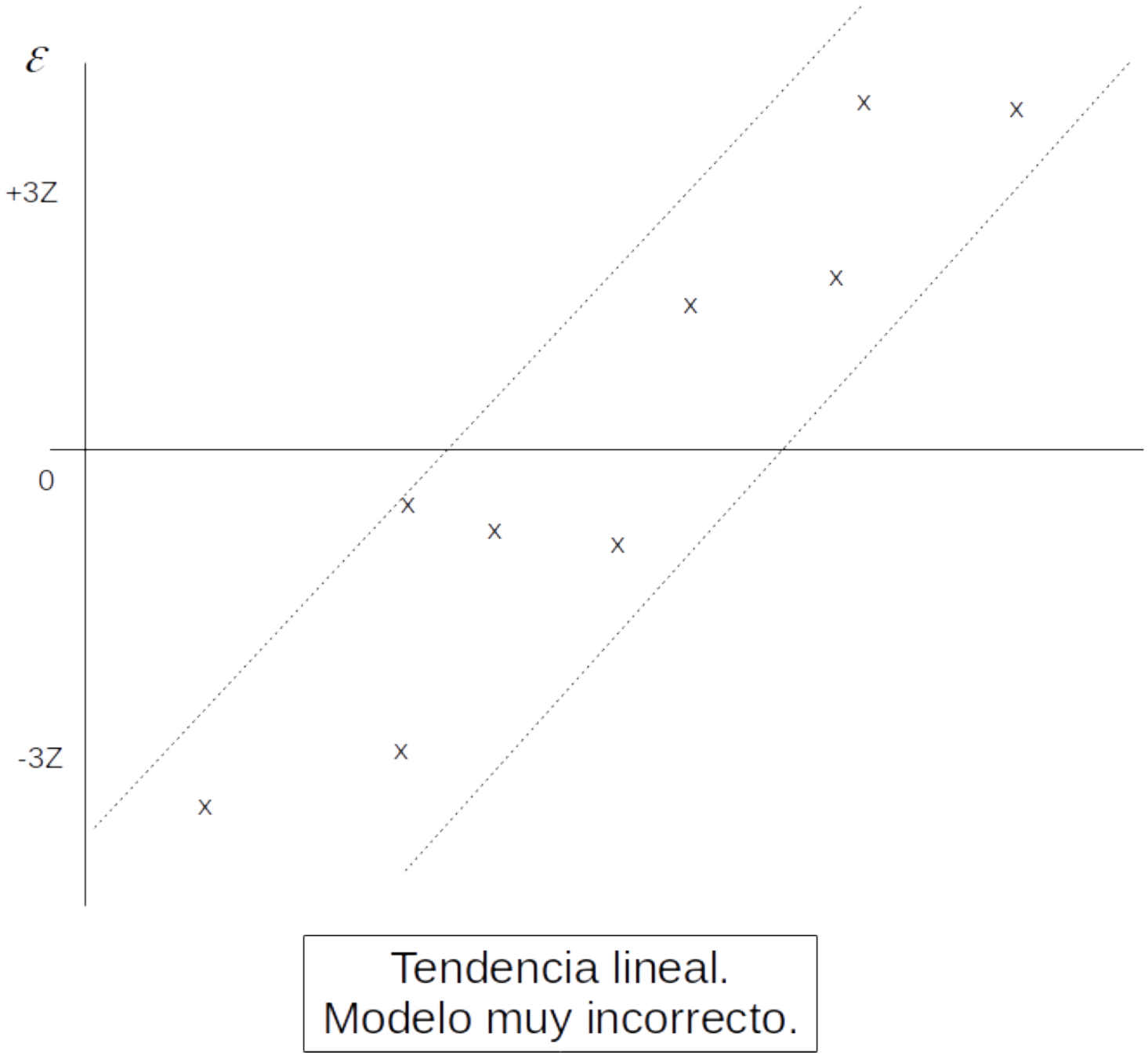
Representación con heterocedasticidad de los Residuos en un modelo de regresión.



Representación con no-linealidad de los Residuos en un modelo de regresión.



Representación con tendencias de los Residuos en un modelo de regresión.



Representación con tendencias lineales de los Residuos en un modelo de regresión.

Anexo sobre cálculo de los estimadores

Para el cálculo de los estimadores y sobre todo para el de dos variables, el método de estimación se puede hacer, como ya se indicó anteriormente, a través del cálculo de diferenciales parciales de la función objetivo (la suma de cuadrados de los residuos) con respecto a cada parámetro.

Como la regresión simple es una simplificación del modelo de regresión múltiple (con varias variables explicativas) y este a su vez del Modelo Lineal General (GLM, en inglés), los cálculos son matriciales. Pero el análisis matemático, también ofrece otro tipo de aproximaciones más "sencillas" para resolver este problema.

Método analítico para más de dos variables

Se genera un sistema de dos ecuaciones partiendo del modelo básico de regresión. A la primera ecuación: se le aplica un sumatorio; y a la segunda: se le multiplica por x y se aplica el sumatorio a la ecuación, quedando:

$$\begin{cases} 1: \sum y & = n \cdot a + b \cdot \sum x \\ 2: \sum xy & = a \cdot \sum x + b \cdot \sum x^2 \end{cases}$$

Nota: Muchas veces, por simplificación, se obvian los subíndices i en las ecuaciones.

1. Si a la primera ecuación la dividimos por n, queda:

$$\frac{\sum y}{n} = \frac{n \cdot a}{n} + \frac{b \cdot \sum x}{n} \Longrightarrow \bar{y} = a + b\ \bar{x} \Longrightarrow \boxed{a = \bar{y} - b\ \bar{x}}$$

2. Si a la segunda ecuación, ahora le restamos el término $-\bar{y} \sum x$, queda:

$$\begin{aligned} \sum xy - \bar{y} \sum x &= a \sum x - \bar{y} \sum x + b \sum x^2\ ; \quad Como\ a = \bar{y} - b\bar{x} \\ Sxy &= (-b\ \bar{x}) \sum x + b \sum x^2 \\ &= b\ (\sum x^2 - \bar{x} \sum x)\ ;\ Como\ (\sum x^2 - \bar{x} \sum x) = Sxx \\ &= b\ Sxx \Longrightarrow \boxed{b = \frac{Sxy}{Sxx}} \end{aligned}$$

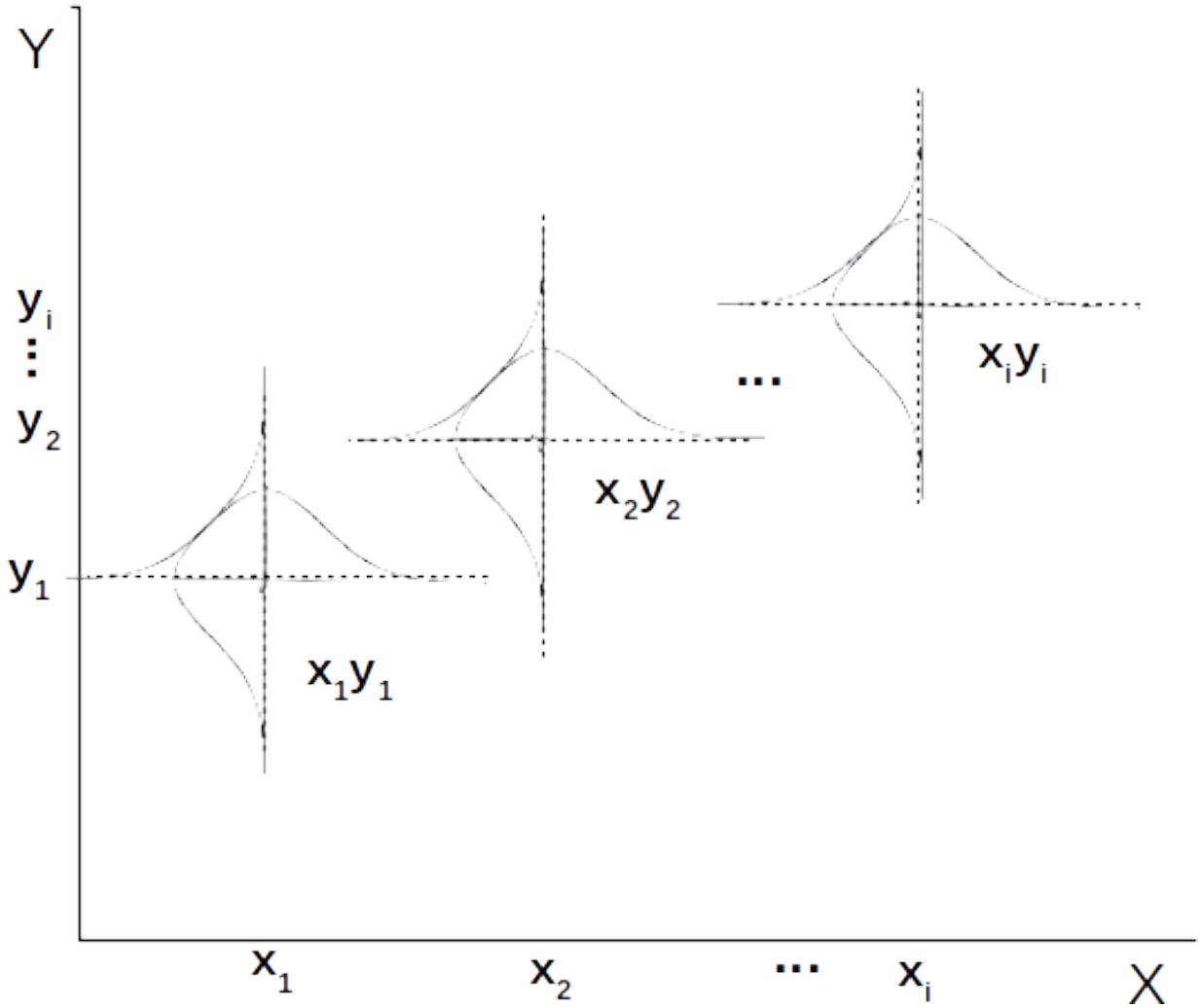
El sistema de ecuaciones anterior, también se puede resolver aplicando la *Regla de Kramer*, quedando:

$$a = \frac{\begin{vmatrix} \sum y & \sum x \\ \sum xy & \sum x^2 \end{vmatrix}}{\begin{vmatrix} n & \sum x \\ \sum x & \sum x^2 \end{vmatrix}}$$

$$b = \frac{\begin{vmatrix} n & \sum y \\ \sum x & \sum xy \end{vmatrix}}{\begin{vmatrix} n & \sum x \\ \sum x & \sum x^2 \end{vmatrix}}$$

11.- Correlación. Efectos variable (II)

La diferencia fundamental entre la regresión (modelo I) de efectos fijos y la correlación (modelo II) de efectos variables es que en este último los pares de valores son aleatorios, i.e. varían y covarían las dos variables a relacionar.



Correlación. Modelo II de efectos variables.

Se puede decir que la correlación mide la fuerza de la relación (covariación) entre un conjunto de variables cuantitativas continuas a través de un estadístico al efecto, que en el caso más sencillo, se denomina: coeficiente de correlación simple (r).

Notas:

- Covariación no implica necesariamente causalidad.
- Ahora sólo hablaremos del coeficiente de correlación de Pearson.
- En el caso de más variables, se hablará de coeficiente de correlación múltiple y de coeficientes de correlación parciales.

La correlación no deja de ser una medida normalizada de la covariación entre dos variables, que está comprendida entre $-1 \leq r \leq +1$.

Se define como:

- **varianza** (s^2)

$$s^2 = \frac{S_{xx}}{n-1} = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

- **covarianza** $cov(x, y)$

$$cov(x, y) = \frac{S_{xy}}{n-1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- **coeficiente de correlación de Pearson** (r)

$$\boxed{r = \frac{cov(x, y)}{s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}}$$