Régression affine

I/ Historique

Francis Galton était un cousin de Charles Darwin, qui s'intéressait à l'évolution des populations notamment humaines. En 1886, il a publié dans le *Journal of the Anthropological Institute of Great Britain and Ireland* un article titré ***Regression Towards Mediocrity in Hereditary Stature*** dans lequel il s'inquiétait que la race humaine perde de sa « pureté » en régressant, par croisement, vers une uniformité qu'il appelait *mediocrity* : « when mid-parents are taller than mediocrity, their children tend to be shorter than they ; when mid-parents are shorter than mediocrity, their children tend to be taller than they [1]». Pour cela il a mesuré les tailles de 930 adultes et de leurs parents[2]. Puis il a reporté les résultats dans un tableau à double entrée :

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

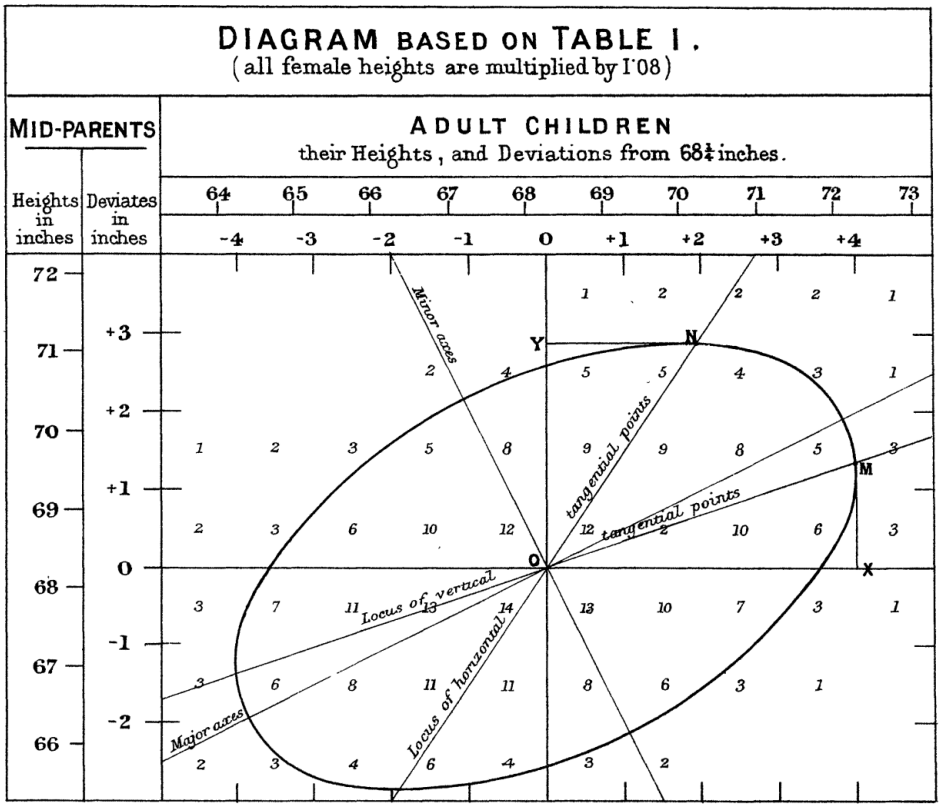
Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

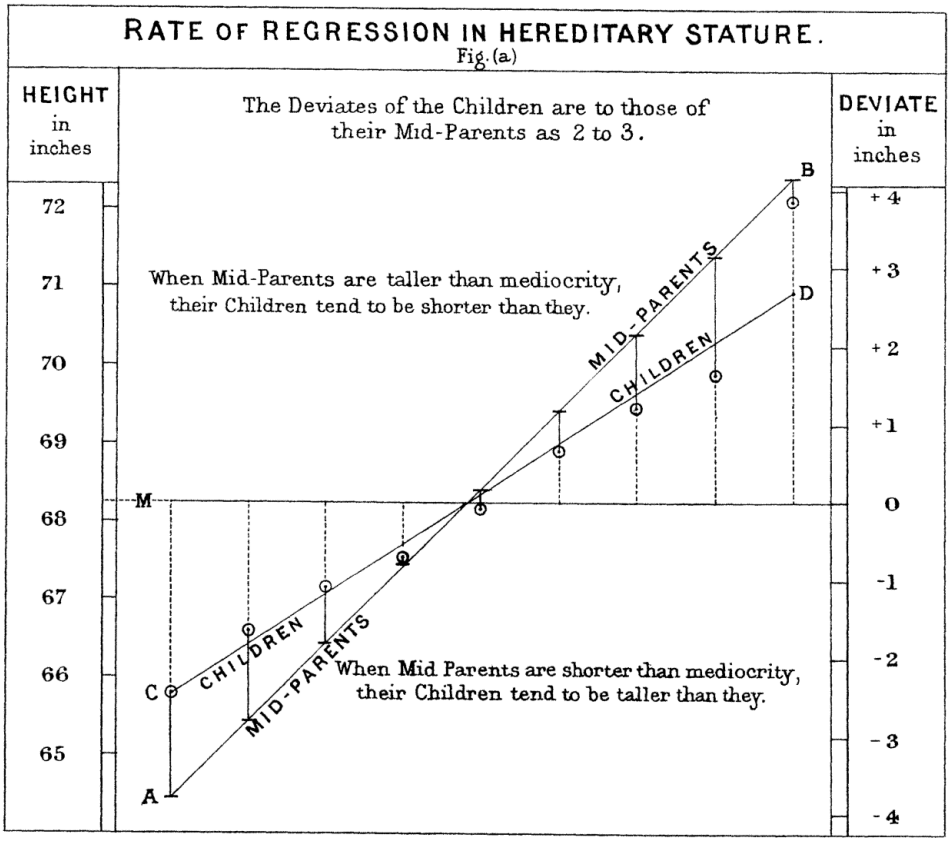
Puis il a tracé un nuage de points avec en abscisse les tailles des enfants et en ordonnée celles de leurs parents :

1 Les enfants des grands tendent à être plus petits que leurs parents, les enfants des petits tendent à être plus grands que leurs parents.

2 Moyenne entre la taille du père et celle de la mère, multipliée par un facteur correctif de 1,08.

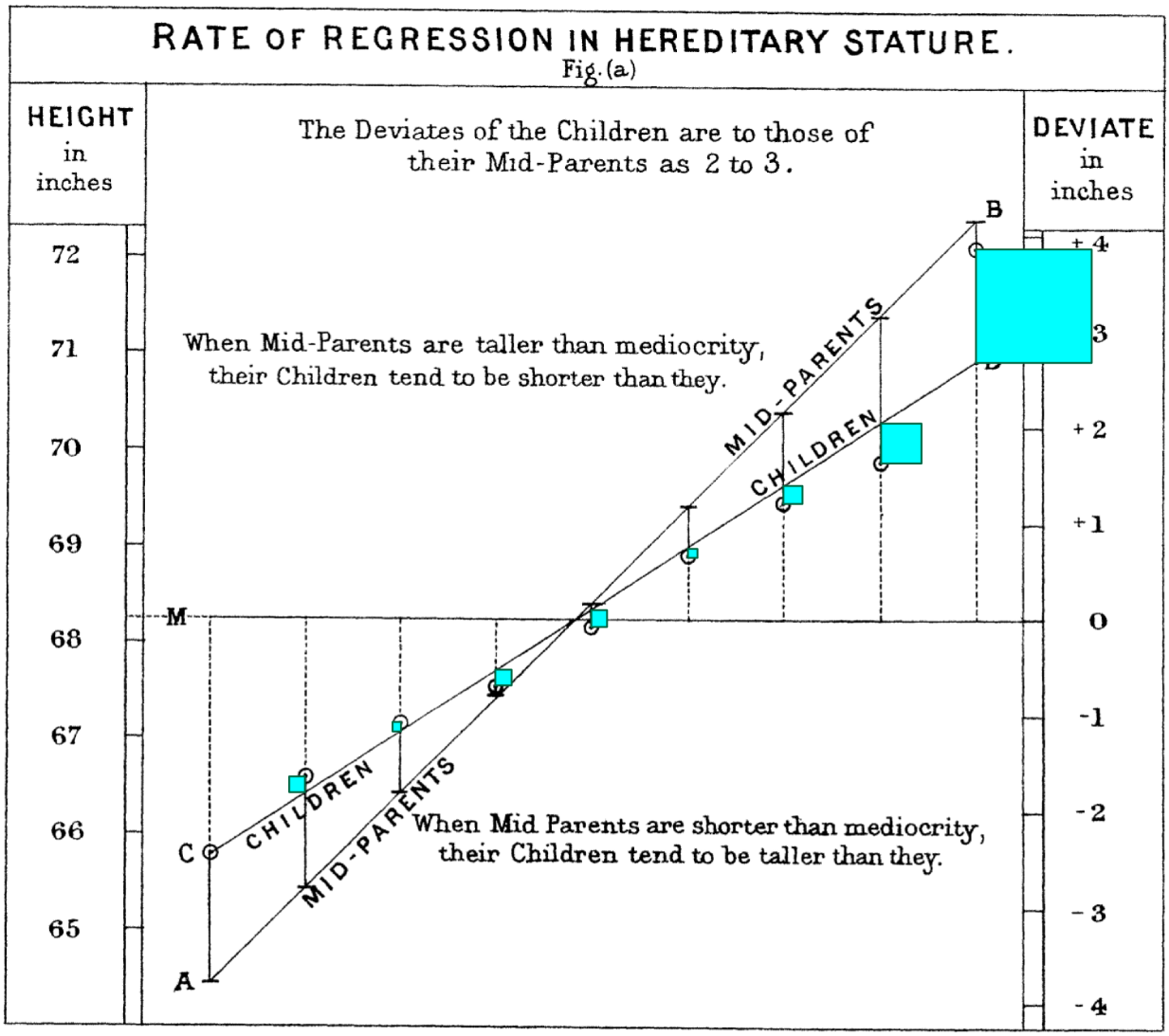


Le nuage de points étant allongé, cela suggère une corrélation entre les tailles des enfants et celles des parents. Galton cherche alors une droite qui passe au plus près du nuage :



(abscisses : écarts par rapport à la taille moyenne en inches ; ordonnées : tailles en inches). Beaucoup de droites ont l'air de passer près du nuage de points, mais une seule minimise l'aire

coloriée en bleu ci-dessous :



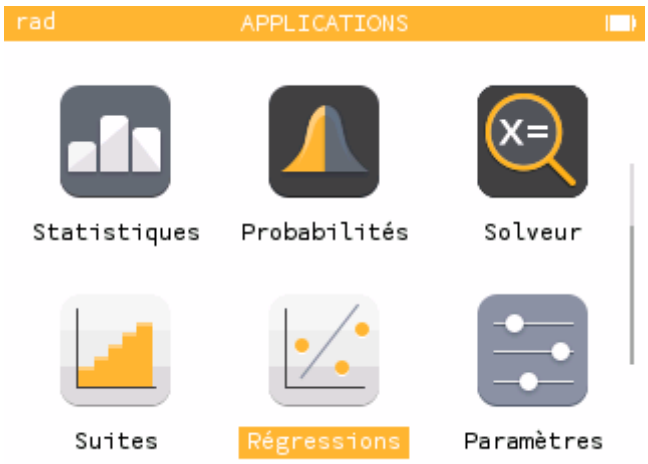
On l'appelle donc ***droite de régression***, et l'algorithme pour calculer ses coefficients est appelé algorithme ***des moindres carrés***.

II/ Calculatrice

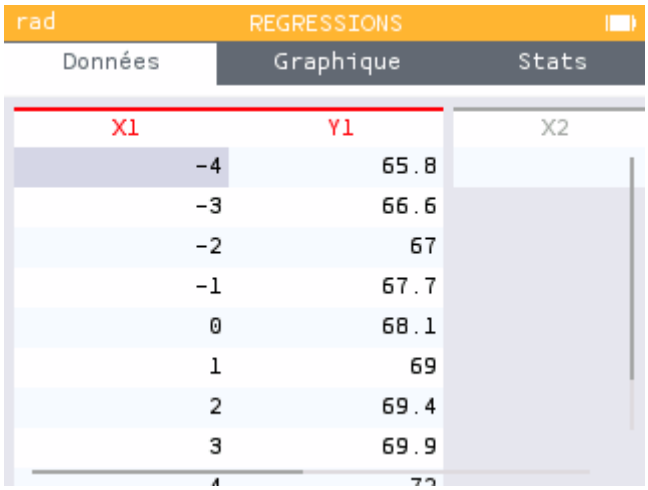
On lit ci-dessus les données

x	-4	-3	-2	-1	0	1	2	3	4
y	65,8	66,6	67	67,7	68,1	69	69,4	69,9	72

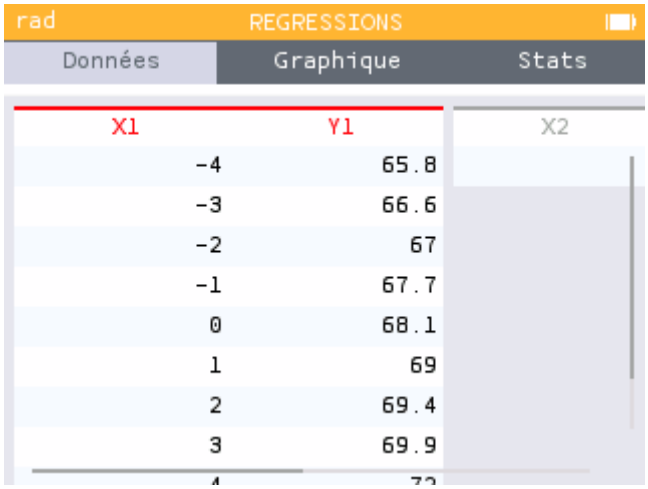
Avec une Numworks, on sélectionne l'application dite « régression » :



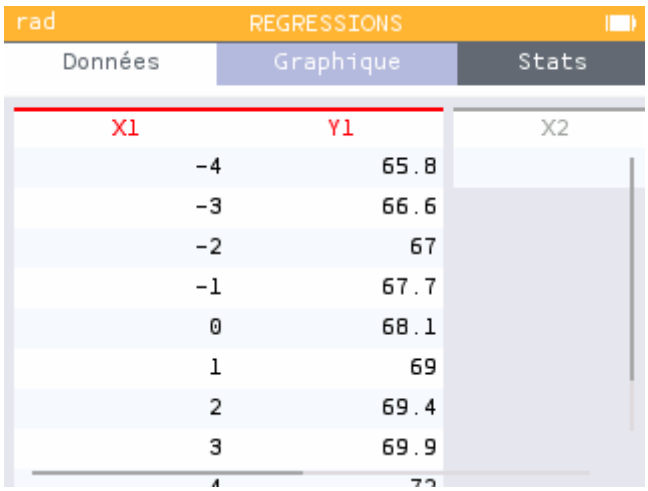
On entre alors les données (une colonne par variable) :



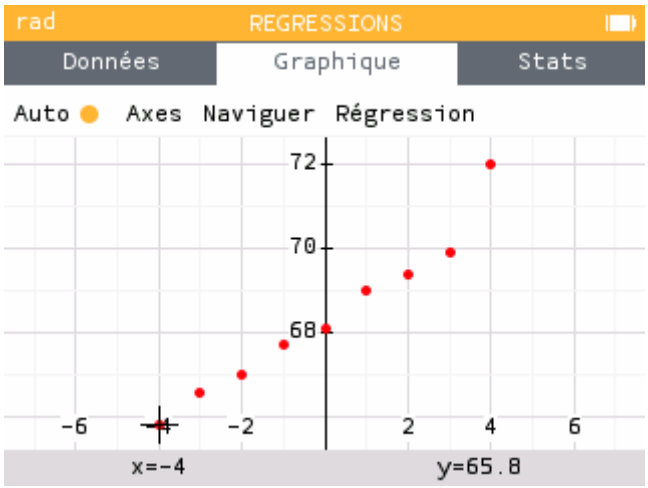
On remonte jusqu'à ce que l'onglet « données » soit en exergue :



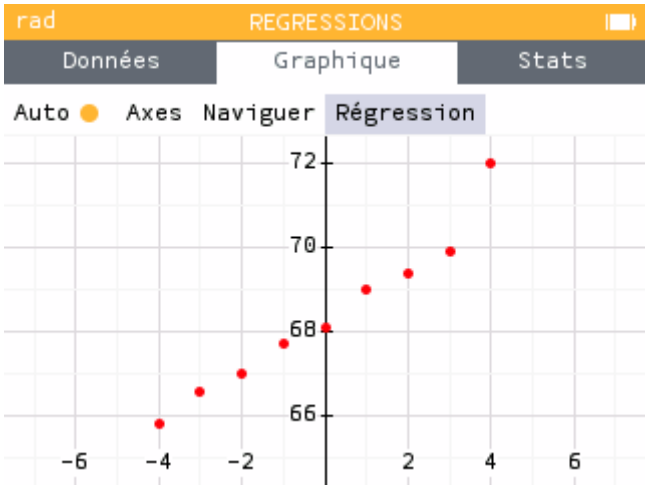
On se déplace sur l'onglet « graphique » :



Un appui sur OK donne alors le graphique, à comparer avec celui de Galton :

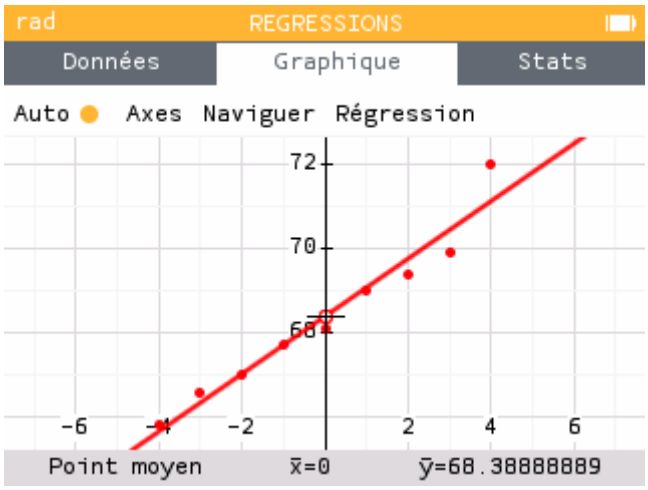


pour avoir la droite des moindres carrés, on va jusqu'à « régression » :

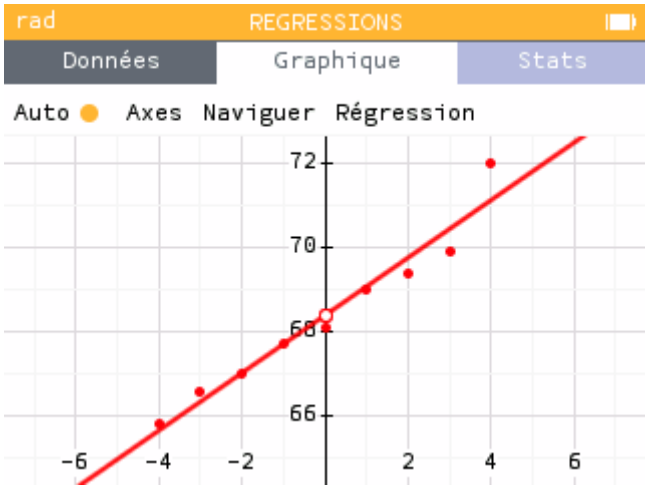


On choisit la régression « linéaire » (en réalité, affine) :

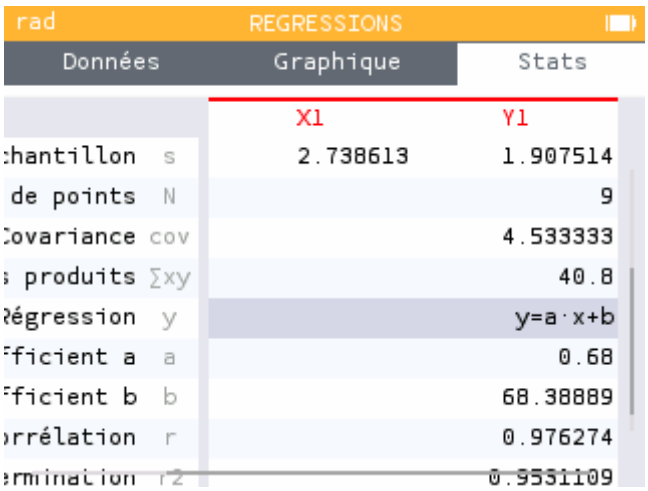
Et on a le dessin de la droite des moindres carrés :



Pour avoir l'équation de la droite, on va vers « Stats » :



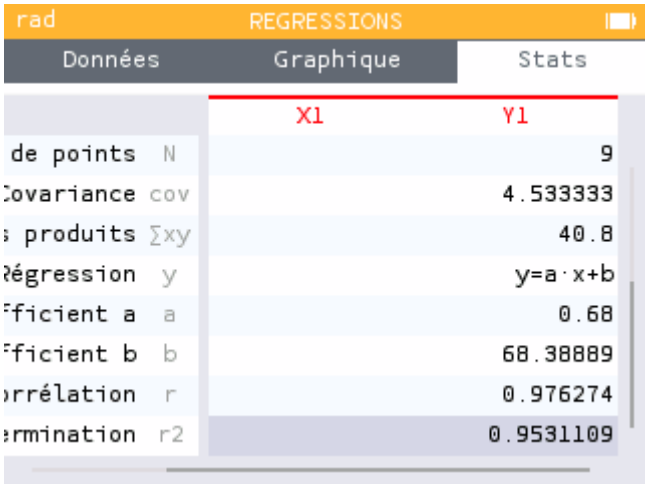
Là, on appuie sur OK, et on a l'affichage de l'équation et des coefficients :



L'équation de la droite de Galton est donc y=0,68×x+1231/9.

III/ Coefficient de corrélation

Le coefficient de corrélation mesure la qualité de l'approximation affine. Il est proche de -1 (fonction affine décroissante) ou de 1 (fonction affine croissante) si les points sont presque alignés. Pour les données de Galton il est 0,976274 donc pas si proche de 1 que cela :



Pour accentuer l'écart par rapport à 1 ou -1, on considère parfois le carré du coefficient de corrélation (noté ci-dessus r2, parfois R). **En mathématiques**, c'est r qui est demandé, **pas son carré**.