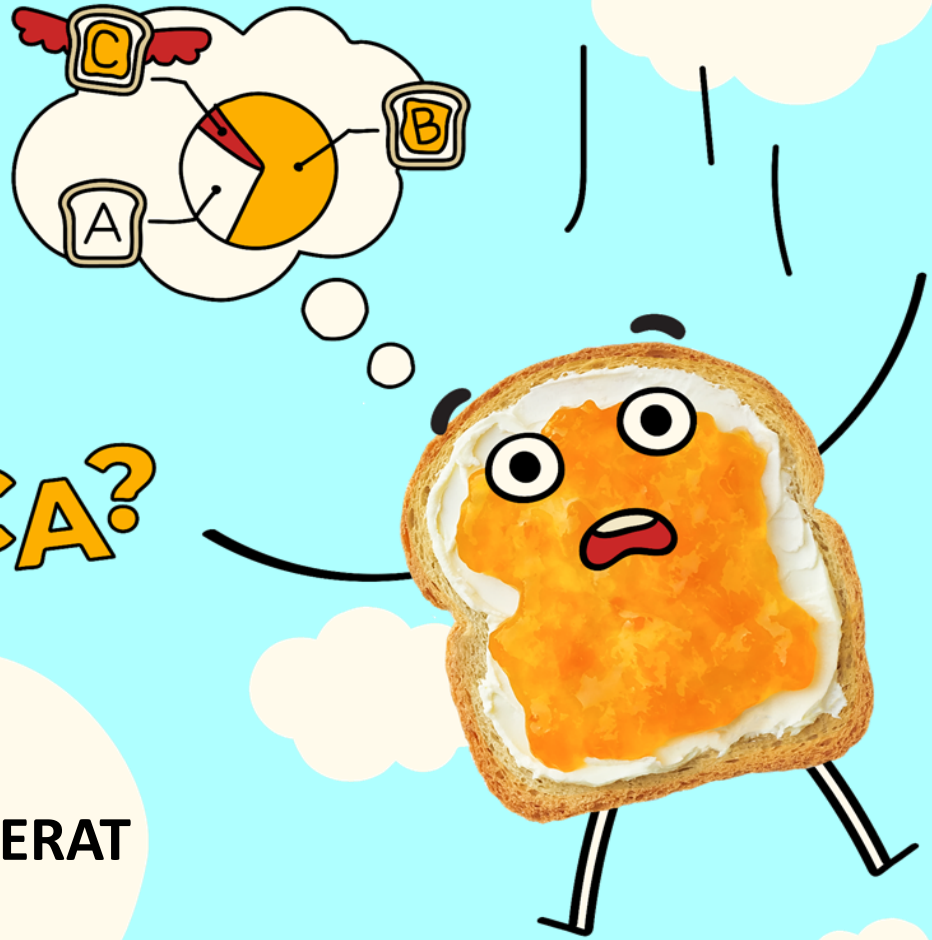
Quarta Jornada de divulgació del món de l'estadística

Per a què serveix L'ESTADÍSTICA?



ESTADíSTICA ADREÇADA A BATXILLERAT

Guadalupe Gómez Melis i Mireia Besalú

UNIVERSITAT DE BARCELONA
GRBIO
LaUB divulga

Les dades s'ajusten a una recta?

Regressió Lineal i Correlació

UNIVERSITAT DE BARCELONA

GRBIO

LaUB divulga

Predició de la potencial propagació d'un incendi

Variables implicades als incendis forestals

- Temperatura (TEMP)
- **Valor Inicial de Propagació** ISI (*Initial Spread Index):* Index que indica la taxa d'ignició i es fa servir per estimar el potencial de propagació d'un incendi forestal. Integra
  - la humitat del combustible (herbes seques, troncs etc)
  - la velocitat del vent.

ISI pren valors entre 0 i 50.

Un valor mes gran de 10 indica una ràpida ignició.

Pregunta

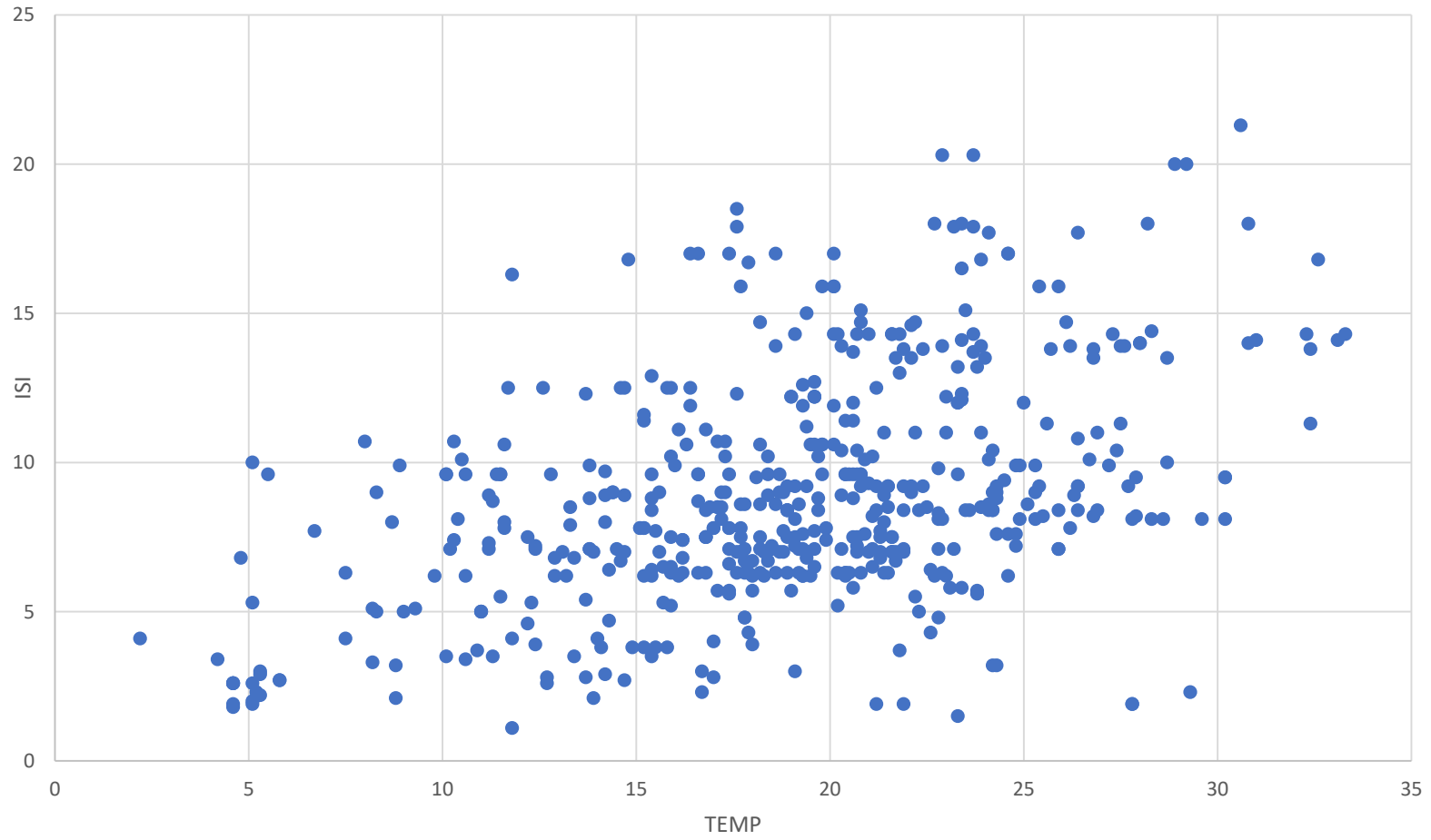
Podem predir el **Valor Inicial de Propagació** (ISI) a partir de la temperatura (TEMP)?

TEMP	ISI
8.2	5.1
18	6.7
14.6	6.7
8.3	9
11.4	9.6
22.2	14.7
24.1	8.5
8	10.7
13.1	7
22.8	7.1
17.8	7.1
19.3	12.6
17	2.8
21.3	7
26.4	9.2
22.9	13.9
15.1	7.8
16.7	3
15.9	6.3

Com es relacionen TEMP i ISI? Podria ser de forma lineal?

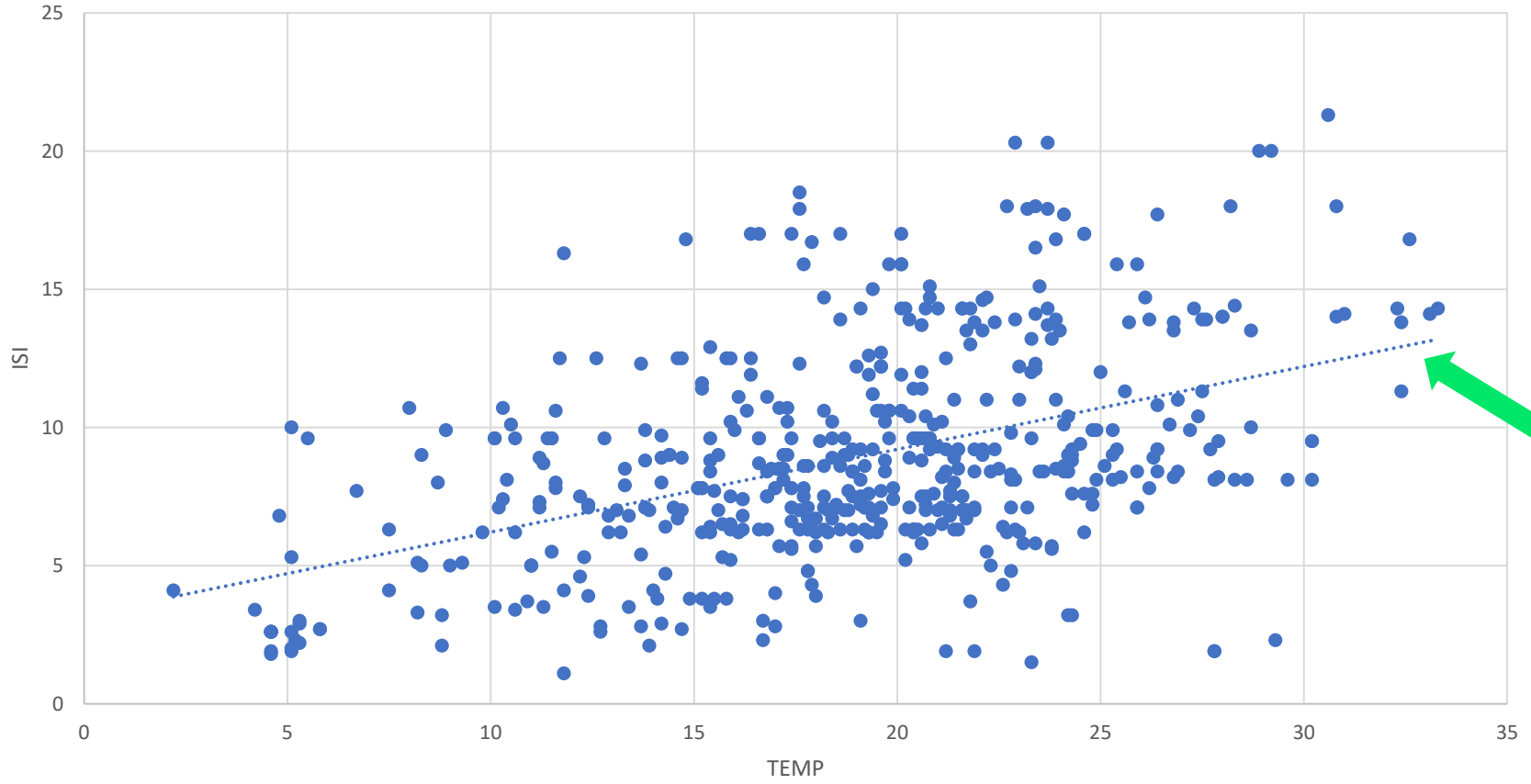
Pas 1: Diagrama de punts. Comencem dibuixant les dades.

A l'eix de les X tenim la TEMP i a l'eix de les Y la variable ISI



Estudiant si TEMP i ISI es relacionen linealment

Pas 2: Afegim una recta y=a+bx de tendència



$$y = a + bx$$

Objectiu: Predir Y a partir de X de forma lineal

Cal obtenir una recta "prou bona" que permeti predir els valors d' ISI a partir de TEMP

Pel gràfic que hem fet estem considerant que:

X = TEMP (Eix X) → variable independent (o explicativa)

Y = ISI (Eix Y) → variable dependent (o resposta)

La línea de tendència serà una recta de la forma

$$y = a + bx$$

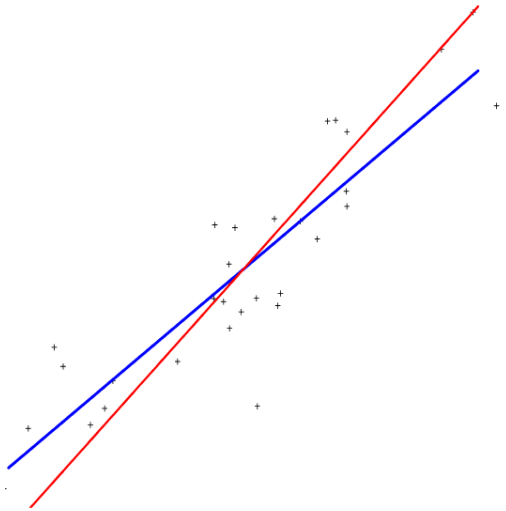
anomenada **recta de regressió**.

Ens preguntem

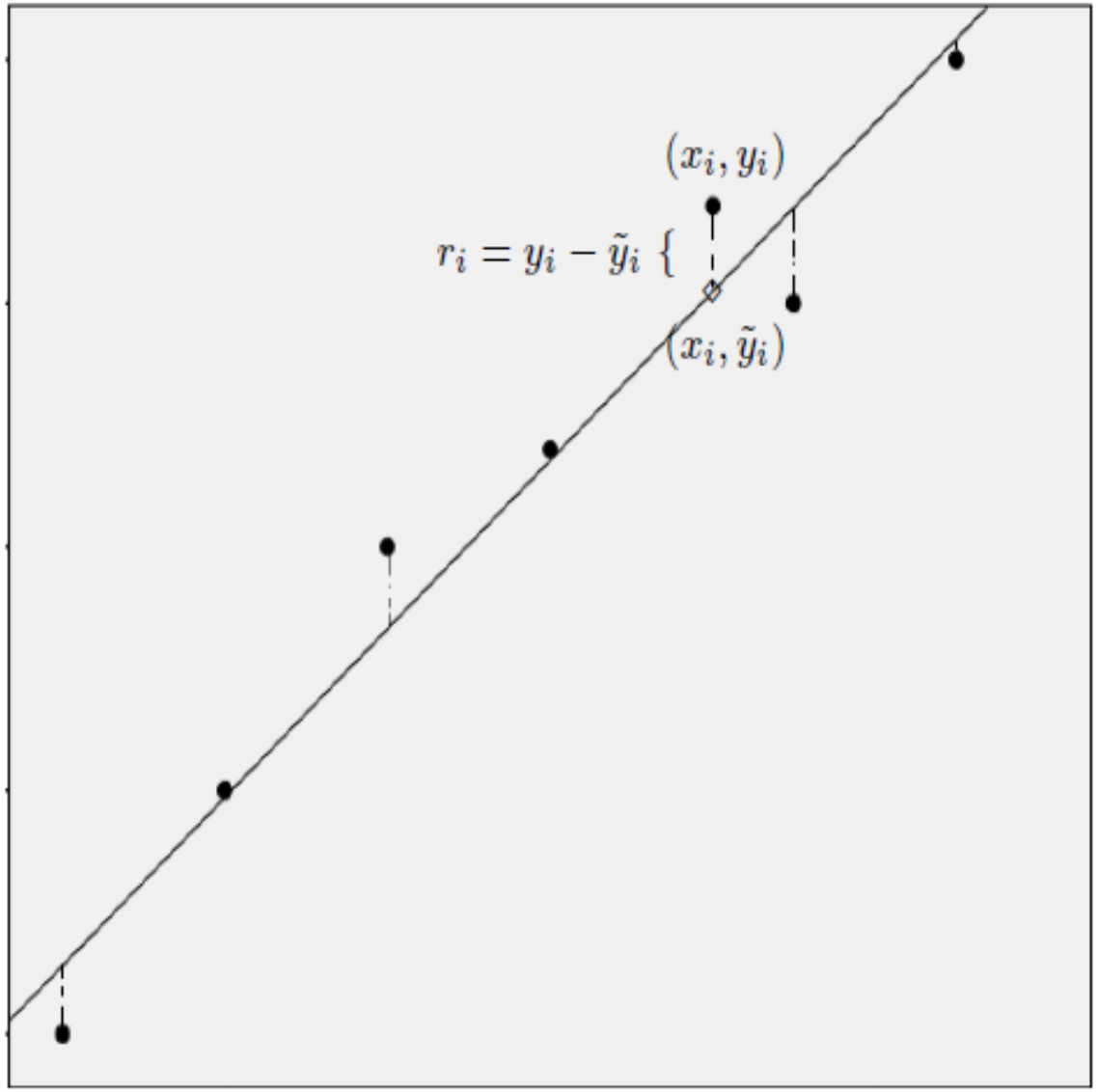
1. Com "triem" la recta de regressió?
2. Com avaluem si la recta de regressió és "prou bona"?

En general, com saber si dues variables contínues tenen una relació lineal i quina és la millor descripció lineal?

Quina és millor:
La blava o la vermella?



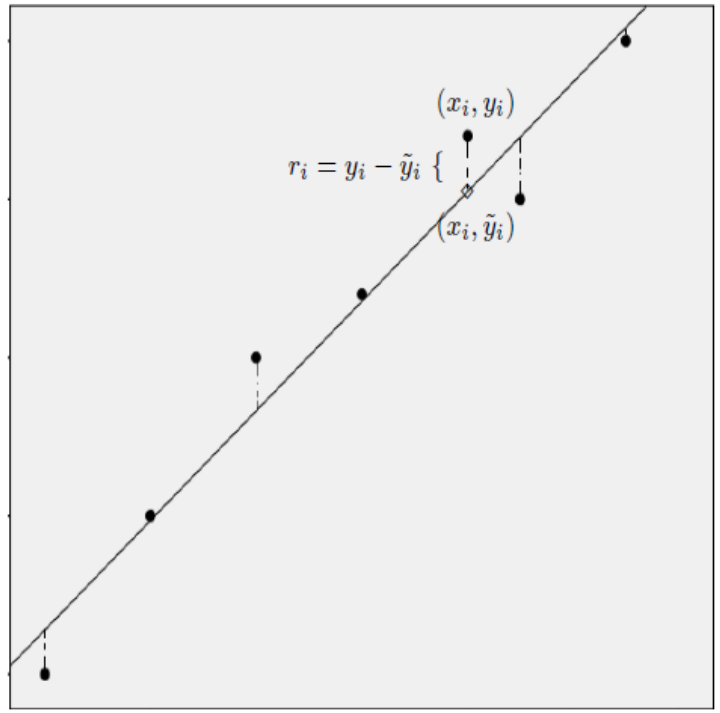
Ajust per mínims quadrats: minimitzen les distàncies verticals



Ajust per mínims quadrats

La recta de regressió per mínims quadrats **minimitza la suma de les distàncies verticals al quadrat**, és a dir, es tracta de buscar els valors per a i b, que anomenarem $\hat{a}\ i\ \hat{b}$, que minimitzin la funció:

$$\sum_{i=1}^{n}(y_i-(\hat{a}+\hat{b}x_i))^2, \qquad a,b\in\Re$$



La solució és:

$$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2};\ \hat{a}=\bar{y}-b\bar{x} \quad \text{amb} \quad \bar{x}=\frac{\sum_{i=1}^{n}x_i}{n} \quad \text{i} \quad \bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}$$

Els residus de l'ajust, r_i són les distàncies dels valors observats als que es predirien amb la recta $r_i=y_i-(\hat{a}+\hat{b}x_i)$

I en els incendis, que podem fer?

Calcular la recta de regressió per ISI (Y) en funció de TEMP (X).

La recta de regressió obtinguda és

$$y = 3,2167 + 0,2996\,x$$

Com l'interpretem?

1. Per cada grau que augmentem la Temperatura, l'ISI augmenta en **0,2996**

$$y = 3,2167 + 0,2996\,x$$

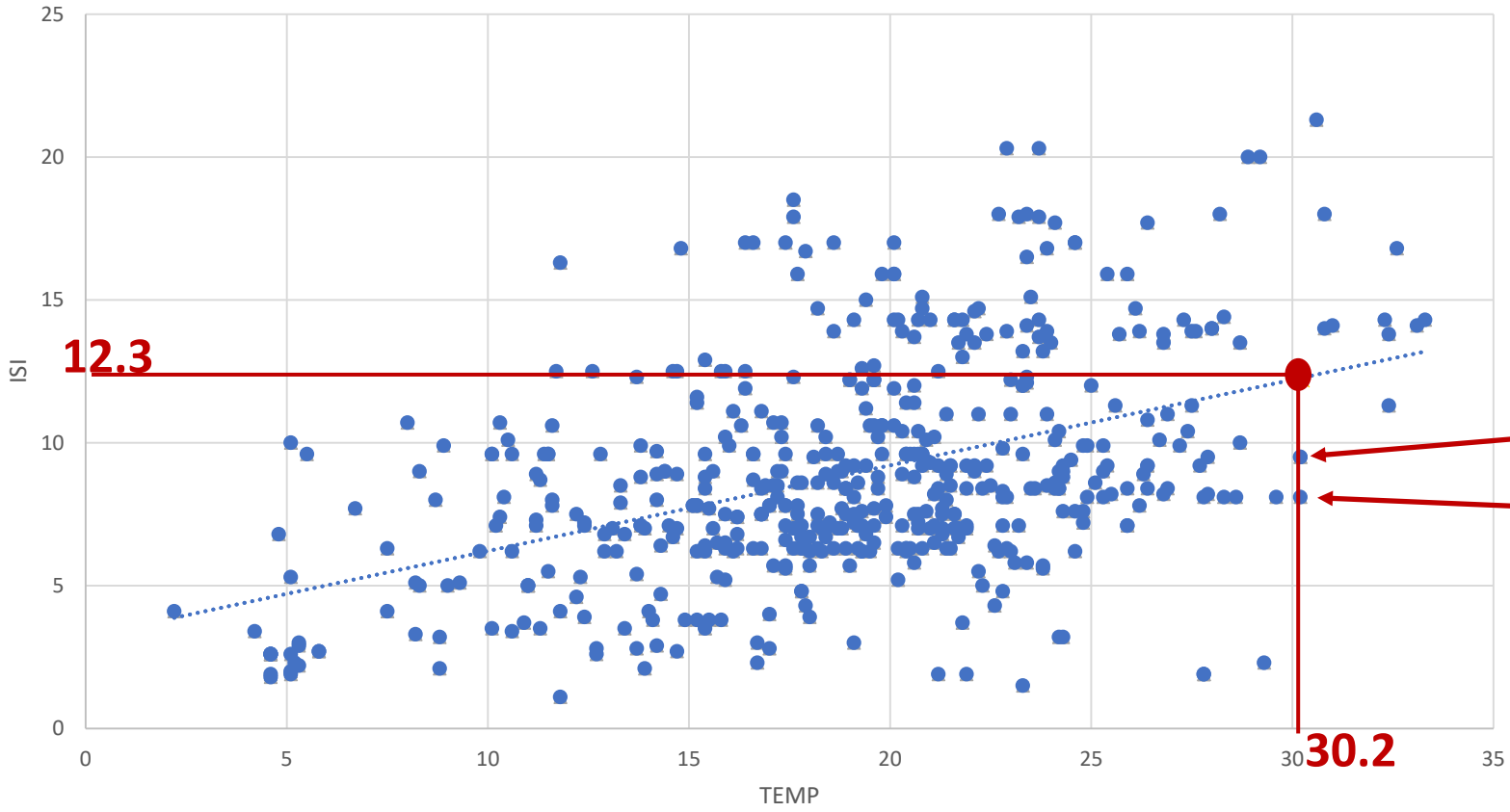
2. El valor d'ISI basal (per una temperatura de 0 graus) seria $3,2167$

$$y = 3,2167 + 0,2996\,x$$

3. Més interesant, fem prediccions.

Per a una temperatura TEMP=30,2 graus, segons el nostre model el valor d'ISI predit és

$$3{,}2167 + 0{,}2996 * \mathbf{30{,}2} = 12{,}3$$



EPPP, hi ha valors ISI observats més petits per la mateixa temperatura.

Que significa aquest 12.3?

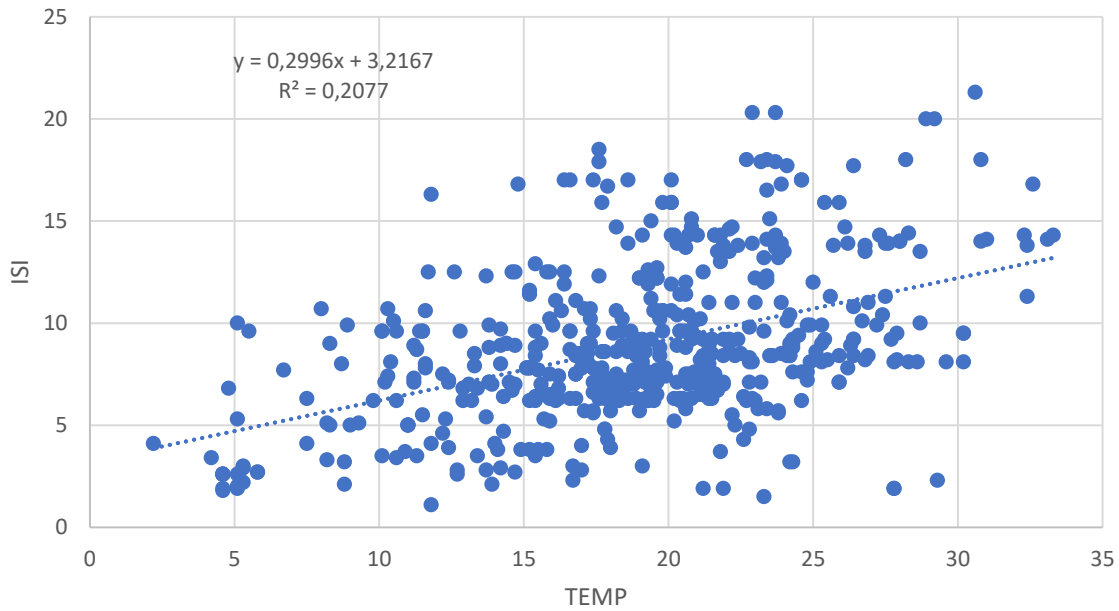
12.3 és la Mitjana d'ISI ajustada pel model

Com determinem si l'ajust és bo?

Mitjançant la proporció de la variabilitat de les dades que queda *explicada* per la recta de regressió. Podem fer servir el coeficient de determinació R^2

$$R^2 = \frac{SS(res)}{SS(total)} = \frac{\sum_{i=1}^{n} r_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Recordem els residus $r_i = y_i - (\hat{a} + \hat{b}x_i)$



R^2 alts representen un millor ajust i $0 \leq R^2 \leq 1$

R^2 es pot escriure en termes de les variàncies de X i Y, ${S^2}_X$ i ${S^2}_Y$ i de la **covariància entre X i Y,** S_{XY} que ara introduirem

$$R^2 = \frac{{S_{XY}}^2}{{S^2}_Y {S^2}_X}$$

Covariància mostral S_{XY}

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x} \cdot \bar{y} \quad \text{on} \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i$$

És una mesura de l'associació lineal entre les variables X i Y

Interpretació:

- Covariància gran positiva → en mitjana X i Y creixen o decreixen juntes
- Covariància gran negativa → en mitjana una creix quan l'altra disminueix

Problema:

Depèn de les unitats de mesura. Canviaria si mesures en metres en comptes de Km

Escapant-nos de la dependència de les unitats de mesura

El Coeficient de **correlació r_{XY}** (Coeficient de correlació de Pearson) ens ajuda a mesurar-ho. Es defineix com

$$r_{XY} = \frac{S_{XY}}{S_Y S_X}$$

Coeficient de correlació

- Grau associació
- Mesura associació lineal

D'on vé el nom?

Co conjuntament

Relació connexió

Coeficient de correlació de Pearson

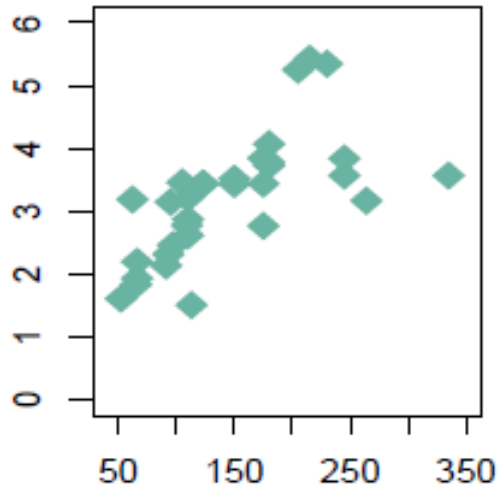
$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

- ✓ Sempre es compleix que: $-1 \leq r_{XY} \leq 1$
- ✓ Només detecta l'associació o dependència lineal
- ✓ Si $r_{XY} = 1$ o $r_{XY} = -1$ la relació lineal entre X i Y és perfecte

Alternativa per fer el càlcul

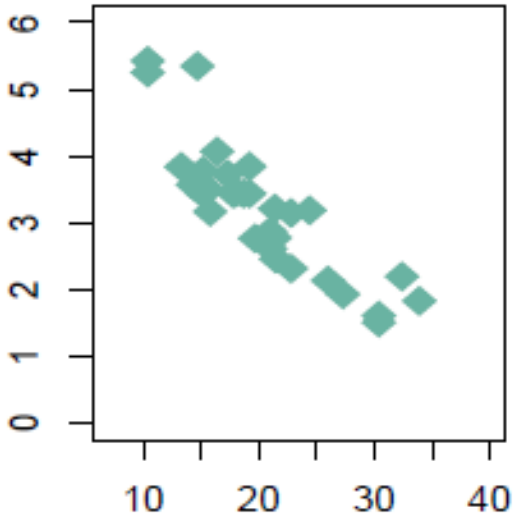
$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i}{S_X S_Y}$$

$r_{XY} = 0.66$



És positiu si quan una variable creix l'altre també

$r_{XY} = -0.87$

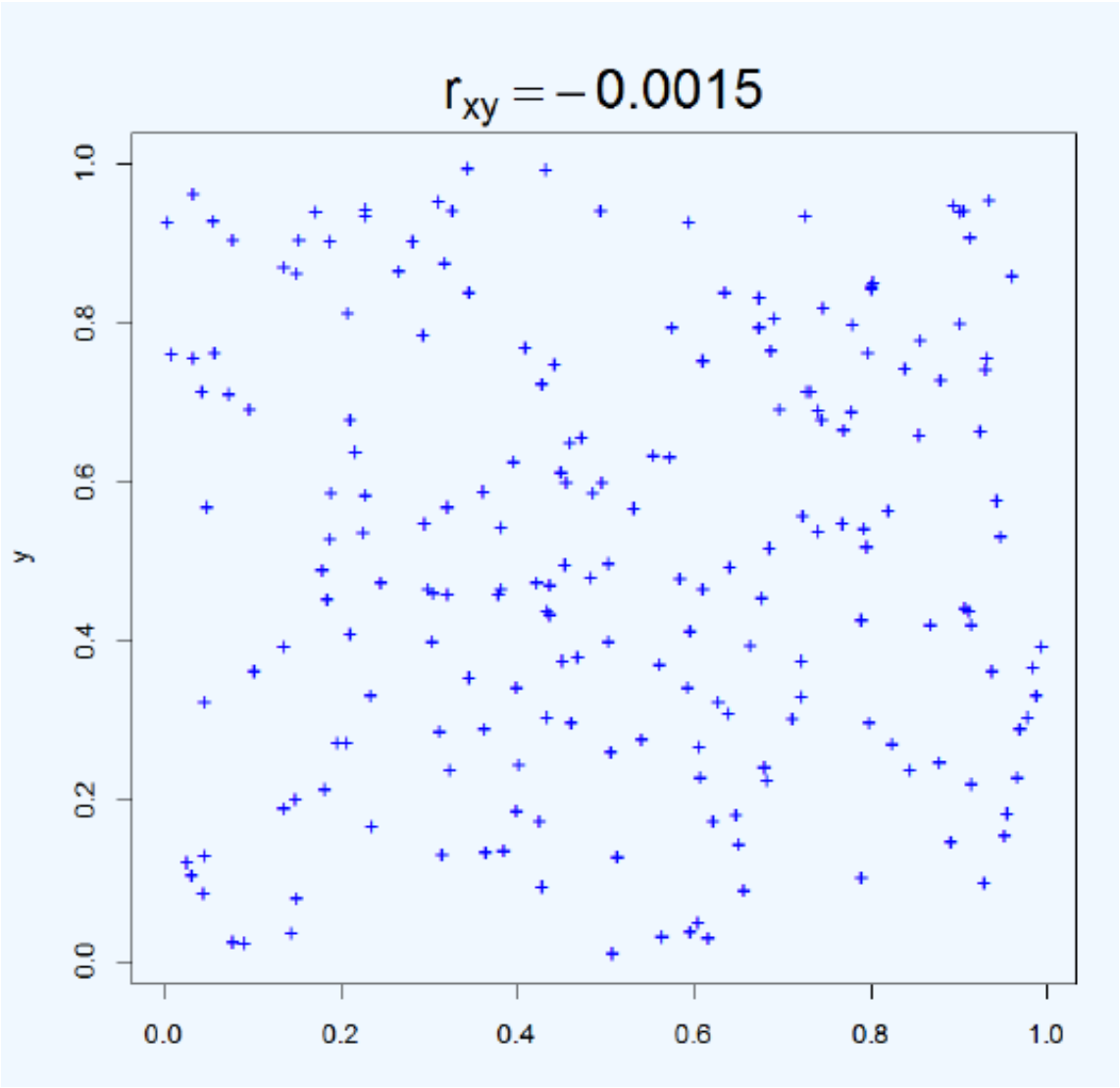


És negatiu si quan una variable creix l'altre decreix

ESTUDIANT DIVERSOS GRAUS DE CORRELACIÓ:

Què mesura r_{XY}?

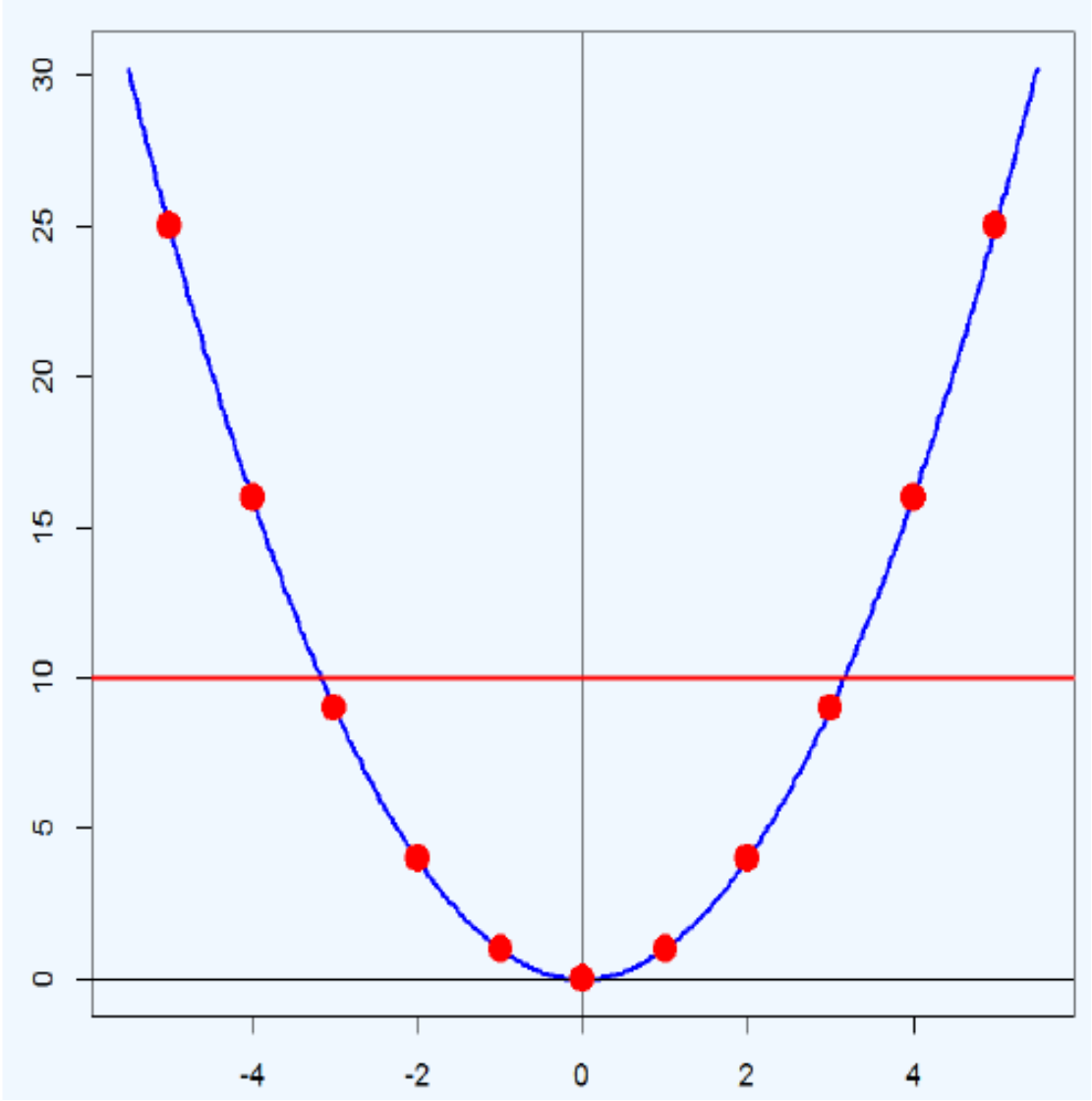
Punts a l'atzar sense relació entre X i Y , tenen r_{XY} pròxim a 0



Punts lligats per una dependència quadràtica de les variables. r_{XY} no detecta l'associació

Y és funció de X però $r_{XY} = 0$

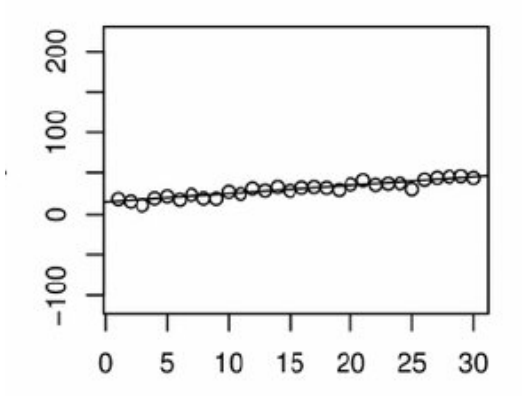
$r_{XY} = 0$



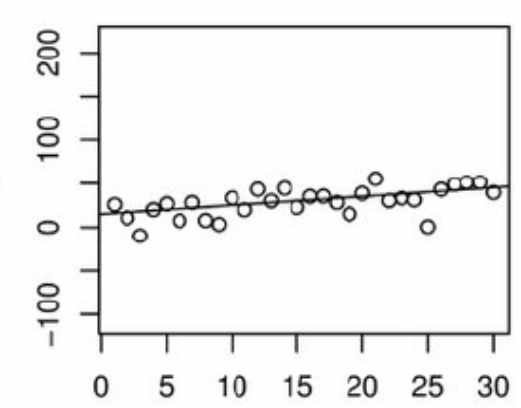
Curiositat

- ✓ Mateixa recta de regressió, però diferentes coeficient de correlació
- ✓ Diferents rectes de regressió, però mateix coeficient de correlació

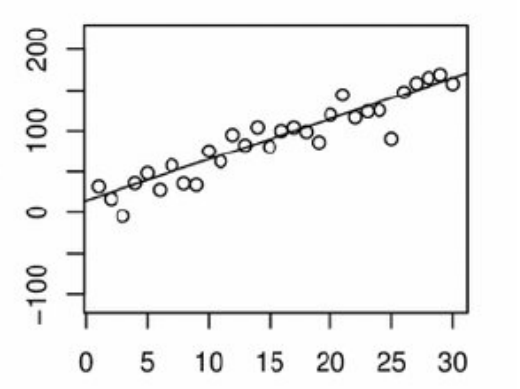
$y = 15 + x; r_{XY} = 0.94$



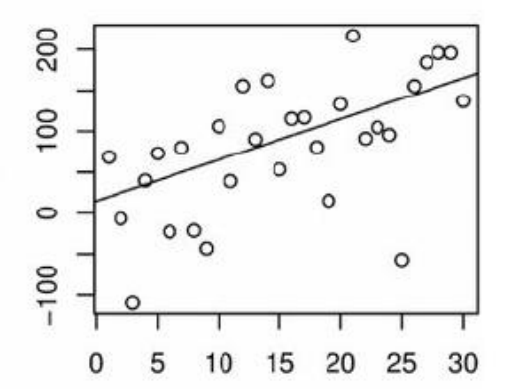
$y = 15 + x; r_{XY} = 0.59$



$y = 15 + 5x; r_{XY} = 0.94$

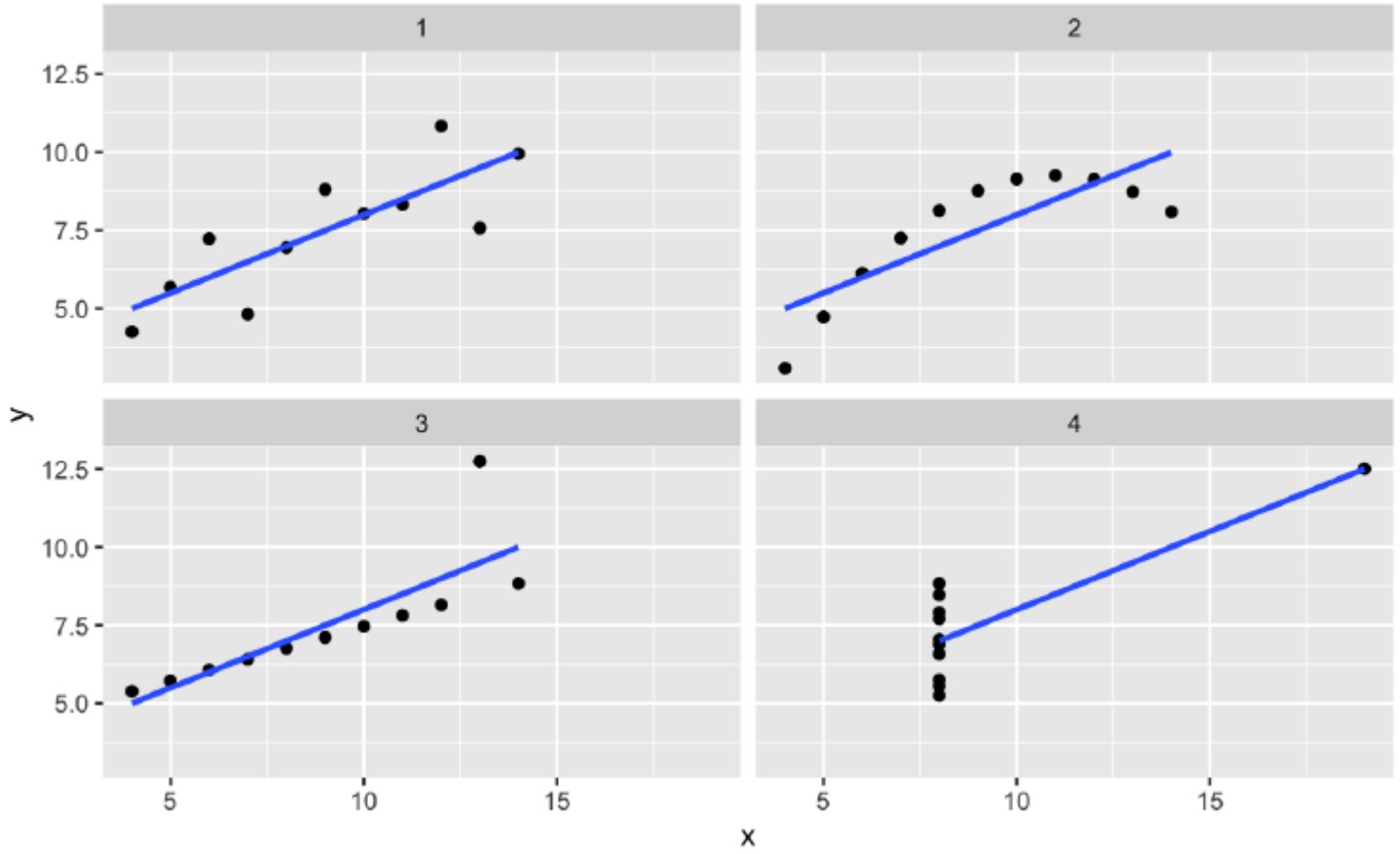


$y = 15 + 5x; r_{XY} = 0.59$



Curiositat: Quartet d'Anscombe

4 conjunts de dades amb la mateixa recta de regressió. Alguna altra semblança?



Els quatre conjunts de dades tenen:

Mitjana de X	9
Variancia de X	11
Mitjana de Y	7.50
Variancia de Y	4.12
Correlació entre X i Y	0.82
Recta de regressió	$y = 3 + 0.5x$

Incendis forestals. Predició ISI a partir de TEMP

ISI: Valor Inicial de Propagació (*Initial Spread Index*)

- ✓ Indica la taxa d'ignició.
- ✓ Pren valors entre 0 i 50
- ✓ Un valor mes gran de 10 indica una ràpida ignició

Càlculs a partir de les dades observades

$$\text{ISI} = 3.2167 + 0.2996 \cdot TEMP$$

Covariància: $S_{XY} = 10.101$

Correlació: $r_{XY} = 0.456$

Coeficient de Determinació:

$$R^2 = r_{XY}^2 = 0.456^2 = 0.2077$$

Prediccions del valor del ISI a partir de TEMP

TEMP=22^{0}C $\quad 3.2167 + 0.2996 \cdot \mathbf{22} = 9.8$

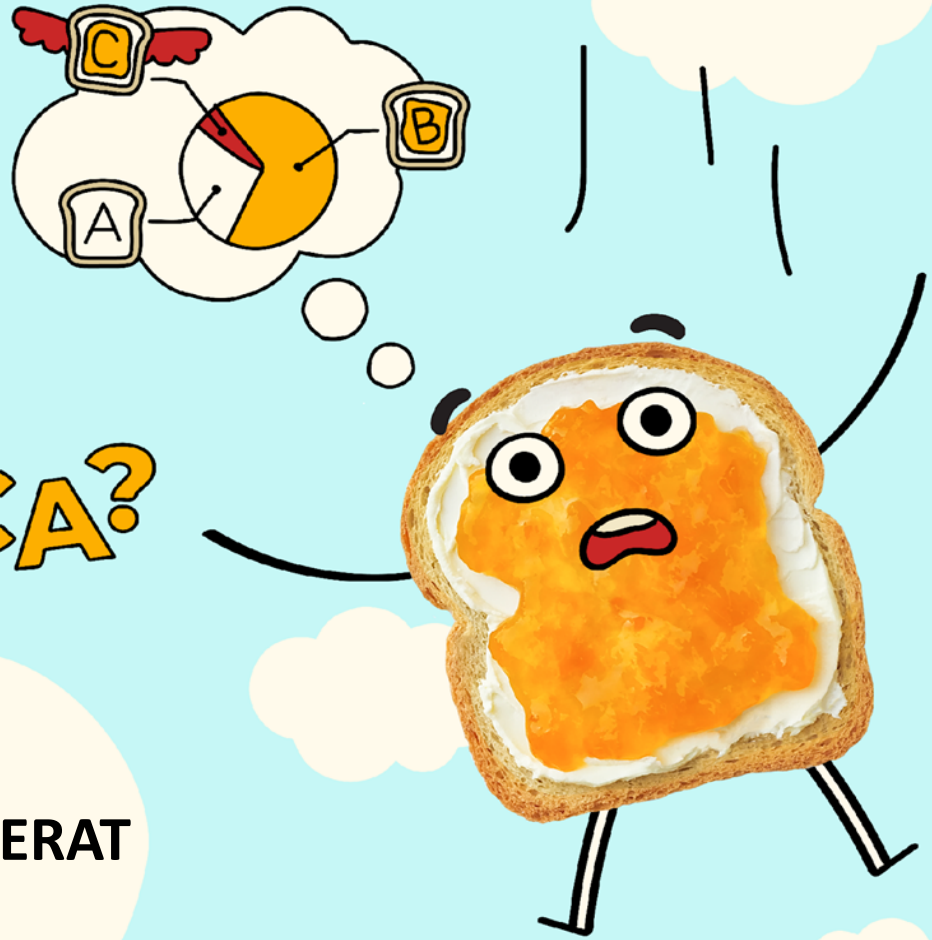
TEMP=25^{0}C $\quad 3.2167 + 0.2996 \cdot \mathbf{25} = 10.7$

TEMP=30^{0}C $\quad 3.2167 + 0.2996 \cdot \mathbf{30} = 12.2$

TEMP=35^{0}C $\quad 3.2167 + 0.2996 \cdot \mathbf{35} = 13.7$

Quarta Jornada de divulgació del món de l'estadística, 4 i 7 de març 2024

Per a què serveix L'ESTADÍSTICA?



ESTADíSTICA ADREÇADA A BATXILLERAT

Activitats part V

UNIVERSITAT DE BARCELONA

GRBIO

LaUB divulga

La resposta correcte s'indica en vermell

Si el coeficient de determinació és 0,81, el coeficient de correlació:

1. És 0,6561
2. Podria ser + 0,9 o - 0,9
3. Ha de ser positiu
4. Ha de ser negatiu

Si el coeficient de correlació és un valor positiu, el pendent de la recta de regressió:

1. També ha de ser positiu
2. Pot ser negatiu o positiu
3. Pot ser zero
4. No pot ser zero

En una anàlisi de regressió entre Y: vendes (en milers d'euros) i X: preu (en euros) va donar com a resultat l'equació següent:

$$Y= 50.000 - 8X$$

L'equació anterior implica que un

1. Augment d'1€ en el preu està associat a una disminució de 8€ en les vendes
2. Augment de 8€ en el preu està associat a un augment de 8.000€ en vendes
3. Augment d'1€ en el preu està associat a una disminució de 42.000€ en vendes
4. Augment d'1€ en el preu està associat amb una disminució de 8000€ en vendes

Si hem calculat la recta de regressió tenint en compte que Y és la variable resposta i X la variable explicativa, i ara ens interessa fer-ho el revés. Tindrem que:

1. La recta de regressió i el coeficient de determinació que obtindrem serà els mateixos.
2. La recta de regressió serà diferent, però el coeficient de determinació serà mateix.
3. La recta de regressió i el coeficient de determinació seran diferents.
4. La recta de regressió serà la mateixa, però el coeficient de determinació serà diferent.