Tema 2: Estadística descriptiva bivariante

Tema 2: Estadística descriptiva bivariante

1. Introducción
2. Tablas de frecuencias bivariantes.
3. Gráficos de dispersión.
4. Medidas de relación lineal
5. La recta de regresión simple.

Tablas de frecuencias bivariantes

Tablas bivariantes

Si tenemos, para cada individuo, dos variables usamos una tabla de doble entrada

Ejemplo: Para cada coche tenemos el número de cilindros y su año de fabricación (cardata.xls)

	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
4	17	12	25	22	28	104
5	1	1	1	0	0	3
6	12	6	2	7	3	30
8	6	10	0	1	0	17
Column Total	36	29	29	30	31	155

Tablas de frecuencias bivariantes

Tablas bivariantes

Si tenemos, para cada individuo, dos variables usamos una tabla de doble entrada

Ejemplo: Para cada coche tenemos el número de cilindros y su año de fabricación (cardata.xls)

	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
4	17	12	25	22	28	104
5	1	1	1	0	0	3
6	12	6	2	7	3	30
8	6	10	0	1	0	17
Column Total	36	29	29	30	31	155

Cada celda: Frecuencias conjuntas, n_{ij}

Tablas de frecuencias bivariantes

Tablas bivariantes

Ejemplo: Para cada coche tenemos el número de cilindros y su año de fabricación (cardata.xls)

	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
4	17	12	25	22	28	104
5	1	1	1	0	0	3
6	12	6	2	7	3	30
8	6	10	0	1	0	17
Column Total	36	29	29	30	31	155

Las frecuencias univariantes:
Frecuencias marginales, n_{io} y n_{oj}

Tablas de frecuencias bivariantes

Tablas bivariantes

Ejemplo: para cada coche tenemos el número de cilindros y su año de fabricación (cardata.xls)

	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
4	17	12	25	22	28	104
5	1	1	1	0	0	3
6	12	6	2	7	3	30
8	6	10	0	1	0	17
Column Total	36	29	29	30	31	155

Cada fila o columna: Frecuencia condicionada

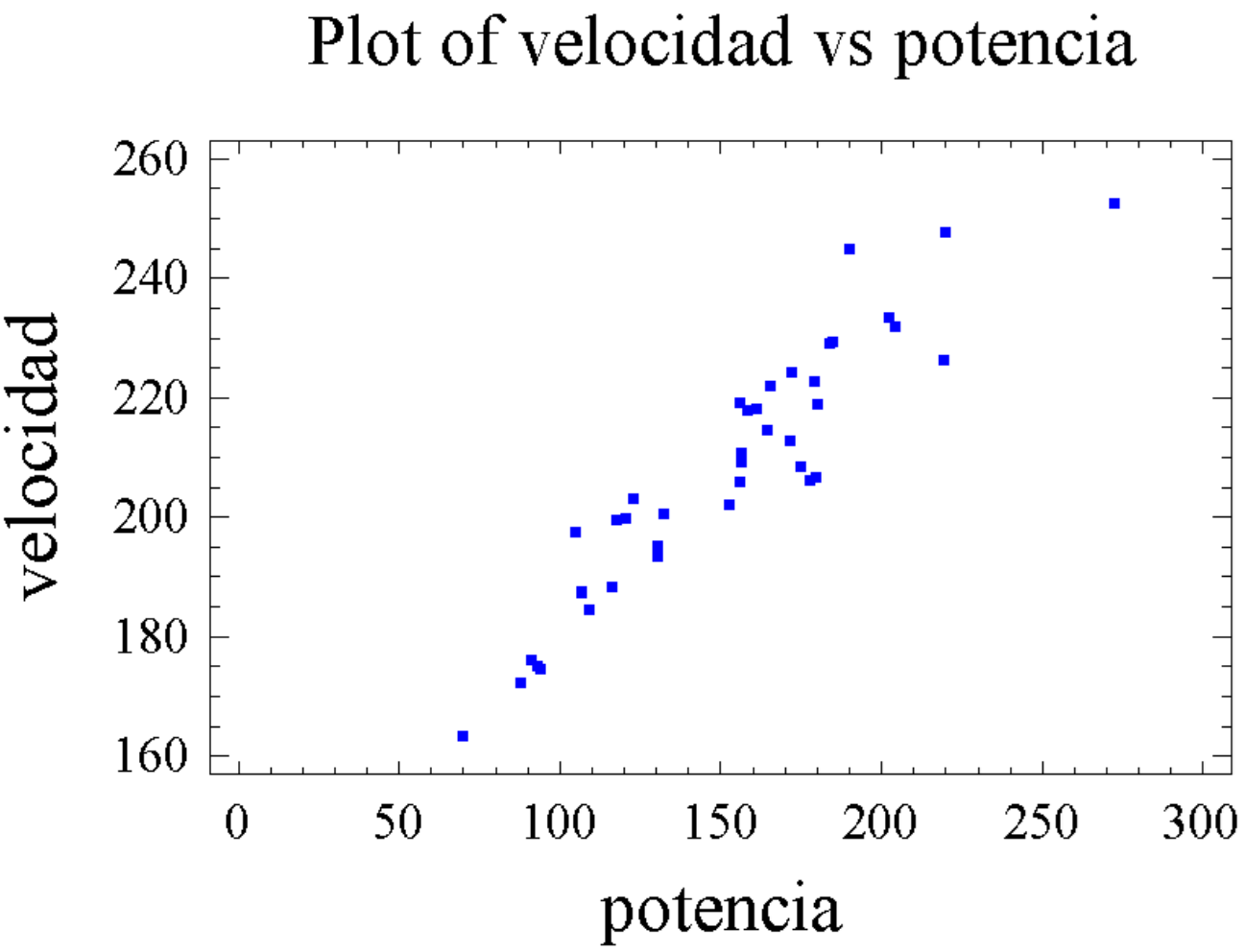
(al valor de la fila o columna), $n_{i|j}$ y $n_{j|i}$

Tema 2: Estadística descriptiva bivariante

1. Introducción
2. Tablas de frecuencias bivariantes.
3. Gráficos de dispersión.
4. Medidas de relación lineal
5. La recta de regresión simple.

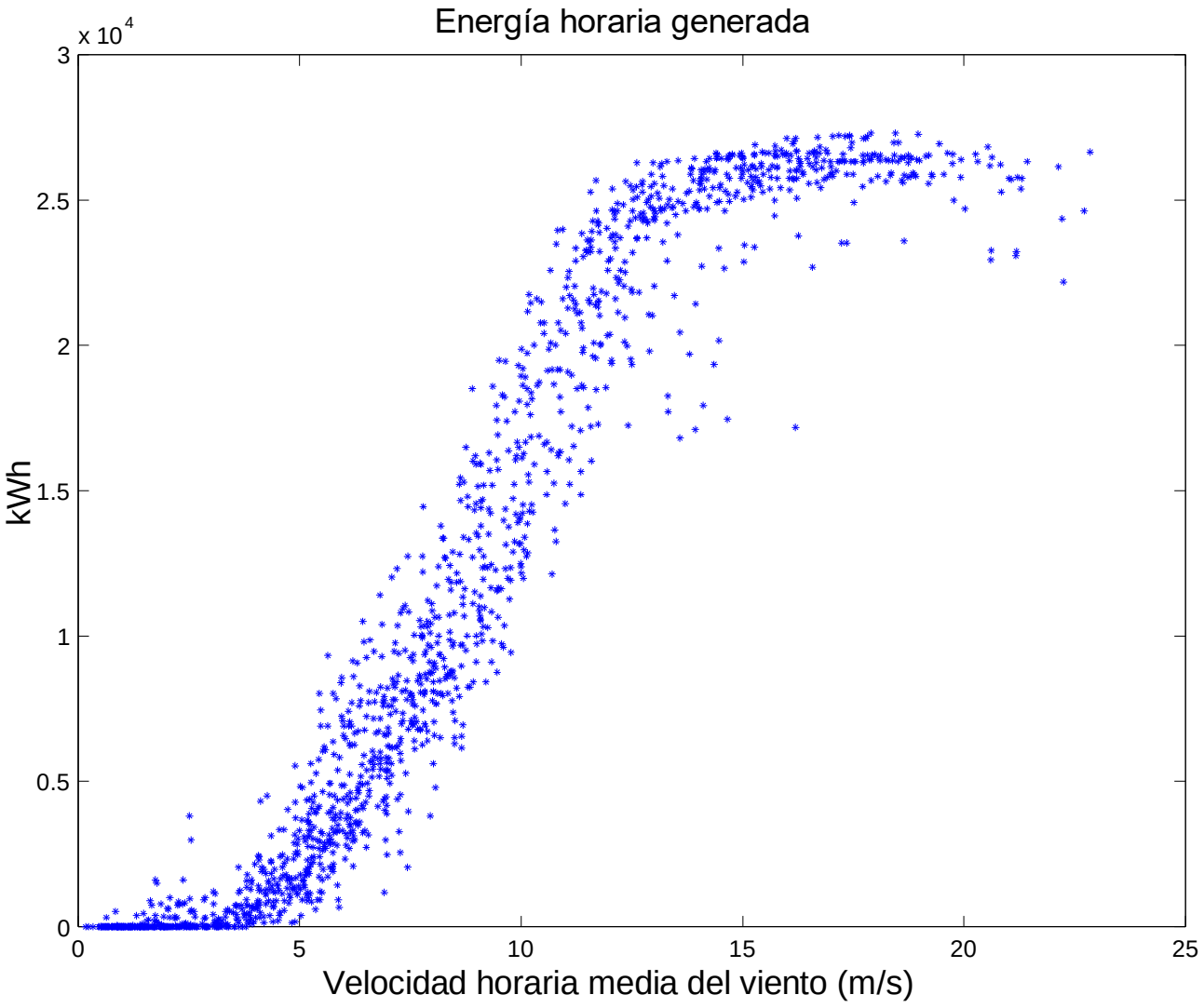
Gráfico de dispersión

Para cada individuo tenemos información de dos variables cuantitativas:



3.5 Gráfico de dispersión

Potencia generada por un molino de viento

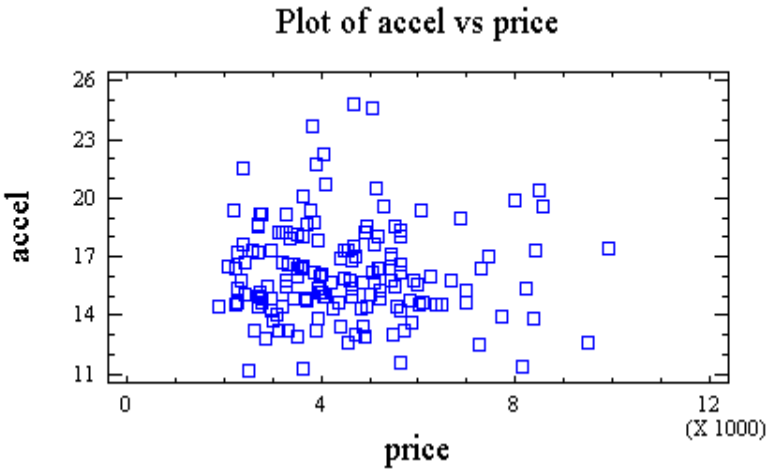


Tema 2: Estadística descriptiva bivariante

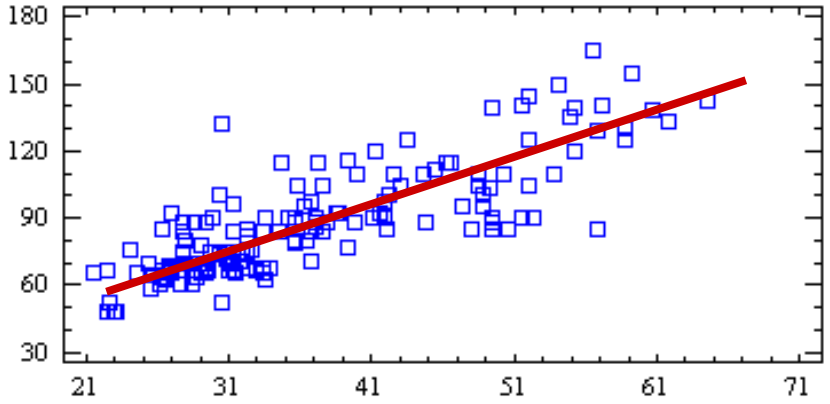
1. Introducción
2. Tablas de frecuencias bivariantes.
3. Gráficos de dispersión.
4. Medidas de relación lineal
5. La recta de regresión simple.

Medidas de dependencia lineal

- Coeficiente de covarianza
- Coeficiente de correlación



Entre estas variables no hay relación lineal



Entre estas variables hay relación lineal

La línea roja podría ser un buen resumen de esa relación

Para n individuos, tenemos datos de dos variables

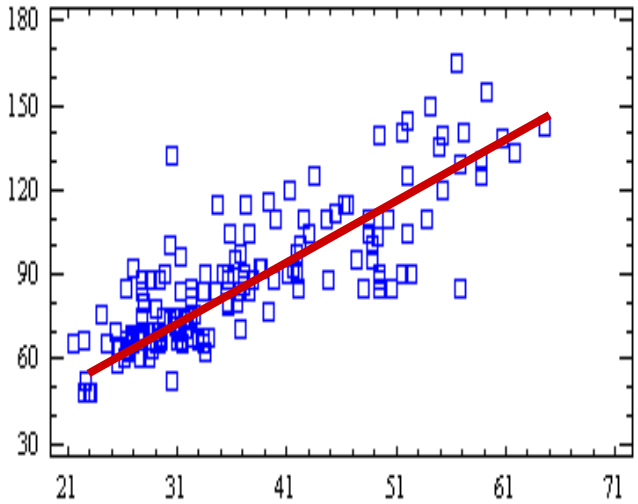
Individuo	**x**	**y**
1	x_1	y_1
2	x_2	y_2
:	:	:
n	x_n	y_n

Covarianza

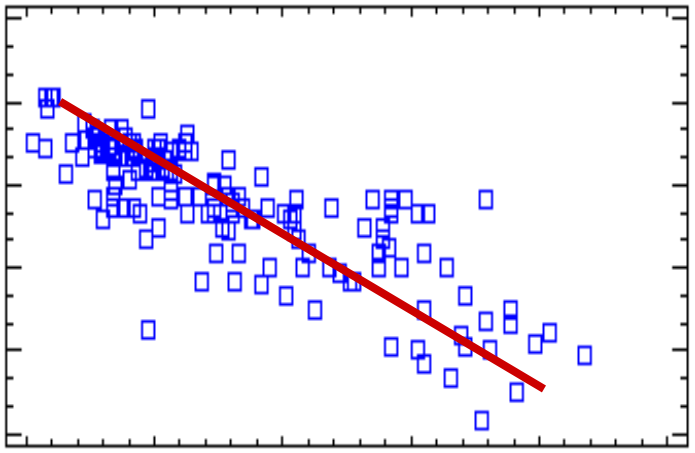
$$cov(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Correlación

$$r = r(x,y) = r_{xy} = \frac{s_{xy}}{s_x s_y}$$



Covarianza y correlación positivas



Covarianza y correlación negativas

Medidas de dependencia lineal

- Coeficiente de covarianza
- Coeficiente de correlación

Una forma habitual de presentar la información es en forma de matrices (simétricas)

Matriz de covarianzas

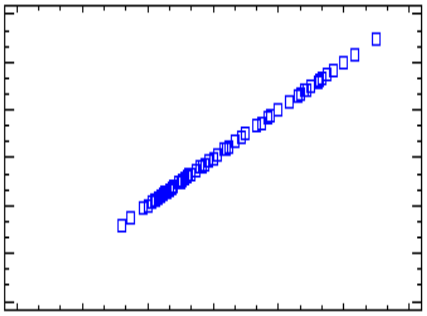
$$M = \begin{bmatrix} s_x^2 & \operatorname{cov}(x,y) \\ \operatorname{cov}(y,x) & s_y^2 \end{bmatrix}$$

Matriz de correlaciones

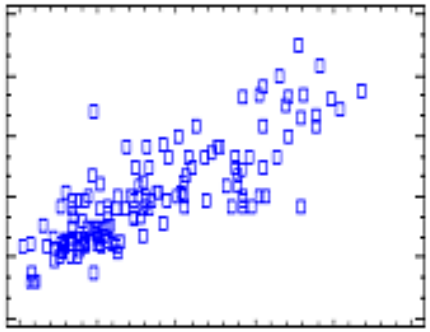
$$R = \begin{bmatrix} 1 & \operatorname{corr}(x,y) \\ \operatorname{corr}(y,x) & 1 \end{bmatrix}$$

$$cov(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \qquad r = r(x,y) = r_{xy} = \frac{s_{xy}}{s_x s_y}$$

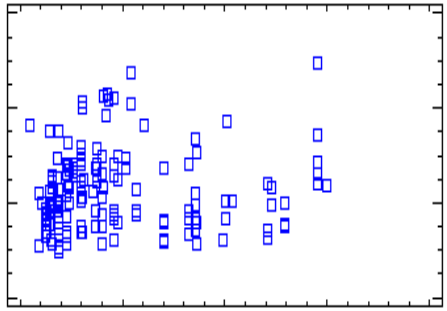
- La covarianza tiene unidades de medida (unidades_x)*(unidades_y)
- La correlación es adimensional. ES MÁS FÁCIL DE INTERPRETAR
- Se puede demostrar que -1 ≤ r ≤ 1



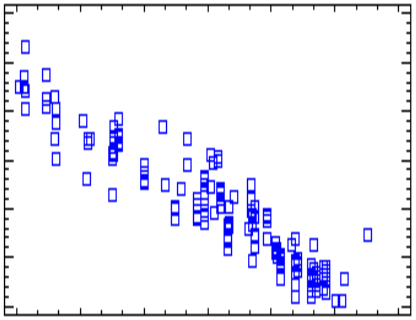
r=1



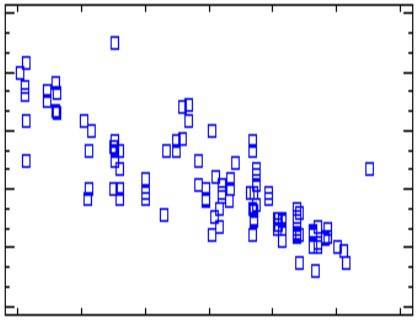
r=0.8



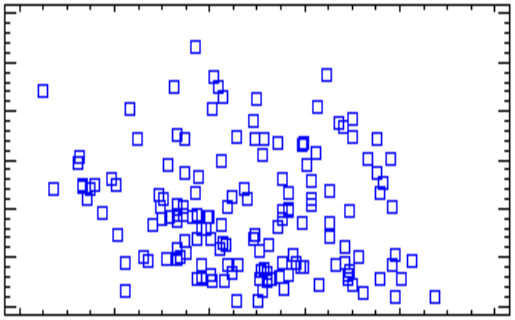
r=0.06



r=-0.94



r=-0.83



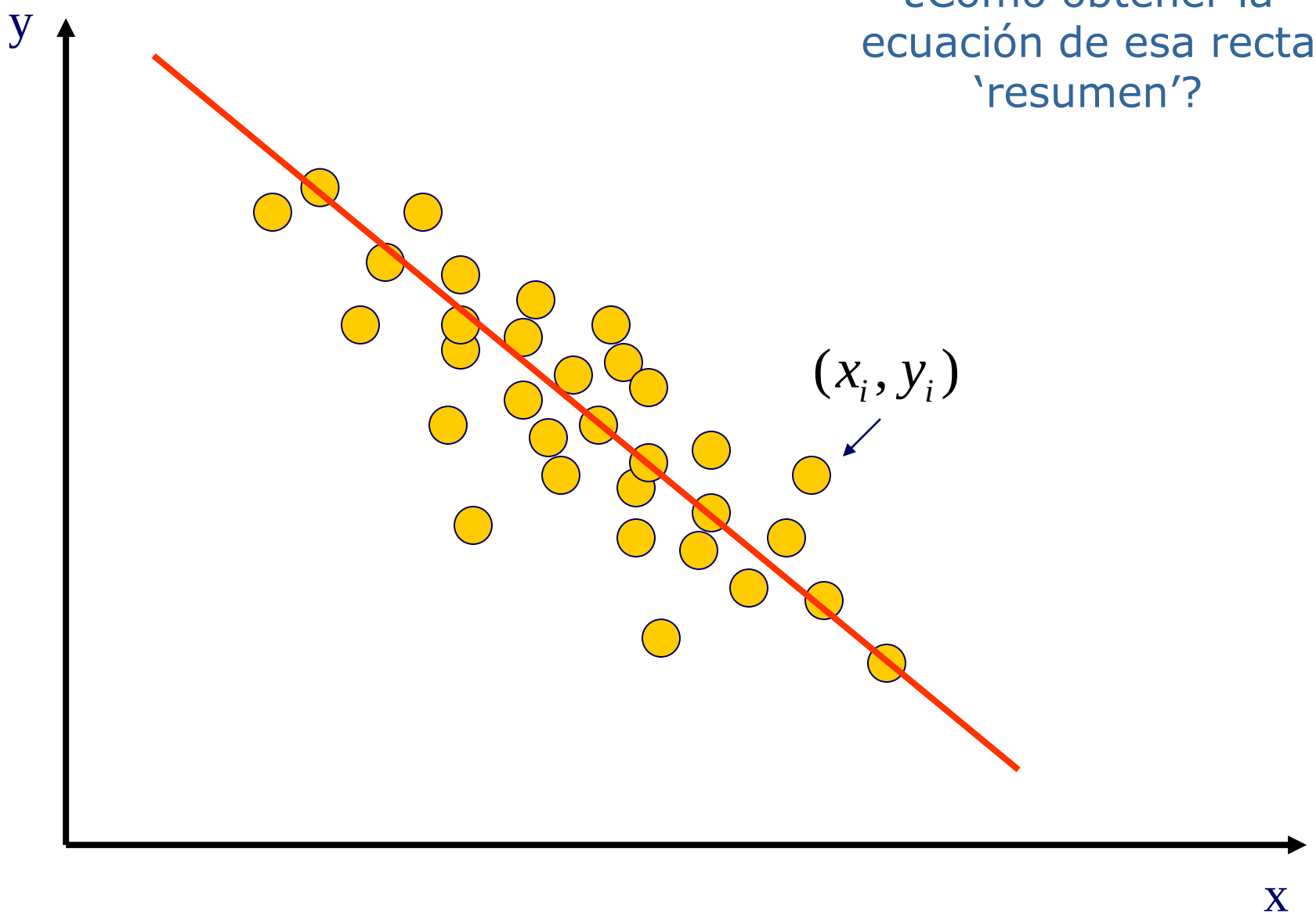
r=-0.08

Tema 2: Estadística descriptiva bivariante

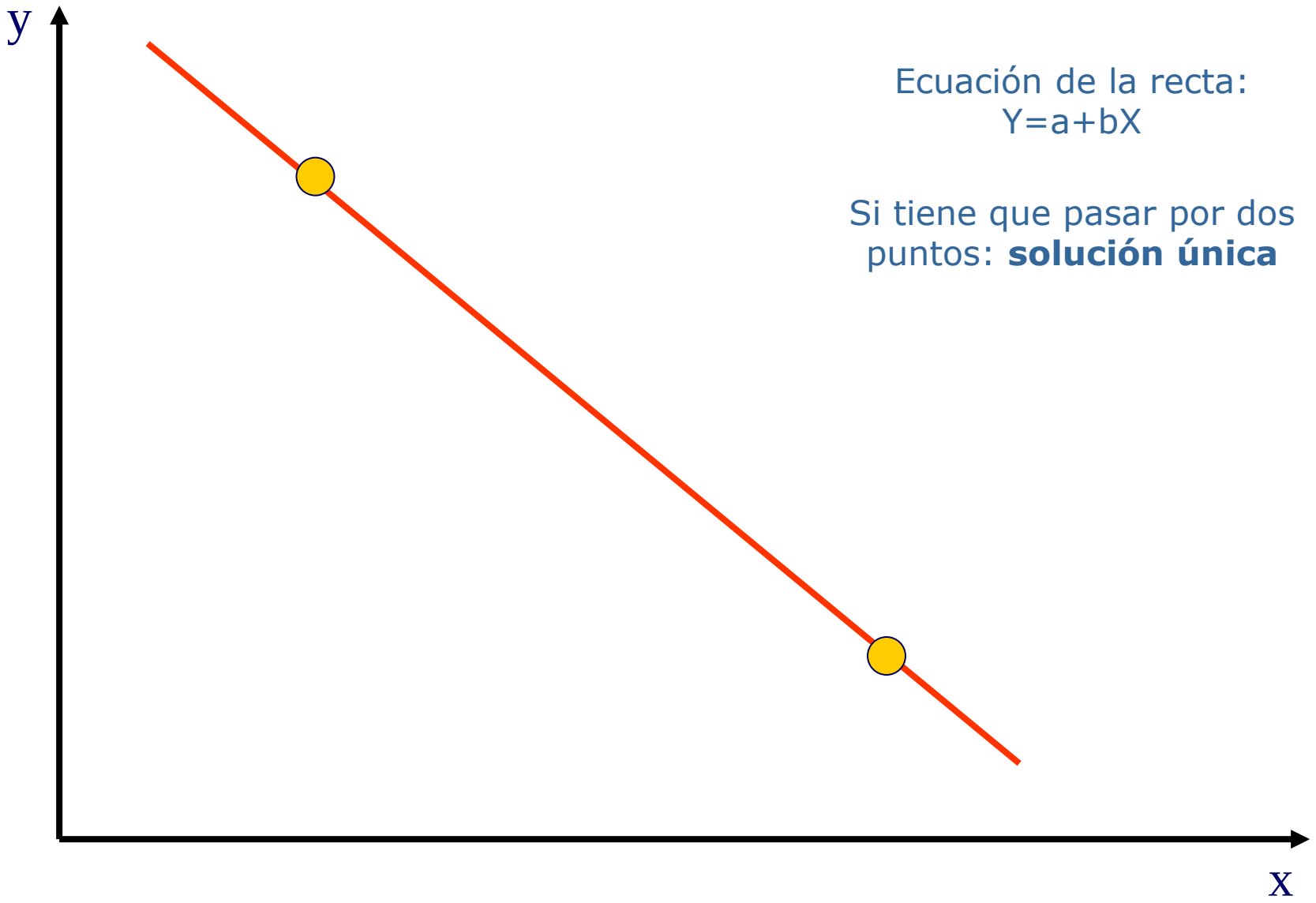
1. Introducción
2. Tablas de frecuencias bivariantes.
3. Gráficos de dispersión.
4. Medidas de relación lineal
5. La recta de regresión simple.

La recta de regresión

¿Cómo obtener la ecuación de esa recta 'resumen'?



La recta de regresión



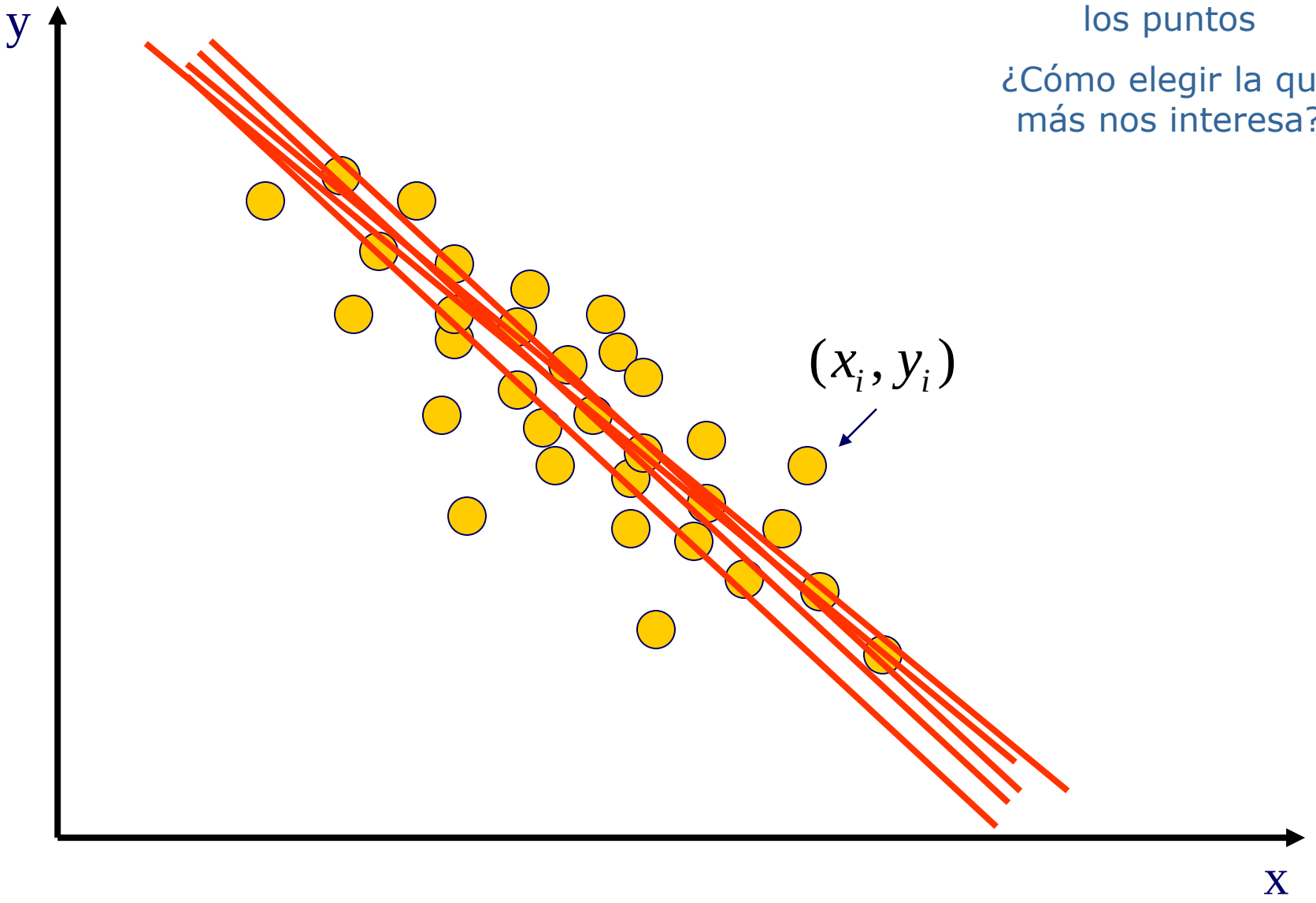
Ecuación de la recta:
Y=a+bX

Si tiene que pasar por dos puntos: **solución única**

La recta de regresión

Es imposible que una recta pase por todos los puntos

¿Cómo elegir la que más nos interesa?



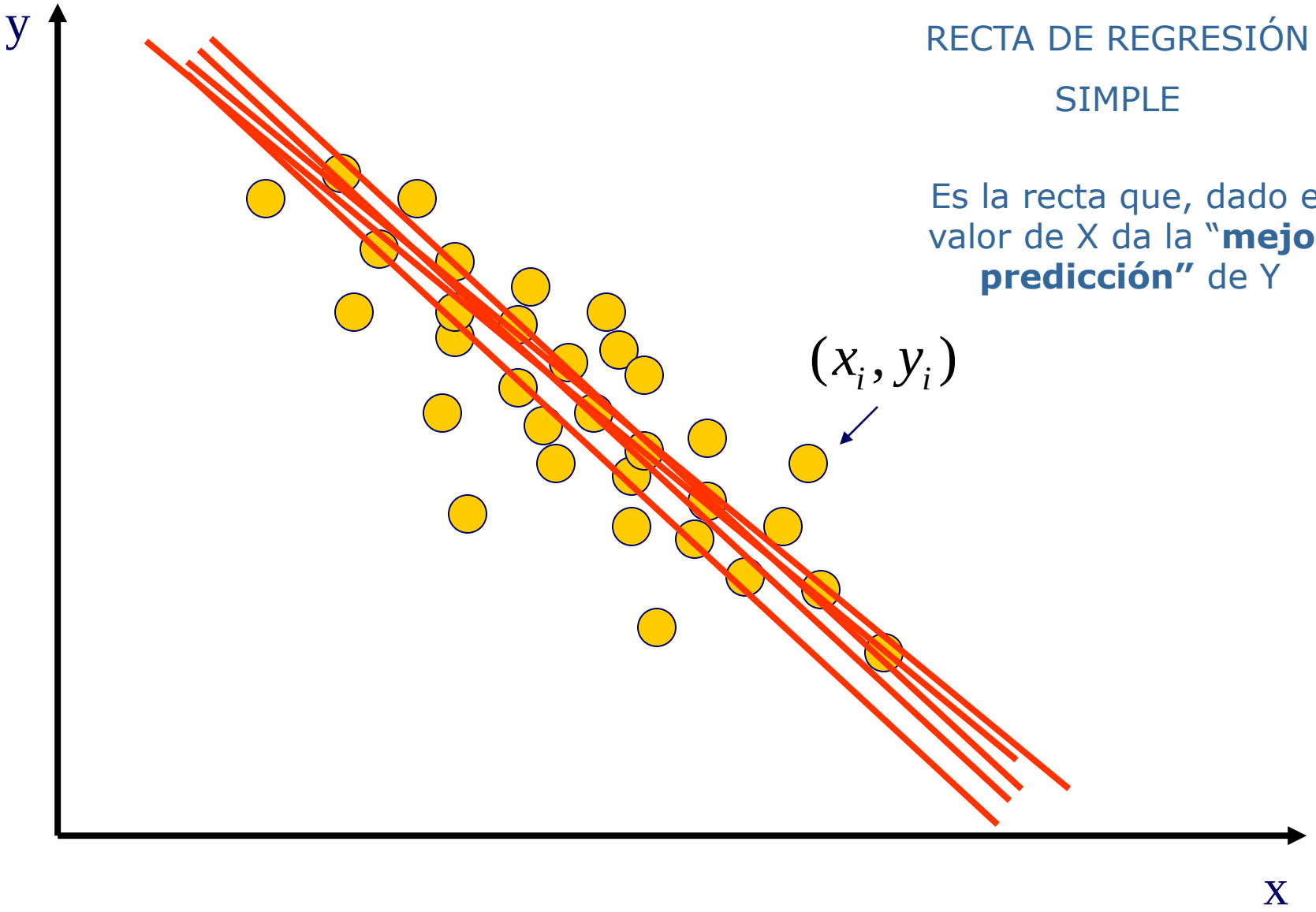
La recta de regresión

Buscamos una recta muy concreta llamada

RECTA DE REGRESIÓN

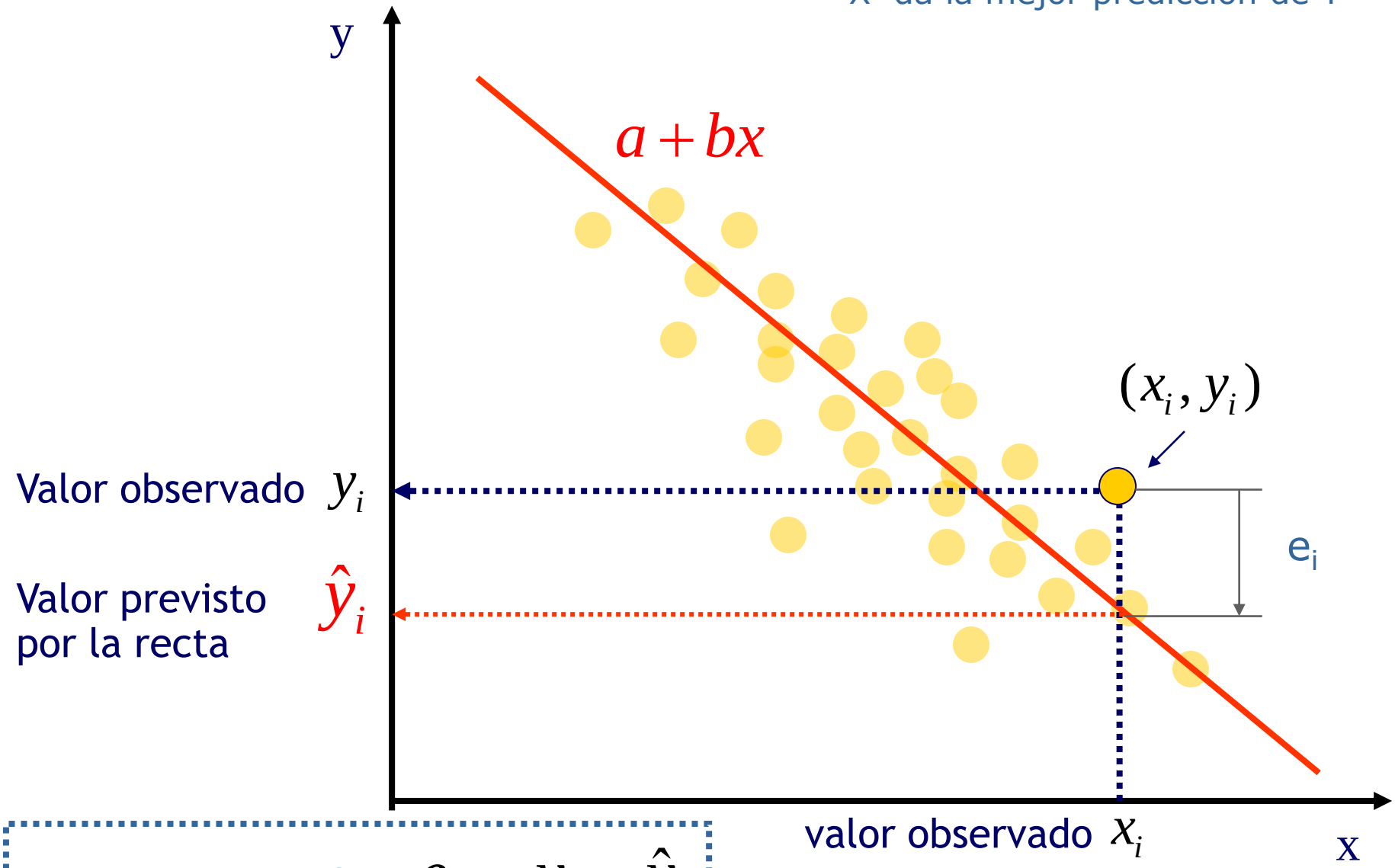
SIMPLE

Es la recta que, dado el valor de X da la "**mejor predicción**" de Y



La recta de regresión

Es la recta que, dado el valor de X da la mejor predicción de Y



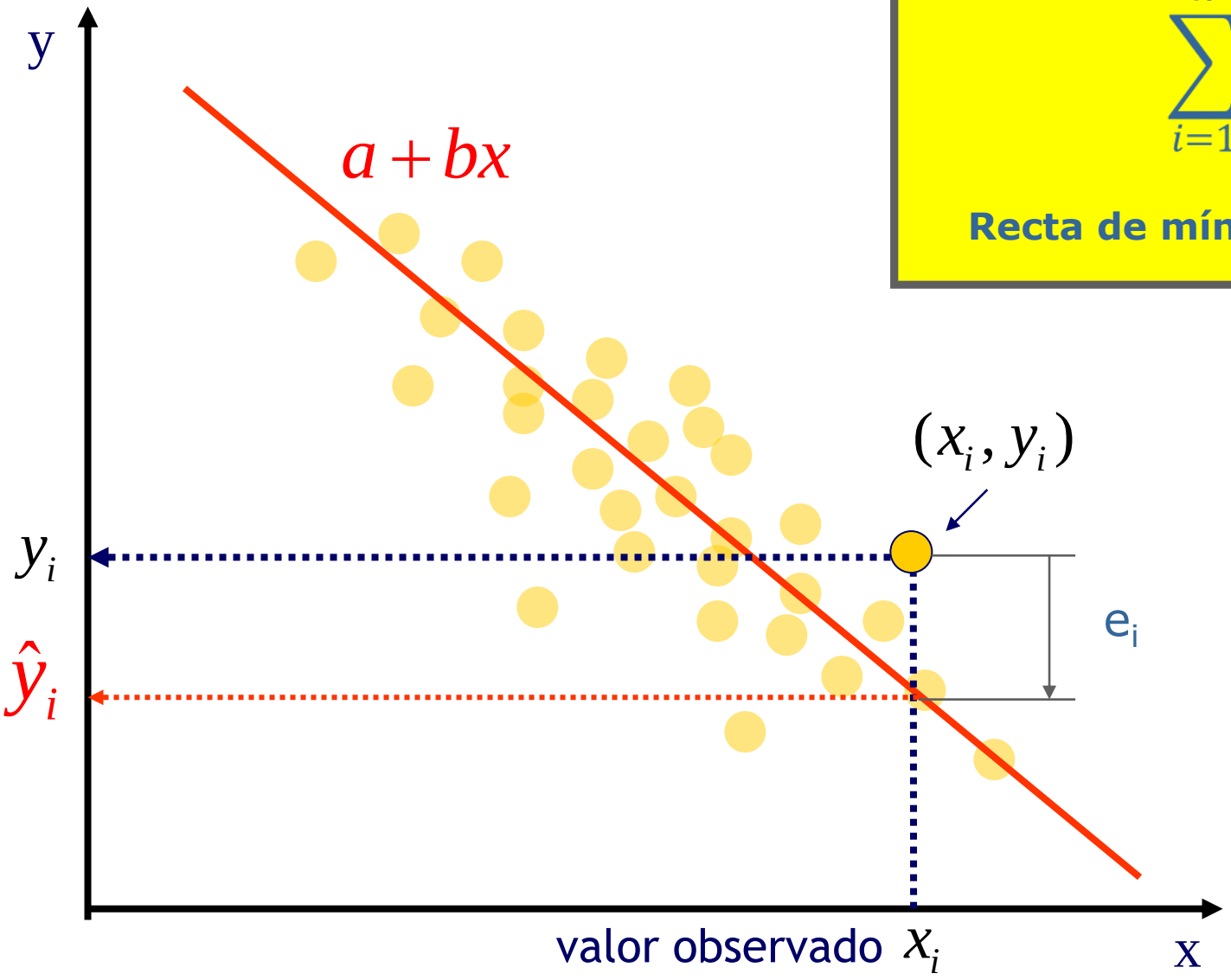
Error de predicción: $e_i = y_i - \hat{y}_i$

La recta de regresión

Buscamos la recta que minimiza los errores de predicción:

$$\sum_{i=1}^{n} e_i^2$$

Recta de mínimos cuadrados

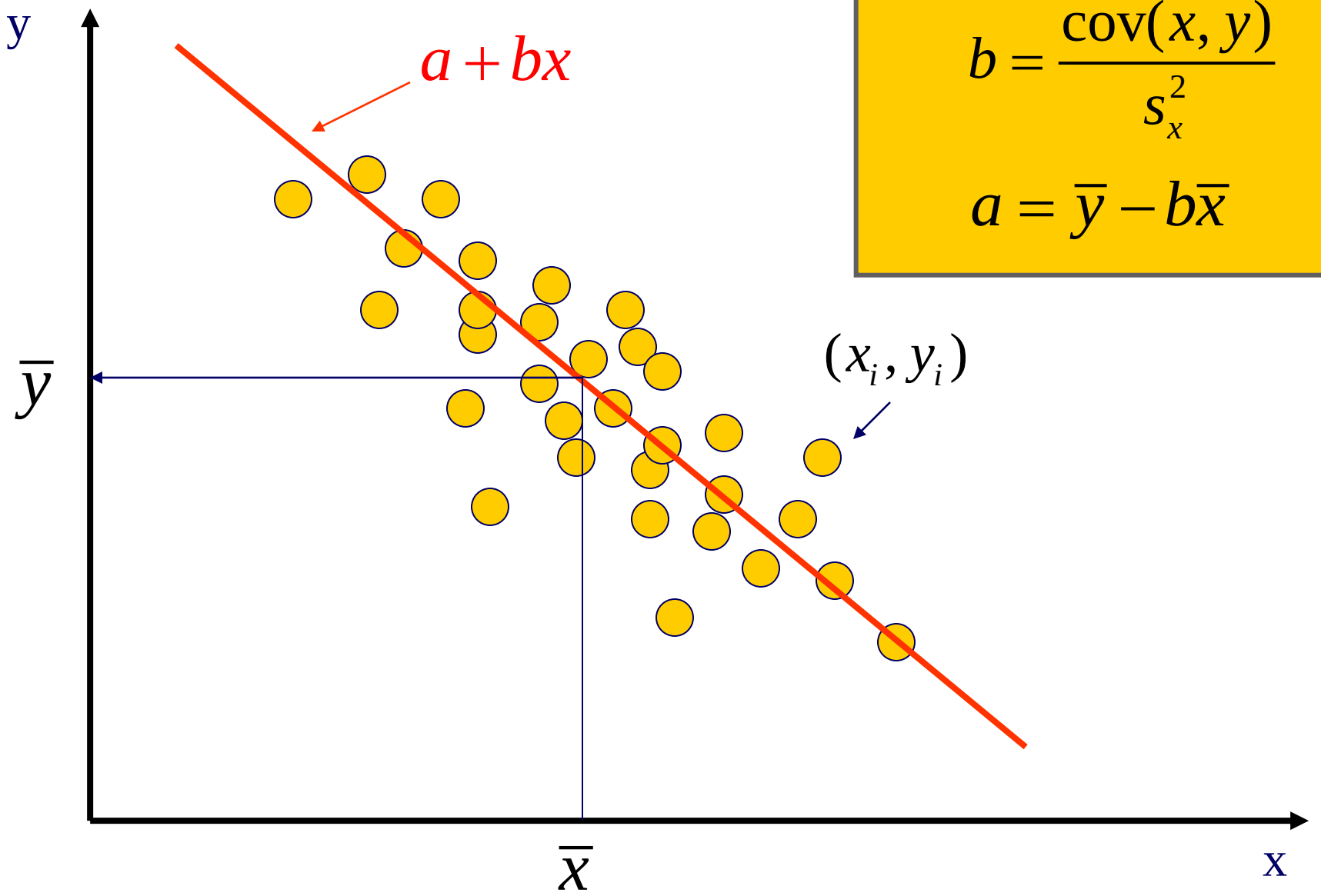


La recta de regresión

SOLUCIÓN

$$b = \frac{\text{cov}(x, y)}{s_x^2}$$

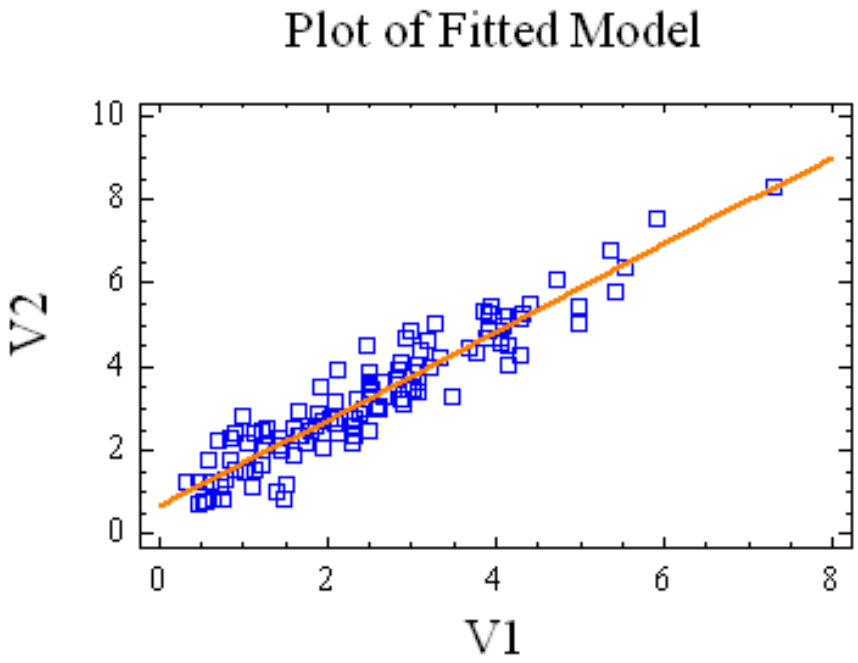
$$a = \bar{y} - b\bar{x}$$



Ejemplo

La variable V1 es la velocidad del viento registrada en la localización 1, mientras que la variable V2 es la velocidad registrada en esos mismos instantes en la localización 2.

Se tiene un total de 115 pares de medidas



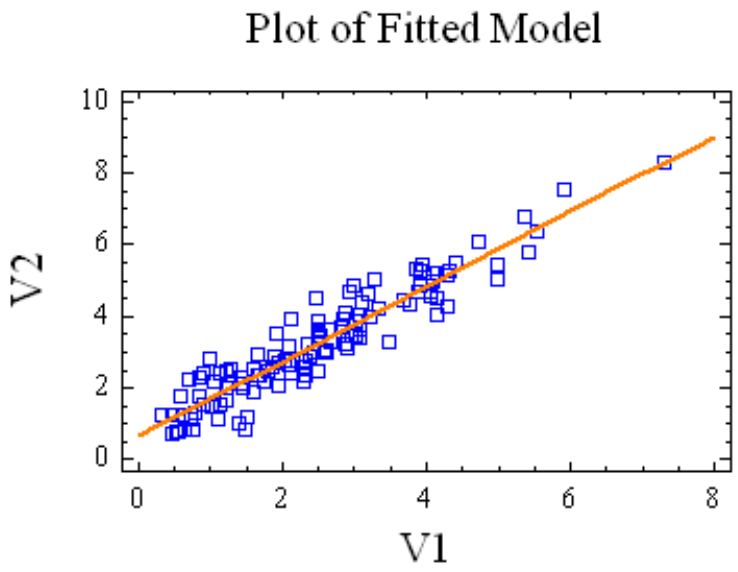
Loc.1:	Loc.2:
media: 2.51	media: 3.28
varianza: 1.91	varianza: 2.36

cov (V1,V2)=1.995

En la localización 1 se va a establecer un sistema informático para la telemedida de la velocidad del viento, pero no para la localización 2.

Se quiere calcular la recta de regresión que permita predecir la velocidad de la localización 2 sabiendo la velocidad de la localización 1

Ejemplo



Loc.1:
media: 2.51
varianza: 1.91

Loc.2:
media: 3.28
varianza: 2.36

cov (V1,V2)=1.995

$$b=\mathrm{cov}(x,y)/\mathrm{var}(x)=1.995/1.91=1.045$$

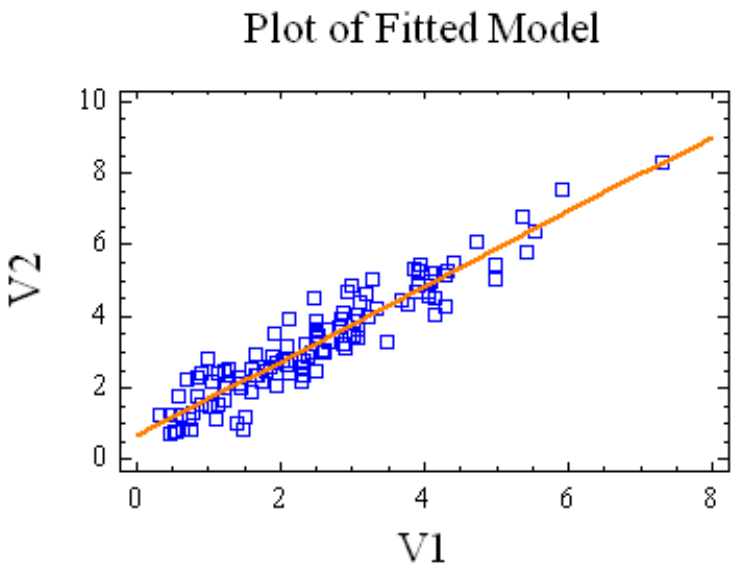
$$a=\bar{y}-b\bar{x}=3.28-1.045\times 2.51=0.657$$

$$\hat{V}_2=0.657+1.045\times V_1$$

Si, por ejemplo, en la Localización 1 se mide una velocidad de viento de 5 m/s, la predicción en la Localización 2 es de un viento de

0.657+1.045x5=5.882 m/s

Ejemplo



$$b=\text{cov}(x,y)/\text{var}(x)=1.995/1.91=1.045$$

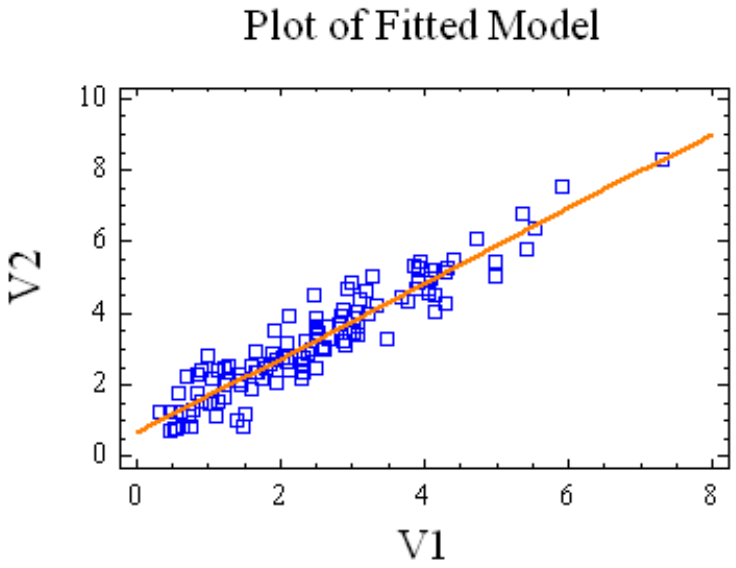
$$a=\bar{y}-b\bar{x}=3.28-1.045\times 2.51=0.657$$

$$\hat{V}_2=0.657+1.045\times V_1$$

Interpretación de b: si **x** aumenta en una unidad, entonces **y** aumenta en b unidades

Si en la localización 1 aumenta la velocidad del viento en 1 m/s, en la localización 2 lo hará en 1.045 m/s

Ejemplo



$$b=\text{cov}(x,y)/\text{var}(x)=1.995/1.91=1.045$$

$$a=\bar{y}-b\bar{x}=3.28-1.045\times 2.51=0.657$$

$$\hat{V}_2=0.657+1.045\times V_1$$

Interpretación de a: si **x** vale 0, entonces **y** toma el valor a

Si en la localización 1 no hay viento, en la localización 2 hay 0.657 m/s, que es un valor pequeño

Evaluación de la regresión

La regresión para predecir **y** a partir de **x** será razonable si:

1. La relación entre **x** e **y** es lineal
2. La relación es fuerte (alta correlación)

Correlación y Coeficiente de determinación R^2

Gráficos

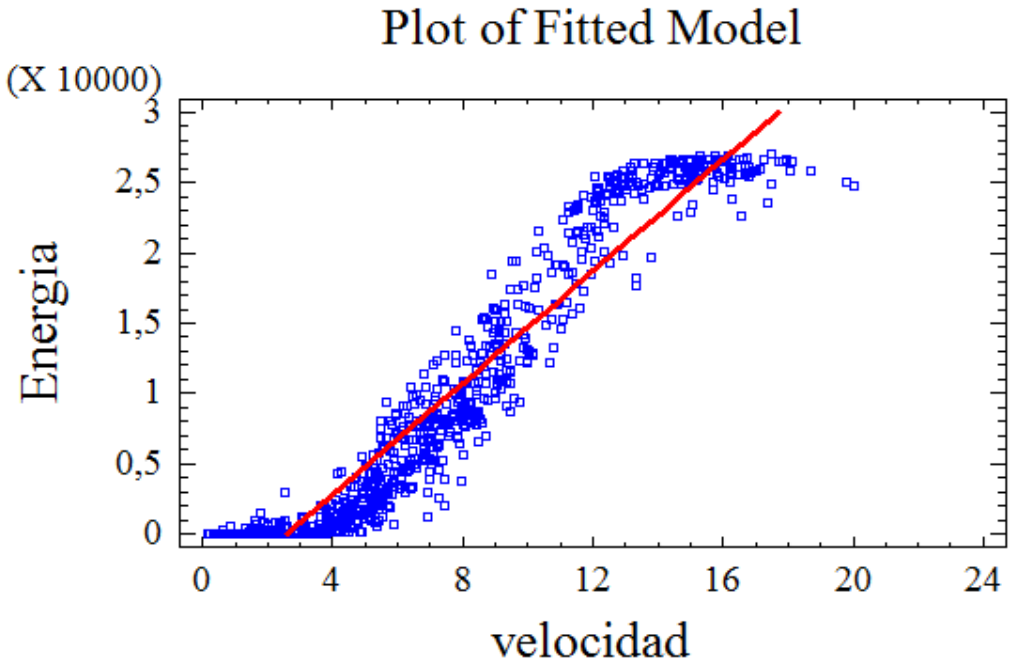
Evaluación de la regresión

La regresión para predecir y a partir de x será razonable si:

1. **La relación entre x e y es lineal**

    1.1 Gráfico xy
    1.2 Gráfico de predicciones vs observaciones
    1.3 Gráfico de residuos vs valores previstos

1.1 Gráfico xy



Este simple gráfico xy nos dice que no hay relación lineal.

La regresión lineal va a predecir mal

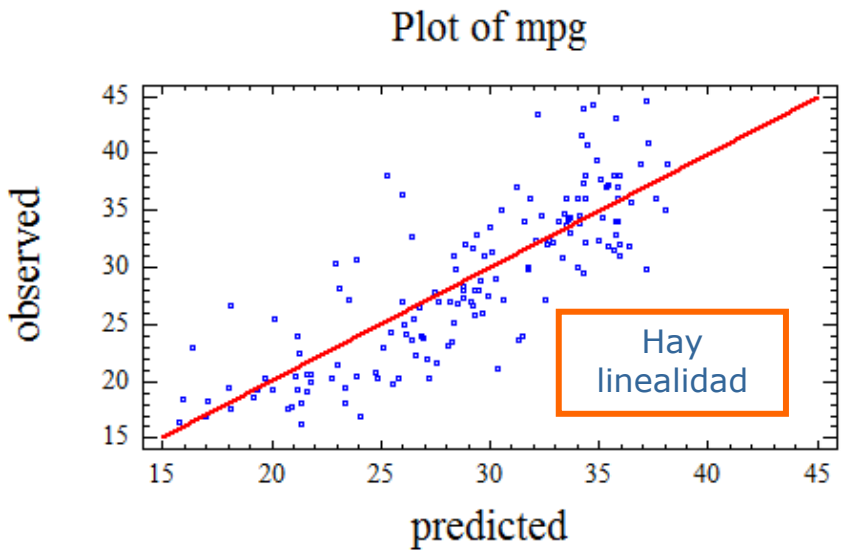
Evaluación de la regresión

La regresión para predecir y a partir de x será razonable si:

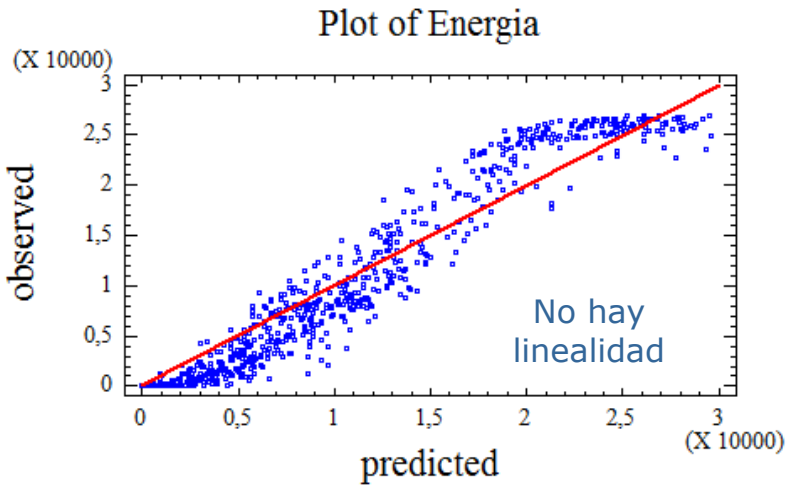
1. **La relación entre x e y es lineal**

   1.1 Gráfico xy
   1.2 Gráfico de predicciones vs observaciones
   1.3 Gráfico de residuos vs valores previstos

1.2 Gráfico de predicciones frente a observaciones



Cardata.xls: queremos explicar mpg (millas por galón) en función del peso (weight)



Parqueeolico.xls: queremos explicar la energía generada en función de la velocidad del viento)

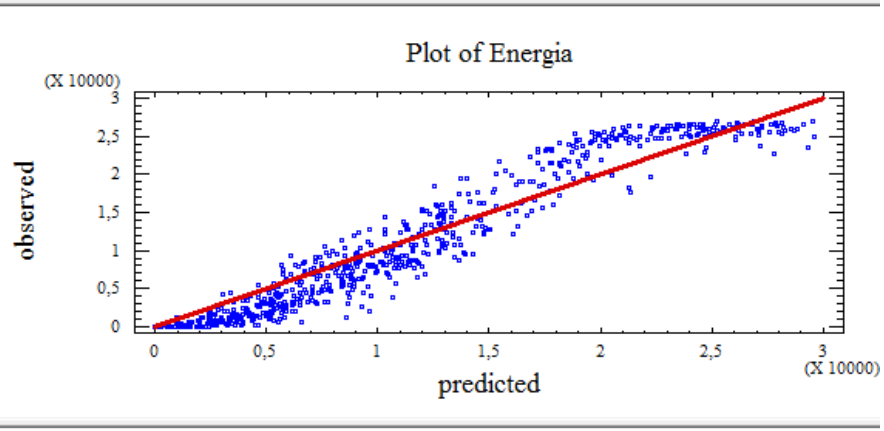
Evaluación de la regresión

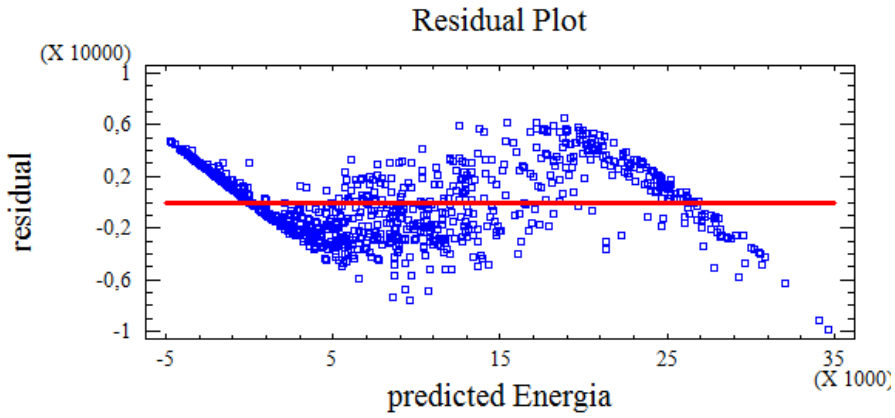
1.3 Gráfico de residuos frente a valores previstos

Es la representación gráfica más importante para evaluar una regresión

parqueeolico1.sf3

	Energia	velocidad	PREDICCION	RESIDUOS
840	25056,	16,2267	27078,8	-2022,77
841	25596,	15,1067	24857,5	738,489
842	25560,	14,0333	22728,7	2831,33
843	26028,	14,3767	23409,7	2618,27
844	24732,	13,4100	21492,5	3239,5
845	25992,	15,0000	24645,9	1346,1
846	26028,	15,1267	24897,2	1130,82
847	26136,	16,2967	27217,6	-1081,6
848	26280,	16,6600	27938,1	-1658,12
849	26604,	16,6000	27819,1	-1215,12
850	26676,	15,7467	26126,8	549,198
851	26424,	14,5200	23693,9	2730,07
852	26424,	14,4133	23482,3	2941,69
853	26892,	15,2767	25194,7	1697,33
854	26856,	15,7333	26100,2	755,774
855	26928,	16,1400	26906,8	21,1793
856	26964,	17,4633	29531,3	-2567,28





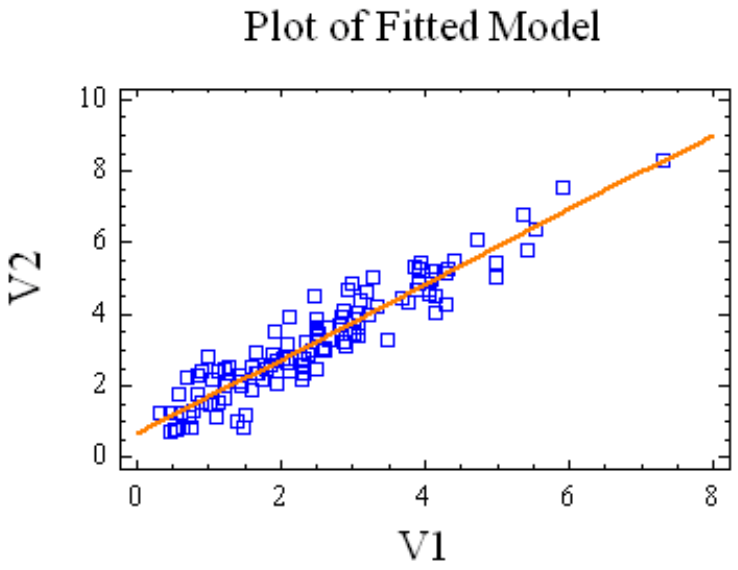
Valores previstos

Valor real - Predicción

La no linealidad es muy clara. La regresión no es adecuada.

Ejemplo

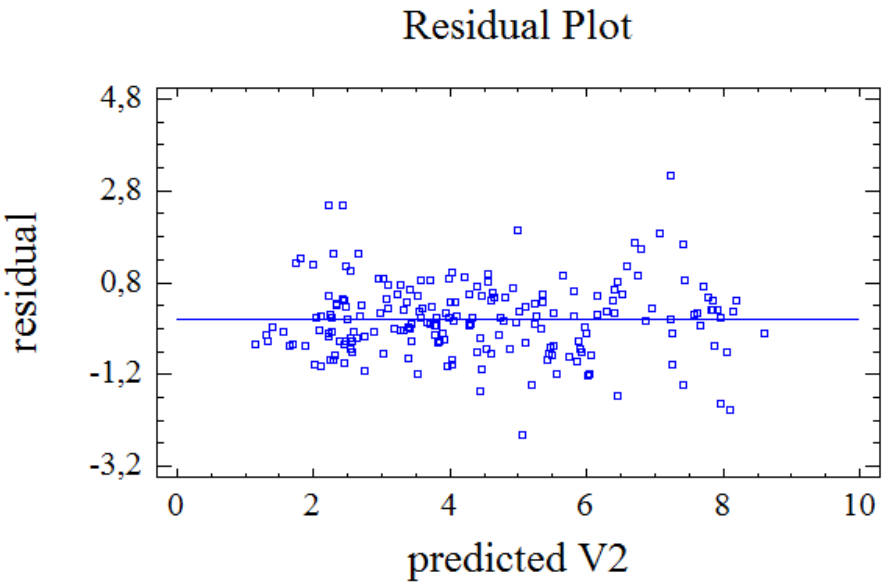
La variable V1 tiene la velocidad del viento registrada en la localización 1, mientras que la variable V2 tiene las velocidades registradas en esos mismos instantes en la localización 2.



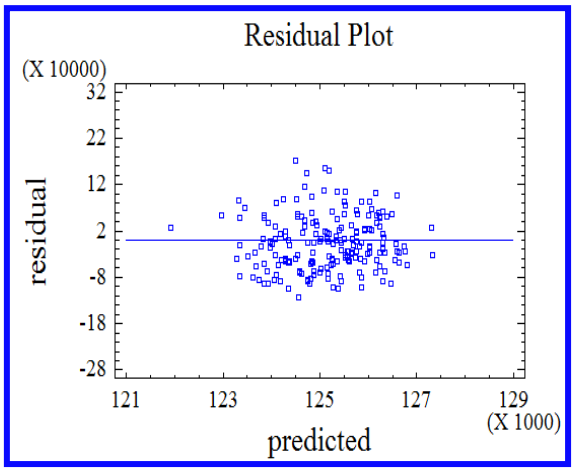
$$\hat{V}_2 = 0.657 + 1.045 \times V_1$$

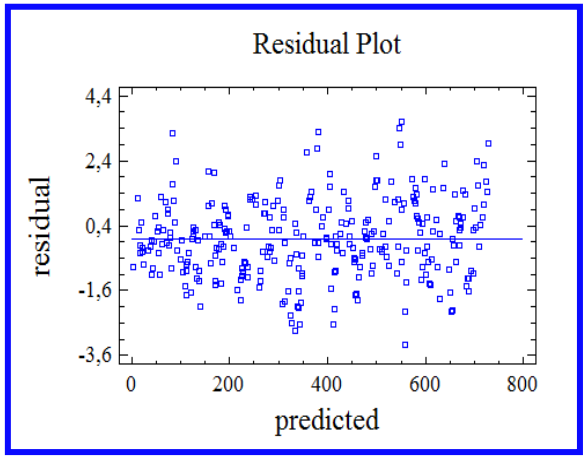
Estos residuos no muestran ninguna estructura evidente.

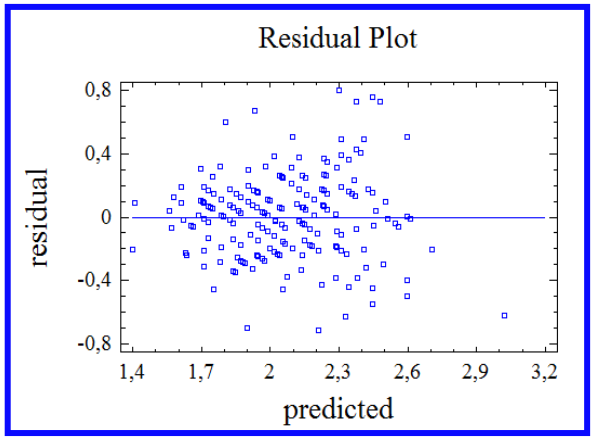
Es señal de que el modelo lineal es adecuado



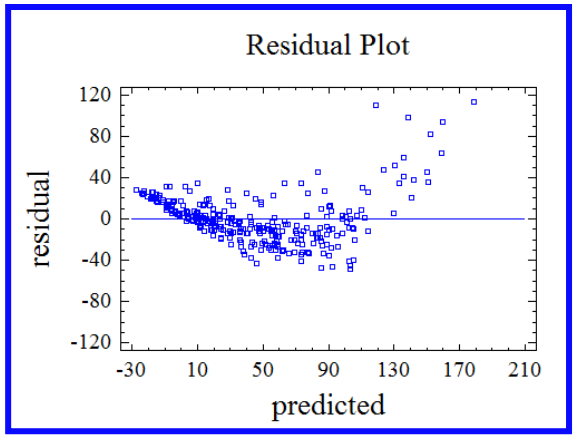
Estos gráficos de residuos son aceptables

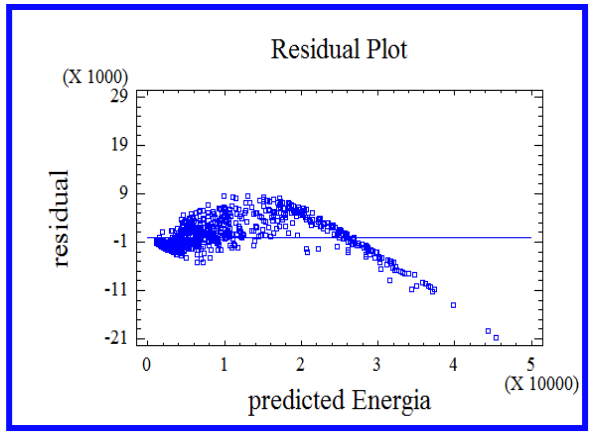


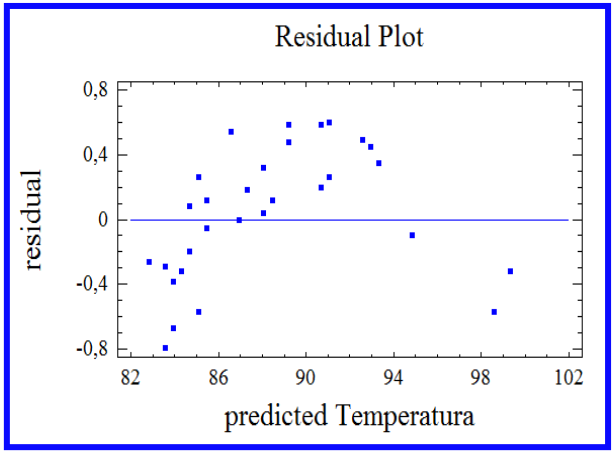




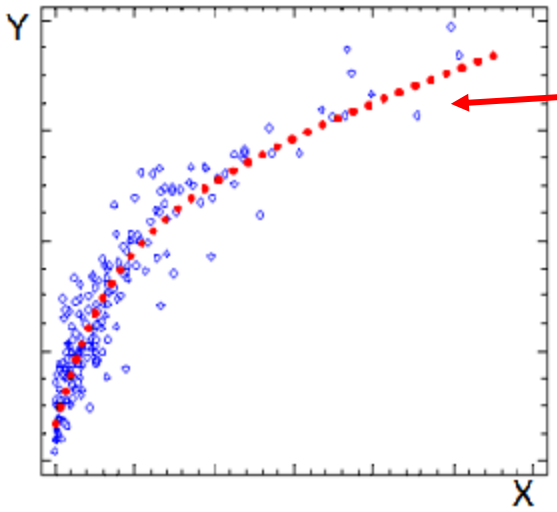
Estos gráficos de residuos NO son aceptables





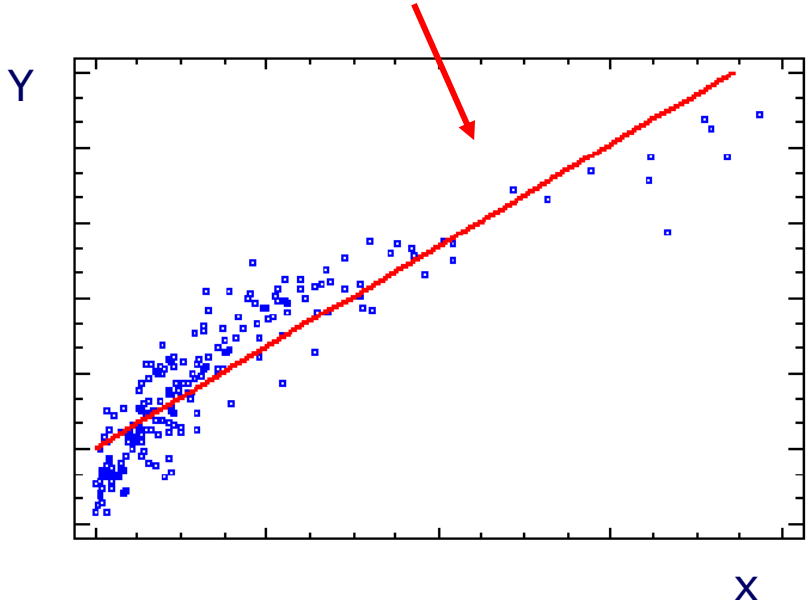


Algunos tipos de no linealidades se pueden corregir transformando la variable



Esta curva es la que nos gustaría usar como resumen de la relación

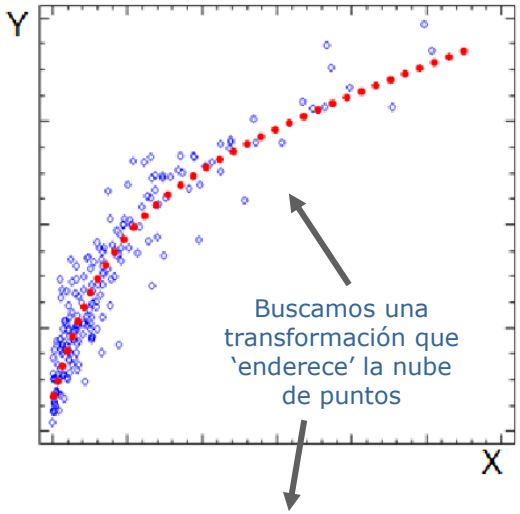
Pero la técnica de regresión simple sólo nos proporciona este tipo de soluciones



Buscamos otras variables

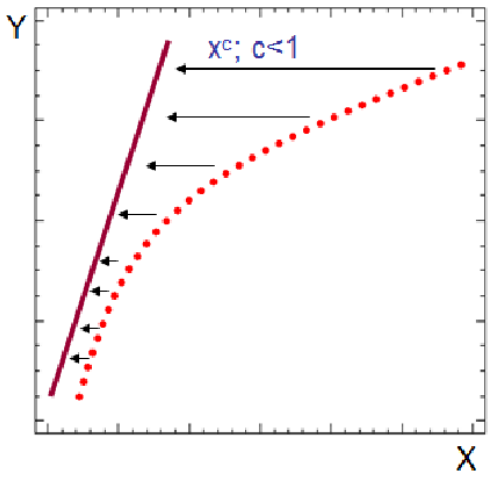
$y^*=f(y)$, $x^*=g(x)$

tales que entre ellas haya relación lineal

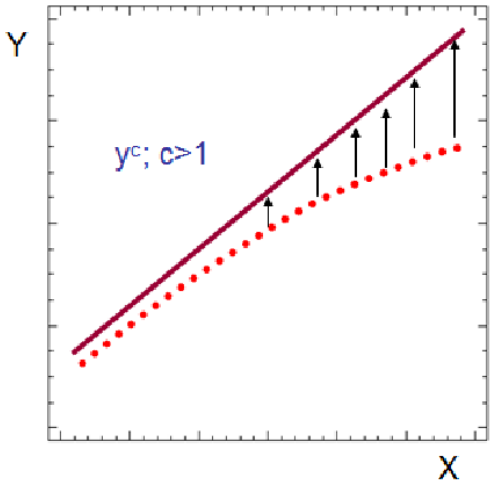


$y=a+bx^c$

$y^c=a+bx$



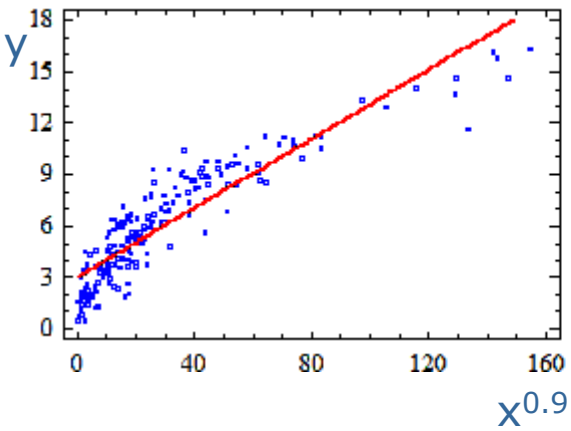
Si c<1, los valores más grandes se comprimen más. En este caso, aplicado sobre x 'enderezamos' la curvatura.



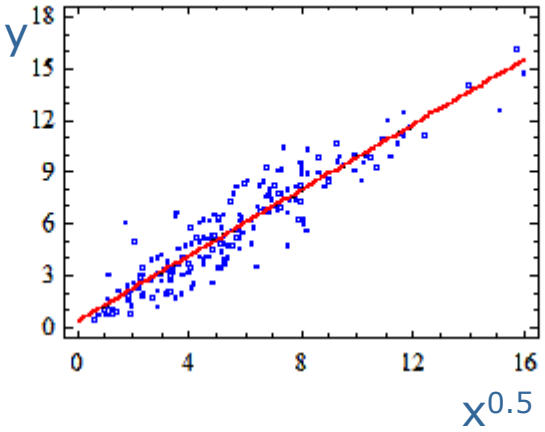
Si c>1, el efecto es el opuesto: los valores más grandes se expanden más. En este caso, aplicado sobre Y también 'enderezamos' la curvatura.

$y=a+bx^c$

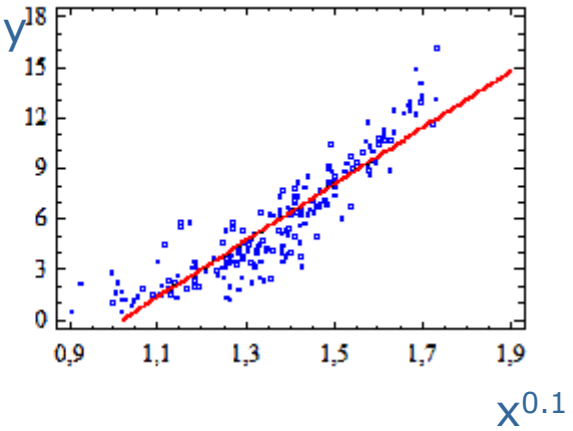
c=0.9 **insuficiente**



c=0.5 **¡perfecto!**



c=0.1 **nos hemos pasado!!**



Evaluación de la regresión

La regresión para predecir **y** a partir de **x** será razonable si:

1. La relación entre x e y es lineal ← Gráficos
2. La relación es fuerte (alta correlación)

Coeficiente de determinación R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

- Toma valores entre 0 y 1
- $R^2 = \mathrm{corr}(x, y)^2$
- El coeficiente de determinación nos dice qué proporción de la varianza de la variable respuesta y viene explicada por la recta de la regresión