

Tema 2

Regresión lineal simple: geometría

(actualizadas el 07-07-2023)

Tema 2. Regresión lineal simple: geometría

2.1 El modelo de regresión lineal simple

2.2 Estimación de los coeficientes por MCO

2.3 Interpretación de los coeficientes

2.4 Propiedades descriptivas del modelo de regresión

2.5 Medidas de bondad de ajuste: coeficiente de determinación

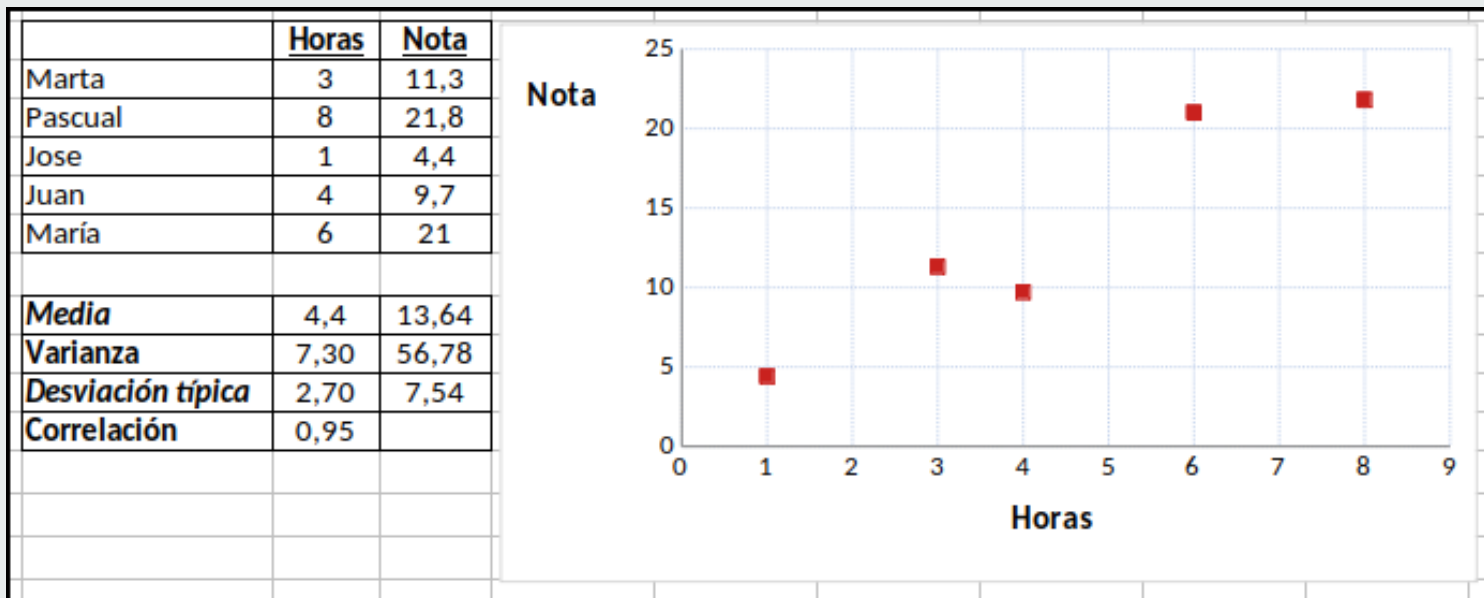
Bibliografía

- Ezequiel Uriel (2013): Capítulo 2 (excepto 2.5)
- Wooldridge (2015): Capítulo 2 (2.1 a 2.3)
- Stock y Watson (2012): Capítulo 4 (4.1 a 4.3)

2.1 El modelo de regresión lineal simple

De forma abreviada MRLS

- La Econometría se ocupa de formular relaciones entre variables económicas, **cuantificarlas** y valorar los resultados obtenidos (AFG).
- Un análisis econométrico empírico suele comenzar con la formulación de una pregunta. Por ejemplo: **¿Cómo afectan las horas de estudio a la nota obtenida en un examen?**
- Supongamos que nos interesa analizar la variable y . Pensamos que y está relacionada con la variable x ; por lo tanto, **trataremos de analizar y cuantificar la relación entre x e y . ¿Cómo?**
- Recogemos datos de x e y . Aquí los tenemos ... **¿y ahora qué?**



Definiendo el MRLS

- Vamos a utilizar el MRLS para analizar y cuantificar la relación entre x e y .
- El MRLS parte de suponer que una variable (y) depende de otra (x): $y = f(x)$
- ¿Qué forma tiene esa relación? El MRLS supone que la **relación entre x e y es lineal**: $y = \beta_1 + \beta_2 x$
- El anterior modelo es **determinista**. Sin embargo las relaciones económicas no son deterministas, siempre hay un cierto grado de incertidumbre o aleatoriedad.
- Por lo tanto en el MRLS se introduce una variable o término adicional (u) llamado **perturbación aleatoria**, de forma que, tendremos un modelo estocástico:

$$y = \beta_1 + \beta_2 x + u$$

Terminología

$$y = \beta_1 + \beta_2 x + u$$

- Nos referiremos a la variable y como variable dependiente, variable a explicar, **regresando** , ...
- Nos referiremos a la variable x como: variable independiente, variable explicativa, **regresor** , ...
- Nos referiremos a la variable u como: término de error, **perturbación** aleatoria, perturbación estocástica, ...
- β_1 y β_2 son **parámetros**, que no conocemos, y queremos estimar

principal objetivo

El objetivo primordial del análisis de regresión, es estimar los parámetros poblacionales (β) partiendo de una muestra dada de datos.

Interpretación del MRLS

- El MRLS supone que cada observación de y es explicado por dos variables: x y la perturbación aleatoria (u)

$$y = \beta_1 + \beta_2 x + u$$

- De forma que podemos pensar que y tiene dos partes:
 - parte determinista o **explicada por x** : $\beta_1 + \beta_2 x$
 - parte estocástica o **no explicada por x** : u

Objetivo inmediato

- Cuantificar la relación entre x e y ; es decir, aproximarnos, **estimar** los parámetros β_1 y β_2 .

¿Cómo lo haremos?

- Utilizando datos y métodos estadísticos (análisis de regresión).

Interpretación de u

- En el MRLS ($y = \beta_1 + \beta_2 x + u$), la perturbación aleatoria representa todos los factores/variables (aparte de x) que afectan o influyen en la variable y .
- Por ejemplo, en el modelo: $\text{salarario} = \beta_1 + \beta_2 \text{educacion} + u$
 - ¿Qué variables pueden estar detrás de u en este ejemplo?
 - Aún así imagine que incluimos en el modelo todas esas variables ¿tendría sentido no incluir u en nuestro modelo?

la perturbación

- La variable (u_i), a la que llamamos perturbación aleatoria aproxima:
 - Variables no incluidas en el modelo
 - errores de medida
 - aleatoriedad intrínseca en todo fenómeno

-
- "No hay modelos correctos, hay modelos útiles". La frase original es de G. Box

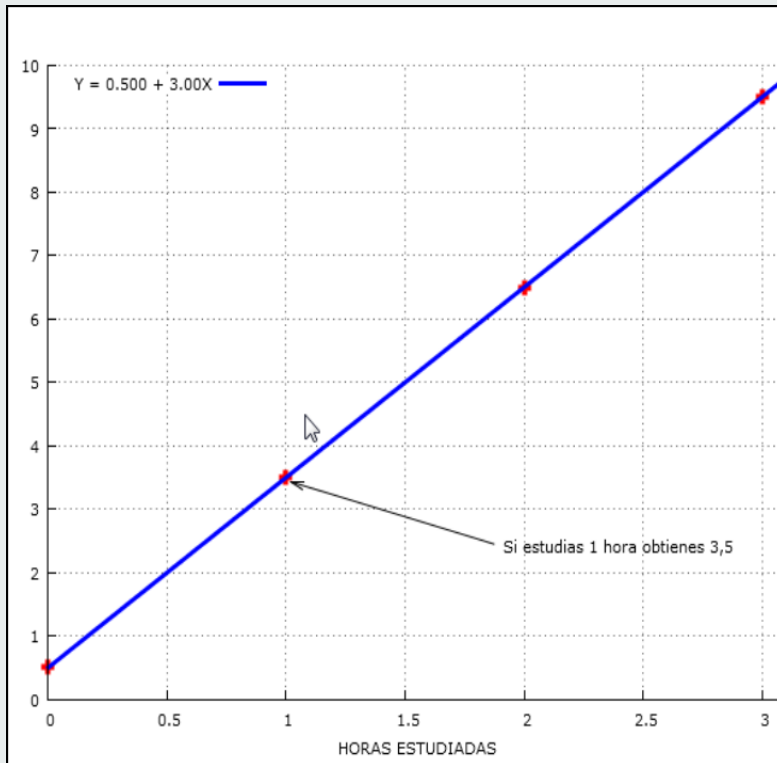
Interpretación de los parámetros

- El modelo $y = \beta_1 + \beta_2 x + u$ define una relación lineal entre x e y
 - β_1 gráficamente es la **ordenada en el origen**
 - β_2 **gráficamente** es la **pendiente**
 - β_2 matemáticamente es la **derivada parcial** de y respecto de x
- Pero, los parámetros no nos interesan, generalmente, por su interpretación gráfica o matemática, sino por su **interpretación económica**:
 - ceteris paribus (si el resto de factores no cambian) β_2 representa el **efecto marginal** de x sobre y
 - al ser el modelo lineal implica que el **efecto marginal de x es constante** (e igual a β_2): si x aumentase en 1 unidad, y aumentaría en β_2 unidades.

-
- Por ejemplo: $nota = 0.5 + 3horas + u$

Por si acaso... otra vez lo mismo

- Supongamos que conocemos los parámetros (en la realidad habrá que estimarlos): $\text{nota} = 0.5 + 3\text{horas} + u$

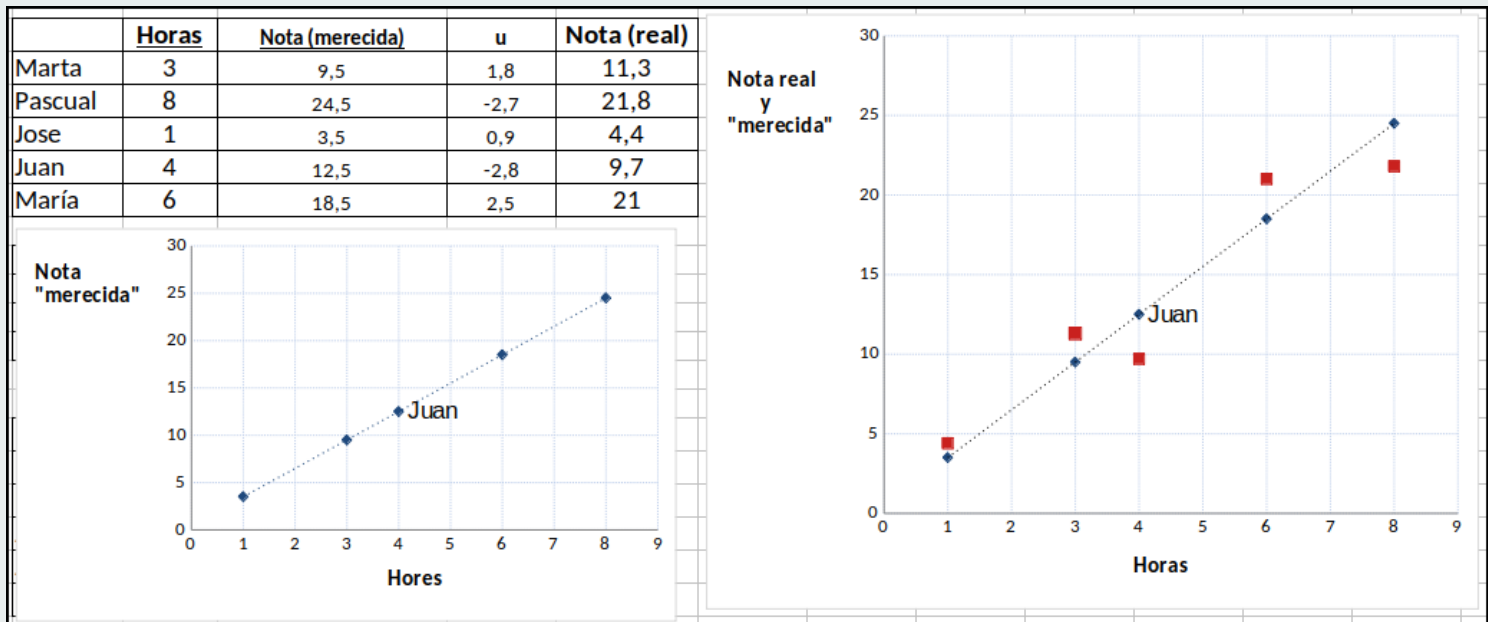


- ¿Qué es β_2 ? ¿Qué representa en nuestro modelo?
- ¿Cómo se interpreta β_1 ? ¿Qué nota sacará una persona que no estudie nada?
- ¿Qué nota sacará una persona que estudie 2 horas?
- ¿Qué podría estar dentro de u_i ?

Modelo Teórico (hipótesis...o racionalizando la procedencia de los datos)

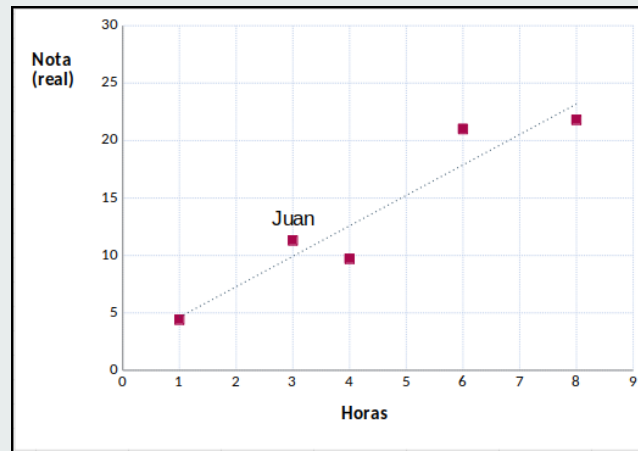
- Supongamos (otra vez) que conocemos los parámetros:

$$nota = 0.5 + 3horas + u$$



En la realidad ...

- En la realidad sólo observaremos y y x . No observaremos por separado sus “dos componentes hipotetizados”, no observaremos ni “la nota que se merece” ni la perturbación (u).
- Gráficamente ... los datos reales no estarán perfectamente alineados



¿Cómo estimar los parámetros?

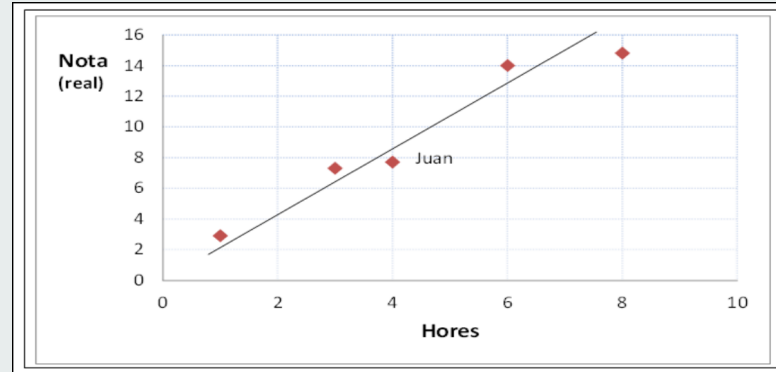
- ¿Cómo hallar/recuperar/**estimar** los valores de β_1 y β_2 ? Podemos pensar que los datos realmente están generados por el modelo $y_i = \beta_1 + \beta_2 x_i + u_i$, y pensar que y es explicado fundamentalmente por x , mientras que el componente aleatorio (u) es de menor importancia.

2.2 Estimación de los coeficientes por MCO

MCO: mínimos cuadrados ordinarios

Estimando los β ¡a ojo!

- Como una primera aproximación se podría trazar una recta que **pasase lo más cerca posible** de las observaciones.... pero no es un criterio muy científico



¿Qué recta elegimos?

- **¿Qué recta elegimos?** En realidad me estoy preguntando cuales son los “mejores” valores para los parámetros β_1 y β_2
- Parece lógico elegir la recta que esté más próxima a los datos. ¿Más próxima?
- **Hay que definir un criterio**

¿Cómo elegimos la recta?

- En realidad ¿cómo estimamos β_1 y β_2 ?
- En abstracto lo tenemos claro: elegiremos como estimaciones de β aquellos valores que seleccionen la recta que se aproxime más a los datos observados; es decir, **la que menos se equivoque**.
- De momento no vamos a estimar sino que vamos a elegir una recta que suponemos la más próxima. A esta recta más próxima a los datos la representamos como: $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$

ahora tenemos 2 modelos

- Fijaros que ahora tenemos 2 ecuaciones, tenemos 2 "modelos":
 - **Modelo teórico:** $y = \beta_1 + \beta_2 x + u$
 - **Modelo estimado:** $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$

Ahora tenemos más elementos

- El **modelo estimado** (o recta de regresión): $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$
 - \hat{y} : la **y-estimada**
 - $\hat{\beta}_1$ y $\hat{\beta}_2$: los "**beta-estimados**"

... y va a aparecer una tercera "ecuación"

- Definimos $\hat{u} = y - \hat{y}$ y lo llamamos **residuo**.

Las 3 "ecuaciones"

- **Modelo teórico:** $y_i = \beta_1 + \beta_2 x_i + u_i$
- **Modelo estimado:** $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$
- **Residuos:** $\hat{u}_i = y_i - \hat{y}_i$

Hay que entenderlo

$\text{Nota}_i = \beta_1 + \beta_2 \text{ horas}_i + u_i$					
$\hat{\text{Nota}}_i = 0.6 + 2.1 \text{ horas}_i$	<u>i</u>	<u>Hores</u>	<u>Nota (real)</u>	<u>Nota estimada</u>	<u>Residus</u>
	Ariadna	3	7.3		
	Gregori	8	14.8	17,4 (=0,6+2,1*8)	-2.4 (= 14,8-17,4)
	Albert	1	2.9		
	Joan	4	7.7		
	Maria	6	14		

Hay que “reconocer” cada uno de los elementos de nuestro esquema

- **Modelo teórico** versus **modelo estimado** ¿Cuál es observable? ¿Cuándo?
- **Y-real** (y) vs. **Y-estimada** (\hat{y}) ¿Cómo se interpretan? ¿Son observables?
- **Perturbación** (u) vs. **residuo** (\hat{u}) ¿Qué son? ¿son observables?
- **Parámetros** (β) vs **estimaciones/estimadores** ($\hat{\beta}$)

Otra vez... esto es básico para entender la materia

- Para cada observación (o individuo i) hay un valor predicho por el modelo estimado (\hat{y}_i) y un error de estimación (\hat{u}_i).
- **No confundir** datos reales (x_i, y_i) con datos predichos (\hat{y}_i)
- **No confundir** perturbaciones (u_i) con residuos (\hat{u}_i).

Ok, pero

- Aún no hemos dicho cómo vamos a obtener el **modelo estimado**, cómo vamos a obtener estimaciones de β_1 y β_2
- Sí, sabemos que tenemos que elegirlos/obtenerlos de forma que la recta estimada se equivoque lo menos posible, que pase **lo más cerca posible** de los datos reales
- Vamos a ello

¿Cómo estimamos los parámetros del modelo teórico?

- ¿Cómo obtenemos los $\hat{\beta}$? Es decir, ¿cómo obtenemos el modelo estimado?
- Una vez que hemos definido los residuos (\hat{u}) ya podemos establecer un criterio: nuestra recta-estimada será aquella que haga lo más pequeños posibles las equivocaciones o residuos.
- ¿Por qué el siguiente criterio no es válido?

$$\text{Min} \sum_{i=1}^N \hat{u}_i = \text{Min} (\hat{u}_1 + \hat{u}_2 + \hat{u}_3 + \dots \hat{u}_N)$$

- El criterio que vamos a utilizar consistirá en **Minimizar la suma de los residuos al cuadrado** (MCO):

$$\text{Min} \sum_{i=1}^N \hat{u}_i^2 = \text{Min} (\hat{u}_1^2 + \hat{u}_2^2 + \hat{u}_3^2 + \dots \hat{u}_N^2)$$

- La suma de los cuadrados de los residuos muchas veces se expresa como **SCR**

Estimación del MRLS con el método MCO (mínimo-cuadrático)

- **Criterio MCO:** $Min \sum \hat{u}_i^2$
 - OK, pero como $\hat{u}_i = y_i - \hat{y}_i$
 - Y como además $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$
 - Al final tenemos que: $\hat{u}_i = y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i) = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i$
- Con lo que al final el criterio MCO puede expresarse como:

$$Min \sum_{i=1}^N (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$$

- Para obtener el mínimo de una expresión hemos de igualar la primera derivada a cero
- ¿Sobre que variables tenemos que derivar la expresión anterior?

Derivando para obtener el mínimo de SCR

Derivamos $\sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$ respecto a $\hat{\beta}_1$ y $\hat{\beta}_2$

las derivadas quedan como

$$\frac{\partial SCR}{\partial \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)$$

$$\frac{\partial SCR}{\partial \hat{\beta}_2} = -2 \sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) x_i$$

Igualamos las derivadas a cero

- $\sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$
- $\sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) x_i = 0$

- Queda arreglar (un poco) para obtener las **ecuaciones normales**

las ecuaciones normales

- $\sum y_i = N\hat{\beta}_1 + \sum \hat{\beta}_2 x_i$
- $\sum y_i x_i = \hat{\beta}_1 \sum x_i + \hat{\beta}_2 \sum x_i^2$

Resolvemos el sistema de ecuaciones normales

- Despejando $\hat{\beta}_1$ de la primera ecuación, obtenemos: $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$
- Sustituyendo $\hat{\beta}_1$ en la segunda ecuación normal, obtenemos:

$$\hat{\beta}_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

- Como veis, hemos obtenido, finalmente, **los estimadores MCO**, y ... podremos obtener estimaciones de los parámetros del modelo a partir de estadísticos básicos de las variables involucradas: medias, varianzas y covarianzas.

Estimadores versus estimaciones

- Los estimadores MCO:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$$\hat{\beta}_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

- Cuando (con una muestra de datos) obtengamos unos valores concretos para $\hat{\beta}_1$ y $\hat{\beta}_2$, entonces tendremos **estimaciones MCO**.

Ejemplo de estimación (a mano!!)

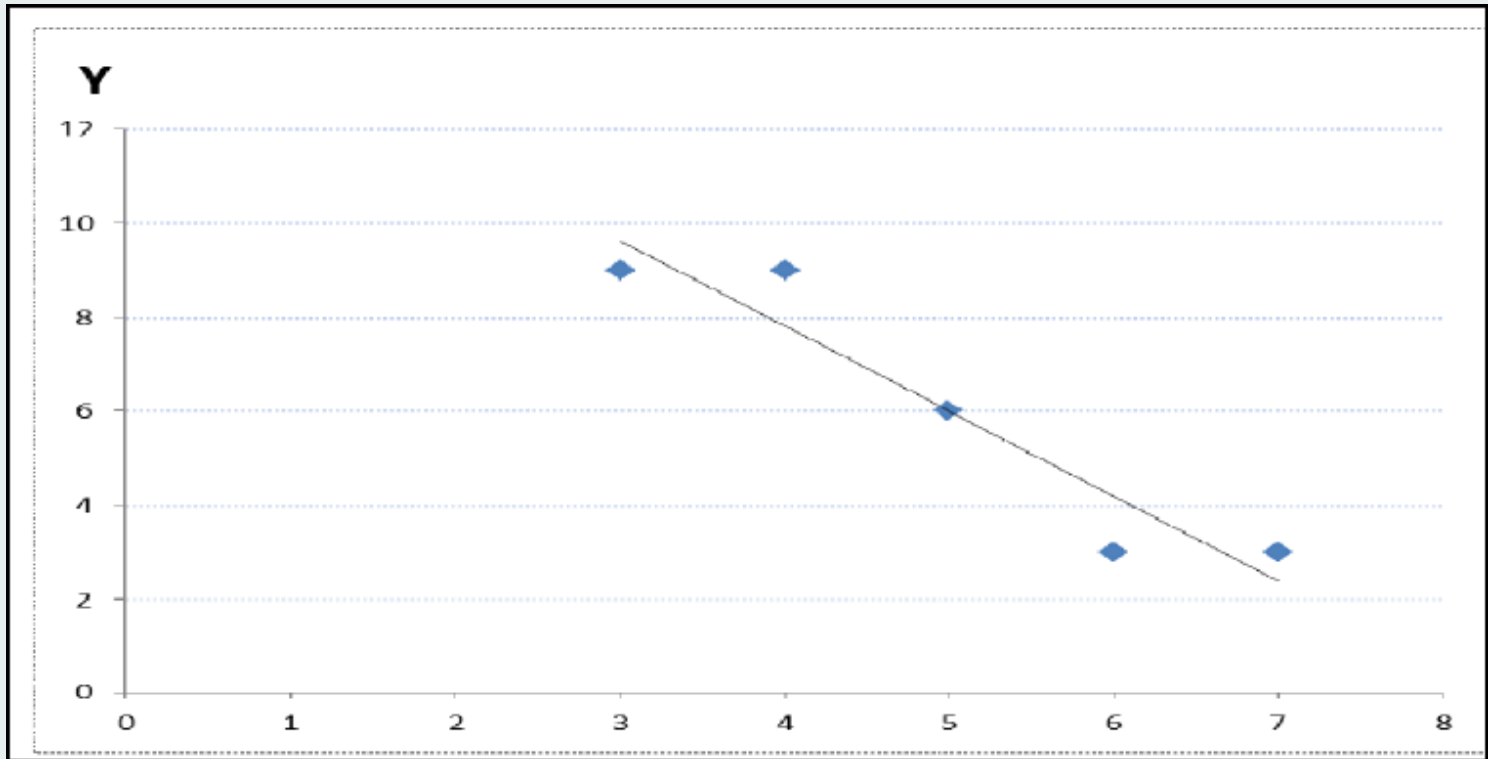
	X	Y	$(X - \bar{X})$	$(X - \bar{X})^2$	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
Juan	5	6	0	0	0	0	0
Carlos	7	3	2	4	-3	9	-6
Susana	4	9	-1	1	3	9	-3
Veronica	6	3	1	1	-3	9	-3
Andrea	3	9	-2	4	3	9	-6
<i>Suma</i>	25	30	0	10	0	36	-18
Media	5	6		2		7,2	-3,6

- ¿podéis obtener los estimadores?
- ¿y las estimaciones?
- ¿podéis obtener la \hat{y}_i ?
- ¿y los residuos (\hat{u}_i)?

Claro que podemos!!!!

Claro que podemos!!!!

- El modelo o recta estimada es: $\hat{y}_i = 15 - 1.8x_i$
- Gráficamente:



Obtenemos los valores de \hat{y}_i y los residuos (\hat{u}_i)

- $\hat{y}_i = 15 - 1.8x_i$
- Los valores de \hat{y}_i y los residuos (\hat{u}_i)

	X	Y	\hat{Y}	\hat{u}
Juan	5	6	6	0
Carlos	7	3	2,4	0,6
Susana	4	9	7,8	1,2
Veronica	6	3	4,2	-1,2
Andrea	3	9	9,6	-0,6

Estimamos el mismo modelo, con los mismos datos, pero ahora con Gretl

Modelo 1: MCO, usando las observaciones 1-5

Variable dependiente: y

	coeficiente	Desv. típica	Estadístico t	valor p	
-----	-----	-----	-----	-----	
const	15,0000	1,80000	8,333	0,0036	***
x	-1,80000	0,346410	-5,196	0,0138	**
Media de la vble. dep.	6,000000	D.T. de la vble. dep.	3,000000		
Suma de cuad. residuos	3,600000	D.T. de la regresión	1,095445		
R-cuadrado	0,900000	R-cuadrado corregido	0,866667		
F(1, 3)	27,00000	Valor p (de F)	0,013847		
Log-verosimilitud	-6,273432	Criterio de Akaike	16,54686		
Criterio de Schwarz	15,76574	Crit. de Hannan-Quinn	14,45040		

- ¿Recordáis cómo se interpretaban las estimaciones? Claro!!

PRÁCTICA: datos_bolmad_95

datos_bolmad_95.gdt

ID #	Nombre de variable	Etiqueta descriptiva
0	const	
1	BOOKVAL	valor contable de la empresa
2	MARKTVAL	valor de mercado de la empresa

gretl: modelo 1

Archivo Editar Contrastes Guardar Gráficos Análisis LaTeX

Modelo 1: MCO, usando las observaciones 1-161
Variable dependiente: MARKTVAL

	coeficiente	Desv. típica	Estadístico t	valor p
const	60,7661	16,8307	3,610	0,0004 ***
BOOKVAL	0,483537	0,0548024	8,823	1,89e-15 ***

Media de la vble. dep.	103,5860	D.T. de la vble. dep.	248,7958
Suma de cuad. residuos	6648589	D.T. de la regresión	204,4872
R-cuadrado	0,328690	R-cuadrado corregido	0,324468
F(1, 159)	77,85022	Valor p (de F)	1,89e-15
Log-verosimilitud	-1084,044	Criterio de Akaike	2172,088
Criterio de Schwarz	2178,251	Crit. de Hannan-Quinn	2174,591

- Interpreta las estimaciones MCO de los parámetros

seguimos con datos_bolmad_95

datos_bolmad_95.gdt		
ID #	Nombre de variable	Etiqueta descriptiva
0	const	
1	BOOKVAL	valor contable de la empresa
2	MARKTVAL	valor de mercado de la empresa

gretl: mostrar datos		
	BOOKVAL	MARKTVAL
Alicante	8,02	29,99
Andalucia	39,96	89,91
Argentaria	474,76	612,44
Atlantico	47,90	64,66
Bankinter	99,53	172,18
BBV	568,67	892,82
Castilla	24,12	47,04
Central Hisp	604,97	429,53
Credito Bale	7,16	10,82
Exterior	232,18	325,06
Galicia	16,54	34,23
Guipuzcuano	20,51	30,97
Pastor	55,92	54,24
Popular	274,91	558,06

gretl: mostrar datos			
Rango de estimación del modelo: 1 - 161			
Desviación típica de la regresión = 204,487			
	MARKTVAL	Estimada	residuo
Alicante	29,99	64,64	-34,65
Andalucia	89,91	80,09	9,82
Argentaria	612,44	290,33	322,11
Atlantico	64,66	83,93	-19,27
Bankinter	172,18	108,89	63,29
BBV	892,82	335,74	557,08 *
Castilla	47,04	72,43	-25,39
Central Hisp	429,53	353,29	76,24
Credito Bale	10,82	64,23	-53,41
Exterior	325,06	173,03	152,03
Galicia	34,23	68,76	-34,53

- Calcula la y-estimada (\hat{y}_i) y el residuo (\hat{u}_i) para Bankinter.

2.3 Interpretación de los coeficientes

Hay que ser consciente de la diferencia conceptual entre parámetros y estimadores/estimaciones

los parámetros (β)

En el **Modelo teórico**: $y = \beta_1 + \beta_2 x + u$

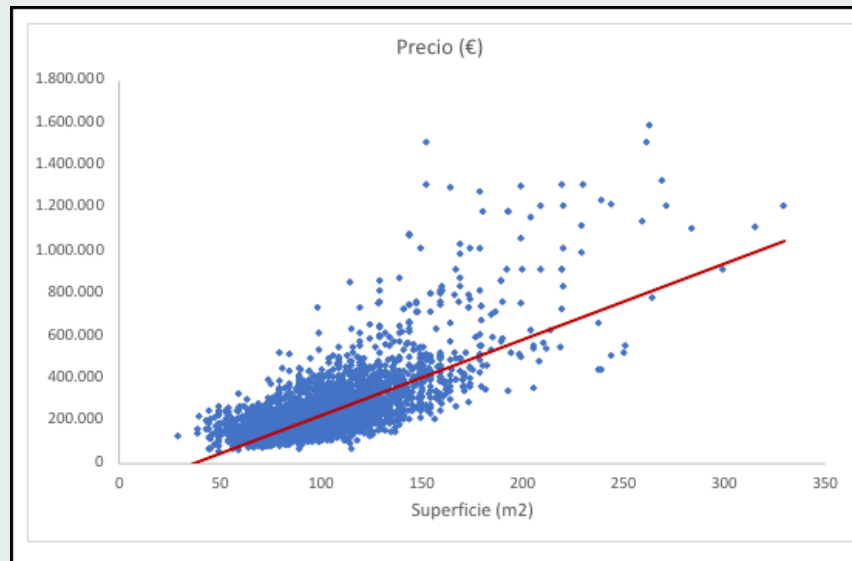
- β_1 es el término independiente.
 - Gráficamente sería la **ordenada en el origen**; es decir sería el valor de y si el resto de variables (x , u) fuesen cero.
 - Generalmente **no tiene** una interpretación con sentido económico o teórico.
- β_2 es el parámetro que acompaña a la variable x .
 - Gráficamente es la **pendiente**.
 - Matemáticamente es la derivada parcial de y respecto de x
 - **Económicamente** representa, o es, el **efecto marginal** de x sobre y ; es decir, indica cuantas unidades aumentaría y si, ceteris paribus, x aumentase en 1 unidad.

¿y las estimaciones ($\hat{\beta}$)?

En el **Modelo estimado**: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

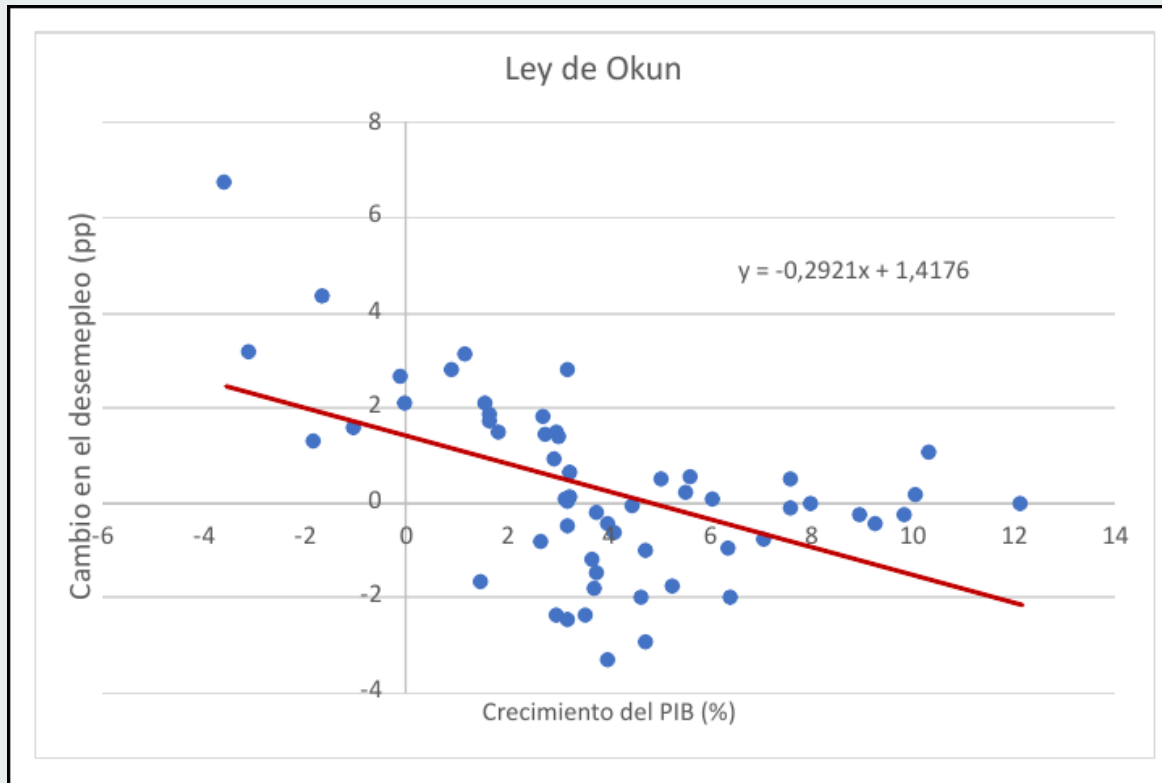
ejemplo para interpretar los $\hat{\beta}$

Modelo: $\text{precio} = \beta_1 + \beta_2 \text{superficie} + u$

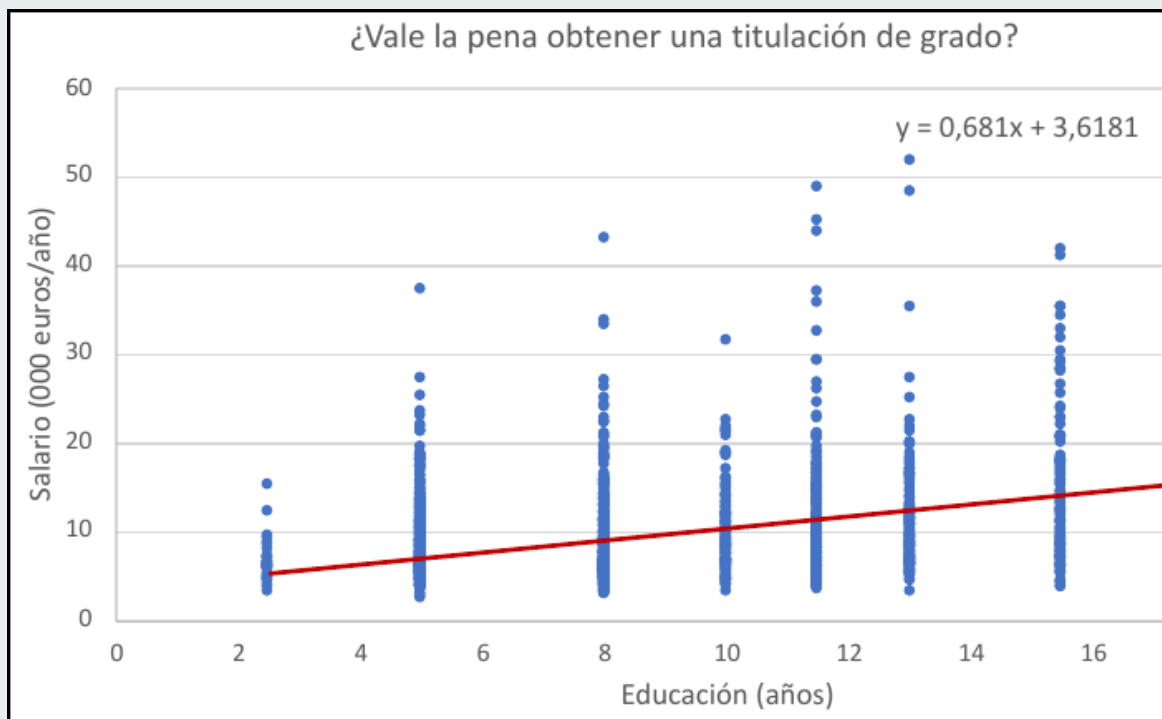


- Modelo estimado: $\hat{\text{precio}} = -135 + 3500 \text{ superficie}$
- la superficie (variable x) está medida en metros cuadrados
- el precio está medido en euros

otro ejemplo para interpretar los $\hat{\beta}$



otro ejemplo



2.4 Propiedades descriptivas del modelo de regresión

No confundir con las propiedades probabilísticas

Propiedades descriptivas

- Si se estima un MRL (con término independiente) por MCO, entonces, **necesariamente** se cumple lo siguiente:
 - 1) La suma de los residuos MCO es cero: $\sum \hat{u}_i = 0$
 - 2) La recta de regresión MCO pasa por el punto de medias muestrales (\bar{y}, \bar{x})
 - 3) La covarianza muestral entre regresor y residuos MCO es cero: $Cov(x, \hat{u}) = 0$
 - 4) La covarianza muestral entre valores ajustados y residuos es cero: $Cov(\hat{y}, \hat{u}) = 0$
- A estas 4 propiedades se las conoce en el contexto del MRL como **propiedades descriptivas**

1ª propiedad descriptiva

- La suma de los residuos MCO es cero :

$$\sum_{i=1}^N \hat{u}_i = 0$$

Demostración

- En la sección 2.2, cuando estábamos obteniendo los estimadores MCO, concretamente en la página 21, teníamos que : $\sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$
- En esta ecuación ya está implícito que $\sum \hat{u}_i = 0$

De la 1ª propiedad se sigue que ...

- La media muestral de los residuos es nula ($\overline{\hat{u}} = 0$)
- $\overline{y} = \overline{\hat{y}}$

2ª propiedad descriptiva

La recta de regresión MCO pasa por el punto de medias muestrales (\bar{y}, \bar{x})

Demostración

- En la sección 2.2, concretamente en la página 22, obtuvimos las ecuaciones normales. la primera ecuación normal es:

$$\sum y_i = N\hat{\beta}_1 + \sum \hat{\beta}_2 x_i$$

- Dividiendo la 1ª ecuación normal por N queda : $\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}$

¿entendéis que quiere decir la 2ª propiedad?

3ª propiedad descriptiva

- La covarianza muestral entre regresor y residuos MCO es cero: $Cov(x, \hat{u}) = 0$
 - como $\overline{\hat{u}} = 0$, el hecho de que $Cov(x, \hat{u}) = 0$, es equivalente a :
$$\sum x_i \hat{u}_i = 0$$

Demostración

- En la sección 2.2, cuando estábamos obteniendo los estimadores MCO, concretamente en la página 21 cuando igualábamos a cero las primeras derivadas, teníamos que : $\sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) x_i = 0$
- Esta ecuación implica que $\sum \hat{u} x = 0$

¿entendéis que quiere decir, que implica, la 3ª propiedad descriptiva?

El MRL plantea que y es explicado por los regresores (x) y la perturbación (u). El modelo consigue explicar una parte del regresando, la \hat{y} pero deja otra parte sin explicar, \hat{u} . La parte de y que no es explicada por x es porque no está relacionada con x ; es decir, los residuos no están correlacionados muestralmente con x

4ª propiedad descriptiva

- La covarianza muestral entre valores ajustados y residuos es cero:
 $Cov(\hat{y}, \hat{u}) = 0$
 - como $\overline{\hat{u}} = 0$, el hecho de que $Cov(\hat{y}, \hat{u}) = 0$, es equivalente a :
 $\sum \hat{y}_i \hat{u}_i = 0$

Demostración

$$\sum \hat{u}_i \hat{y}_i = \sum \hat{u}_i (\hat{\beta}_1 + \hat{\beta}_2 x_i) = \hat{\beta}_1 \sum \hat{u}_i + \hat{\beta}_2 \sum \hat{u}_i x_i = 0 + 0 = 0$$

¿entendéis que quiere decir, que implica, la 4ª propiedad descriptiva?

De la definición de residuo se sigue que $y_i = \hat{y}_i + \hat{u}_i$; es decir, al estimar un MRL por MCO descomponemos y en dos componentes: esos dos componentes están incorrelados en la muestra.

2.5 Medidas de bondad de ajuste: coeficiente de determinación

Coeficiente de determinación o R^2

Bondad del ajuste

- Podemos pensar que el propósito del análisis de regresión es explicar el comportamiento de la variable dependiente o regresando (y).
- Una vez estimamos por MCO un MRL, nos interesaría tener una medida que nos informase del grado de ajuste entre el modelo y los datos: un estadístico que nos informe de la bondad del ajuste.

Construyendo R^2 . (Parte explicada vs. No-explicada)

- El modelo estimado no explica o predice perfectamente las observaciones de y , si no que comete errores. Explica una parte (\hat{y}) pero deja una parte sin explicar, llamada residuos (\hat{u})
- Es decir, tras la estimación, el modelo estimado descompone cada valor de la variable endógena (y) en dos partes, la parte explicada por el modelo (\hat{y}) y la parte que el modelo no consigue explicar: el residuo (\hat{u}). Además, **la parte explicada y la no explicada están incorreladas en la muestra** (4a propiedad descriptiva).

la varianza de una suma

- $y = \hat{y} + \hat{u}$
- $Var(y) = Var(\hat{y} + \hat{u}) = Var(\hat{y}) + Var(\hat{u}) + 2Cov(\hat{y}, \hat{u})$
- Como $Cov(\hat{y}, \hat{u}) = 0$, entonces: $Var(y) = Var(\hat{y}) + Var(\hat{u})$

Definiendo R^2

$$R^2 = \frac{Var(\hat{y})}{Var(y)}$$

- R^2 se interpreta como el porcentaje (en tanto por uno) de la varianza total de y que es explicada por el modelo.
- Para entenderlo, imagina que $Var(y) = 100$ y que $Var(\hat{y}) = 70$. En ese caso $R^2 = 0.7$; lo que indica que el modelo explica el 70% de la varianza de y .
- En el ejemplo anterior, si $Var(y) = 100$ y $Var(\hat{y}) = 70$, entonces la varianza residual ($Var(\hat{u})$) tendrá un valor de 30 ; de forma que el R^2 también se puede calcular como: $R^2 = 1 - \frac{Var(\hat{u})}{Var(y)}$

Generalmente en los libros ...

- Generalmente las expresiones que parecen en los libros de Econometría para R^2 emplean, en lugar de las varianzas, los numeradores de las varianzas.

- Estos numeradores son:

- $SCT = \sum (y_i - \bar{y})^2$

- $SCE = \sum (\hat{y}_i - \bar{\hat{y}})^2$

- $SCR = \sum (\hat{u}_i - \bar{\hat{u}})^2$. Como resulta que $\bar{\hat{u}} = 0$, entonces la suma de residuos al cuadrado finalmente queda como: $SCR = \sum \hat{u}_i^2$

- De forma que, si en lugar de las varianzas, usamos los numeradores de las varianzas:

$$R^2 = \frac{SCE}{SCT} , \text{ o (si usamos la segunda expresión) } R^2 = 1 - \frac{SCR}{SCT}$$

- En cualquier caso la interpretación de R^2 no cambia.

Interpretación y propiedades de R^2

- $(R^2 * 100)$ es el porcentaje de de la variación muestral de y que viene explicada por el modelo; es decir, si R^2 fuese 0,8, el modelo explicaría el 80% de la varianza de y .
- R^2 mide hasta qué punto el modelo o “recta” de regresión ajusta bien a los datos muestrales.
- El coeficiente de determinación R^2 permite valorar la capacidad explicativa del modelo.
- El valor de R^2 está acotado en $[0,1]$.
- si $R^2 = 1$ la recta de regresión se ajusta perfectamente a los datos, por lo tanto todos los residuos son cero. **¿Qué tiene que pasar para que $R^2 = 1$?**
- si $R^2 = 0$, o cerca de cero, hay un pobre ajuste: la variación de y está "poco" representada o explicada por la recta de regresión. **¿Qué tiene que pasar para que $R^2 = 0$?**
- Si el MRL que hemos estimado no tuviese término independiente, entonces R^2 puede tomar valores negativos

Más sobre R^2

- R^2 es el cuadrado del coeficiente de correlación entre x e y , pero **sólo** en el modelo de regresión lineal simple.
- Como queremos modelos con alto poder explicativo, a igualdad de otros factores elegiremos modelos con elevados R^2 ; es decir, R^2 se puede usar para comparar modelos pero solo si tienen la misma variable endógena.
- Un R^2 bajo no significa que el modelo estimado no proporcione estimaciones fiables de los efectos parciales.
- La bondad del ajuste no es la única característica que debemos valorar en una ecuación de regresión
- R^2 no disminuye cuando añadimos variables explicativas, normalmente aumenta (lógico?!)
- Por lo tanto, R^2 no es de utilidad para decidir si incluimos variables. Incluiremos más variables en el modelo si el efecto parcial de esa variable es no nulo (contraste de hipótesis).
- R^2 no varía al hacer cambios de escala u origen en x o y .