



K L E O P A T R A

PTOLOMEO

ΠΤΟΛΕΜΑΙΣ



*Nombre del rey Ptolomeo en español, en griego y jeroglífico*

La clave para descifrar la **pedra Rosetta** fueron los glifos de los nombres de Cleopatra y Ptolomeo.

## Aprendiendo idiomas X

María Jesús Vázquez Gallo



1. **Introducción.**
2. **Hipótesis en regresión simple.**
3. **Análisis del modelo de regresión simple.**
4. **Para saber más...**

## 1. Introducción

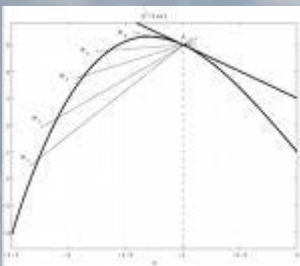
En **Aprendiendo idiomas IX** vimos que un modelo de **regresión simple** representa el valor medio de una variable respuesta **y** en términos de otra variable explicativa **x**.

**Ejemplo:** Queremos estudiar el sueldo de un directivo en función de su edad. Para ello, tomamos como variable respuesta **y**: sueldo anual en euros del directivo; y como variable explicativa **x**: la edad del directivo.

**Hipótesis fundamental:**

Se supone que la dependencia de **y** con respecto a **x** es **lineal**.

**Pero ¿una relación entre dos variables siempre es lineal?**



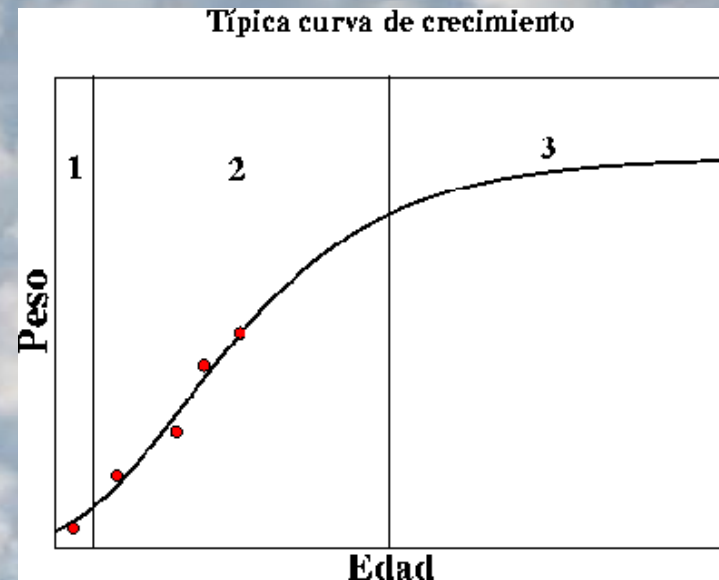
NO, pero una relación lineal es una aproximación sencilla de una relación más compleja (igual que una recta es un objeto más “sencillo” que una parábola o que cualquier otra curva en general)

## 1. Introducción

**Observación:** puede que  $y$  dependa linealmente de  $x$  en cierto rango de valores de  $x$  pero no en todos.

**Ejemplo:** Estudiamos el peso en función de la edad, tomando como variable respuesta  $y$ : peso de un individuo; y como variable explicativa  $x$ : su edad.

Para edades pequeñas, la gráfica es aproximadamente una recta pero, a partir de cierta edad, la gráfica ya no se parece tanto a una recta: la relación deja de ser lineal.



## 2. Hipótesis en regresión simple

Un modelo de regresión con variable explicativa  $x$  y variable respuesta  $y$  asume las siguientes hipótesis:

- Para cada  $x$ , la distribución de  $y$  es una **normal**. (¿Recuerdas la campana de Gauss de **Aprendiendo idiomas III**?).
- Cuando  $x$  varía, la media de dicha distribución normal crece **linealmente** con  $x$ .
- Se supone que hay otra serie de factores que influyen –poco- en la variable  $y$ . Estos factores se agrupan en una **perturbación aleatoria** con distribución normal y **esperanza nula**.
- Se supone que la **varianza** de la perturbación es constante y no depende de  $x$ . Se dice entonces que la **perturbación** es **homocedástica**. (Recuerda que la varianza es una medida de dispersión de una variable aleatoria, cuya raíz cuadrada se llama **desviación típica**, como vimos en **Aprendiendo idiomas I**).



\* En una fórmula:

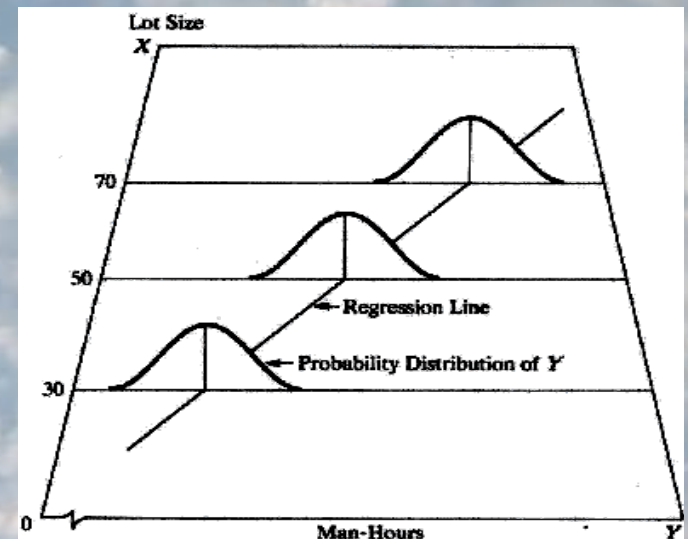
$$y = \beta_0 + \beta_1 x + u$$

con  $u$  la perturbación aleatoria y  $\beta_0$ ,  $\beta_1$  parámetros que queremos estimar.

¿Se te ocurre qué factores puede reflejar la perturbación aleatoria en los ejemplos que hemos visto hasta ahora?

\* En una imagen:

Globalmente, lo que se supone es que, para cada valor de  $x$ , la distribución de la variable respuesta  $y$  es normal, con media que crece linealmente con  $x$  y varianza constante.

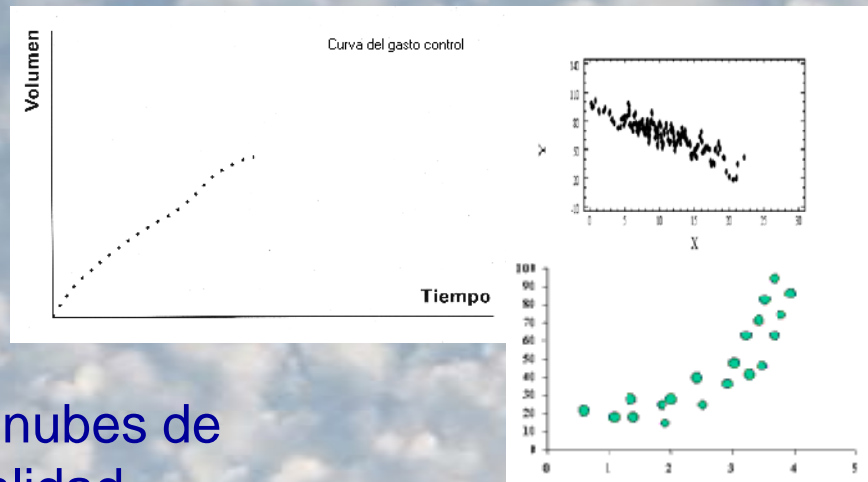


Es importante **comprobar las hipótesis** realizadas en el modelo:

¿Cómo hacerlo?

### 1) Linealidad.

- Si conocemos un único valor de  $y$  para cada  $x$ : representamos la nube de puntos  $(x,y)$  y comprobamos visualmente que se asemeja a una recta.
- Si tenemos varios valores de  $y$  para cada  $x$ : realizamos un contraste de hipótesis como vimos en **Aprendiendo idiomas IV y V**.



**Ejemplo:** una de estas nubes de puntos no sugiere linealidad...

## 2. Hipótesis en regresión simple

¿Qué hacemos si hay indicios de NO LINEALIDAD?

Se puede intentar transformar las variables, como veremos...

### 2) Homocedasticidad.

**Ejemplo:** Si estudiamos la inversión en bolsa como variable respuesta  $y$  explicada por la cantidad de ingresos familiares  $x$ :

- Para  $x$  baja,  $y$  será nulo y no habrá variabilidad entre familias.
- Para  $x$  alta,  $y$  variará según las familias.

Por tanto, la varianza de cada distribución normal de  $y$ , en este caso, depende de  $x$  y no se tendría homocedasticidad.

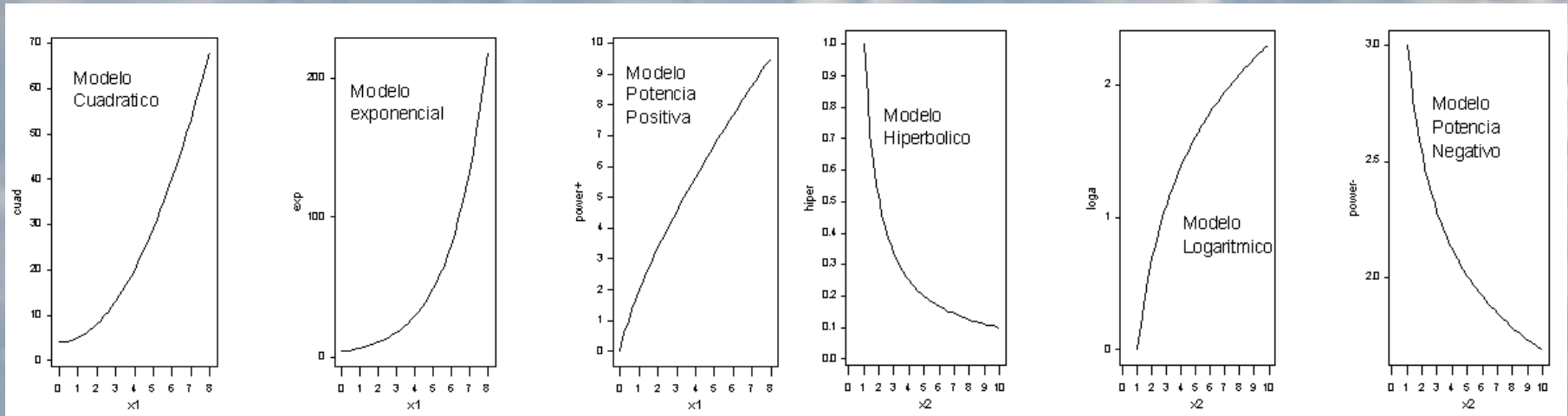
Las varianzas se analizan a través de las llamadas **tablas ANOVA** (analysis of variance) o **ADEVA** (análisis de varianza).

¿Qué hacemos si hay indicios de HETEROCEDASTICIDAD?

Se puede intentar transformar las variables.

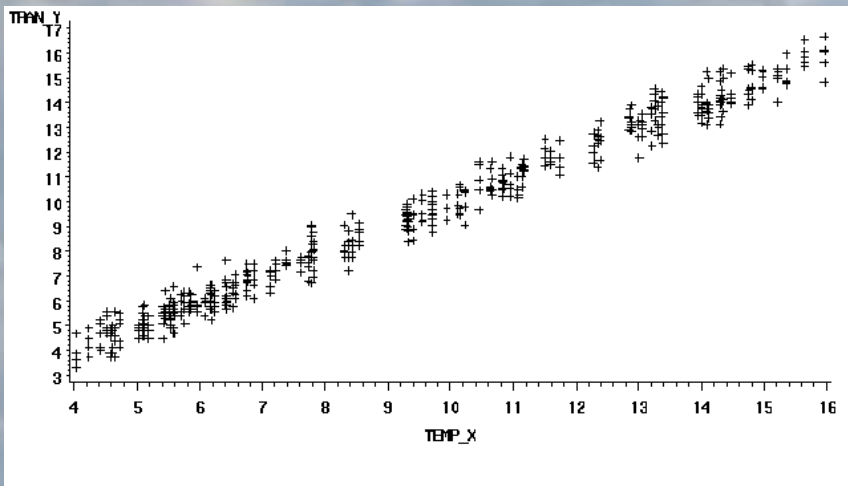
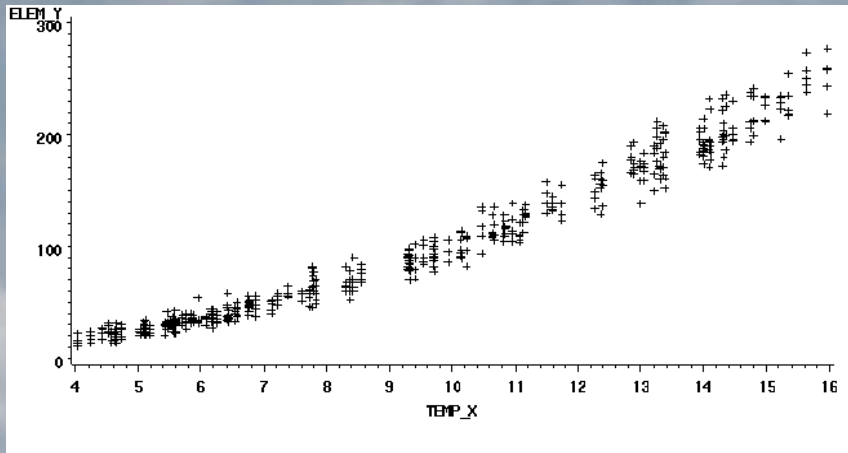
## Transformaciones típicas en regresión lineal simple

Nombre del modelo	Ecuación del Modelo	Transformación	Modelo Linealizado
Exponencial	$Y = \alpha e^{\beta X}$	$Z = \text{Log} Y \quad X = X$	$Z = \text{Log} \alpha + \beta X$
Logarítmico	$Y = \alpha + \beta \text{Log} X$	$Y = Y \quad W = \text{Log} X$	$Y = \alpha + \beta W$
Doblemente Logarítmico o Potencia	$Y = \alpha X^{\beta}$	$Z = \text{Log} Y \quad W = \text{Log} X$	$Z = \text{Log} \alpha + \beta W$
Hiperbólico	$Y = \alpha + \beta/X$	$Y = Y \quad W = 1/X$	$Y = \alpha + \beta W$
Doblemente Inverso	$Y = 1/(\alpha + \beta X)$	$Z = 1/Y \quad X = X$	$Z = \alpha + \beta X$





### Ejemplo



Esta nube de puntos con aspecto no lineal, se convierte en la nube claramente lineal de más abajo, al aplicar la transformación  $z = \sqrt{y}$

## 2. Hipótesis en regresión simple

### 3) Perturbaciones independientes entre sí.

La falta de independencia suele darse al trabajar con variables aleatorias que se observan a lo largo del tiempo, es decir, cuando se trabaja con **series temporales**.

Una primera solución a este problema consiste en **aleatorizar** la recogida muestral.

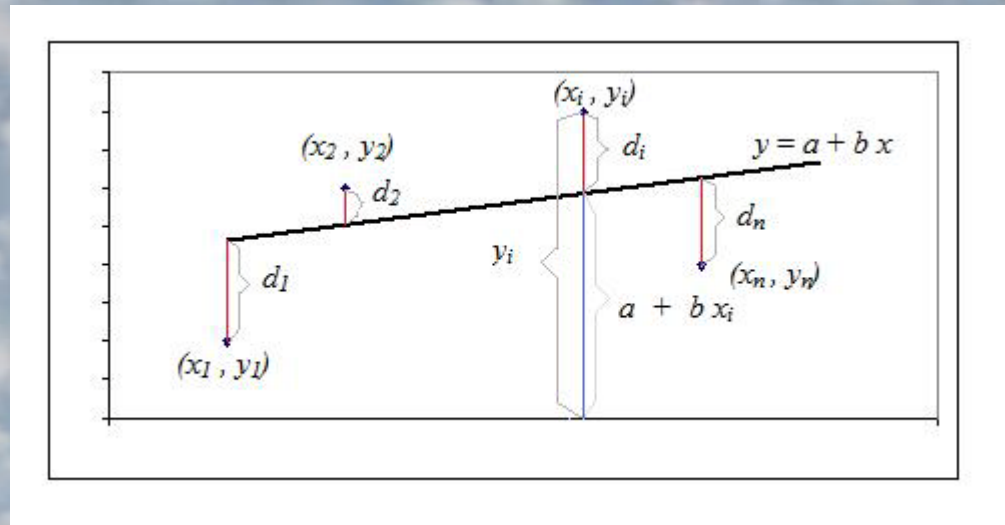
La ausencia de aleatoriedad es especialmente grave ya que puede invalidar por completo las conclusiones del análisis estadístico.

**Ejemplo:** Si se estudia la relación entre el consumo de combustible y el PIB en distintos países en un año concreto, podemos esperar independencia en el resto de factores que afectan a dicho consumo, pero sí estudiamos la misma relación en un único país a lo largo del tiempo, los factores recogidos en la perturbación pueden evolucionar con el PIB y no ser independientes...

Contrastadas las hipótesis, se estiman los parámetros  $\beta_0$ ,  $\beta_1$  del modelo, como vimos en [Aprendiendo idiomas IX](#)

$$y = \beta_0 + \beta_1 x + u$$

Tendremos entonces la recta que “mejor” se ajusta a nuestra nube de puntos:

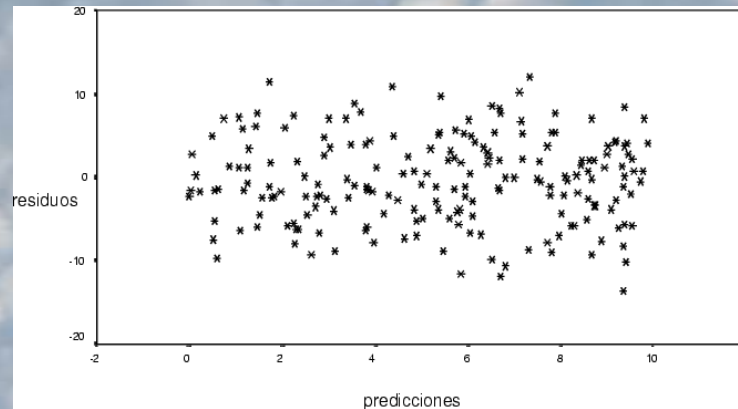


Los parámetros  $\beta_0$ ,  $\beta_1$  son los que minimizan las distancias verticales de los puntos de la nube a una recta. En otras palabras, minimizan la suma de los cuadrados de los **residuos** del modelo.

## ¿Qué son los residuos del modelo?

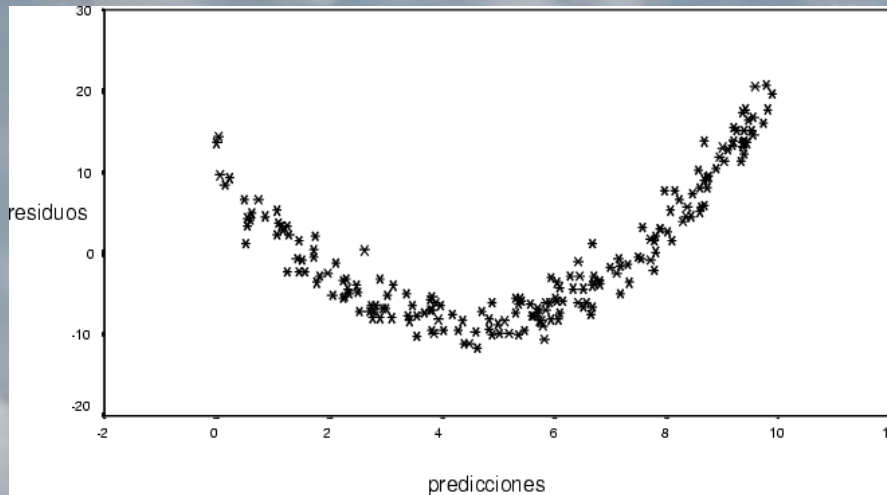
Son las diferencias entre los valores observados (la altura o coordenada  $y$  de los puntos) y los previstos (la  $y$  correspondiente en la recta de regresión).

El análisis de los residuos puede detectar problemas de no linealidad o heterocedasticidad. Si disponemos en un gráfico los residuos frente a los valores previstos:



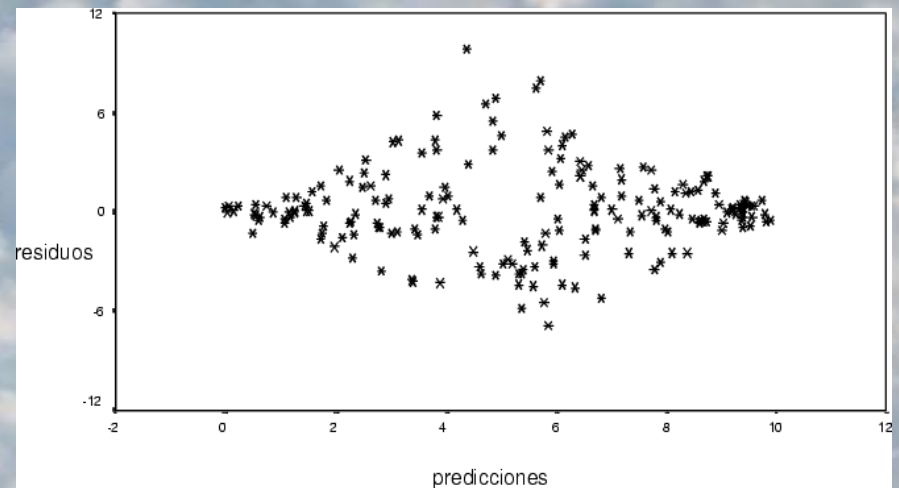
Aquí, los residuos se distribuyen aleatoriamente y esto indica que no hay problemas.

### 3. Análisis del modelo de regresión simple



Aquí, los residuos indican claramente falta de linealidad.

Aquí, los residuos indican heterocedasticidad, puesto que su variabilidad no es constante.



Otra forma de analizar un modelo de regresión es calcular su coeficiente de determinación que, en el caso de la regresión lineal simple que estamos tratando es el **coeficiente  $r$  de correlación** lineal de Pearson (lo vimos en **Aprendiendo idiomas VIII**).

**¿Recuerdas...?** Una  $r$  con tamaño cercano a 1 suele indicar que la relación estudiada está cercana a ser lineal. Una  $r$  cercana a 0 suele indicar que la relación está lejos de ser lineal.

**¡OJO!** Suele utilizarse para comparar modelos pero, modelos con la misma recta de regresión y el mismo coeficiente  $r$  pueden corresponder a datos cuya nube de puntos es aproximadamente lineal o no...

**Conclusión:** La mejor estrategia para juzgar un modelo es contrastar sus hipótesis y analizar sus residuos. Recordar siempre que una correlación alta no implica causalidad y que una correlación baja indica falta de relación lineal, pero puede existir otro tipo de relación.



Si quieres profundizar en los modelos de regresión simple y en los modelos de regresión general, puedes consultar, entre otros, el libro:

**Estadística. Modelos y métodos. 2, Modelos lineales y series temporales / Daniel Peña Sánchez de Rivera.**

Madrid : Alianza, 1989 Edición 2ª ed.