Diagrames de dispersió i regressió lineal. Ús de MINITAB

Víctor Mañosa
Dept. Matemàtica Aplicada III
Universitat Politècnica de Catalunya

1. Objectius

Donada una col·lecció de dades aparellades $\{(x_i,y_i), i=1...,n\}$ volem preguntar-nos:

1) Hi ha relació entre la variable X={x1,...xn} i la variable Y={y1,..yn}?

2) En cas afirmatiu:

a) Quin tipus de relació hi ha?

b) Puc trobar un model matemàtic que les relacioni y=f(x) ?

C) Quin “grau” de relació hi ha?

2. El diagrama de dispersió. Donada la taula de dades aparellades

X	Y
x1	y1
x2	y2
xn	yn

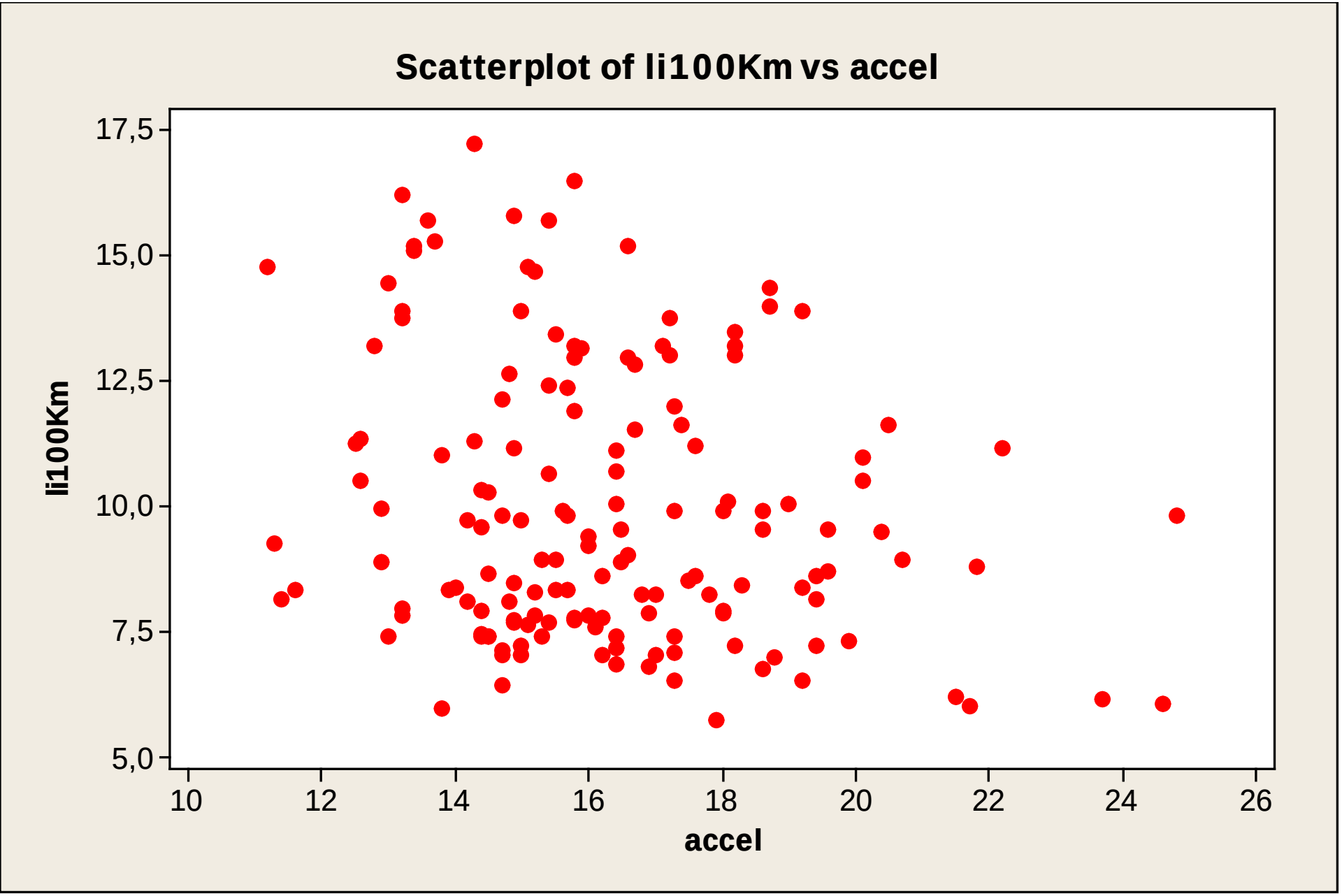
Podem dibuixar el diagrama de dispersió. En els següents exemples tenim

Exemple 1: X= acceleració màxima d'un vehicle, Y= consum. NO HI HAURÀ CORRELACIÓ

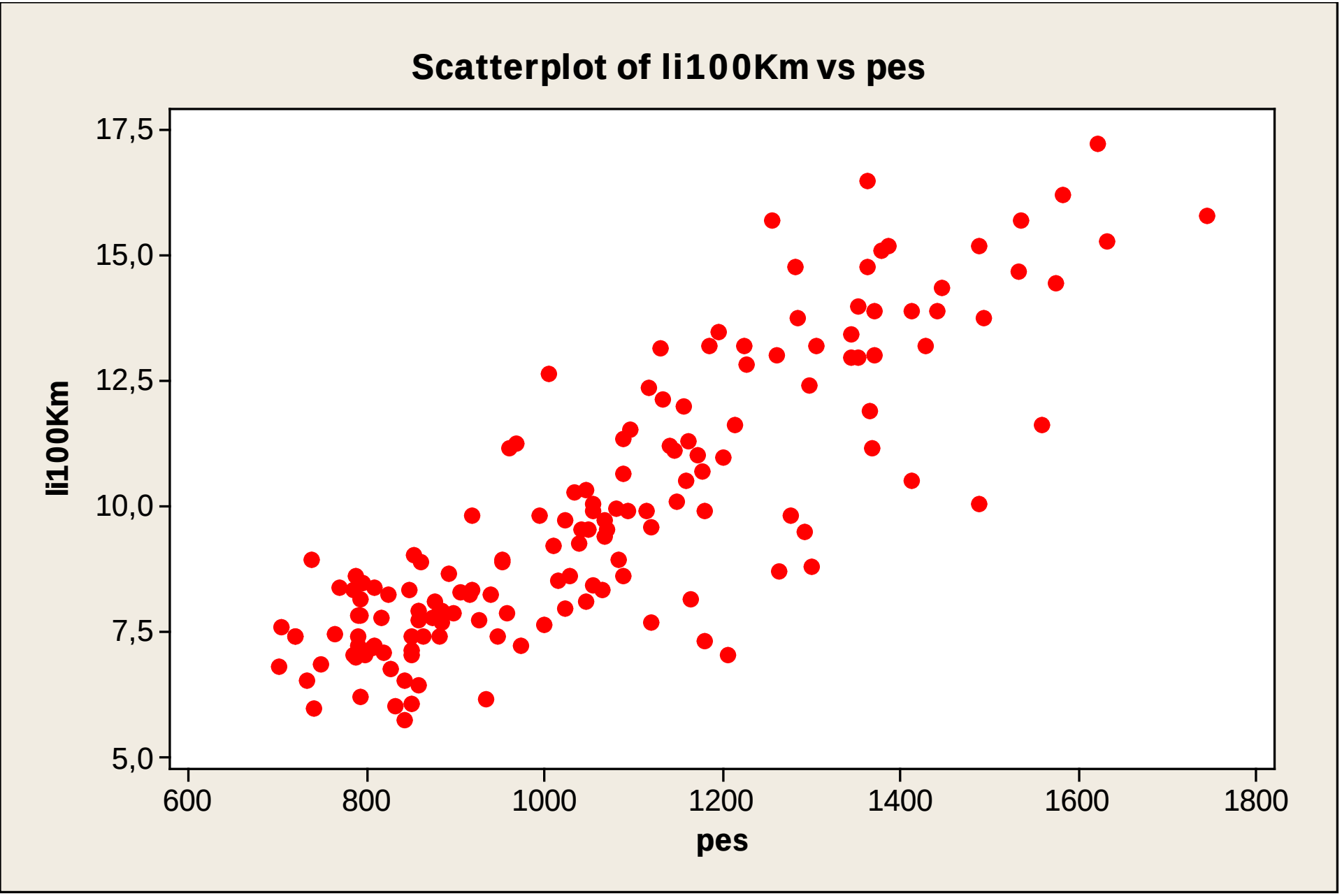
Exemple 2: X= pes d'un vehicle, Y= consum. **HI HAURÀ CORRELACIÓ LINEAL**

Exemple 3: X= preu d'un vehicle, Y= consum. NO HI HAURÀ CORRELACIÓ

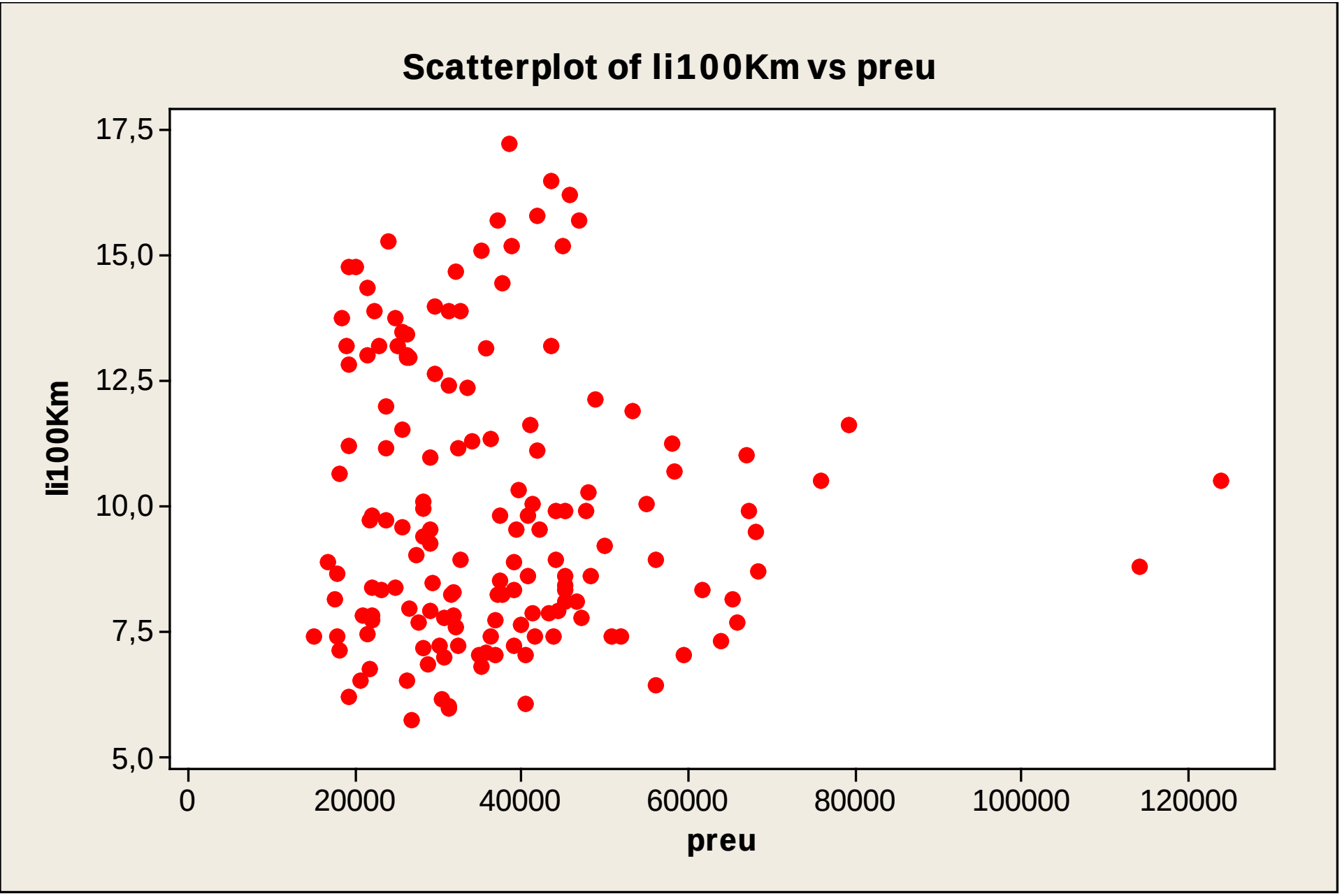
Exemple 1: No hi ha correlació.



Exemple 2: Hi ha correlació lineal.



Exemple 3: No hi ha correlació.



3. Recordem que la recta de regressió ve donada per:

$$y = \frac{\sigma_{XY}}{\sigma_X^2}\left(x - \bar{X}\right) + \bar{Y}$$

On

$$\bar{X} = \frac{\sum_{i=1} x_i}{n}$$

$$\bar{Y} = \frac{\sum_{i=1} y_i}{n}$$

$$\sigma_{XY} = S_{XY} = \frac{\sum_{i=1} x_i y_i}{n} - \bar{X}\bar{Y}$$

$$\sigma_X^2 = S_{n,X}^2 = \frac{\sum_{i=1} x_i^2}{n} - \bar{X}^2$$

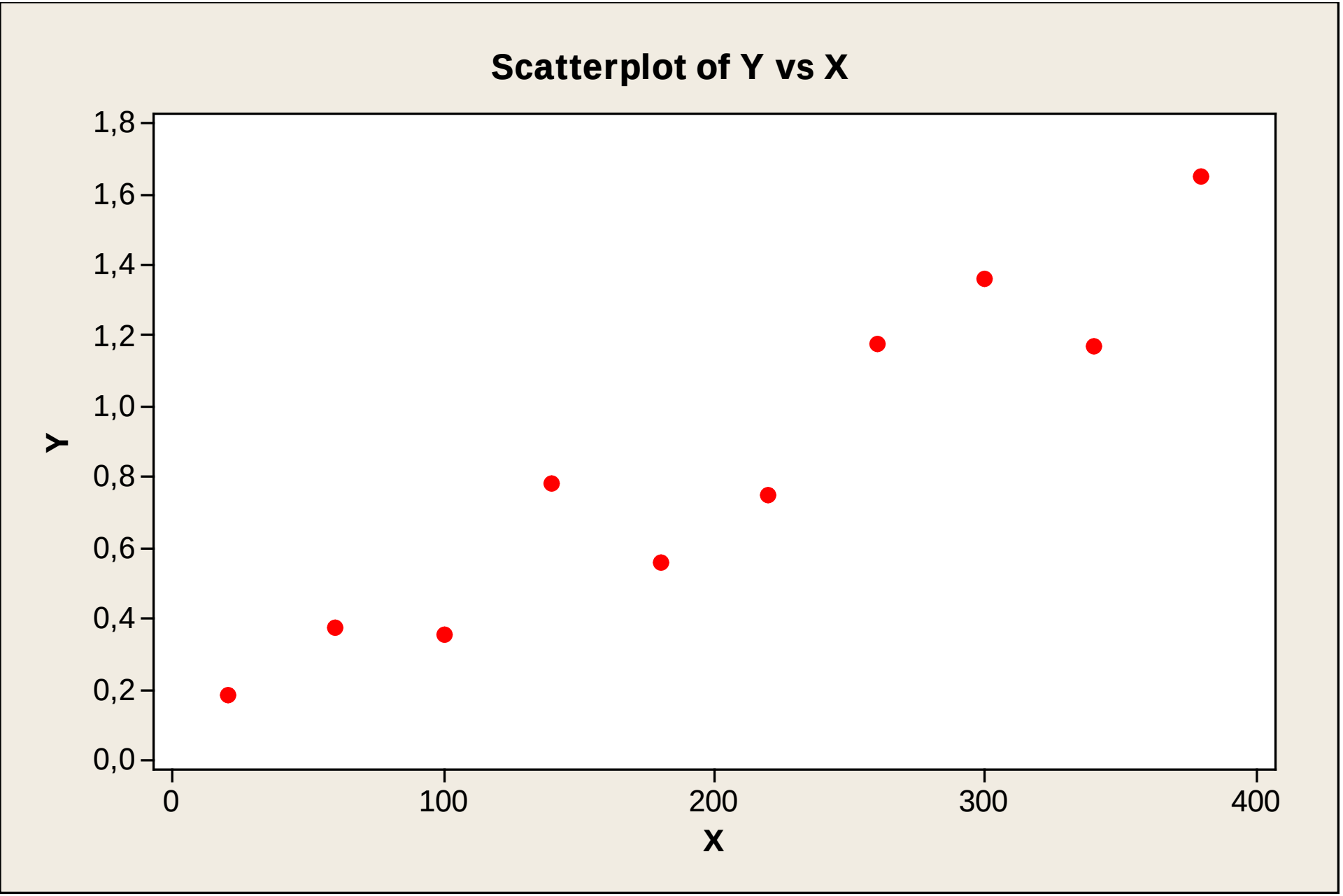
$$\sigma_Y^2 = S_{n,Y}^2 = \frac{\sum_{i=1} y_i^2}{n} - \bar{Y}^2$$

4. Considerem un exemple: El següent és un exemple amb dades proveïdes pel Prof. Josep Gibergans del Departament Matemàtica Aplicada III, UPC.

Les següents dades són les mesures de la velocitat de l'aire i del coeficient d'evaporació de les gotetes de combustible en una turbina de propulsió:

X: Velocitat de l'aire (cm/s)	Y: Coeficient d'evaporació (mm^2/s)
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36
340	1.17
380	1.65

El diagrama de dispersió és:



Si hem de treballar a mà fem:

	X	Y	X^2	Y^2	XY
	20	0,18	400	0,0324	3,6
	60	0,37	3600	0,1369	22,2
	100	0,35	10000	0,1225	35
	140	0,78	19600	0,6084	109,2
	180	0,56	32400	0,3136	100,8
	220	0,75	48400	0,5625	165
	260	1,18	67600	1,3924	306,8
	300	1,36	90000	1,8496	408
	340	1,17	115600	1,3689	397,8
	380	1,65	144400	2,7225	627
SUMA	2000	8,35	532000	9,1097	2175,4

D'aquesta manera obtenim els següents càlculs:

$$n = 10$$

$$\bar{X} = \frac{\sum_{i=1} x_i}{n} = \frac{2000}{10} = 200$$

$$\bar{Y} = \frac{\sum_{i=1} y_i}{n} = \frac{8.35}{10} = 0.835$$

$$\sigma_{XY} = S_{XY} = \frac{\sum_{i=1} x_i y_i}{n} - \bar{X}\bar{Y} = \frac{2175.5}{10} - 200 \cdot 0.835 = 50.54$$

$$\sigma_X^2 = S_{n,X}^2 = \frac{\sum_{i=1} x_i^2}{n} - \bar{X}^2 = \frac{53200}{10} - 200^2 = 13200$$

$$\sigma_Y^2 = S_{n,Y}^2 = \frac{\sum_{i=1} y_i^2}{n} - \bar{Y}^2 = \frac{9.1097}{10} - 0.835^2 = 0.213745$$

Per tant tenim:

$$y = \frac{\sigma_{XY}}{\sigma_X^2}(x - \bar{X}) + \bar{Y} = \frac{50.54}{13200}(x - 200) + 0.835$$

$$= 0.00383(x - 200) + 0.835 = 0.00383x + 0.069$$

... i la recta de regressió és doncs:

$$y = 0.00383x + 0.069$$

El coeficient de determinació:

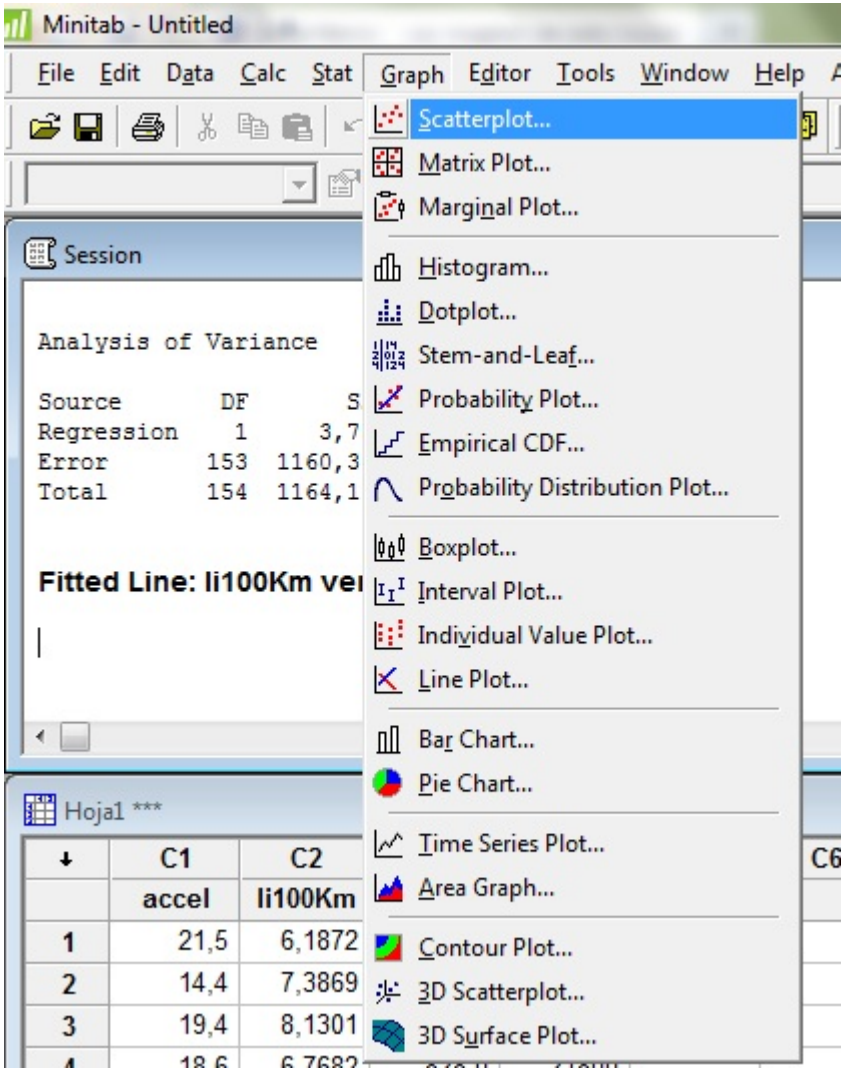
$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} = \frac{50.54^2}{13200 \cdot 0.213745} = 0.9053$$

El que es pot interpretar com un 90,53% de correlació lineal.

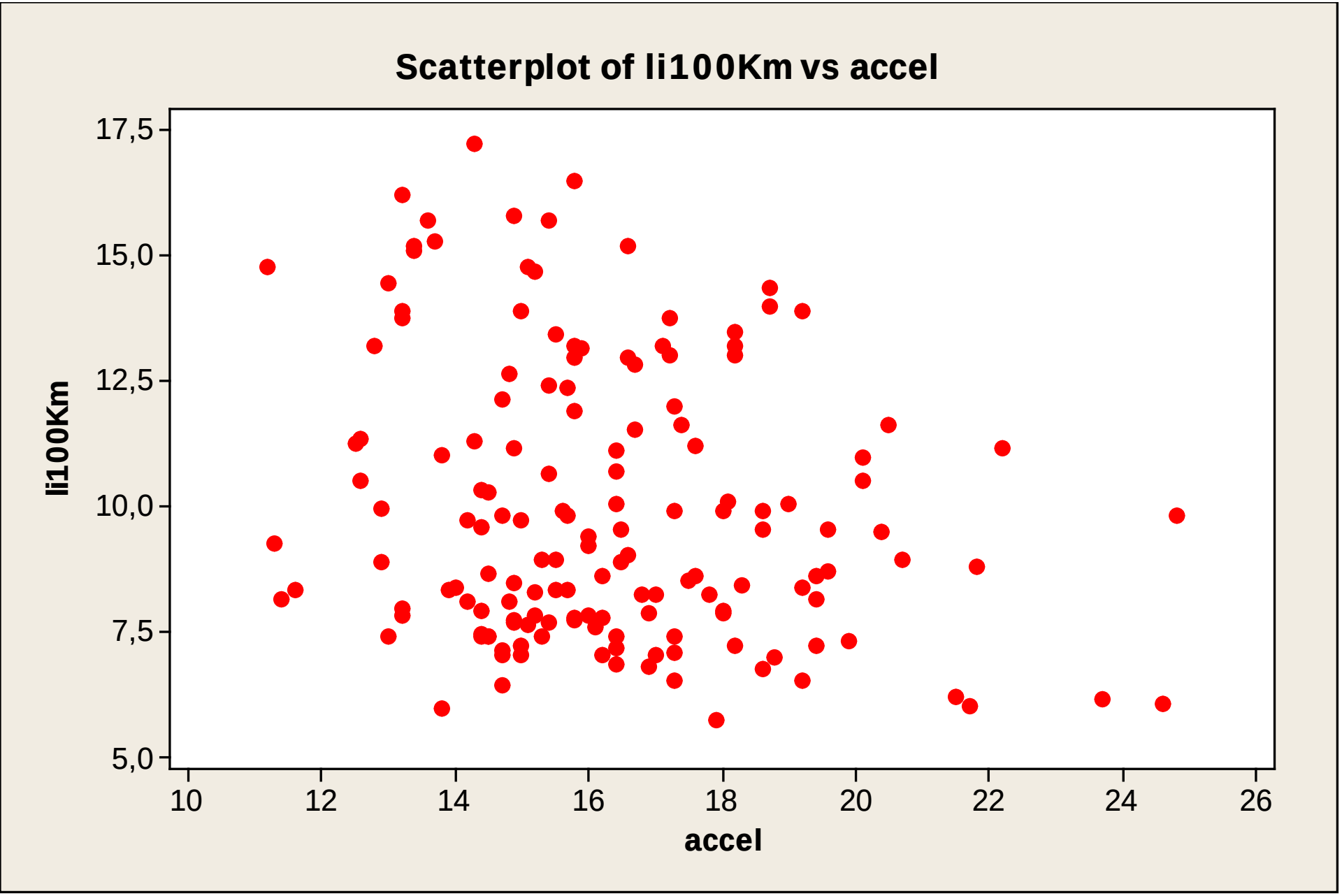
5. Com fer el diagrama de dispersió (Scatterplot) amb MINITAB:

Primerament carregarem les dades d'un arxiu Excel, per exemple l'arxiu *cotxes.xls*, després farem:

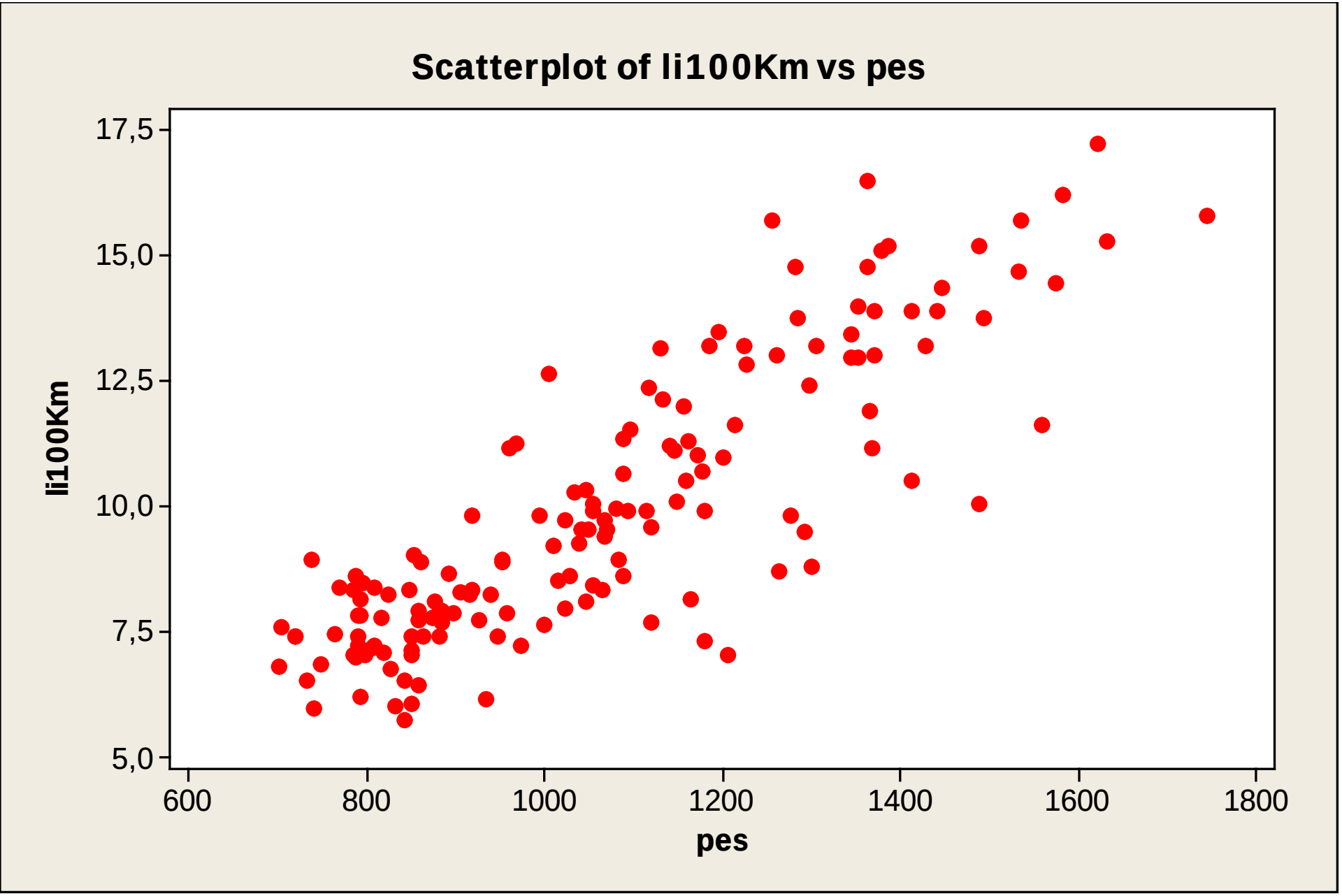
Fixeu-vos



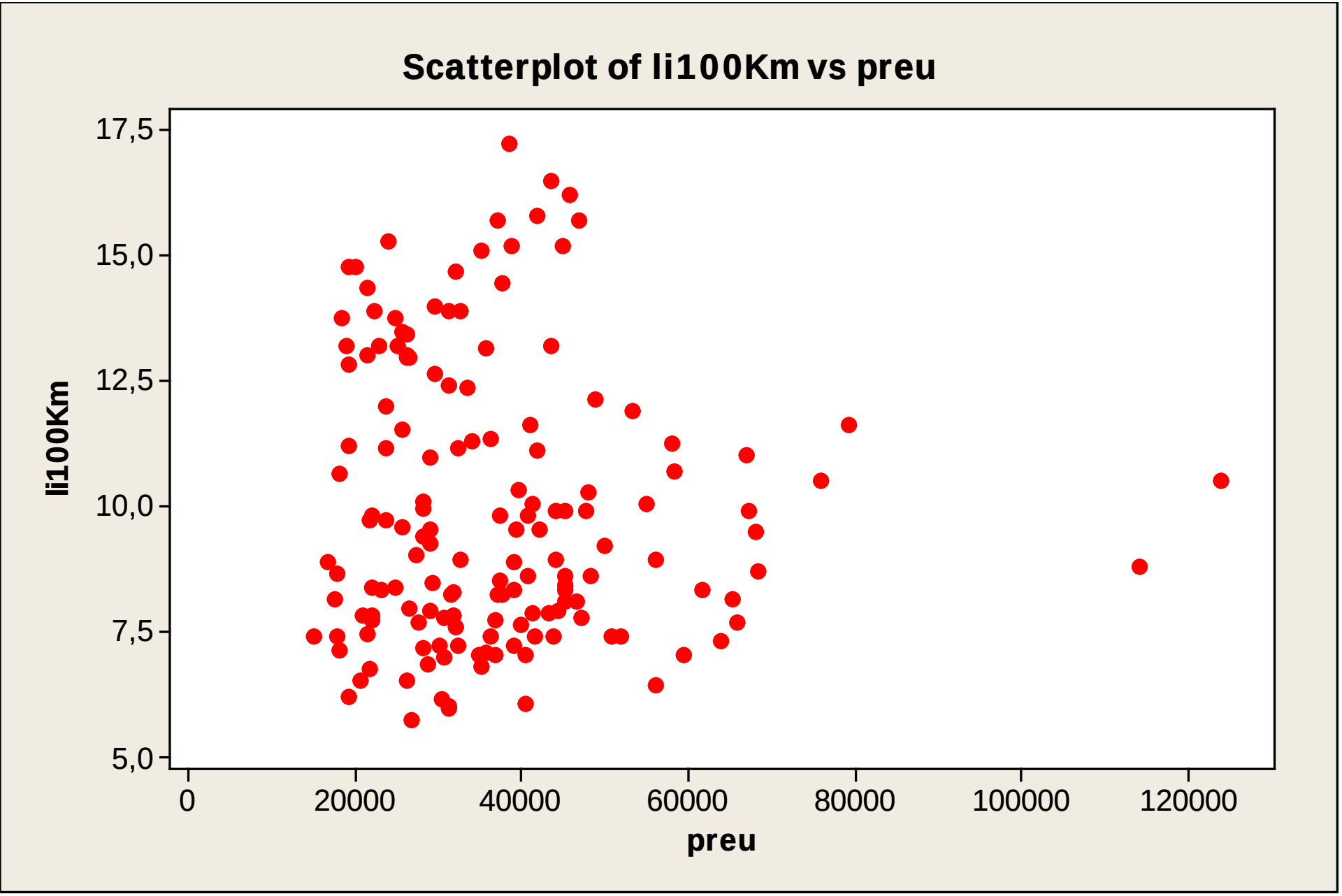
Exemple 1: No hi ha correlació.



Exemple 2: Hi ha correlació.

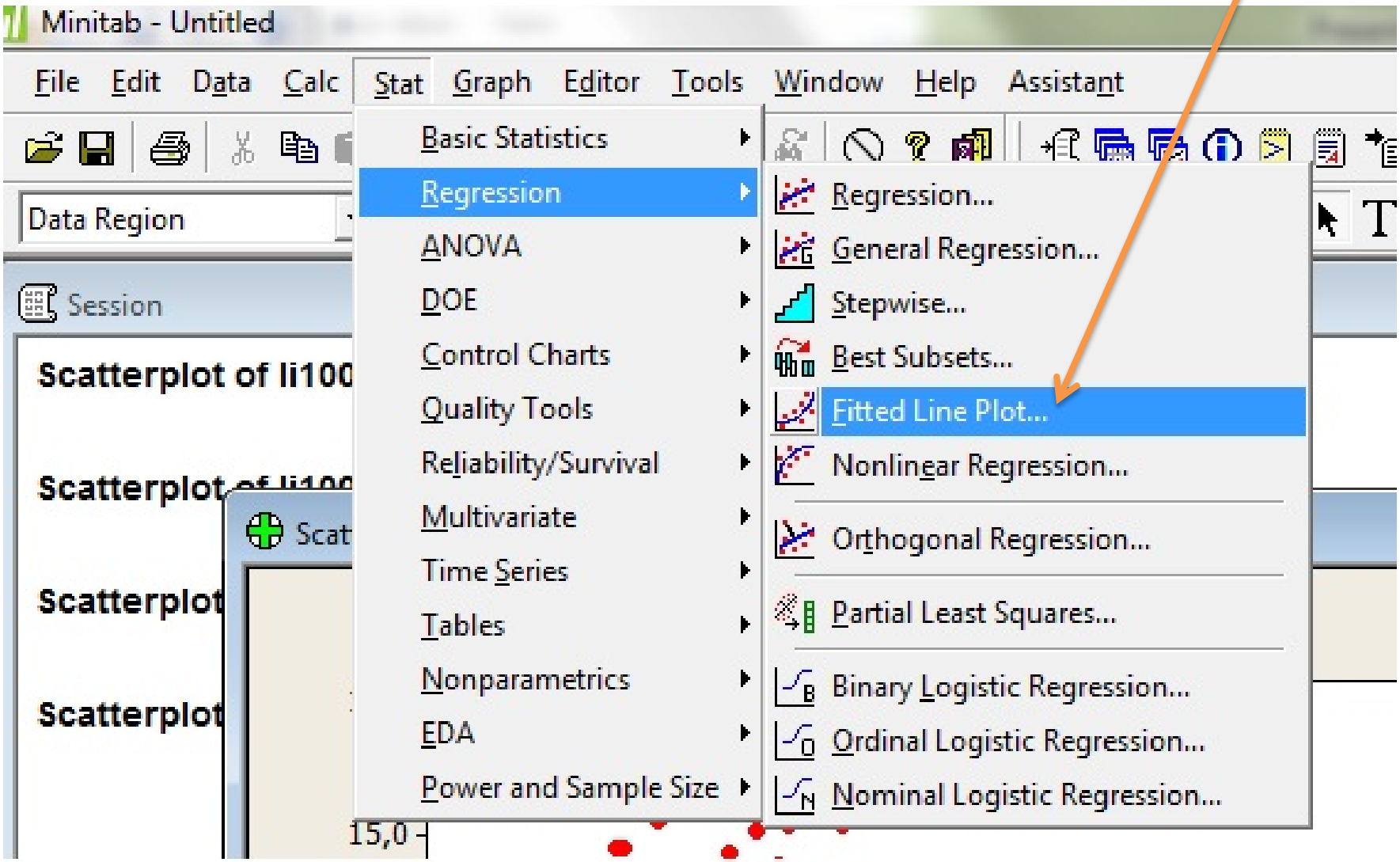


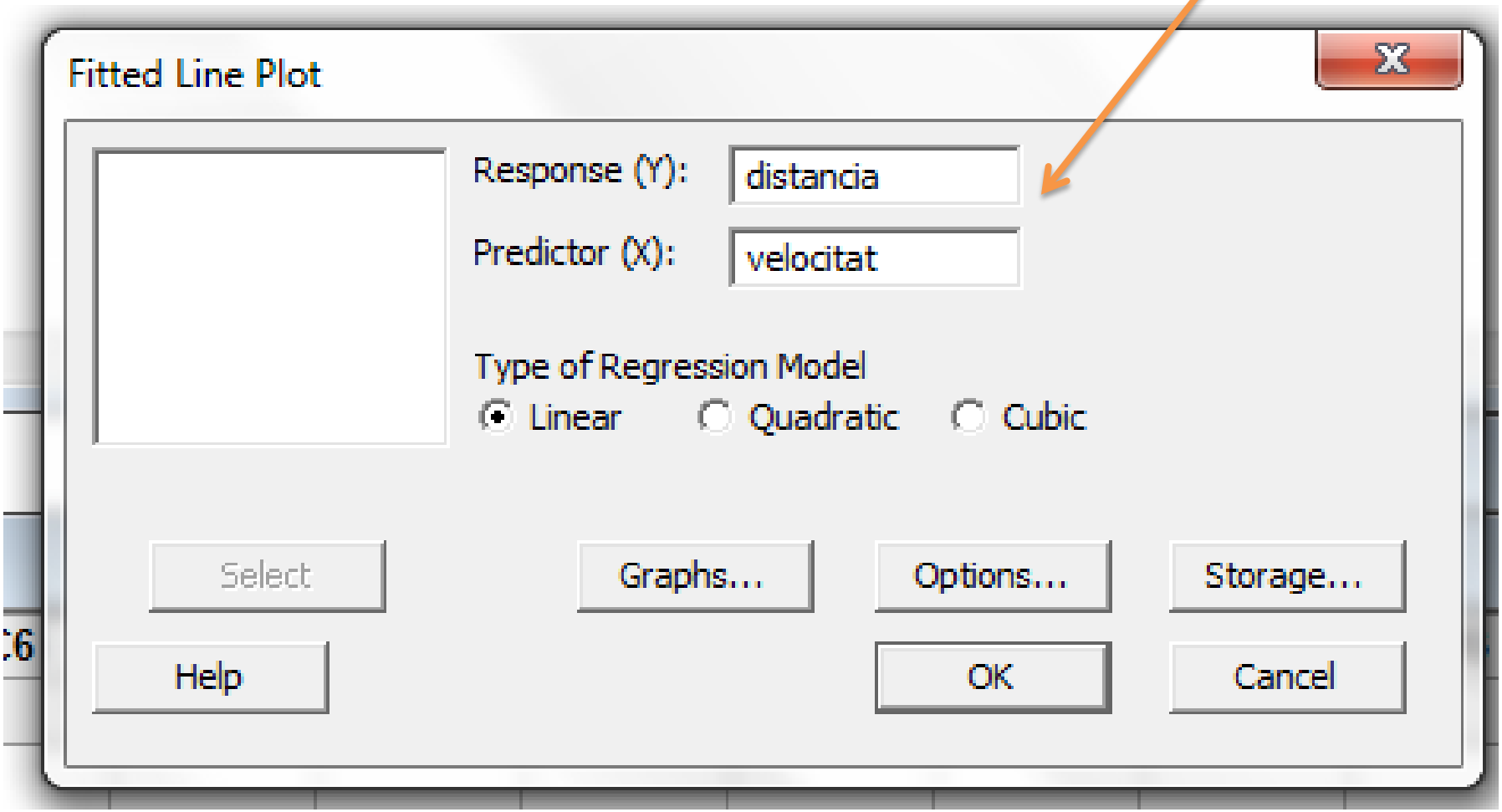
Exemple 3: No hi ha correlació.

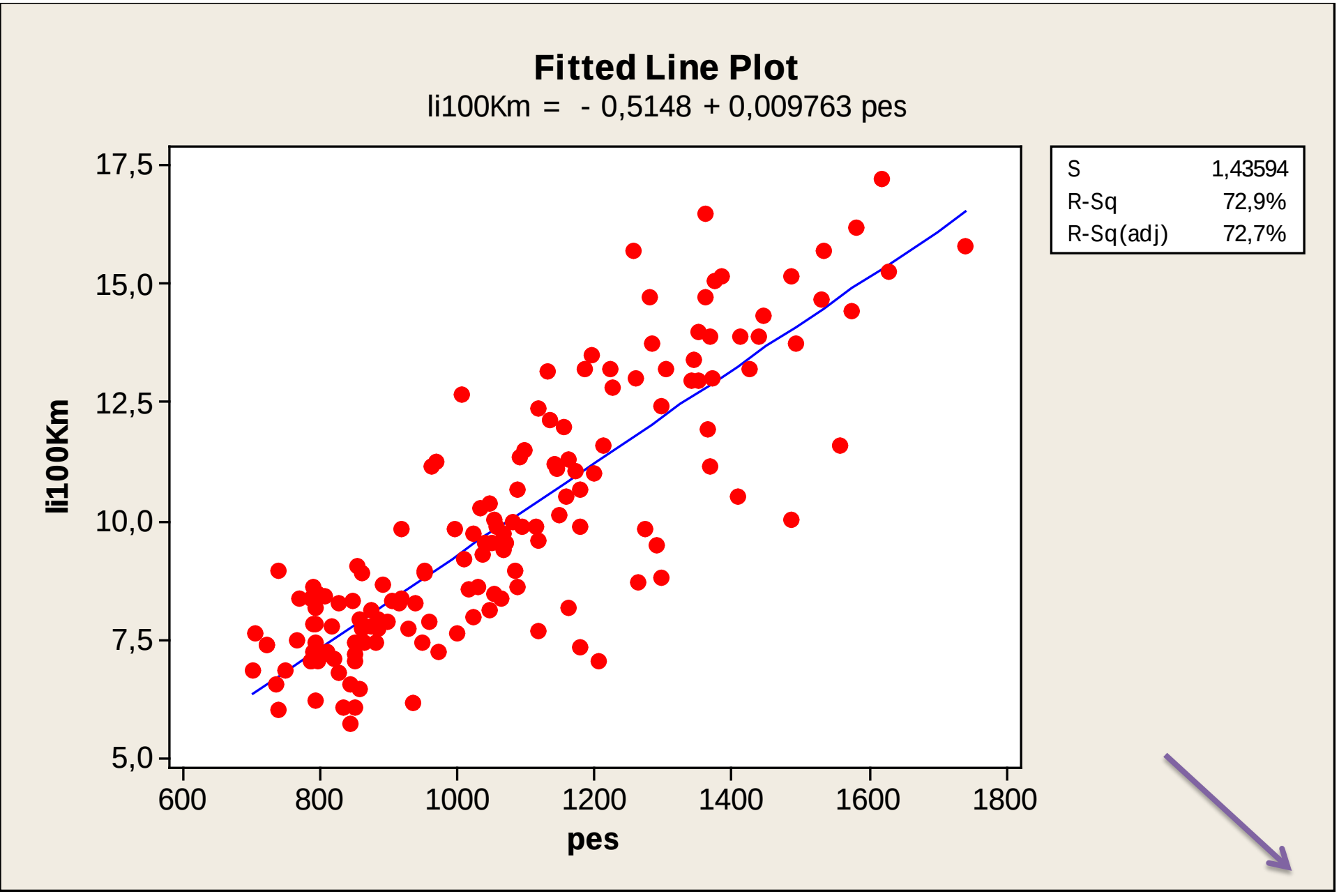


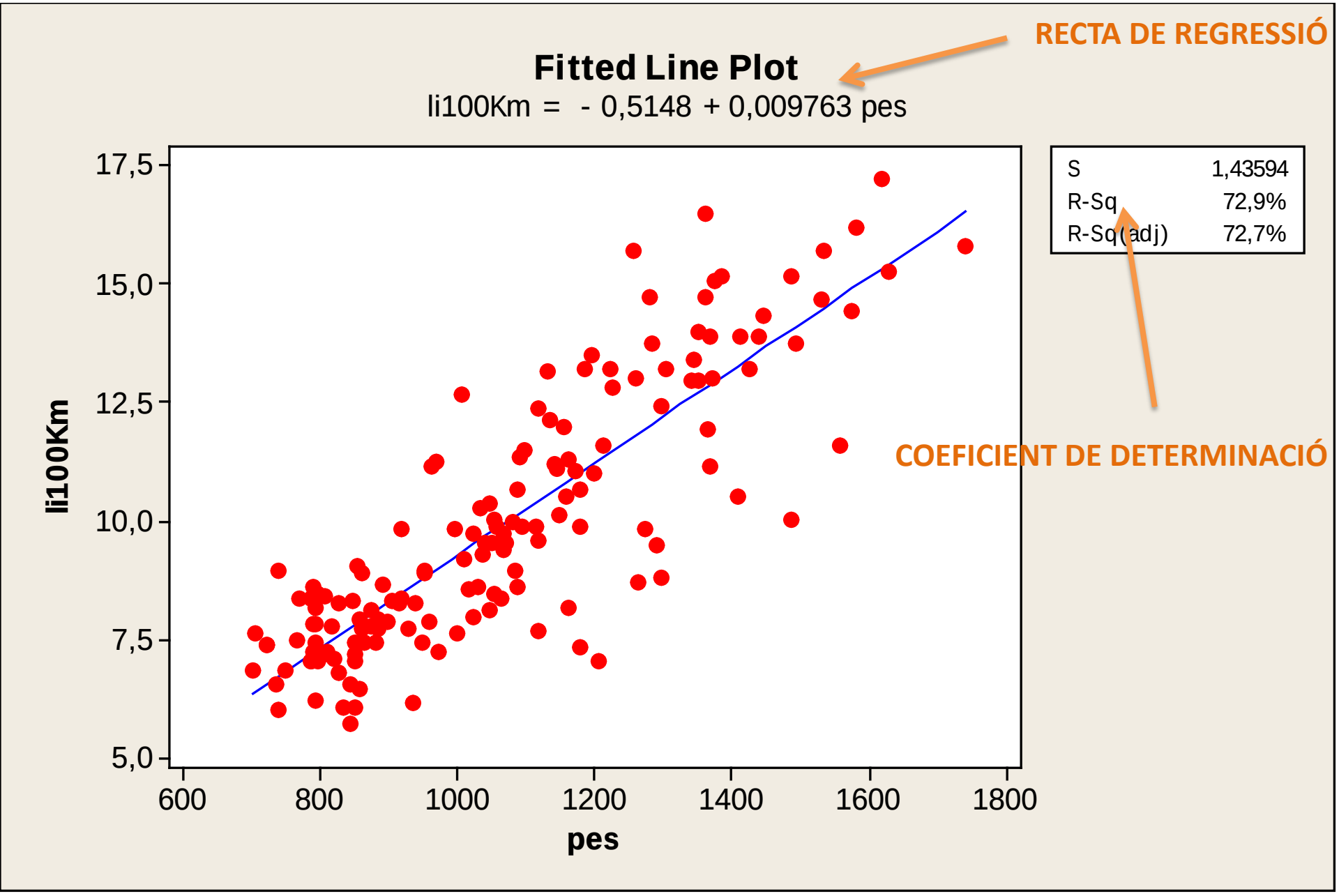
6. Obtenció de la recta de regressió

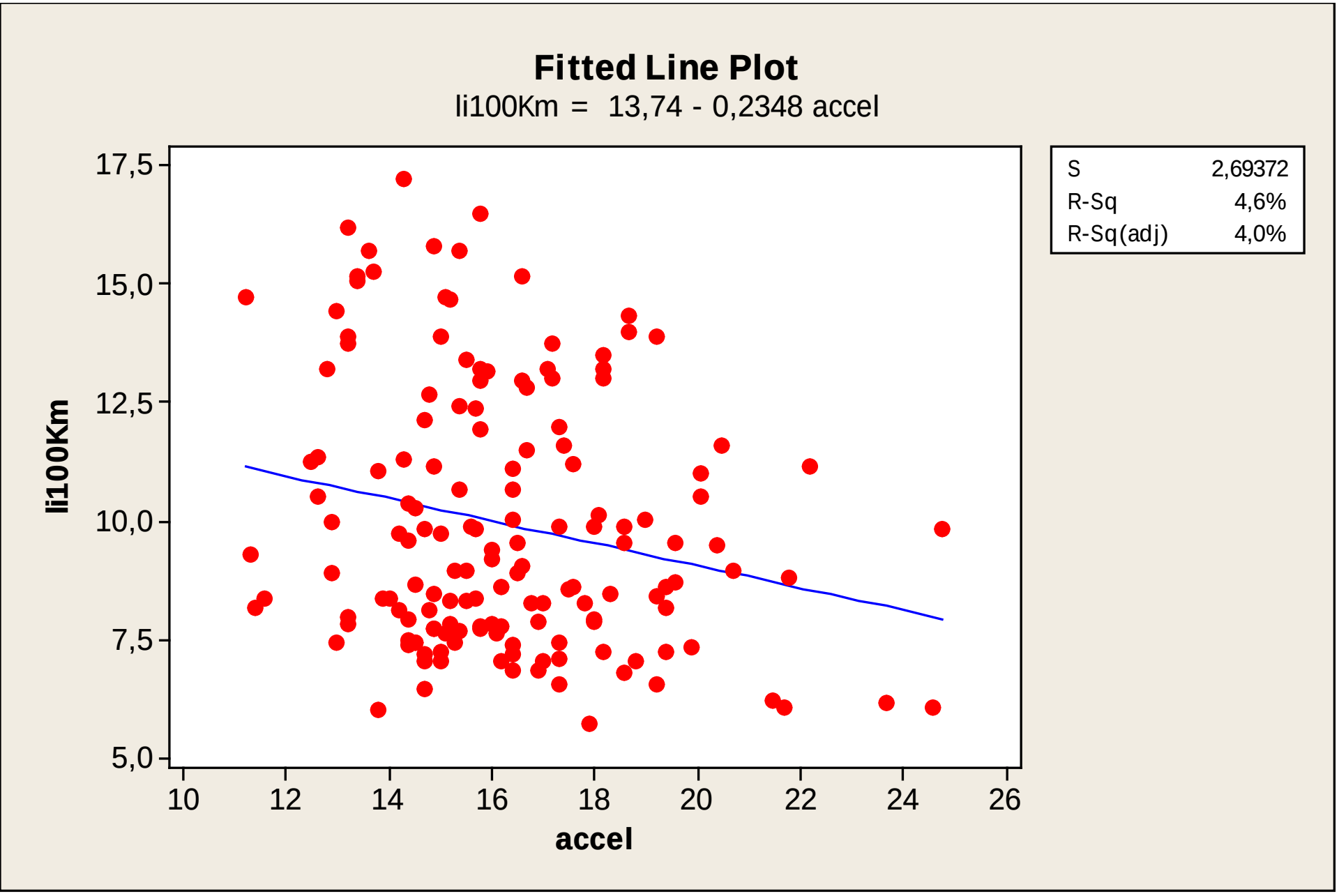
Fixeu-vos

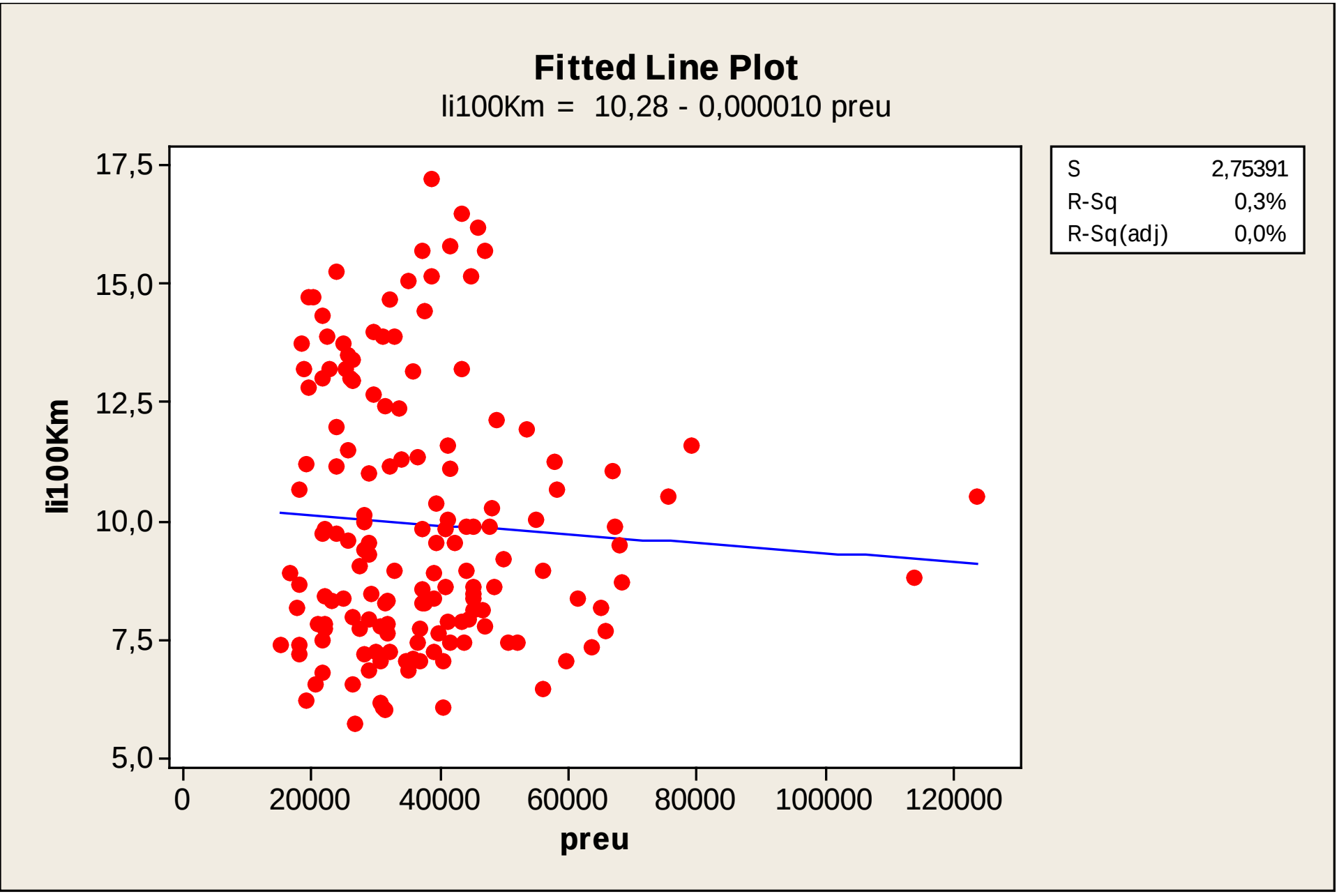



Fixeu-vos
Fitted Line Plot
Response (Y):
distancia
Predictor (X):
velocitat
Type of Regression Model
Linear
Quadratic
Cubic
Select
Graphs...
Options...
Storage...
Help
OK
Cancel


Fitted Line Plot
li100Km = - 0,5148 + 0,009763 pes
S 1,43594
R-Sq 72,9%
R-Sq(adj) 72,7%
li100Km
17,5
15,0
12,5
10,0
7,5
5,0
600
800
1000
1200
1400
1600
1800
pes


RECTA DE REGRESSIÓ
Fitted Line Plot
li100Km = - 0,5148 + 0,009763 pes
S 1,43594
R-Sq 72,9%
R-Sq(adj) 72,7%
COEFICIENT DE DETERMINACIÓ
li100Km
17,5
15,0
12,5
10,0
7,5
5,0
600
800
1000
1200
1400
1600
1800
pes


Fitted Line Plot
li100Km = 13,74 - 0,2348 accel
S 2,69372
R-Sq 4,6%
R-Sq(adj) 4,0%
li100Km
17,5
15,0
12,5
10,0
7,5
5,0
10
12
14
16
18
20
22
24
26
accel

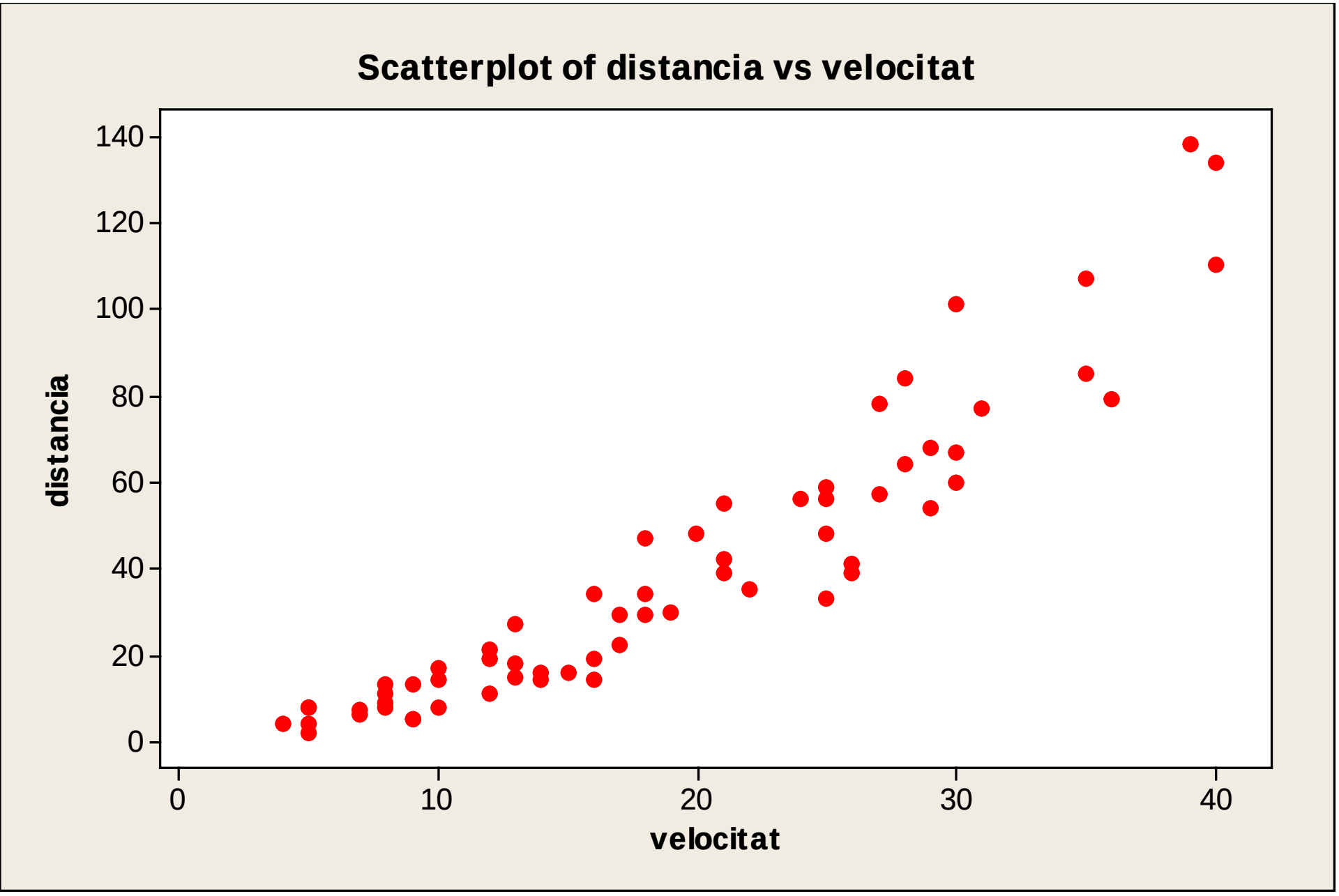

Fitted Line Plot
li100Km = 10,28 - 0,000010 preu
S 2,75391
R-Sq 0,3%
R-Sq(adj) 0,0%
li100Km
17,5
15,0
12,5
10,0
7,5
5,0
0
20000
40000
60000
80000
100000
120000
preu

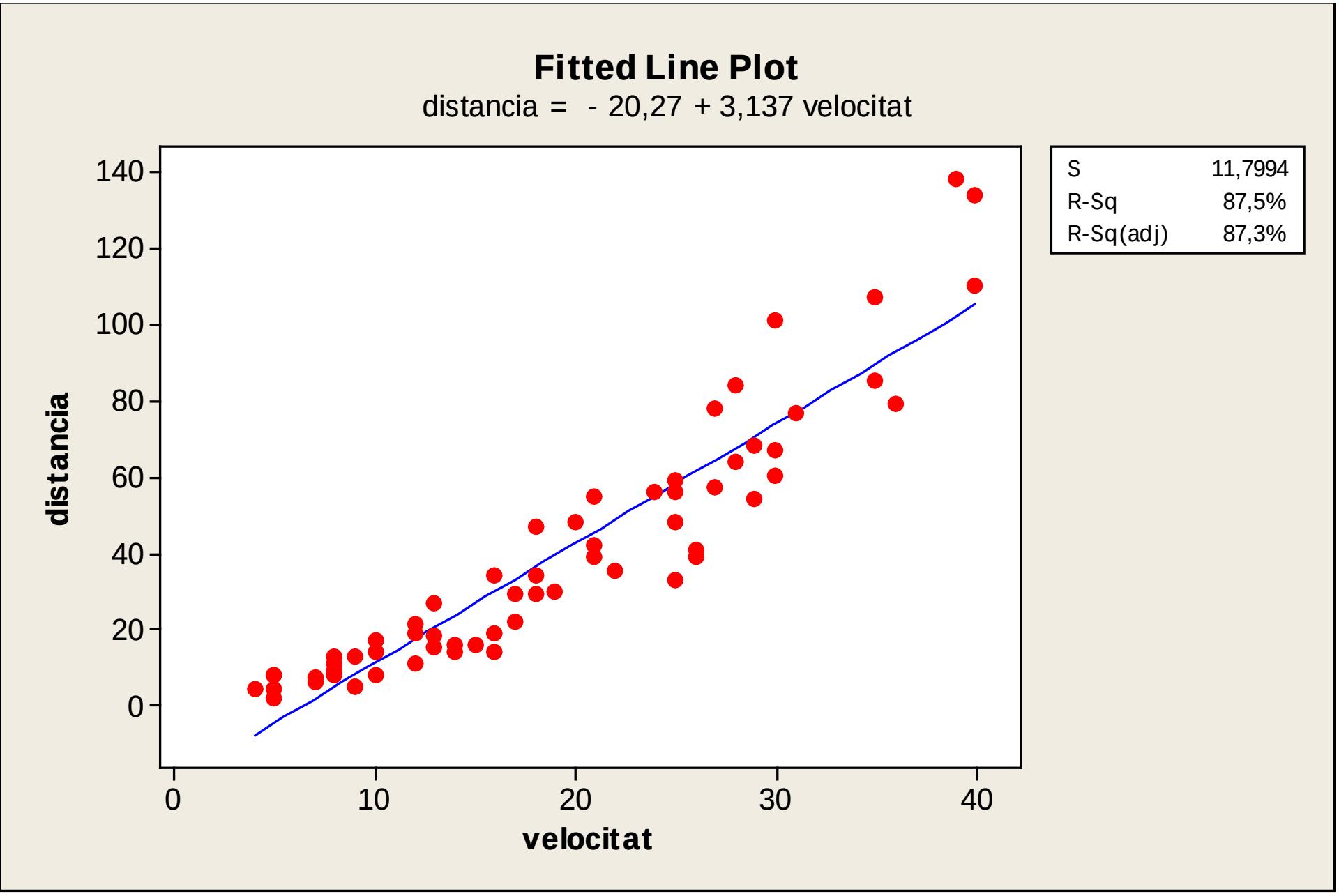
7. Regressió no lineal

POT PASSAR QUE UN MODEL **NO LINEAL** SIGUI MILLOR

Carregarem les dades de l'arxiu *distfrenada.xls*,

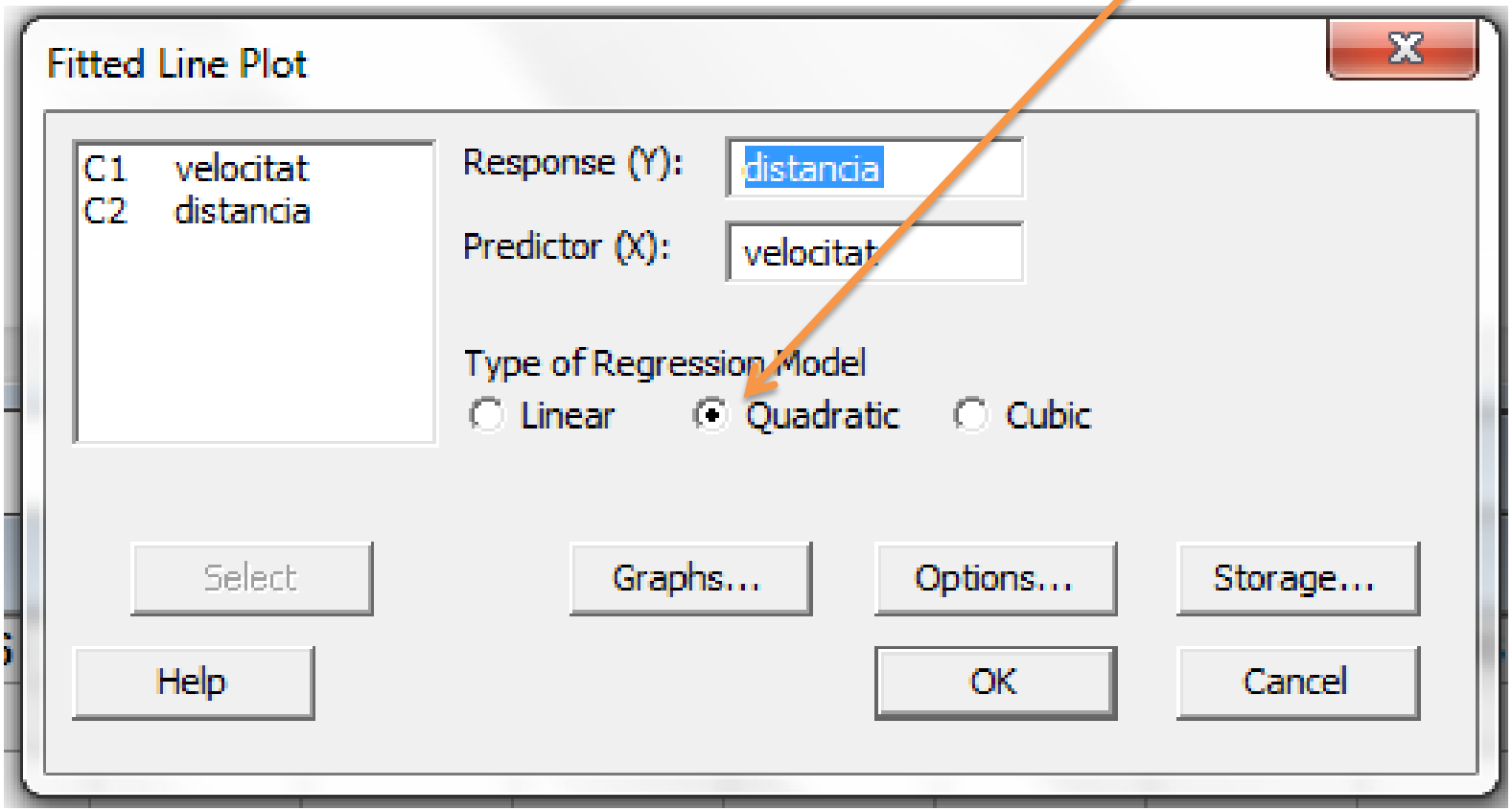
Conté dades sobre la distancia de frenada contra la velocitat d'un cert vehicle.

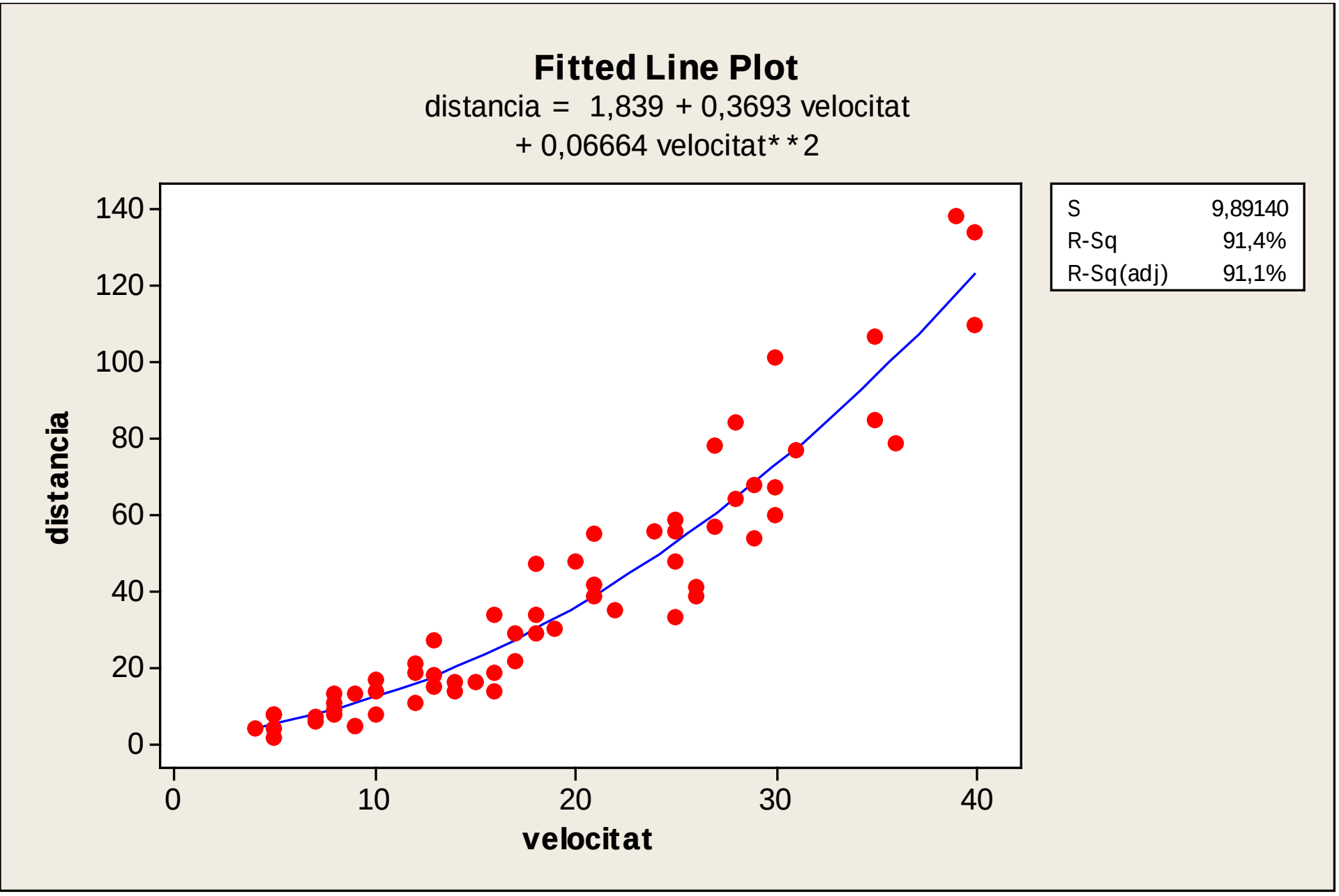

Scatterplot of distancia vs velocitat
distancia
velocitat
0
20
40
60
80
100
120
140
0
10
20
30
40


Fitted Line Plot
distancia = - 20,27 + 3,137 velocitat
S 11,7994
R-Sq 87,5%
R-Sq(adj) 87,3%
distancia
velocitat
0
20
40
60
80
100
120
140
0
10
20
30
40

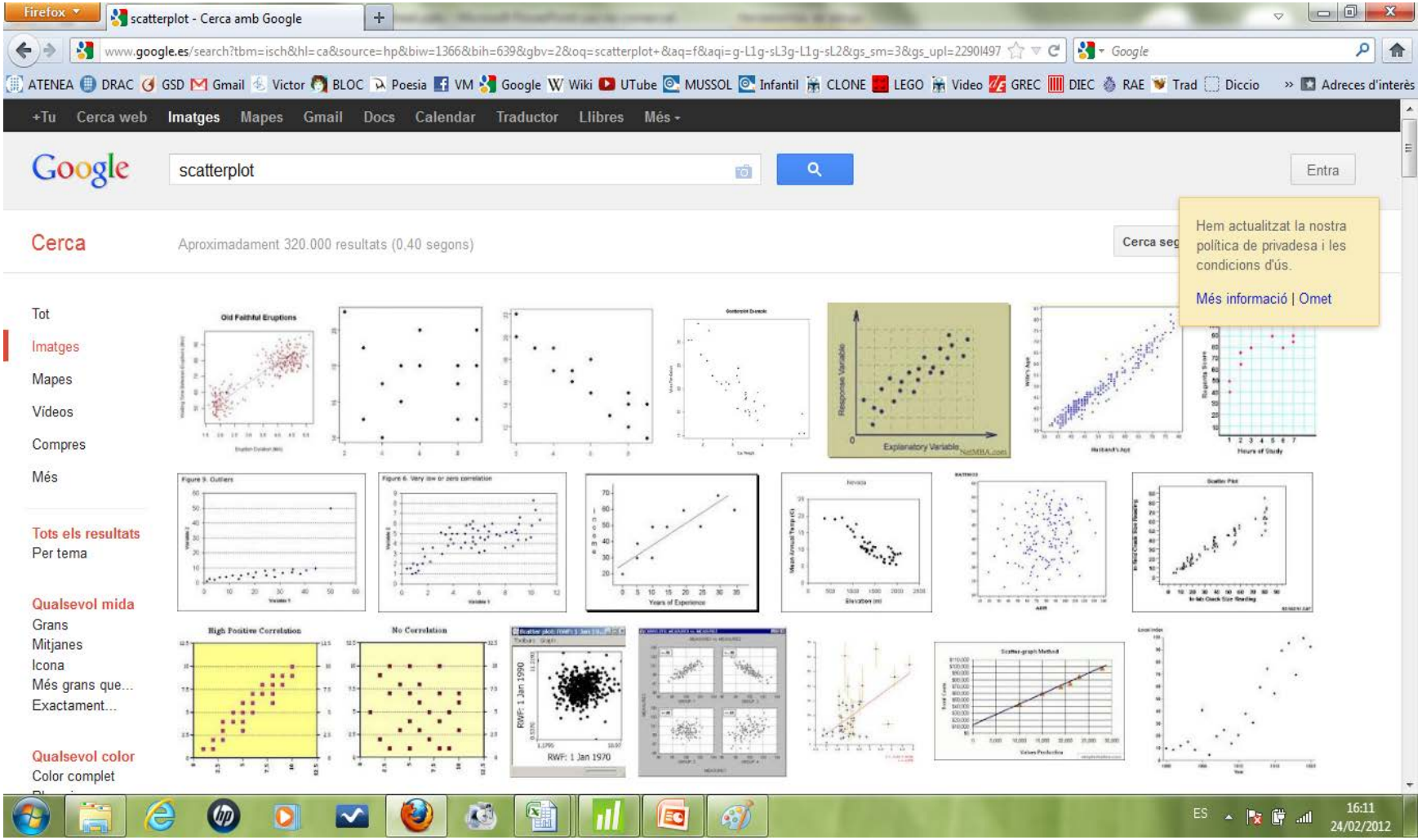
Però...

Fixeu-vos



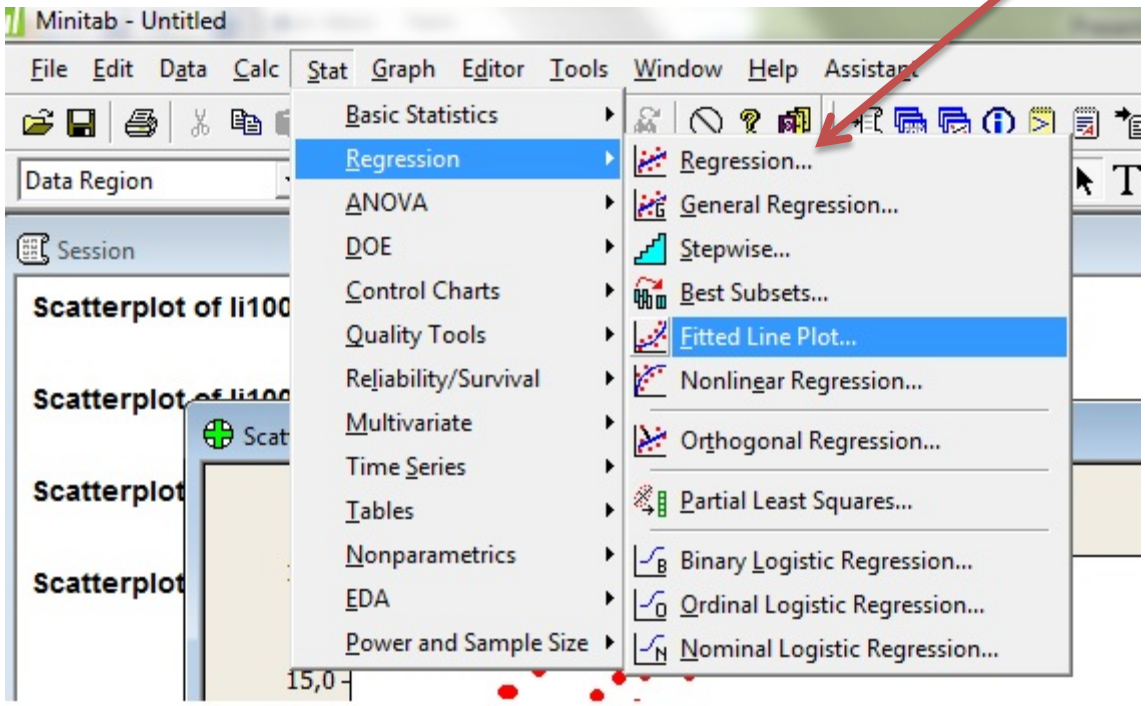

Fitted Line Plot
distancia = 1,839 + 0,3693 velocitat
+ 0,06664 velocitat**2
S 9,89140
R-Sq 91,4%
R-Sq(adj) 91,1%
distancia
velocitat
0
20
40
60
80
100
120
140
0
10
20
30
40

Si voleu tenir més exemples, poseu “scatterplot” en un buscador:



8. ANÀLISI DELS RESIDUS

SI ESCOLLIU "REGRESSION"

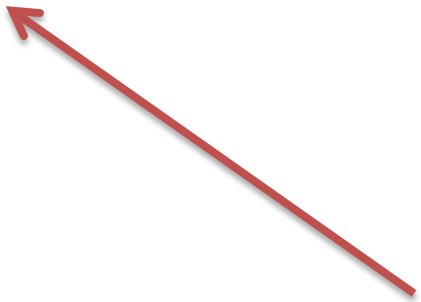


Regression Analysis: distancia versus velocitat

The regression equation is
distancia = - 20,3 + 3,14 velocitat

Predictor	**Coef**	SE Coef	T	P
Constant	**-20,273**	3,238	-6,26	0,000
velocitat	**3,1366**	0,1517	20,68	0,000

S = 11,7994 **R-Sq = 87,5%** R-Sq(adj) = 87,3%



Això és la nostra ☏

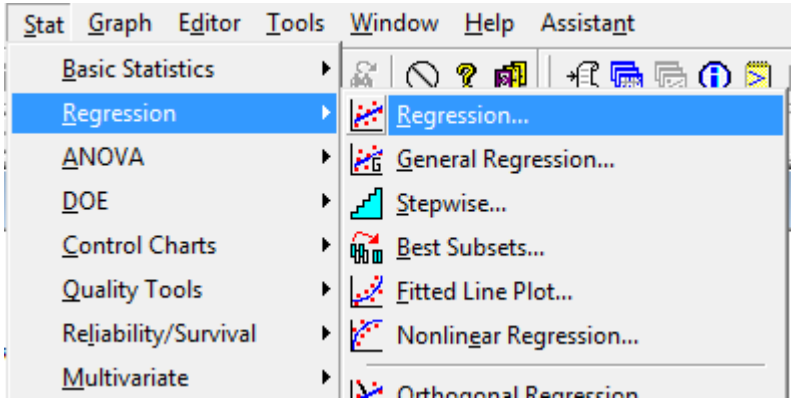
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	59540	59540	427,65	0,000
Residual Error	61	**8493**	139		
Total	62	68033			

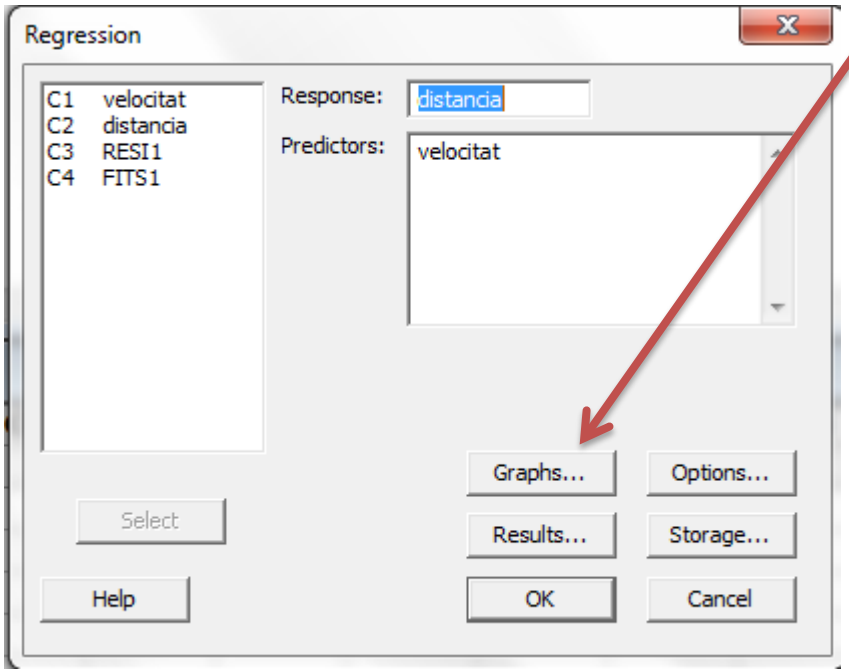
Unusual Observations

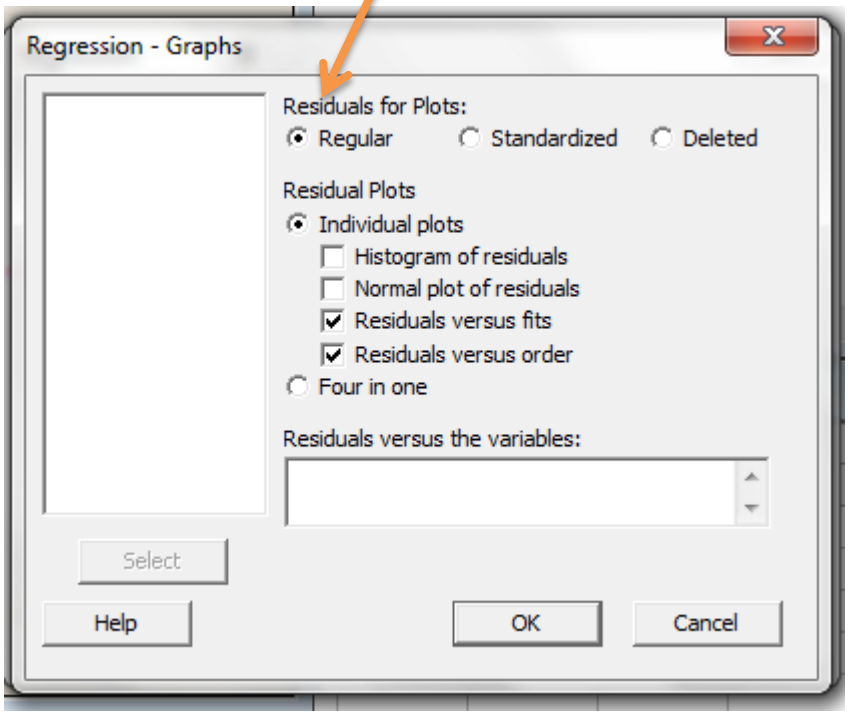
Obs	velocitat	distancia	Fit	SE Fit	Residual	St Resid
24	25,0	33,00	58,14	1,75	-25,14	-2,15R
42	30,0	101,00	73,82	2,24	27,18	2,35R
59	39,0	138,00	102,05	3,38	35,95	3,18R
63	40,0	134,00	105,19	3,52	28,81	2,56R

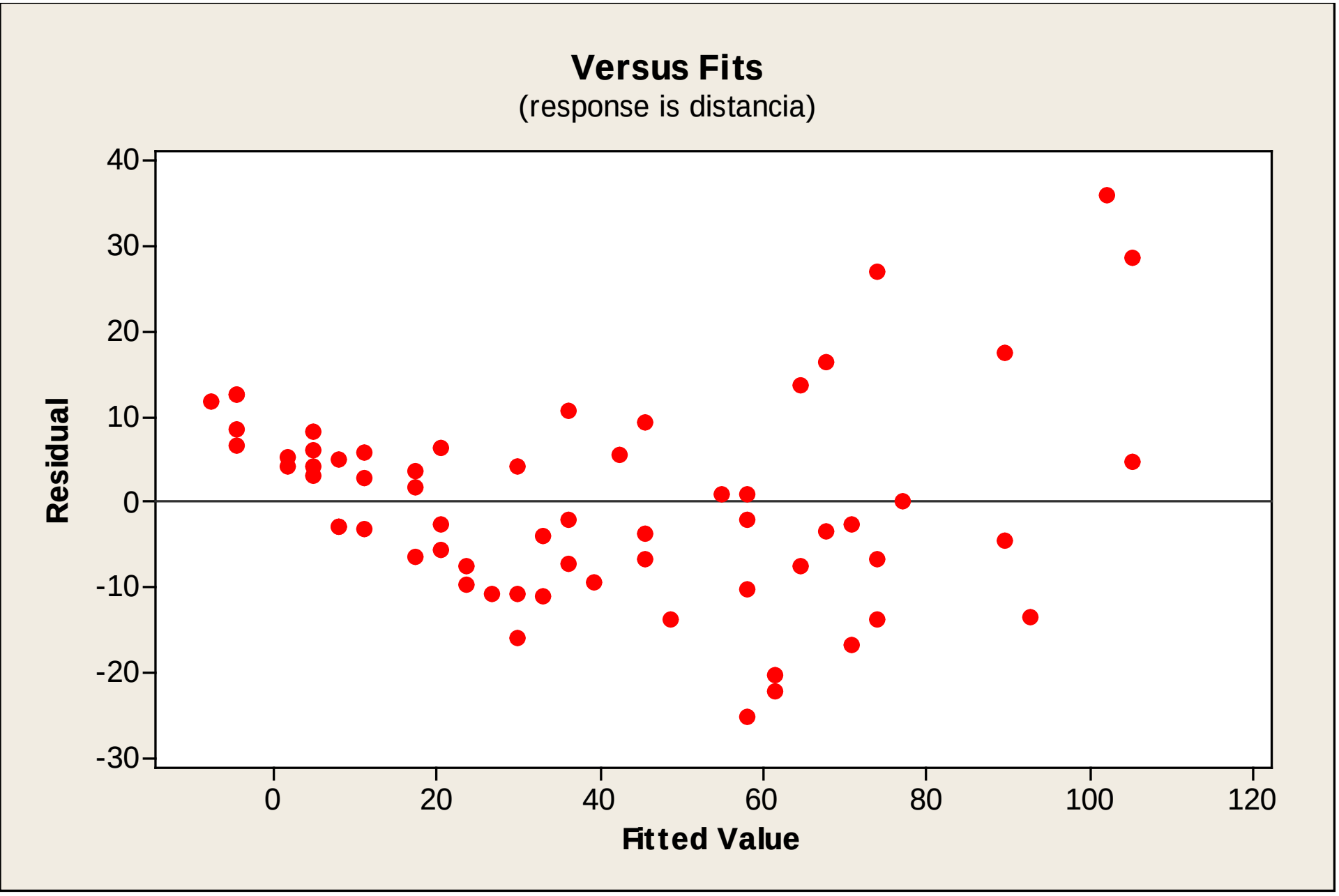
R denotes an observation with a large standardized residual.

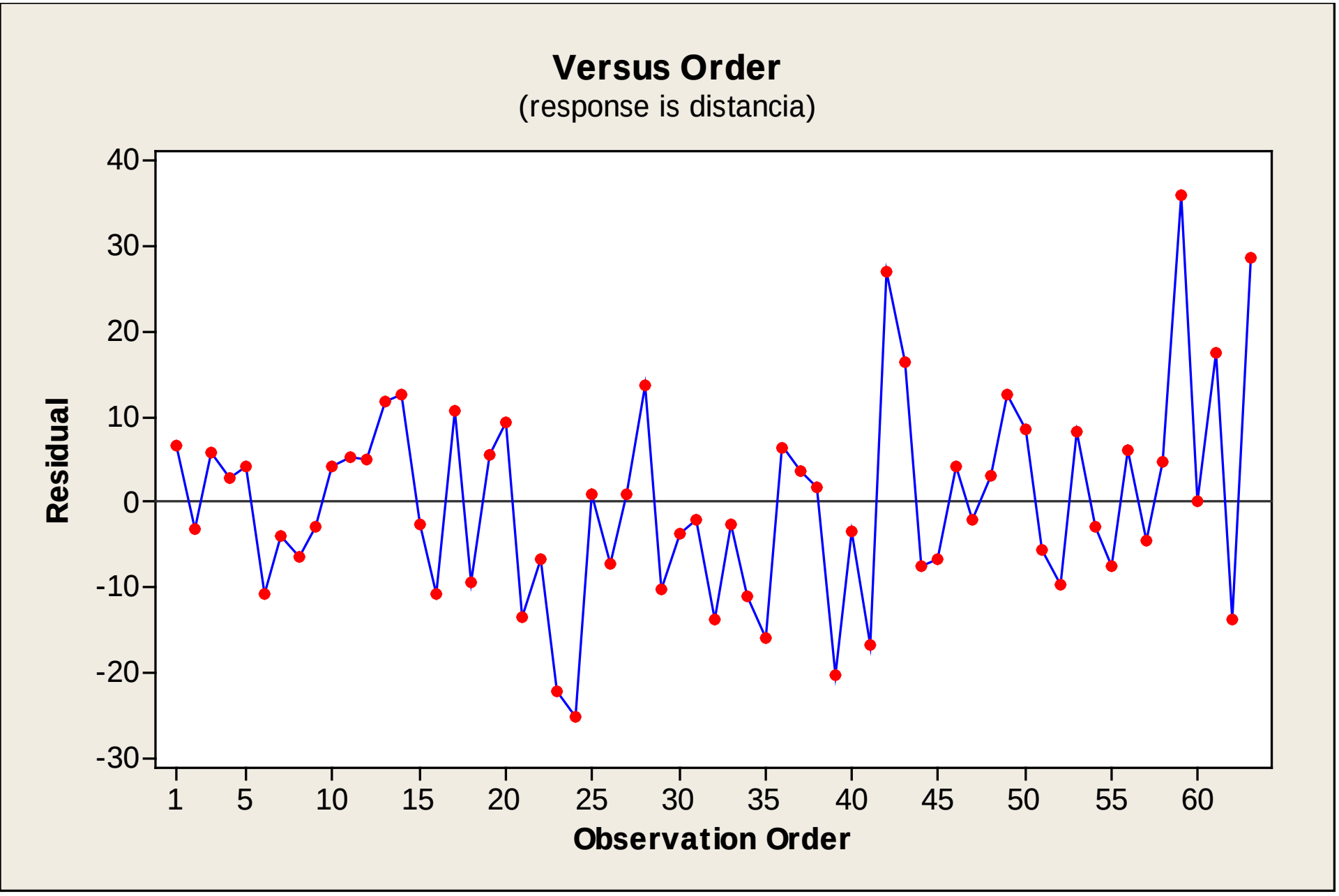

Stat
Graph
Editor
Tools
Window
Help
Assistant
Basic Statistics
Regression
ANOVA
DOE
Control Charts
Quality Tools
Reliability/Survival
Multivariate
Regression...
General Regression...
Stepwise...
Best Subsets...
Fitted Line Plot...
Nonlinear Regression...
Orthogonal Regression

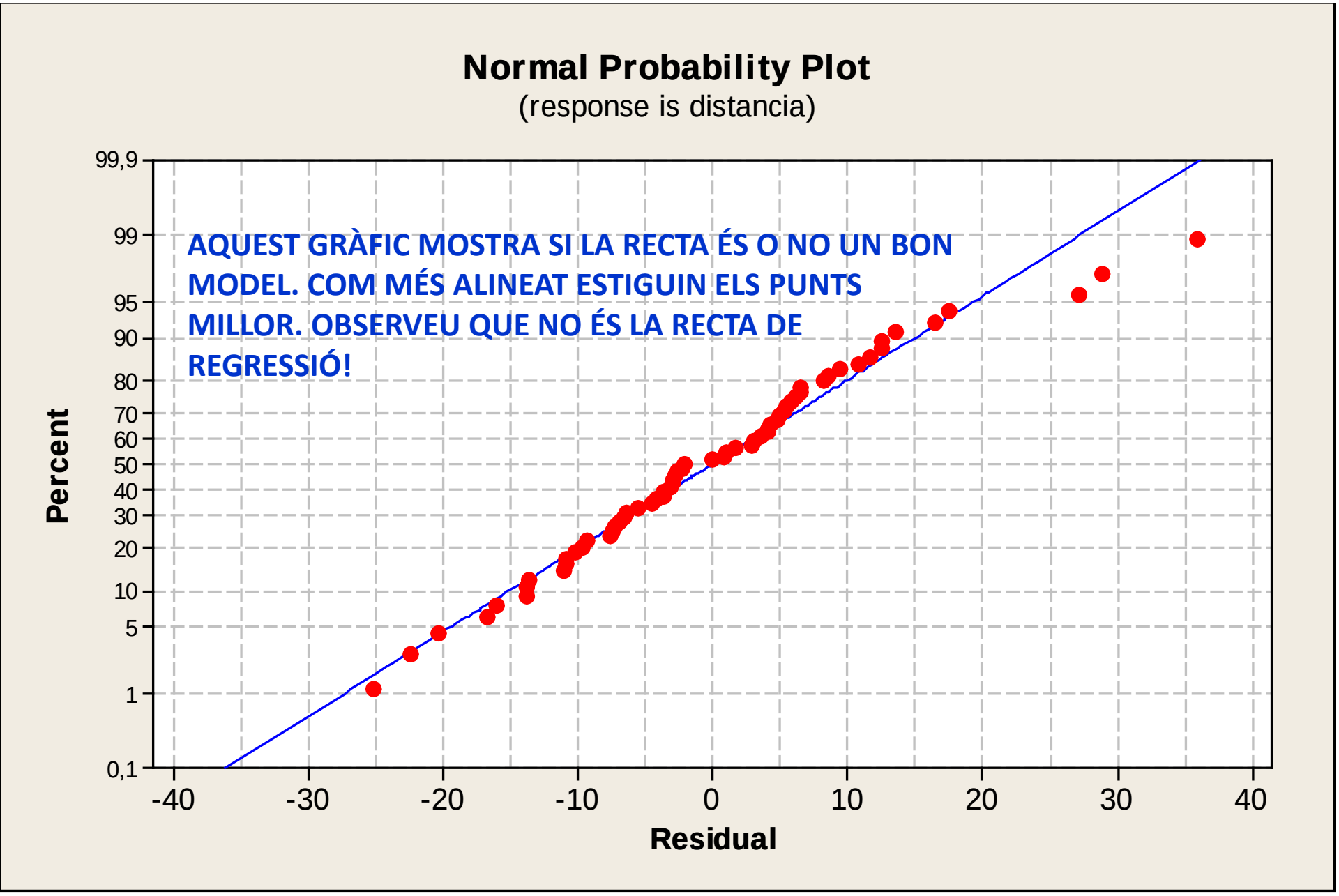
Fixeu-vos

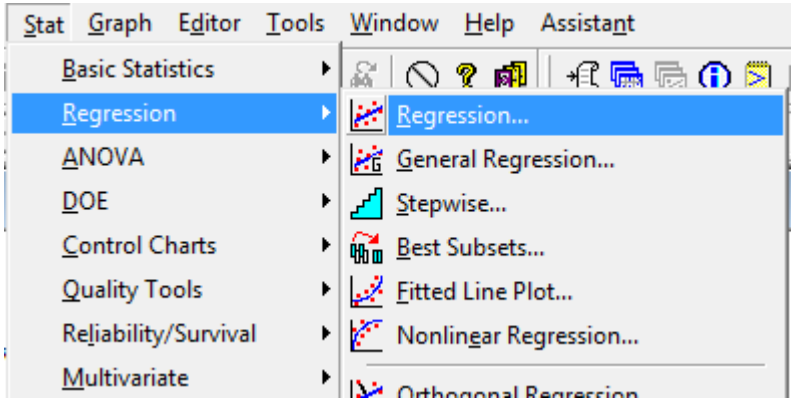

Regression
C1 velocitat
C2 distancia
C3 RESI1
C4 FITS1
Response:
distancia
Predictors:
velocitat
Graphs...
Options...
Results...
Storage...
Select
Help
OK
Cancel


Regression - Graphs
Residuals for Plots:
Regular
Standardized
Deleted
Residual Plots
Individual plots
Histogram of residuals
Normal plot of residuals
Residuals versus fits
Residuals versus order
Four in one
Residuals versus the variables:
Select
Help
OK
Cancel


Versus Fits
(response is distancia)
Residual
Fitted Value
40
30
20
10
0
-10
-20
-30
0
20
40
60
80
100
120

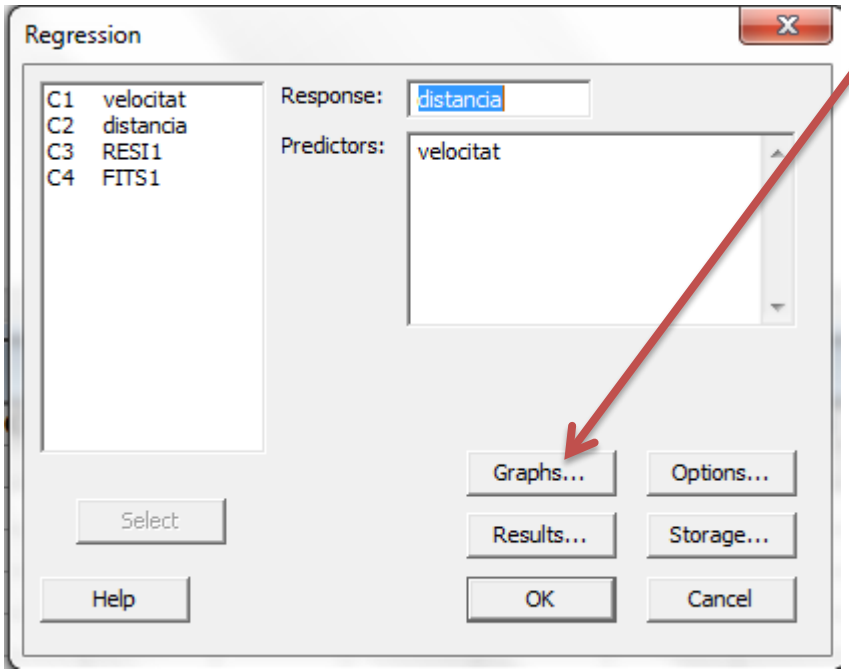

Versus Order
(response is distancia)
Residual
40
30
20
10
0
-10
-20
-30
1
5
10
15
20
25
30
35
40
45
50
55
60
Observation Order

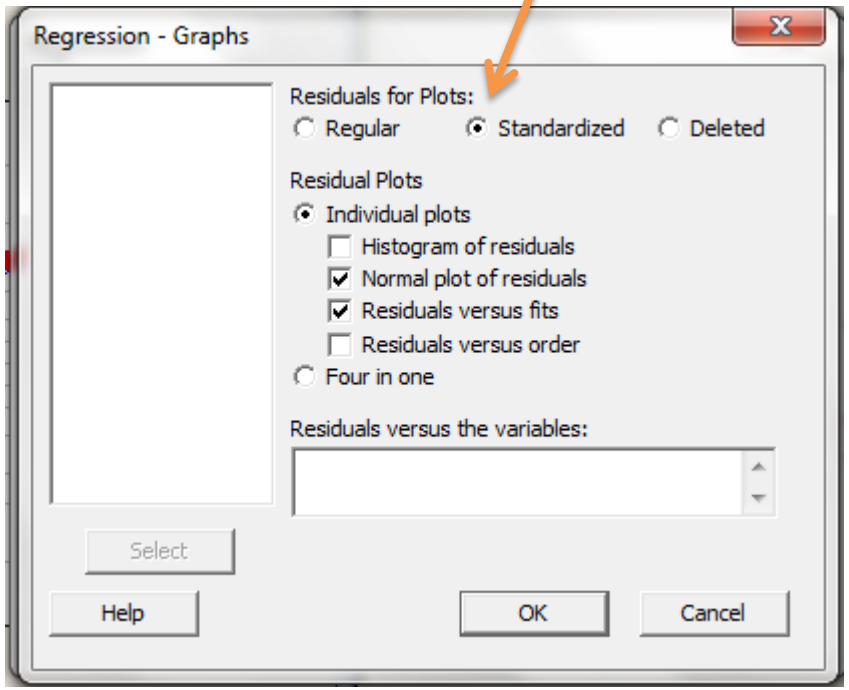

Normal Probability Plot
(response is distancia)
AQUEST GRÀFIC MOSTRA SI LA RECTA ÉS O NO UN BON MODEL. COM MÉS ALINEAT ESTIGUIN ELS PUNTS MILLOR. OBSERVEU QUE NO ÉS LA RECTA DE REGRESSIÓ!
Percent
99,9
99
95
90
80
70
60
50
40
30
20
10
5
1
0,1
-40
-30
-20
-10
0
10
20
30
40
Residual

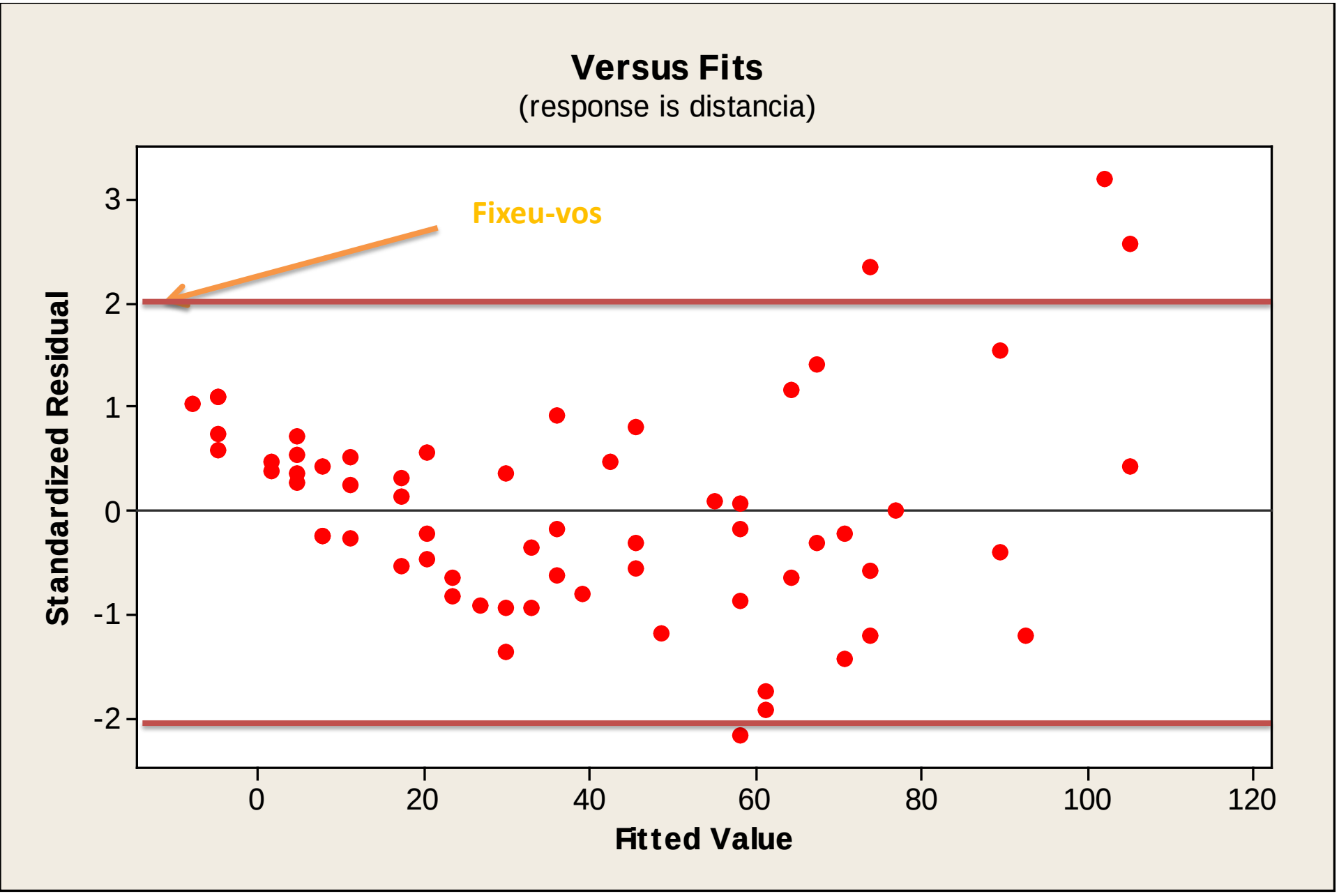


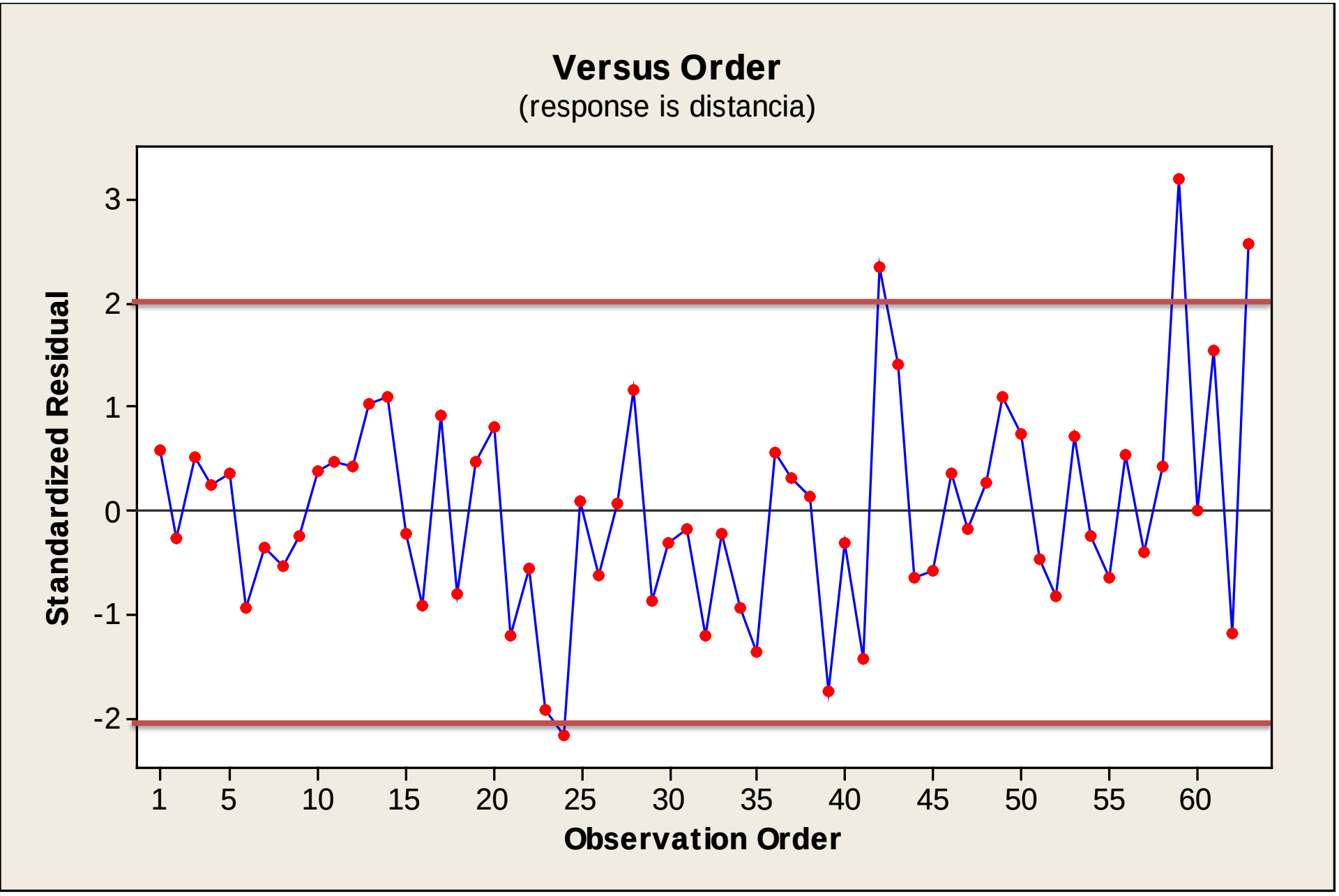
També es poden "estandaritzar" els residus

Fixeu-vos

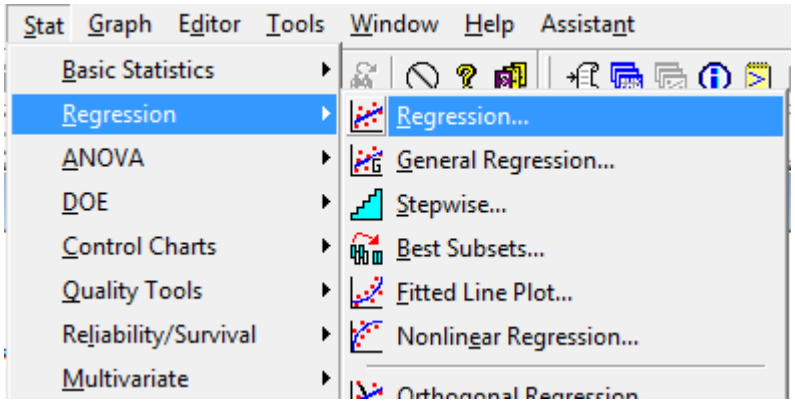


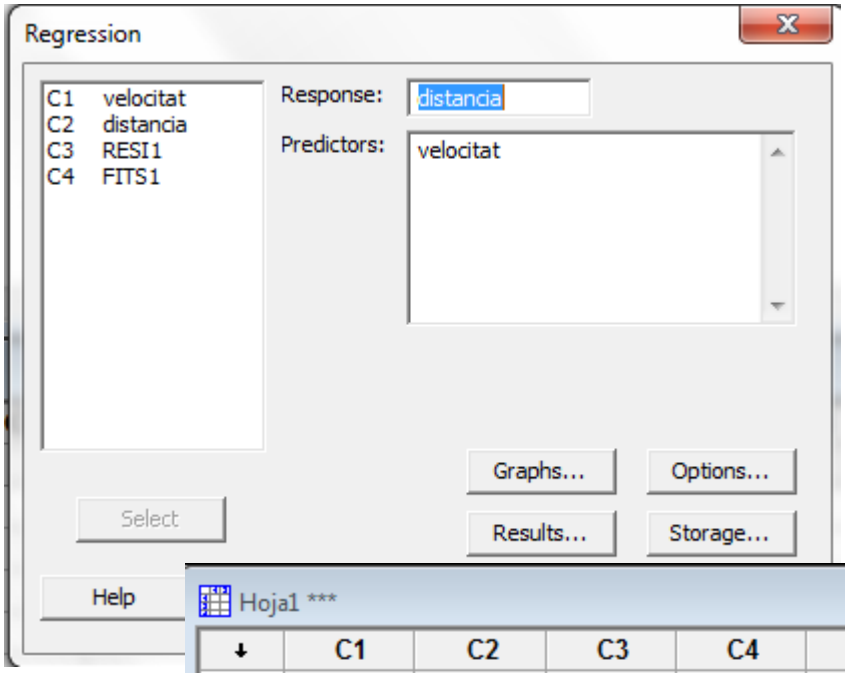



Versus Fits
(response is distancia)
Fixeu-vos
Standardized Residual
3
2
1
0
-1
-2
0
20
40
60
80
100
120
Fitted Value

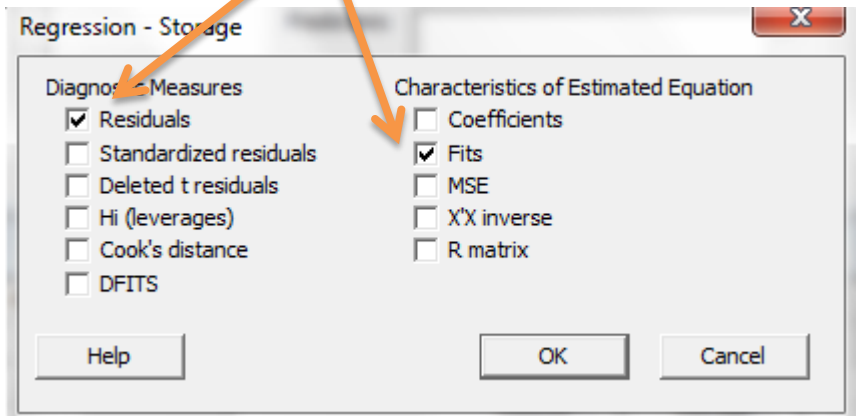

Versus Order
(response is distancia)
Standardized Residual
3
2
1
0
-1
-2
1
5
10
15
20
25
30
35
40
45
50
55
60
Observation Order

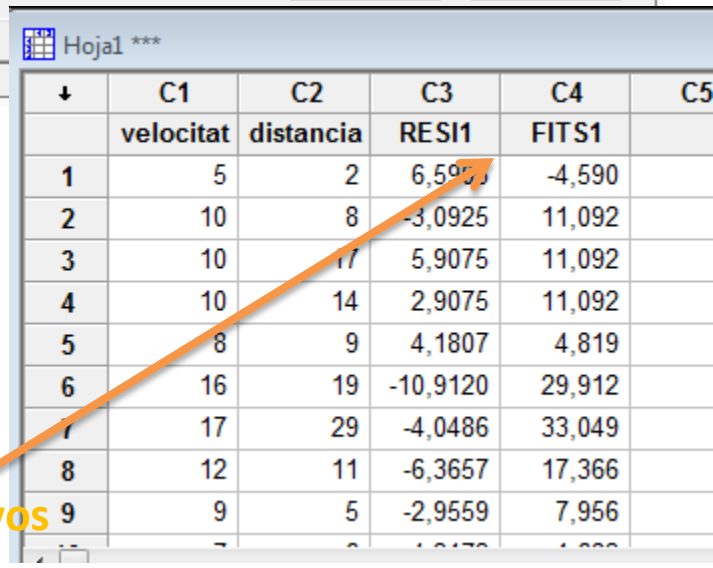
9. COM FER PREDICCIONS? SI VOLEM CALCULAR ELS $\hat{y}_i = \hat{y}(x_i)$





Fixeu-vos

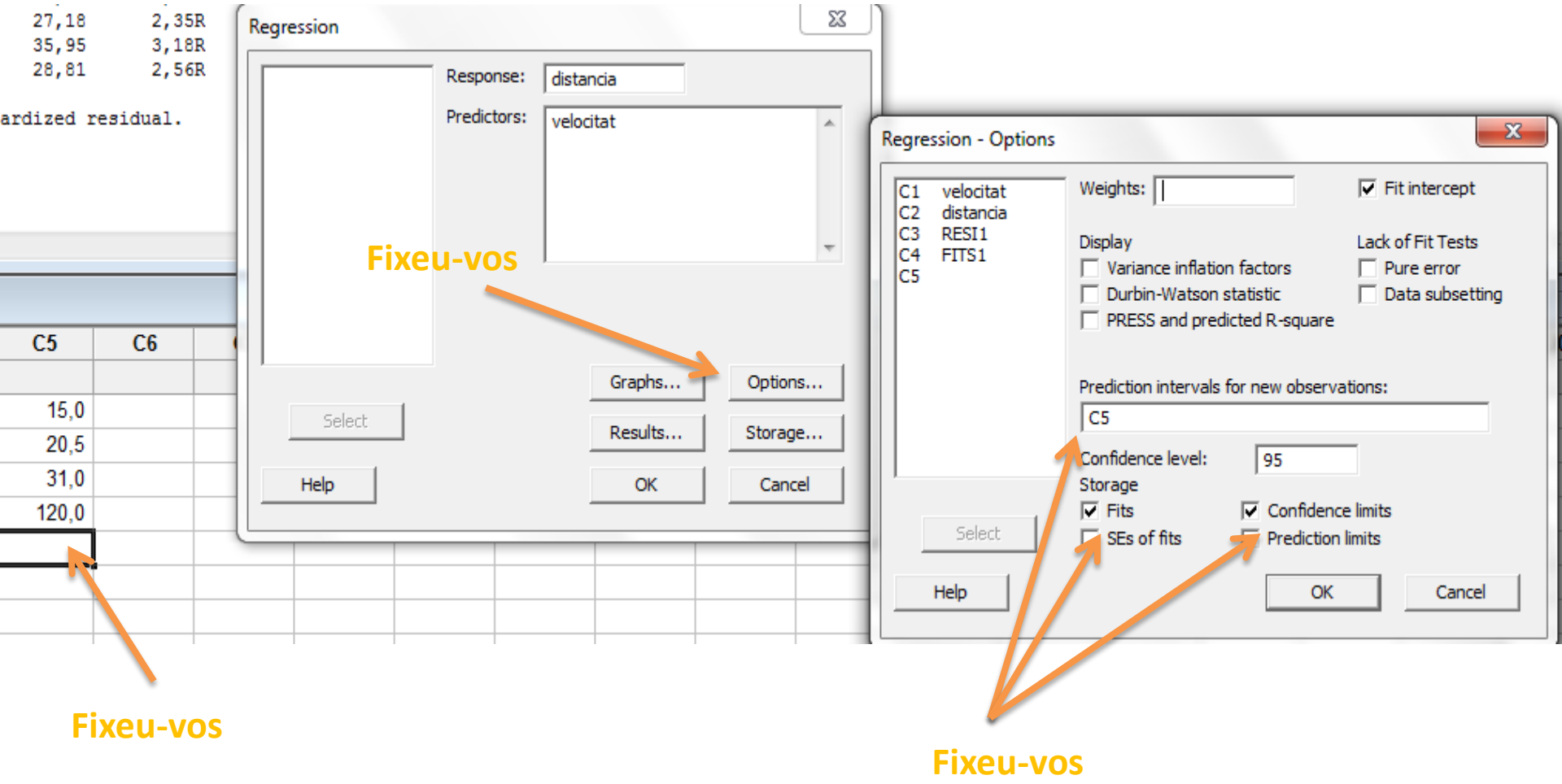




	C1	C2	C3	C4	C5
	velocitat	distancia	RESI1	FITS1	
1	5	2	6,59[illegible]	-4,590	
2	10	8	-3,0925	11,092	
3	10	[illegible]	5,9075	11,092	
4	10	14	2,9075	11,092	
5	8	9	4,1807	4,819	
6	16	19	-10,9120	29,912	
[illegible]	17	29	-4,0486	33,049	
8	12	11	-6,3657	17,366	
9	9	5	-2,9559	7,956	

Fixeu-vos

SI VOLEM CALCULAR ALTRES VALORS, **OMPLIM UNA COLUMNA, PER EXEMPLE C5**



Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	**26,78**	1,60	**(23,57; 29,98)**	(2,96; 50,59)
2	**44,03**	1,50	**(41,02; 47,04)**	(20,24; 67,81)
3	**76,96**	2,35	**(72,25; 81,67)**	(52,90; 101,02)
4	**356,12**	15,40	**(325,33; 386,90)**	(317,33; 394,90)**XX**

XX denotes a point that is an extreme outlier in the predictors.

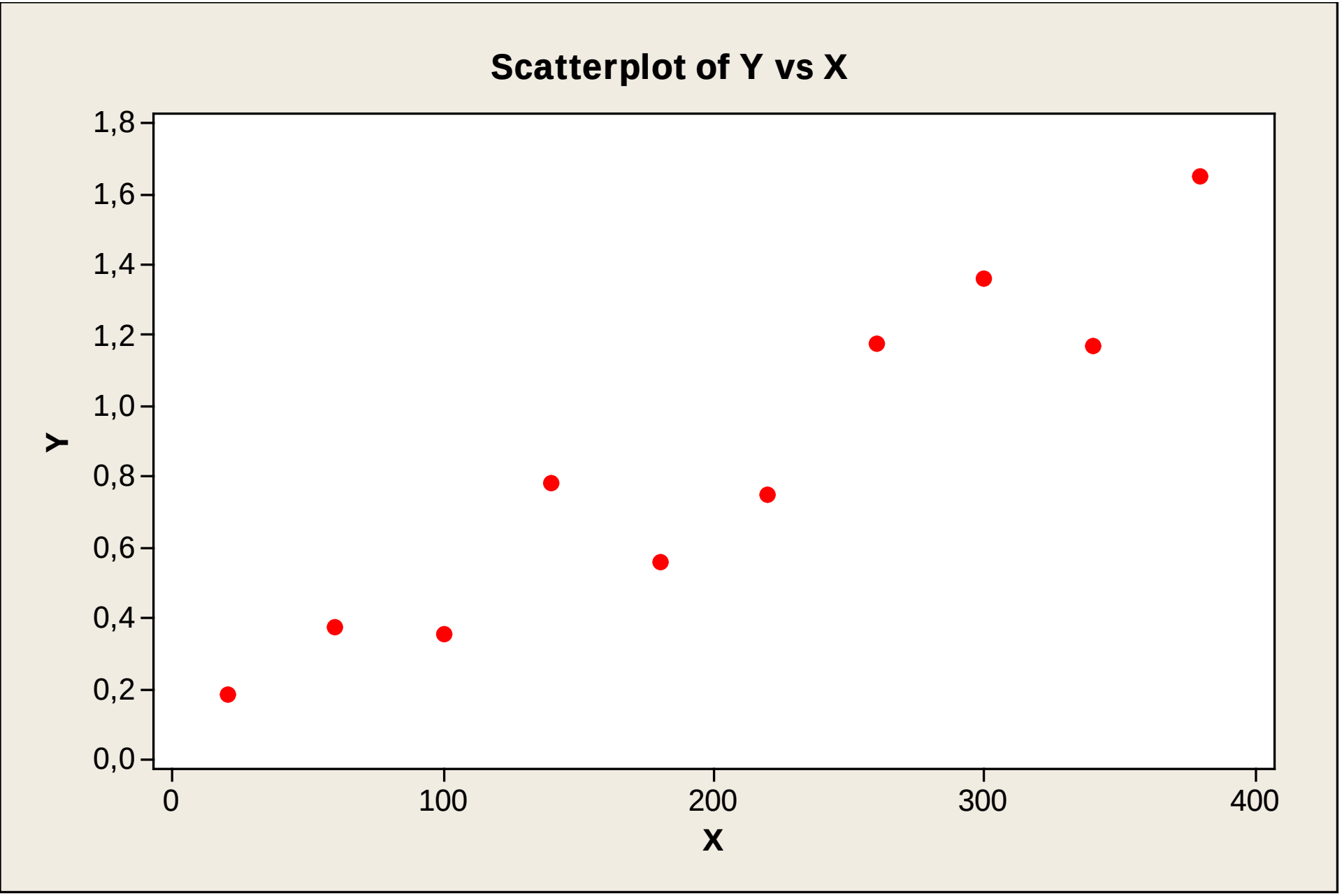
C5	C6	C7	C8	C9	C10
	RESI2	FITS2	PFIT2	CLIM2	CLIM3
15,0	6,5905	-4,590	26,775	23,568	29,982
20,5	-3,0925	11,092	44,027	41,018	47,035
31,0	5,9075	11,092	76,961	72,254	81,668
120,0	2,9075	11,092	356,117	325,331	386,903

10. Reprenem l'exemple de l'apartat 4 amb dades proveïdes pel Prof. Josep Gibergans del departament Matemàtica Aplicada III, UPC.

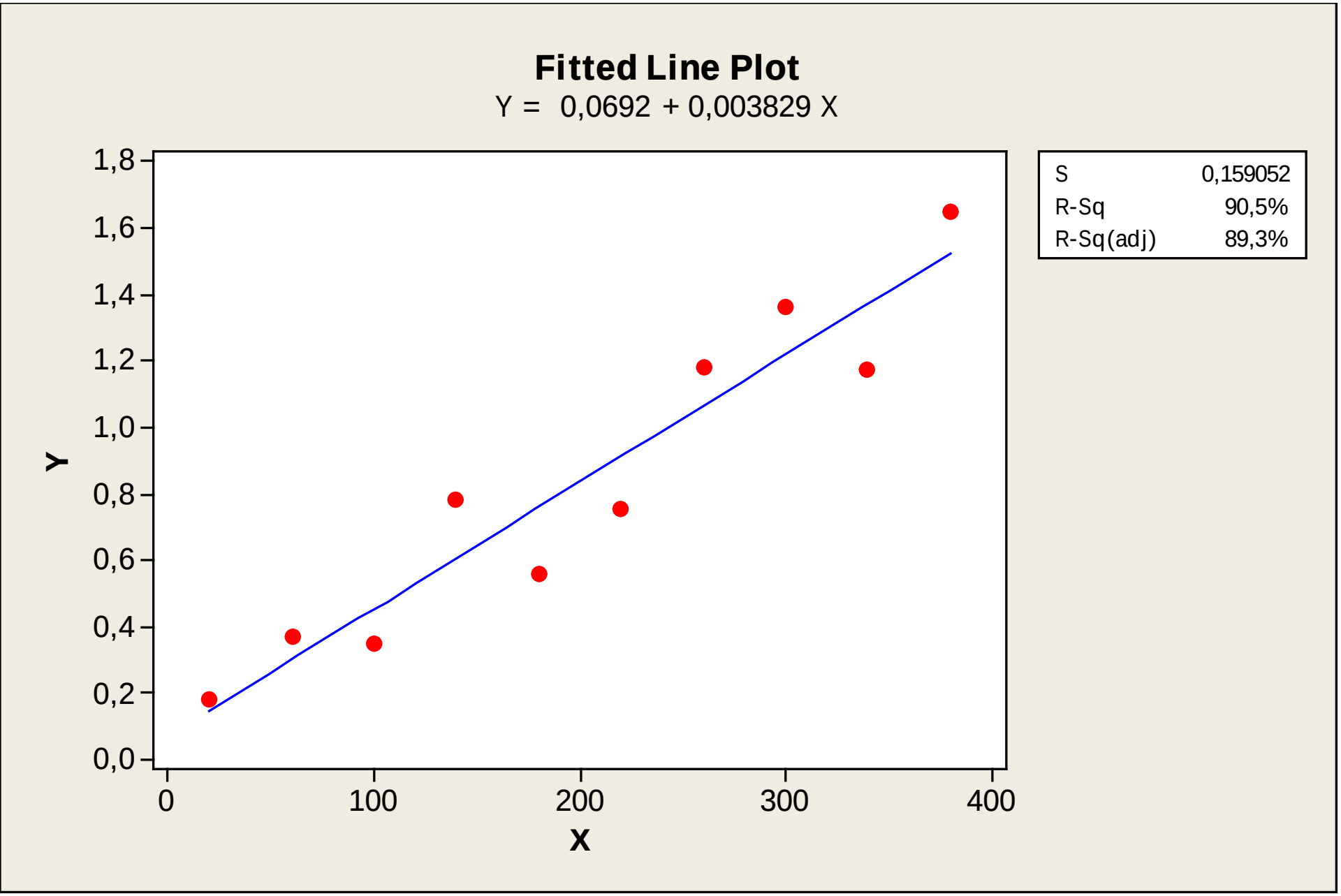
Les següents dades són les mesures de la velocitat de l'aire i del coeficient d'evaporació de les gotetes de combustible en una turbina de propulsió:

X: Velocitat de l'aire (cm/s)	Y: Coeficient d'evaporació (mm^2/s)
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36
340	1.17
380	1.65

El diagrama de dispersió obtingut amb MINITAB és:



En MINITAB obtenim:



Regression Analysis: Y versus X

The regression equation is
Y = 0,069 + 0,00383 X

Predictor	Coef	SE Coef	T	P
Constant	0,0692	0,1010	0,69	0,512
X	0,0038288	0,0004378	8,75	0,000

S = 0,159052 **R-Sq = 90,5%** R-Sq(adj) = 89,3%

Analysis of Variance

Source	DF	**SS**	MS	F	P
Regression	1	1,9351	1,9351	76,49	0,000
Residual Error	8	**0,2024**	0,0253		
Total	9	2,1374			

Els residus estandaritzats mostren que no hi ha observacions anòmales.