variables.

2. Si $r_{x,y} = -1$, on dit qu'il y a une parfaite corrélation linéaire négative entre les deux variables.
3. Si $r_{x,y} = 0$, on dit qu'il y a absence de corrélation linéaire entre les deux variables.
4. On dit qu'il y a une forte corrélation linéaire entre les deux variables (ou forte dépendance linéaire) si $r_{x,y}$ est proche de ± 1.
5. En revanche, si $r_{x,y}$ est proche de zéro (0), on dit qu'il y a une faible corrélation linéaire entre les deux variables.

3.3 Ajustement d'un nuage de points

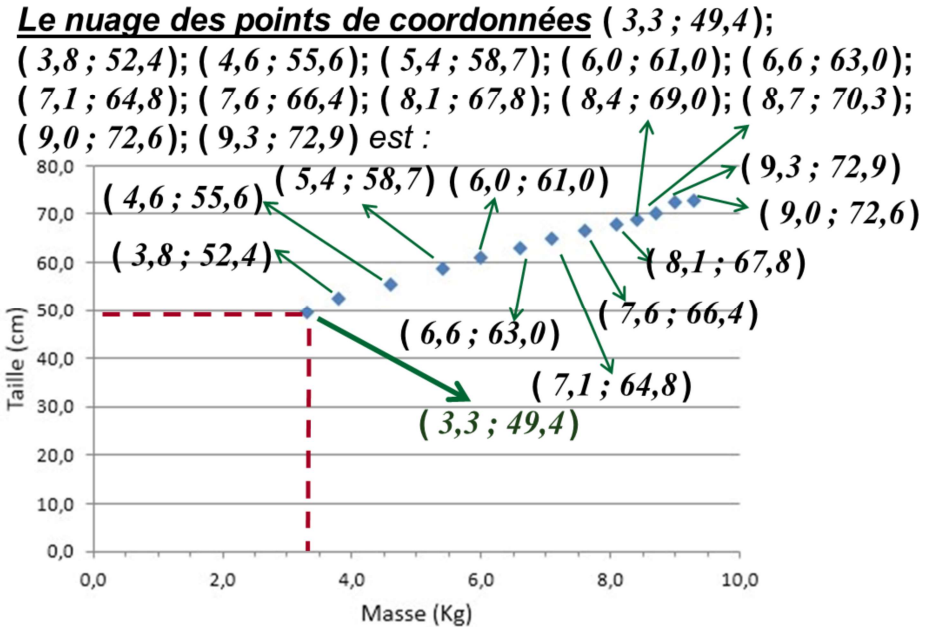
Dans toute la suite on considère N observations sur les deux variables X et Y.

3.3.1 Nuage de points

Ensemble de points isolés représentés dans un graphique cartésien ; c'est-à-dire des points $M_1, M_2, \ldots, M_n$ de coordonnées (x_1, y_1) ; (x_2, y_2); … ; (x_n, y_n)

Exemple 3.3.1 On considère le tableau suivant, relatif à une population associé à deux variables mesurées sur 13 bébés tels que, X = « le poids du bébé » et Y = « la taille du bébé »

Masse (kg)	3,3	3,8	4,6	5,4	6,0	6,6	7,1	7,6	8,1	8,4	8,7	9,0	9,3
Taille (cm)	49,4	52,4	55,6	58,7	61,0	63,0	64,8	66,4	67,8	69,0	70,3	72,6	72,9



3.3.2 Ajustement linéaire d'un nuage de points

On considère N observations sur les deux variables X et Y, donc ;

1. Ces observations peuvent être représentées par un nuage de points.
2. Notre but est d'exprimer Y en fonction de X.
3. La représentation du nuage de points peut nous renseigner sur l'allure de la courbe de régression.

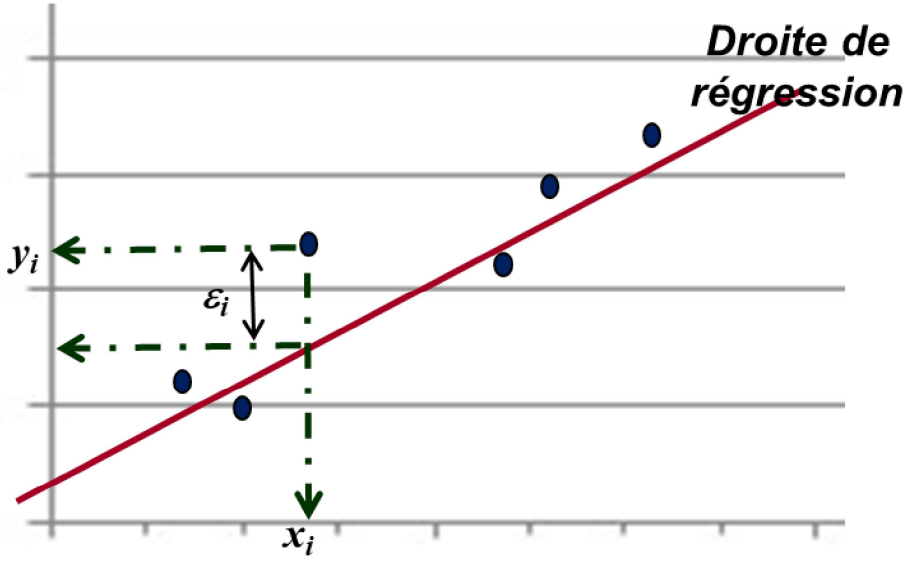
Remarque 3.3.1

1. L'ajustement linéaire consiste à trouver l'équation d'une droite du type $y = ax + b$, appelée droite de régression. Cette droite donne l'évolution de la variable Y (variable expliquée) en fonction de la variable explicative X.

2. La méthode d'ajustement que nous allons exposer est appelée « méthode des Moindres Carrés Ordinaires » ou simplement **« MCO ».**

- **La méthode des Moindres Carrés Ordinaires**

Considérons N couples d'observations (x_i, y_i), leurs nuage est :



Donc les couples (x_i, y_i) vérifient :

$$y_i = (ax_i + b) + \varepsilon_i \quad \forall i \in \{1, \cdots, N\}$$

où ε_i représente le résidu du couple (x_i, y_i). On peut alors écrire :

$$\varepsilon_i = y_i - (ax_i + b)$$

Remarque 3.3.2 La méthode des **MCO** consiste à minimiser $\sum_{i=1}^{i=N} \varepsilon_i^{\,2}$ tels que :

$$\sum_{i=1}^{i=N} \varepsilon_i^{\,2} = \sum_{i=1}^{i=N} (y_i - ax_i - b)^2 = f(a,b)$$

Les deux conditions de premier ordre de la minimisation de cette fonction f par rapport à a et à b sont :

$$\frac{\partial\left(\sum_{i=1}^{i=N} \varepsilon_i^{\,2}\right)}{\partial a} = 0 \quad \text{et} \quad \frac{\partial\left(\sum_{i=1}^{i=N} \varepsilon_i^{\,2}\right)}{\partial b} = 0$$

$$\Rightarrow \frac{\partial\left(\sum_{i=1}^{i=N} \varepsilon_i^{\,2}\right)}{\partial a} = 2\sum_{i=1}^{i=N} (y_i - ax_i - b)(-x_i) = 0 \Rightarrow \sum_{i=1}^{i=N} (y_i - ax_i - b)(x_i) = 0 \qquad (1)$$

et $\frac{\partial\left(\sum_{i=1}^{i=N} \varepsilon_i^{\,2}\right)}{\partial b} = 2\sum_{i=1}^{i=N} (y_i - ax_i - b)(-1) = 0 \Rightarrow \sum_{i=1}^{i=N} (y_i - ax_i - b) = 0 \qquad (2)$

$$(1) \Rightarrow \sum_{i=1}^{i=N} (y_i x_i - a x_i^{\,2} - b x_i) = \sum_{i=1}^{i=N} y_i x_i - a\sum_{i=1}^{i=N} x_i^{\,2} - b\sum_{i=1}^{i=N} x_i = 0 \quad (3)$$

$$(2) \Rightarrow \sum_{i=1}^{i=N} (y_i - ax_i - b) = \sum_{i=1}^{i=N} y_i - a\sum_{i=1}^{i=N} x_i - Nb = 0 \qquad (4)$$

En divisant les deux membres de l'équation (**4**) par N, on obtient :

$$\frac{1}{N}\sum_{i=1}^{i=N} y_i - \frac{a}{N}\sum_{i=1}^{i=N} x_i - \frac{Nb}{N} = 0$$

Sachant que $\bar{x} = \frac{1}{N}\sum_{i=1}^{i=N} x_i$ et $\bar{y} = \frac{1}{N}\sum_{i=1}^{i=N} y_i$ donc l'équation devient :

$$\bar{y} - a\bar{x} - b = 0 \quad (5) \Leftrightarrow b = \bar{y} - a\bar{x}$$

En remplaçant, dans l'équation (**3**), b par $\bar{y} - a\bar{x}$, d'après l'équation (*5*) on obtient

$$\sum_{i=1}^{i=N} y_i x_i - a\sum_{i=1}^{i=N} x_i^{\,2} - (\bar{y} - a\bar{x})\sum_{i=1}^{i=N} x_i = 0$$

$$\Leftrightarrow \sum_{i=1}^{i=N} y_i x_i - a\sum_{i=1}^{i=N} x_i^{\,2} - \bar{y}\underbrace{\sum_{i=1}^{i=N} x_i}_{N.\bar{x}} + a\bar{x}\underbrace{\sum_{i=1}^{i=N} x_i}_{N.\bar{x}} = 0$$

$$\Leftrightarrow \sum_{i=1}^{i=N} y_i x_i - a\sum_{i=1}^{i=N} x_i^{\,2} - N\bar{x}.\bar{y} + aN\bar{x}^2 = 0$$

$$\Leftrightarrow \sum_{i=1}^{i=N} y_i x_i - N\bar{x}.\bar{y} = a\left(\sum_{i=1}^{i=N} x_i{}^2 - N\bar{x}^2\right)$$

Ainsi, on obtient la valeur estimée de la pente de la droite de régression :

$$\hat{a} = \frac{\sum_{i=1}^{i=N} x_i y_i - N\bar{x}.\bar{y}}{\sum_{i=1}^{i=N} x_i{}^2 - N\bar{x}^2} \Rightarrow \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

Donc l'équation de la droite de régression est :

$$y = \hat{a}x + \hat{b}$$

Remarque 3.3.3 On peut aussi calculer la valeur estimée de la pente de la droite de régression en utilisant l'une de ces deux expressions suivantes :

$$\hat{a} = \frac{Cov(x,y)}{V(x)} \quad \text{et} \quad \hat{a} = \frac{\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{i=N}(x_i - \bar{x})^2} \text{ car}$$

$$\hat{a} = \frac{\sum_{i=1}^{i=N} y_i x_i - N\bar{x}.\bar{y}}{\sum_{i=1}^{i=N} x_i{}^2 - N\bar{x}^2} = \frac{\frac{1}{N}\left(\sum_{i=1}^{i=N} y_i x_i - N\bar{x}.\bar{y}\right)}{\frac{1}{N}\left(\sum_{i=1}^{i=N} x_i{}^2 - N\bar{x}^2\right)} = \frac{\frac{1}{N}\sum_{i=1}^{i=N} y_i x_i - \bar{x}.\bar{y}}{\frac{1}{N}\sum_{i=1}^{i=N} x_i{}^2 - \bar{x}^2}$$

$$\hat{a} = \frac{Cov(x,y)}{V(x)} \quad \text{et} \quad \hat{a} = \frac{\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{i=N}(x_i - \bar{x})^2}$$

Remarque 3.3.4 La droite de régression passe par le point moyen de coordonnées $(\bar{x}, \bar{y})$.

En effet

$$\hat{b} = \bar{y} - \hat{a}\bar{x} \Rightarrow \bar{y} = \hat{a}\bar{x} + \hat{b}.$$

Exercice d'application

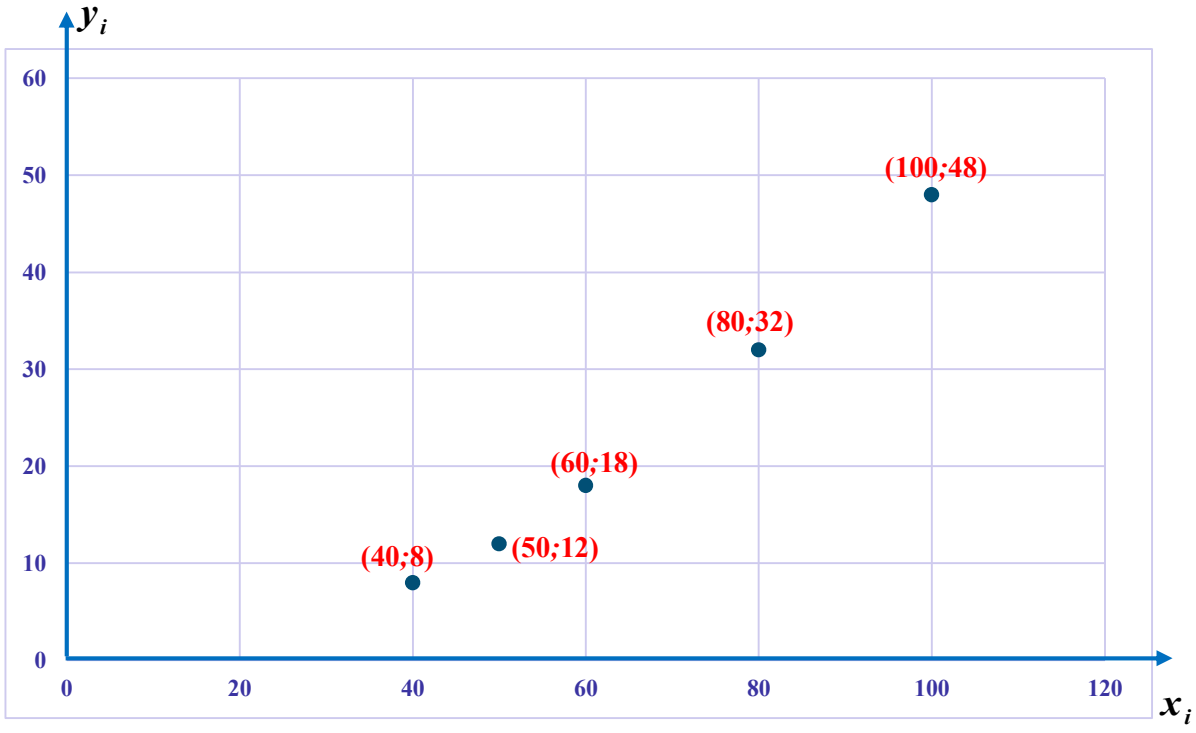
Le tableau suivant donne la distance de freinage d'un véhicule automobile sur une route sèche, en fonction de sa vitesse.

Vitesse en Km/h (x_i)	*Distance en m* (y_i)
40	8
50	12
60	18
80	32
100	48

1. Construire le nuage des points.
2. Calculer la covariance entre la vitesse X et la distance Y.
   Que peut on déduire sur la relation entre X et Y.
3. Calculer le coefficient de corrélation linéaire.
   Conclure sur l'intensité de la liaison entre X et Y.
4. Déterminer, en utilisant la méthode des moindres carrés, l'équation de la droite de régression permettant d'estimer la distance de freinage en fonction de la vitesse du véhicule.
5. Interpréter la pente et la constante de l'équation de la droite obtenue.
6. A combien peut-on estimer la distance de freinage d'un véhicule roulant à 120 km/h.
7. Déterminer cette même droite sachant qu'une sixième mesure a donné pour :
   $x_i = 0$; $y_i = 0$.

Solution de l'exercice

1. Le nuage des points (x_i, y_i)



2. La covariance entre la vitesse X et la distance Y.

$$Cov(x,y)=\left(\frac{1}{5}\sum_{i=1}^{i=5}x_i y_i\right)-(\bar{x}.\bar{y}) \text{ avec } \bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i \text{ et } \bar{y}=\frac{1}{N}\sum_{i=1}^{N}y_i$$

i	x_i	y_i	$x_i y_i$
1	40	8	320
2	50	12	600
3	60	18	1080
4	80	32	2560
5	100	48	4800
Total	330	118	9360

$$\Rightarrow \bar{x}=\frac{330}{5}=66;\ \bar{y}=\frac{118}{5}=23{,}6$$

$$\Rightarrow Cov(x,y)=\frac{9360}{5}-66\times 23{,}6=1872-1557{,}6$$

$$\Rightarrow Cov(x,y)=314{,}4.$$

Conclusion

Comme $Cov(x, y) > 0$ alors la relation entre la vitesse et la distance de freinage est positive et les deux variables varient dans le même sens.

3. Le coefficient de corrélation linéaire.

Conclure sur l'intensité de la liaison entre X et Y.

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	40	8	320	1600	64
2	50	12	600	2500	144
3	60	18	1080	3600	324
4	80	32	2560	6400	1024
5	100	48	4800	10000	2304
Total	330	118	9360	24100	3860

$$r_{x,y}=\frac{Cov(x,y)}{\sqrt{V(x)V(y)}} \text{ avec } V(x)=\left(\frac{1}{N}\sum_{i=1}^{i=r}x_i^{\,2}\right)-\bar{x}^2 \text{ et } V(y)=\left(\frac{1}{N}\sum_{i=1}^{i=r}y_i^{\,2}\right)-\bar{y}^2$$

$$\Rightarrow V(x)=\frac{24100}{5}-(66)^2=4820-4356\Rightarrow V(x)=464.$$

Et $V(y)=\frac{3860}{5}-(23{,}6)^2=772-556{,}96\Rightarrow V(y)=215{,}04$

$$\Rightarrow r_{x,y} = \frac{314{,}4}{\sqrt{464 \times 215{,}04}} \Rightarrow r_{x,y} = 0{,}99$$

Conclusion

Les variables varient dans le même sens. La valeur de $r_{x,y}$, proche de **1**, cela traduit une forte corrélation linéaire entre les deux variables.

4. L'équation de la droite de régression permettant d'estimer la distance de freinage en fonction de la vitesse en utilisant la méthode des moindres carrés ordinaire.

$$\hat{a} = \frac{\sum_{i=1}^{i=N} x_i y_i - N\bar{x}.\bar{y}}{\sum_{i=1}^{i=N} x_i{}^2 - N\bar{x}^2} \Rightarrow \hat{a} = \frac{9360 - 5 \times 66 \times 23{,}6}{24100 - 5 \times 66^2} \Rightarrow \hat{a} = 0{,}67$$

De plus l'ordonnée à l'origine est égale à :

$$\hat{b} = \bar{y} - \hat{a}\bar{x} = 23{,}6 - 0{,}67 \times 66 \Rightarrow \hat{b} = -20{,}62$$

donc L'équation s'écrit : $y = 0{,}67x - 20{,}62$

5. Interprétation de la pente et la constante de l'équation de la droite obtenue :

- Lorsque la vitesse augmente de 1 ***km/h*** la distance de freinage augmente de $\hat{a} = 0{,}67m$.
- La constante $\hat{b} = -20{,}62$ Indique qu'à l'arrêt le véhicule est en retard d'une distance de $20{,}62\text{m}$.

6. L'estimation de la distance de freinage d'un véhicule roulant à 120 ***km/h.***

L'équation étant : $y = 0{,}67x - 20{,}62$.

En remplaçant x par 120 on obtient :

$y = 0{,}67 \times 120 - 20{,}62 = 59{,}78$.

Donc la distance de freinage d'un véhicule roulant à 120 ***km/h*** **est** $y = 59{,}78$ ***m.***

7. Détermination de la même droite sachant qu'une sixième mesure a donné pour :

$x_i = 0$; $y_i = 0$.

C'est-à-dire l'équation des moindres carrés avec $(x_i = 0 \; ; \; y_i = 0)$

Il suffit de refaire les calculs avec les mêmes sommes mais en divisant par le nouveau nombre d'observations = effectif total qui est égal à $N = 6$.

$$\Rightarrow \bar{x} = \frac{330}{6} = 55; \quad \bar{y} = \frac{118}{6} = 19{,}67$$

$$\Rightarrow \hat{a} = \frac{\sum_{i=1}^{i=N} y_i x_i - N\bar{x}.\bar{y}}{\sum_{i=1}^{i=N} x_i^{\,2} - N\bar{x}^2} = \frac{9360 - 6\times 55\times 19{,}67}{24100 - 6\times 55^2} \Rightarrow \hat{a} = 0{,}48$$

$$\Rightarrow \hat{b} = \bar{y} - \hat{a}\bar{x} = 19{,}67 - 0{,}48\times 55 \Rightarrow \hat{b} = -6{,}74$$

donc L'équation s'écrit : $y = 0{,}48x - 6{,}74$

3.3.3 Ajustement non linéaire d'un nuage de points

On considère N observations sur les deux variables X et Y.
Dans le cas général, la relation entre X et Y semble être plutôt non linéaire, c'est-à-dire n'est pas de la forme $y = ax + b$.
En fait lorsque le nuage de points manifeste en tendance courbe et que le coefficient de corrélation linéaire n'est pas proche de 1 en valeur absolue, l'ajustement de ce nuage par une droite est hasardeux et aboutira à des estimations de mauvaise qualité. Dans ce cas, on peut tenter d'utiliser un des modèles proposés dans ce paragraphe. En fait, chacun de ces modèles utilise le principe d'ajustement par la méthode des moindres carrés (donc ils utilisent tous une droite) mais en "transformant" au préalable les données pour obtenir un modèle linéaire à partir du modèle non-linéaire considéré.
On va voir les deux modèles d'ajustements non linéaires suivants :

- Ajustement hyperbolique.
- Ajustement par une fonction puissance.

1er modèle : L'ajustement hyperbolique

Les N points (x_i, y_i) ne sont pas alignés, mais plutôt proches d'une courbe représentant une fonction hyperbolique de la forme :

$$y = \frac{b}{x^a} \text{ avec } a > 0, \quad b > 0$$

Dans ce cas le nuage aura l'allure suivante