Droite de régression linéaire par la méthode des moindres carrés.

Le plan est muni du repère $(O;\vec{i};\vec{j})$

Soit un couple de caractères (X,Y) dont le coefficient de corrélation linéaire vérifie $|\rho(X,Y)|\simeq 1$

On veut approximer Y par une fonction linéaire de X définie par $f(X)=aX+b$ de telle façon que $g(a,b) = \frac{1}{n}\sum_{i=1}^{i=n}(f(X_i)-Y_i)^2 = \frac{1}{n}\sum_{i=1}^{i=n}(aX_i+b-Y_i)^2$, moyenne des écarts au carré entre les Y_i et $f(X_i)$, soit minimum. C'est à dire trouver trouver la droite d, d'équation $y = ax+b$, qui passe « le plus près des points du nuage ».

1 Cas particulier $\bar{X} = E(X) = 0 \quad et \quad \bar{Y} = E(Y) = 0$

$$\frac{\partial g(a,b)}{\partial a} = \frac{2}{n}\sum_{i=1}^{i=n} X_i(aX_i+b-Y_i)= \frac{2}{n}\sum_{i=1}^{i=n} a(X_i)^2+bX_i-X_iY_i$$
$$= 2\left(a\frac{1}{n}\sum_{i=1}^{i=n}(X_i)^2+b\frac{1}{n}\sum_{i=1}^{i=n}X_i-\frac{2}{n}\sum_{i=1}^{i=n}X_iY_i\right)$$
$$= 2\,(aV(X)+b\,E(X)-cov(X;Y))$$
$$= 2\,(a\,V(X)-cov(X;Y))$$

$$\frac{\partial g(a,b)}{\partial b} = \frac{2}{n}\sum_{i=1}^{i=n}(aX_i+b-Y_i)= 2\text{a}\frac{1}{n}\sum_{i=1}^{i=n}X_i+b\frac{1}{n}\sum_{i=1}^{i=n}b-\frac{1}{n}\sum_{i=1}^{i=n}Y_i$$
$$= 2\,(aE(X)+b-E(Y)) = b$$

Une condition nécessaire pour que (a,b) soit un minimum est que les dérivées partielles soient nulles.

$$\frac{\partial g(a,b)}{\partial a} = 0 \Rightarrow a = \frac{cov(X;Y)}{V(X)}$$
$$\frac{\partial g(a,b)}{\partial b} = 0 \Rightarrow b = 0$$

La matrice hessienne des dérivées secondes partielles est $\begin{pmatrix} V(X) & 0 \\ 0 & 1\end{pmatrix}$, $V(X)\geqslant 0 \;\; et \;\; 1>0$ donc ce point est un minimum.

$$y = \frac{cov(X;Y)}{V(X)}x$$

2 Cas général.

On effectue le changement d'origine $O\rightarrow G$ où $G(E(X);E(Y))$ est le point moyen.

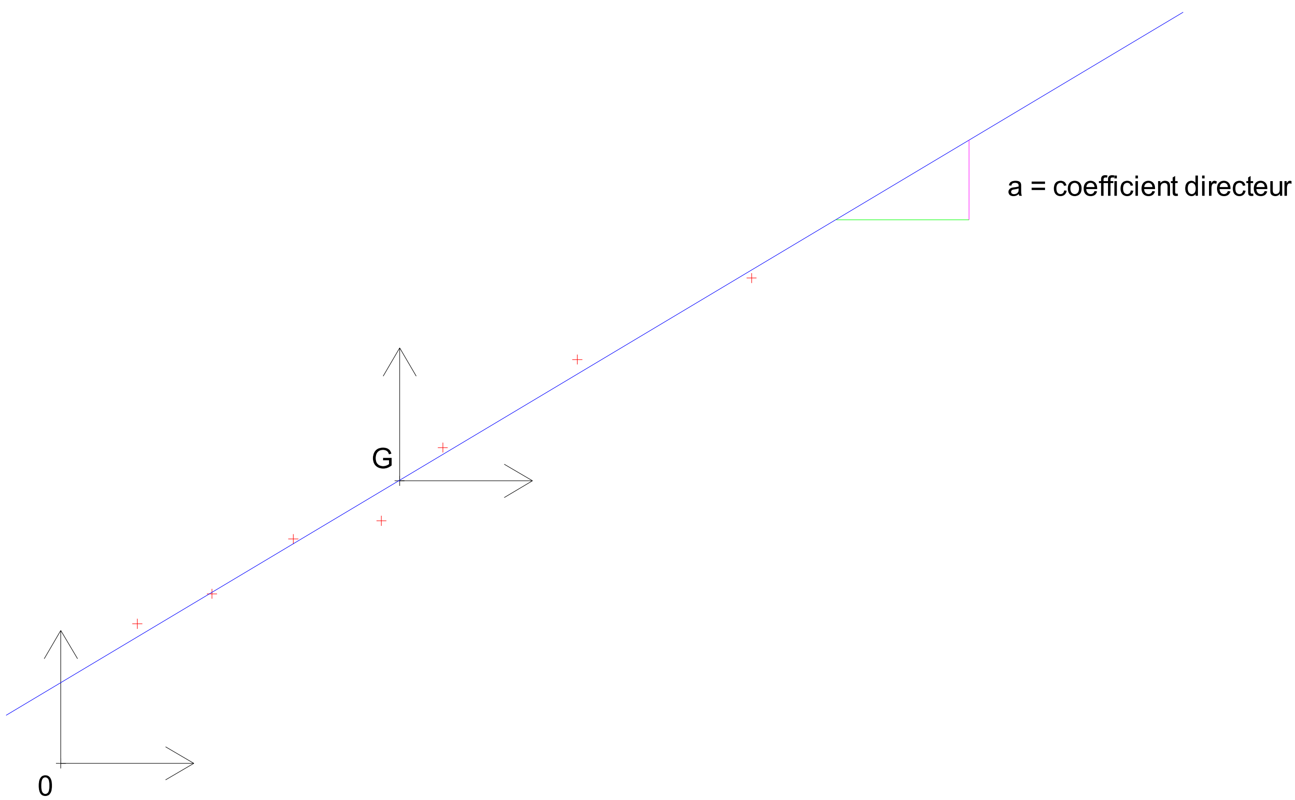
Dans le nouveau repère les points du nuage ont pour coordonnées $M_i(X_i-E(X);Y_i-E(Y))$ donc correspondent aux caractères Z et T tels que :

$$Z = X - E(X) \quad et \quad T = Y - E(Y) \Rightarrow E(Z) = E(Y) = 0$$

Dans ce repère la droite de régression linéaire a pour équation : $y = \dfrac{cov(Z;T)}{V(Z)} x$

On sait que : $V(X - E(X)) = V(X) \;\; et \;\; cov(X - E(X); Y - E(Y)) = cov(X;Y)$ donc l'équation s'écrit :

$$y = \frac{cov(X;Y)}{V(X)} x \quad .$$



Un changement d'origine conserve le coefficient directeur, a, d'une droite donc dans le repère $(O; \vec{i}; \vec{j})$ l'équation de la droite de régression linéaire est :

$$y = \frac{cov(X;Y)}{V(X)} x + b \quad .$$

Cette droite passe par le point moyen $G(E(X); E(Y))$:

$$E(Y) = \frac{cov(X;Y)}{V(X)} E(X) + b \quad \text{donc} \quad b = E(Y) - \frac{cov(X;Y)}{V(X)} E(X)$$