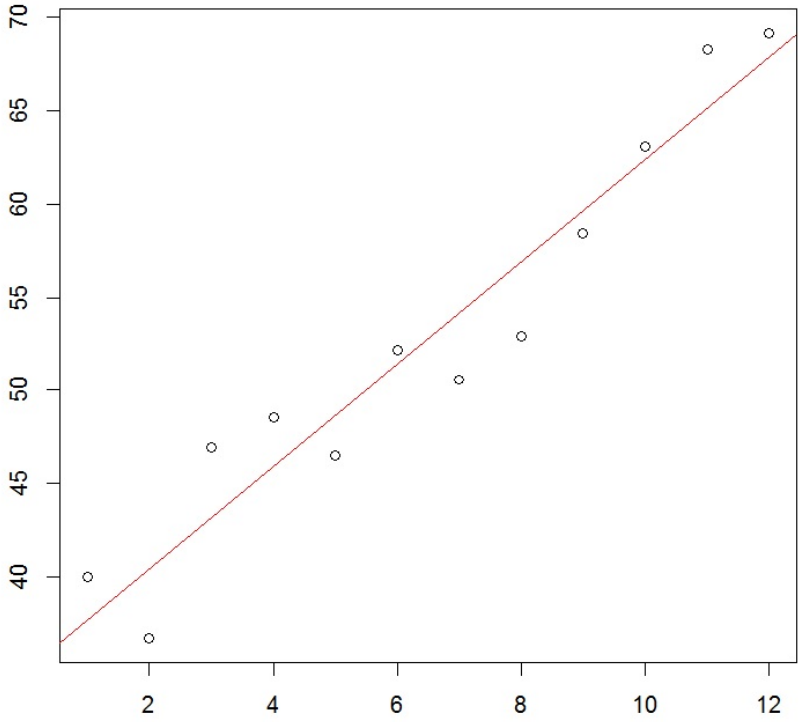
Université de Caen

Ajustement d'un nuage de points

Christophe Chesneau

https://chesneau.users.lmno.cnrs.fr/



Caen, le 08 Janvier 2018

Table des matières

1 Contexte statistique **5**

2 Méthode des points observés **13**

3 Méthode des points moyens **17**

4 Méthode des moindres carrés **23**

5 Pour s'entraîner **31**

6 Quelques compléments **33**

~ Note ~

Ce document résume les principales méthodes d'ajustement d'un nuage de points abordées dans les filières appliquées (Terminale STMG, BTS CGO, Licence 1...).

Des exemples et des graphiques viennent illustrer ces méthodes.

Je vous invite à me contacter pour tout commentaire :

christophe.chesneau@gmail.com

Bonne lecture !

1 Contexte statistique

Point de départ

On souhaite prévoir et/ou expliquer les valeurs d'une variable numérique Y à partir des valeurs d'une variable numérique X. Pour ce faire, on dispose de données qui sont n valeurs du couple de variables (X, Y) notées $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$. Elles se présentent généralement sous la forme d'un tableau :

x_i	x_1	x_2	$\ldots$	x_n
y_i	y_1	y_2	$\ldots$	y_n

Ainsi, quand X vaut x_1, on a mesuré la valeur y_1 pour Y, quand X vaut x_2, on a mesuré la valeur y_2 pour Y ...

Exemples

Exemple 1. Une étude a été menée auprès de 12 étudiants afin d'expliquer le score à un examen de mathématiques à partir du temps consacré à la préparation de cet examen. Pour chaque étudiant, on dispose :

- du temps de révision en heures (variable X),
- du score obtenu sur 800 points (variable Y).

Les résultats sont :

x_i	4	9	10	14	4	7	12	1	3	8	11	5
y_i	390	580	650	730	410	530	600	350	400	590	640	450

Ainsi, avec une préparation de 4 heures, le premier étudiant a obtenu le score de 390 à l'examen, avec une préparation de 9 heures, le deuxième étudiant a obtenu le score de 580 à l'examen...

Exemple 2. On étudie l'évolution du nombre d'inscriptions à un jeu en ligne au cours du temps. Pour chaque mois de l'année 2016, on dispose :

- du rang du mois (variable X ; janvier est rang 1, février est le rang 2...),
- du nombre d'inscriptions en milliers (variable Y).

Les résultats sont :

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	37	43	41	40	51	47	48	54	56	64	66	73

Ainsi, au moins de janvier 2016, il y a eut 37000 inscriptions au jeu, en Février 2016 il y a eut 43000 inscriptions au jeu...

Nuage de points

Les observations peuvent être représentées sur le repère orthonormé (O, I, J) par n points :

Points	M_1	M_2	...	M_n
Coordonnées	$(x_1; y_1)$	$(x_2; y_2)$	...	$(x_n; y_n)$

L'ensemble de ces points est appelé nuage de points. La silhouette de ce nuage de points est une indication précieuse sur la nature de la relation entre Y et X.

Ajustement affine du nuage de points

Si la silhouette du nuage de points est étirée dans une direction, une relation affine/linéaire entre Y et X est envisageable : on suppose l'existence de deux coefficients réels inconnus α et β tels que

$$Y = \alpha + \beta X$$

plus un terme d'erreur secondaire "de valeur moyenne nulle" et "indépendant de X" représentant une somme des petites variations aléatoires (erreurs de mesures, effets non prévisibles...). Telle est la forme générique d'un modèle statistique connu : *le modèle de régression linéaire simple.*

Pour toute valeur x de X, une valeur estimée y de Y est donnée par :

$$y = a + bx,$$

où a désigne une valeur estimée de α et b désigne une valeur estimée de β, toutes deux calculées à l'aide des données.

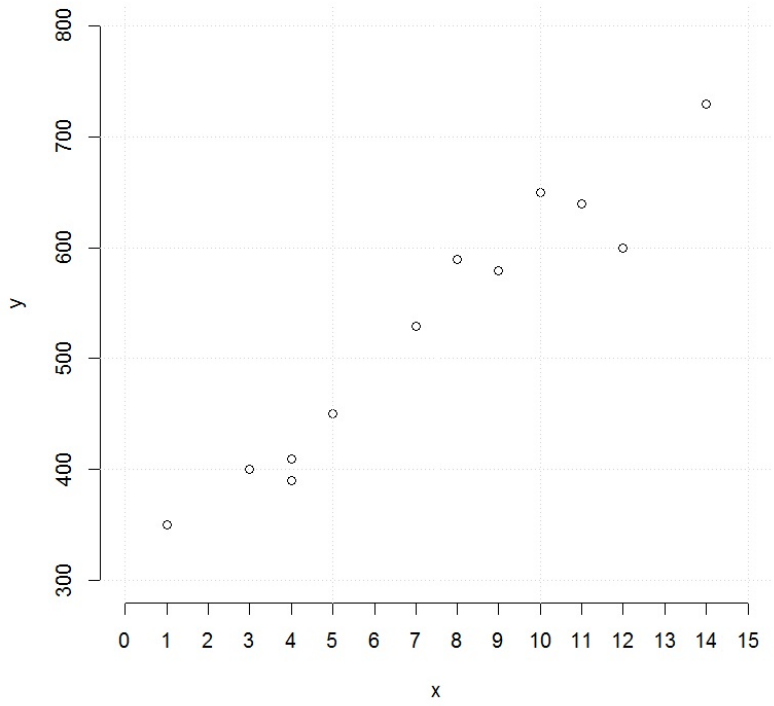
Ainsi, à partir des valeurs x de X, estimer avec précision les valeurs de Y correspondantes revient à déterminer a et b de sorte à ce que la droite d'équation $y = a + bx$ ajuste au mieux le nuage de points.

Exemples

Retour sur l'exemple 1. Une étude a été menée auprès de 12 étudiants afin d'expliquer le score à un examen de mathématiques à partir du temps consacré à la préparation de cet examen. Pour chaque étudiant, on dispose du temps de révision en heures (variable X) et du score obtenu sur 800 points (variable Y). Les résultats sont :

x_i	4	9	10	14	4	7	12	1	3	8	11	5
y_i	390	580	650	730	410	530	600	350	400	590	640	450

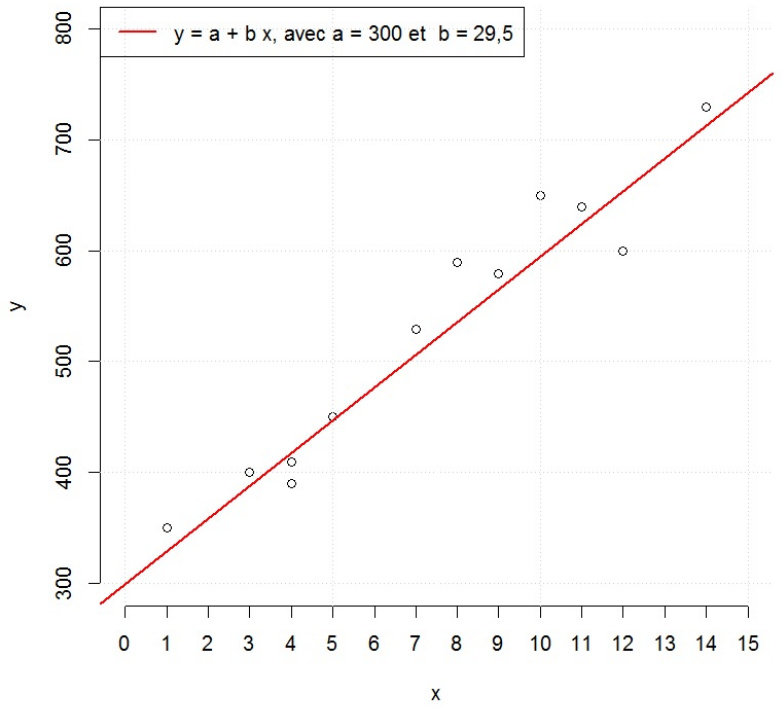
Le nuage de points associé est :



Par exemple, le deuxième point du nuage en partant de la gauche correspond à l'étudiant numéro 9 : le point M_9 correspondant est de coordonnées $(3; 400)$.

La silhouette du nuage de points est étirée dans une direction, une relation affine entre Y et X est envisageable. Ainsi, à partir des valeurs x de X, estimer avec précision les valeurs de Y correspondantes revient à déterminer a et b de sorte à ce que la droite d'équation $y = a + bx$ ajuste au mieux le nuage de points.

Après plusieurs essais graphiques "à l'œil", en utilisant la calculatrice (ou autre), on constate que la droite suivante ajuste "pas trop mal" le nuage de points :



Ainsi, avec cette méthode "au jugé", on propose les coefficients $a = 300$ et $b = 29,5$, pour une droite d'équation : $y = a + bx$. Avec cette équation, on peut alors faire des prévisions. Par exemple, une valeur estimée du score d'un étudiant ayant consacré 16 heures de préparation à l'examen est :

$$y = a + bx = 300 + 29,5 \times 16 = 772.$$

Commentaire : Ce score est en fait une valeur estimée de la moyenne de tous les scores des étudiants ayant fait une préparation de 16 *heures, valeur que l'on attribue à tous ces étudiants.*

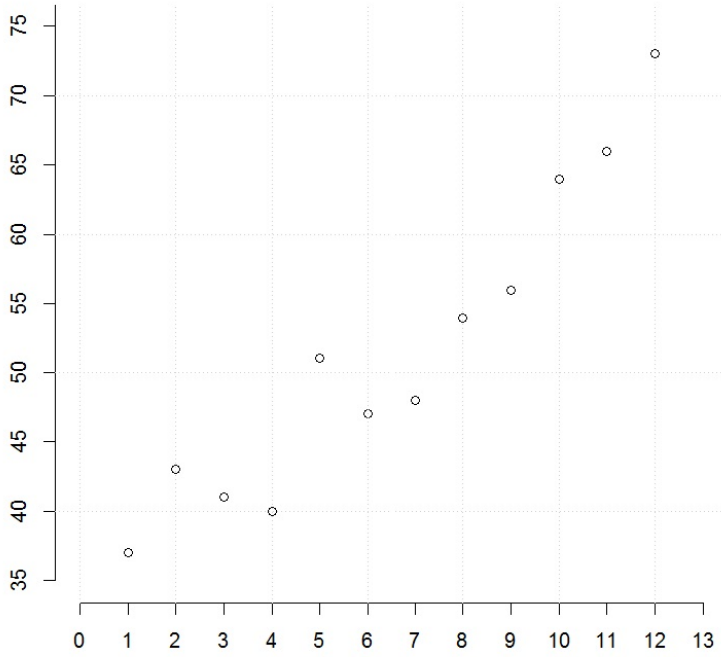
Aussi, avec cet ajustement, un étudiant peut espérer avoir la moyenne, donc un score de plus de 400 sur 800, en ayant fait une préparation de plus de x heures, avec x vérifiant :

$$y \geq 400 \quad \Leftrightarrow \quad 300 + 29,5 \times x \geq 400 \quad \Leftrightarrow \quad x \geq \frac{400 - 300}{29,5} = 3.389831.$$

Retour sur l'exemple 2. On étudie l'évolution du nombre d'inscriptions à un jeu en ligne au cours du temps. Pour chaque mois de l'année 2016, on dispose du rang du mois (variable X ; janvier est rang 1, février est le rang 2…) et du nombre d'inscriptions en milliers (variable Y). Les résultats sont :

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	37	43	41	40	51	47	48	54	56	64	66	73

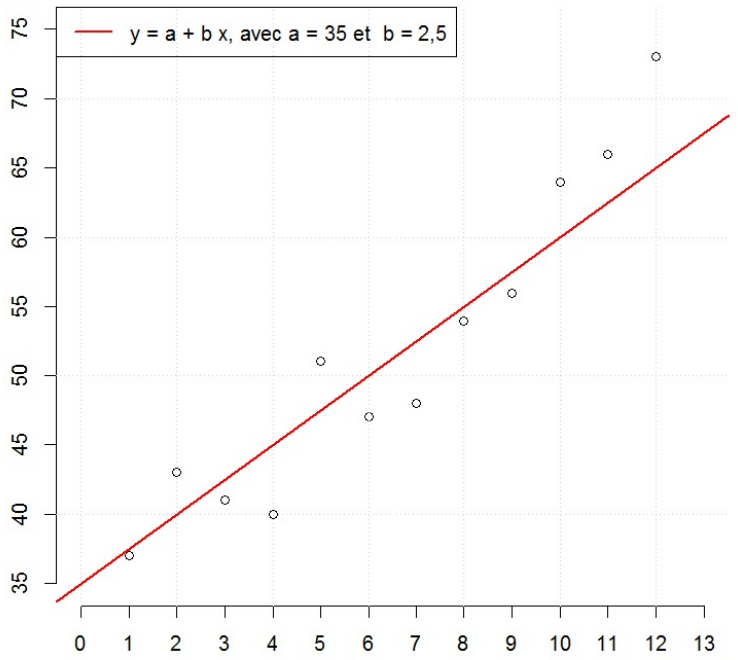
Le nuage de points associé est :



Par exemple, le quatrième point du nuage en partant de la gauche correspond au rang 4 Avril : le point M_4 correspondant est de coordonnées $(4; 40)$.

La silhouette du nuage de points est étirée dans une direction, une relation affine entre Y et X est envisageable. Ainsi, à partir des valeurs x de X, estimer avec précision les valeurs de Y correspondantes revient à déterminer a et b de sorte à ce que la droite d'équation $y = a + bx$ ajuste au mieux le nuage de points.

De nouveau, après plusieurs essais graphiques "à l'œil", en utilisant la calculatrice (ou autre), on constate que la droite suivante ajuste "pas trop mal" le nuage de points :



Ainsi, avec cette méthode "au jugé", on propose les coefficients $a = 35$ et $b = 2,5$, pour une droite d'équation : $y = a+bx$. Avec cette équation, on peut alors faire des prévisions. Par exemple, au rang 13 correspondant au mois de janvier 2017, une valeur estimée du nombre d'inscriptions au jeu en milliers est :

$$y = a + bx = 35 + 2,5 \times 13 = 67,5.$$

Ainsi, en janvier 2017, on prévoit 67500 inscriptions.

Aussi, avec cet ajustement, on peut espérer que le nombre d'inscriptions au jeu dépasse 80000 au rang x, avec x vérifiant :

$$y \geq 80 \quad \Leftrightarrow \quad 35 + 2,5 \times x \geq 80 \quad \Leftrightarrow \quad x \geq \frac{80 - 35}{2,5} = 18.$$

Cela correspond à Juin 2017.

Méthodes

La méthode "au jugé" dépend de l'utilisateur et donne donc des prévisions subjectives ; le choix de a et b ne repose sur aucun socle théorique. Plusieurs autres méthodes existent. Il y a notamment :

- la méthode des points observés,
- la méthode des points moyens,
- la méthode des moindres carrés.

Ces méthodes amènent des estimations de a et b différentes. Elle sont présentées ci-après.

2 Méthode des points observés

Résultat central : Équation d'une droite passant par deux points

Soient A et B deux points sur le repère orthonormé (O, I, J) de coordonnées respectives $(x_A; y_A)$ et $(x_B; y_B)$. Alors la droite passant par les points A et B a pour équation $y = a + bx$, avec

$$b = \frac{y_B - y_A}{x_B - x_A}, \qquad a = y_A - bx_A.$$

Méthode des points observés

La méthode des points observés propose d'ajuster le nuage de points par une droite passant par le point M_j de coordonnées $(x_j; y_j)$ et le point M_k de coordonnées $(x_k; y_k)$ choisis parmi $M_1, M_2, \ldots, M_n$. Cette droite est d'équation $y = a + bx$, avec

$$b = \frac{y_k - y_j}{x_k - x_j}, \qquad a = y_j - bx_j.$$

Une idée est de choisir M_j et M_k tels que la droite qui y passent ajuste "visiblement bien" le nuage de points.

Méthode des points extrêmes

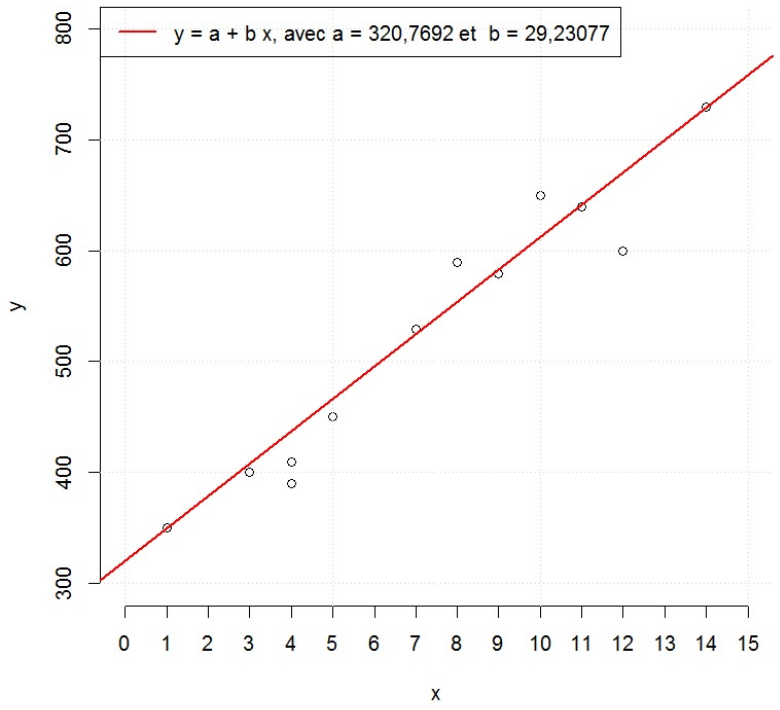
La méthode des points extrêmes est un cas particulier de la méthode des points observés. Elle propose d'ajuster le nuage de points par une droite passant par le point M_j situé le plus à gauche et le point M_k situé le plus à droite.

Exemples

Retour sur l'exemple 1. Une étude a été menée auprès de 12 étudiants afin d'expliquer le score à un examen de mathématiques à partir du temps consacré à la préparation de cet examen. Pour chaque étudiant, on dispose du temps de révision en heures (variable X) et du score obtenu sur 800 points (variable Y). Les résultats sont :

x_i	4	9	10	14	4	7	12	1	3	8	11	5
y_i	390	580	650	730	410	530	600	350	400	590	640	450

La méthode des points extrêmes propose la droite suivante :



On a alors considéré le point situé le plus à gauche du nuage de points et le point situé le plus à droite. Le premier point étant M_8 de coordonnées $(1; 350)$ et le deuxième point étant M_4 de coordonnées $(14; 730)$. En utilisant ces coordonnées, l'équation de la droite est $y = a + bx$, avec

$$b = \frac{730 - 350}{14 - 1} = 29,23077, \qquad a = 350 - 29,23077 \times 1 = 320,7692.$$

Avec cette équation, on peut alors faire des prévisions. Par exemple, une valeur estimée du score d'un étudiant ayant consacré 16 heures de préparation à l'examen est :

$$y = a + bx = 320,7692 + 29,23077 \times 16 = 788,4615.$$

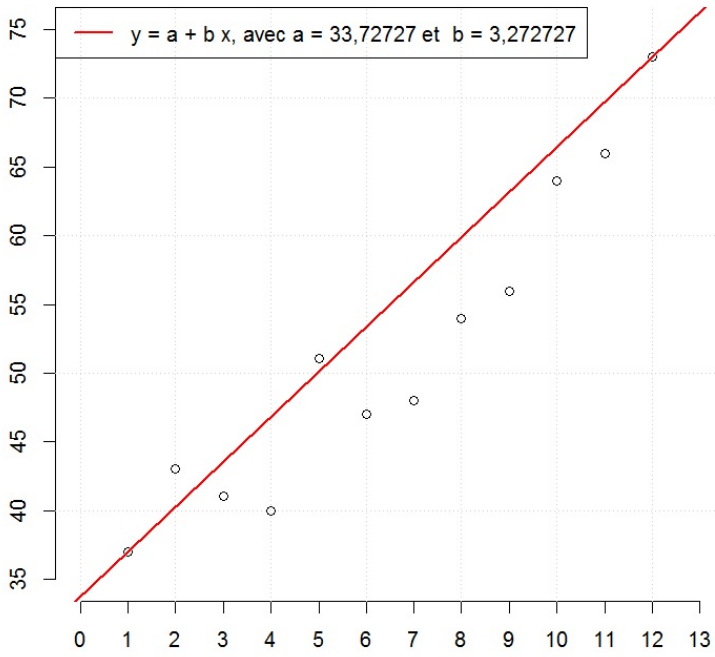
Ainsi, on prévoit un score de 789 pour un tel étudiant.

Retour sur l'exemple 2. On étudie l'évolution du nombre d'inscriptions à un jeu en ligne au cours du temps.

Pour chaque mois de l'année 2016, on dispose du rang du mois (variable X ; janvier est rang 1, février est le rang 2...) et du nombre d'inscriptions en milliers (variable Y). Les résultats sont :

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	37	43	41	40	51	47	48	54	56	64	66	73

La méthode des points extrêmes propose la droite suivante :



On a alors considéré le point situé le plus à gauche du nuage de points et le point situé le plus à droite. Le premier point étant M_1 de coordonnées $(1; 37)$ et le dernier point étant M_{12} de coordonnées $(12; 73)$. En utilisant ces coordonnées, l'équation de cette droite est $y = a + bx$, avec

$$b = \frac{73 - 37}{12 - 1} = 3,272727, \qquad a = 37 - 3,272727 \times 1 = 33,72727.$$

Avec cette équation, on peut alors faire des prévisions. Par exemple, au rang 13 correspondant au mois de janvier 2017, une valeur estimée du nombre d'inscriptions au jeu en milliers est :

$$y = a + bx = 33,72727 + 3,272727 \times 13 = 76,27272.$$

Ainsi, on prévoit 76500 inscriptions en janvier 2017.

3 Méthode des points moyens

Point moyen

Le point moyen d'un ensemble de points est un point G de coordonnées la moyenne des coordonnées des points de cet ensemble. Par exemple, le point moyen du nuage de points formé de $M_1, M_2, \ldots, M_n$ (de coordonnées respectives $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$) est le point G de coordonnées $(\overline{x}; \overline{y})$, où $\overline{x}$ et $\overline{y}$ désignent les moyennes :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Méthode des points moyens (ou méthode de Mayer)

La méthode des points moyens propose d'ajuster le nuage de points par une droite passant par les deux points moyens G_1 et G_2 de deux ensembles de points du nuage, l'un formé des points les plus à gauche, et l'autre formé des points les plus à droite. Ainsi, ces deux ensembles forment une partition du nuage de points et contiennent le même nombre de points (plus un pour l'un si n est impair).

Ainsi, pour G_1 de coordonnées $(\overline{x}_1; \overline{y}_1)$ et G_2 de coordonnées $(\overline{x}_2; \overline{y}_2)$, la méthode des points moyens propose la droite d'équation $y = a + bx$, avec

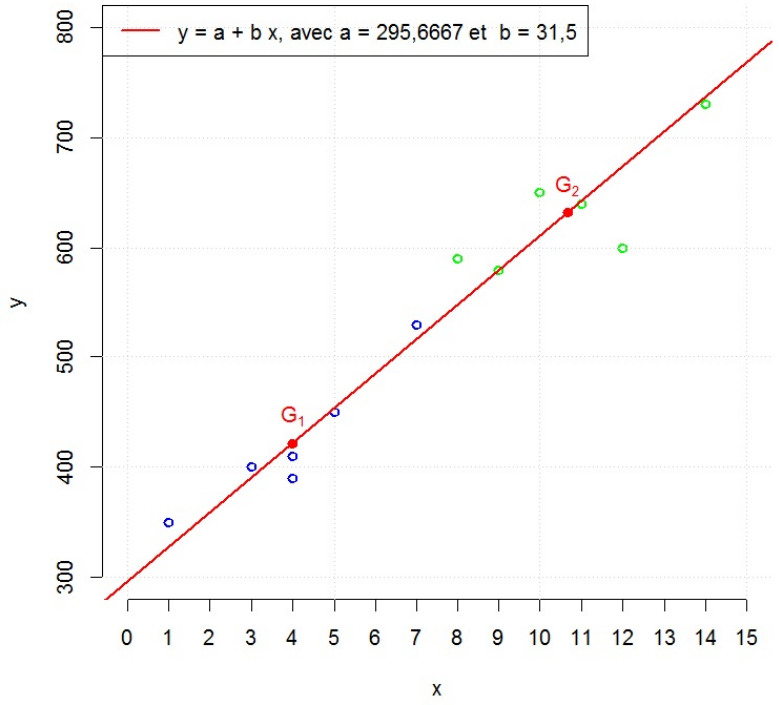
$$b = \frac{\overline{y}_2 - \overline{y}_1}{\overline{x}_2 - \overline{x}_1}, \qquad a = \overline{y}_1 - b\overline{x}_1.$$

Exemples

Retour sur l'exemple 1. Une étude a été menée auprès de 12 étudiants afin d'expliquer le score à un examen de mathématiques à partir du temps consacré à la préparation de cet examen. Pour chaque étudiant, on dispose du temps de révision en heures (variable X) et du score obtenu sur 800 points (variable Y). Les résultats sont :

x_i	4	9	10	14	4	7	12	1	3	8	11	5
y_i	390	580	650	730	410	530	600	350	400	590	640	450

La méthode des points moyens propose la droite suivante :



On a alors considéré deux ensembles de points du nuage. L'un est formé des points les plus à gauche (en bleue) :

M_8	M_9	M_1	M_5	M_{12}	M_6
$(1;350)$	$(3;400)$	$(4;390)$	$(4;410)$	$(5;450)$	$(7;530)$

L'autre est formé des points les plus à droite (en vert) :

M_{10}	M_2	M_3	M_{11}	M_7	M_4
$(8;590)$	$(9;580)$	$(10;650)$	$(11;640)$	$(12;600)$	$(14;730)$

On a déterminé les points moyens G_1 et G_2 de ces ensembles.

Ainsi, G_1 est de coordonnées :

$$(\overline{x}_1; \overline{y}_1) = \left(\frac{1+3+4+4+5+7}{6}; \frac{350+400+390+410+450+530}{6} \right) = (4; 421,6667)$$

et G_2 est de coordonnées :

$$(\overline{x}_2; \overline{y}_2) = \left(\frac{8+9+10+11+12+14}{6}; \frac{590+580+650+640+600+730}{6} \right) = (10,66667; 631,66667).$$

En utilisant ces coordonnées, l'équation de la droite passant par G_1 et G_2 est $y = a + bx$, avec

$$b = \frac{631,66667 - 421,6667}{10,66667 - 4} = 31,5, \qquad a = 421,6667 - 31,5 \times 4 = 295,6667.$$

Avec cette équation, on peut alors faire des prévisions. Par exemple, une valeur estimée du score d'un étudiant ayant consacré 16 heures de préparation à l'examen est :

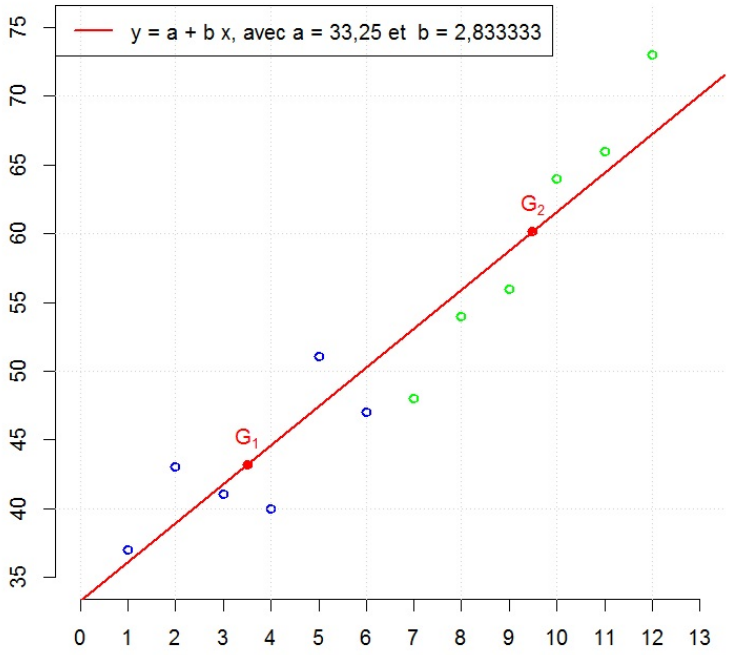
$$y = a + bx = 295,6667 + 31,5 \times 16 = 799,6667.$$

Ainsi, on prévoit un score de 800 pour un tel étudiant.

Retour sur l'exemple 2. On étudie l'évolution du nombre d'inscriptions à un jeu en ligne au cours du temps. Pour chaque mois de l'année 2016, on dispose du rang du mois (variable X ; janvier est rang 1, février est le rang 2...) et du nombre d'inscriptions en milliers (variable Y). Les résultats sont :

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	37	43	41	40	51	47	48	54	56	64	66	73

La méthode des points moyens propose la droite suivante :



On a alors considéré deux ensembles de points du nuage. L'un est formé des points les plus à gauche (en bleue) :

M_1	M_2	M_3	M_4	M_5	M_6
$(1;37)$	$(2;43)$	$(3;41)$	$(4;40)$	$(5;51)$	$(6;47)$

L'autre est formé des points les plus à droite (en vert) :

M_7	M_8	M_9	M_{10}	M_{11}	M_{12}
$(7;48)$	$(8;54)$	$(9;56)$	$(10;64)$	$(11;66)$	$(12;73)$

On a déterminé les points moyens G_1 et G_2 de ces ensembles.

Ainsi, G_1 est de coordonnées :

$$(\overline{x}_1;\overline{y}_1) = \left(\frac{1+2+3+4+5+6}{6}; \frac{37+43+41+40+51+47}{6}\right) = (3,5; 43,16667)$$

et G_2 est de coordonnées :

$$(\overline{x}_2;\overline{y}_2) = \left(\frac{7+8+9+10+11+12}{6}; \frac{48+54+56+64+66+73}{6}\right) = (9,5; 60,16667).$$

En utilisant ces coordonnées, l'équation de la droite passant par G_1 et G_2 est $y = a + bx$, avec

$$b = \frac{60,16667 - 43,16667}{9,5 - 3,5} = 2,833333, \qquad a = 43,16667 - 2,833333 \times 3,5 = 33,25.$$

Avec cette équation, on peut alors faire des prévisions. Par exemple, au rang 13 correspondant au mois de janvier 2017, une valeur estimée du nombre d'inscriptions au jeu en milliers est :

$$y = a + bx = 33,25 + 2,833333 \times 13 = 70,08333.$$

Ainsi, on prévoit 70080 inscriptions en janvier 2017.

4 Méthode des moindres carrés

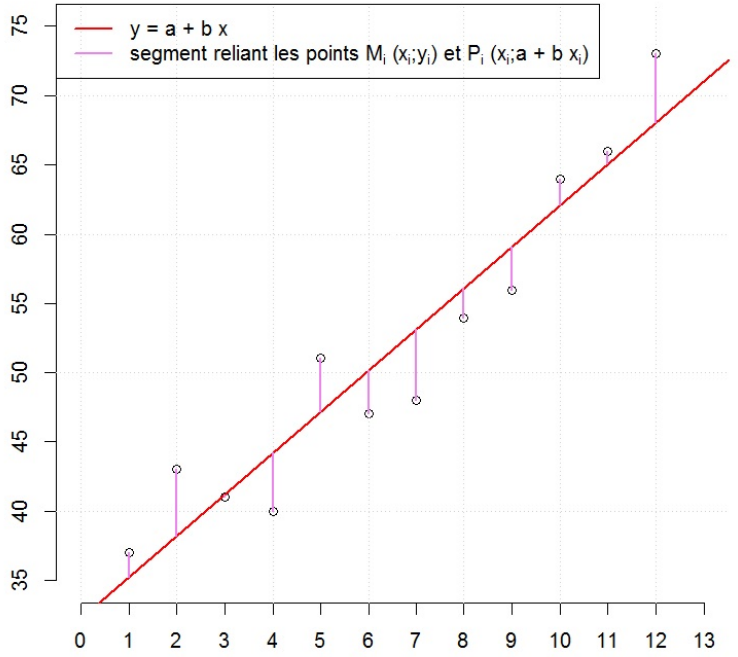
Méthode des moindres carrés

La méthode des moindres carrés propose d'ajuster le nuage de points par une droite d'équation $y = a + bx$, avec a et b qui rendent minimale la somme des carrés : $\sum_{i=1}^{n}(y_i - (a + bx_i))^2$. Cette droite, que l'on suppose unique, est appelée droite de régression.

L'idée de cette méthode est de déterminer une droite qui minimise une mesure totale des écarts entre les points du nuage et les points de mêmes abscisses se trouvant sur la droite. Ainsi, plus cette mesure est petite, plus la droite est proche de tous les points du nuage, meilleur est l'ajustement.

Illustration

Dans le graphique ci-dessous, chaque segment violet relie un point du nuage et le point de même abscisse se trouvant sur la droite d'équation $y = a + bx$:



Ainsi, pour tout $i \in \{1, \dots, n\}$, le i-ème segment relie le point M_i de coordonnées $(x_i; y_i)$ et le point P_i de coordonnées $(x_i; a+bx_i)$. La longueur de ce segment correspond à la distance $d_i = |y_i - (a+bx_i)|$. On cherche donc à minimiser la somme des carrés de ces distances : $\sum_{i=1}^{n} d_i^2$.

Notations

Dorénavant, on pose : $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$,

$$\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} \times \overline{y},$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2}, \qquad \sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2}.$$

Équation de la droite de régression

On peut montrer que la droite de régression est la droite d'équation $y = a + bx$, avec

$$b = \frac{\sigma_{x,y}}{\sigma_x^2}, \qquad a = \overline{y} - b\overline{x}.$$

Notons que $a = \overline{y} - b\overline{x}$ implique $\overline{y} = a + b\overline{x}$, ce qui signifie que la droite de régression passe par le point moyen G du nuage de points (de coordonnées $(\overline{x}; \overline{y})$).

Utilisation de la calculatrice

La plupart des calculatrices graphiques ont une option "régression linéaire" qui donne directement les valeurs de a et b. Il faut alors faire attention à ce que les coefficients "a" et "b" de l'option correspondent bien à ceux de l'équation : $y = a + bx$.

Commentaire : Par exemple, dans les calculatrices de marque `Casio`*, l'équation considérée est* $y = ax + b$ *; il faut donc inverser les rôles de* a *et* b *pour avoir les bonnes valeurs.*

Droite de régression avec une calculatrice `Texas Instrument`

Pour calculer les coefficients a et b de la droite de régression d'équation $y = a + bx$ avec une calculatrice de marque `Texas Instrument` :

- Aller au menu *Stat*, puis `Edit`. Appuyer sur Entrée.
- Rentrer les valeurs de X dans `L1` et les valeurs de Y dans `L2`.
- Aller au menu *Stat*, puis `Calc`, puis `Lin-Reg(a+bx)`, indiquer `L1` et `L2`.
- Appuyer sur Entrée 2 fois.

Coefficient de corrélation linéaire

On appelle coefficient de corrélation linéaire le réel r défini par

$$r = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}.$$

Ajustement affine et coefficient de corrélation linéaire

Dans un premier temps, remarquons que

$$b = \frac{\sigma_{x,y}}{\sigma_x^2} = \frac{\sigma_y}{\sigma_x} \times \frac{\sigma_{x,y}}{\sigma_x \sigma_y} = \frac{\sigma_y}{\sigma_x} r.$$

Comme $\sigma_x > 0$ et $\sigma_y > 0$, le coefficient directeur b de la droite de régression et r sont de même signe (à une droite de régression croissante correspond un r positif...). Dès lors, on peut deviner le signe de r avec la silhouette du nuage de points. D'autre part, en utilisant l'équation de la droite de régression, on peut montrer que

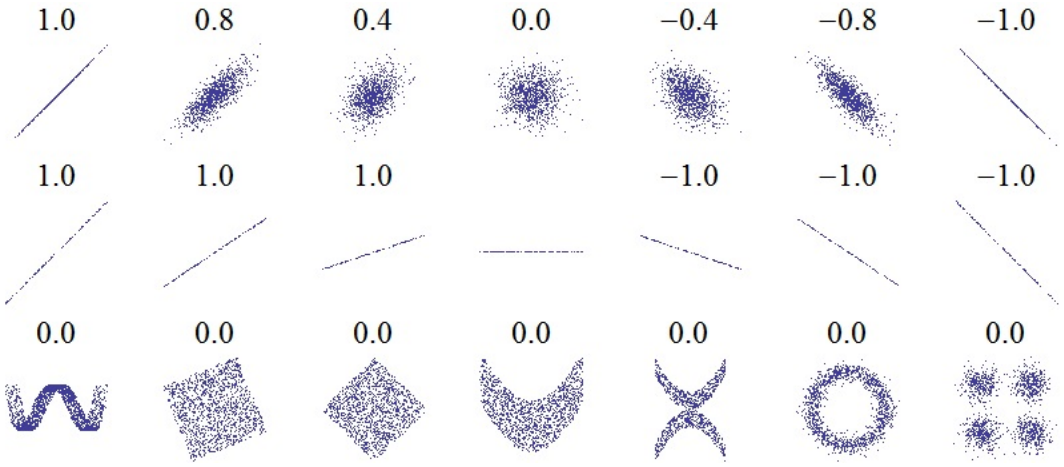
$$|r| = \sqrt{1 - \frac{1}{n\sigma_y^2} \sum_{i=1}^{n} (y_i - (a + bx_i))^2}.$$

Cela entraîne que $-1 \leq r \leq 1$. De plus, on a $|r| = 1$ si et seulement si

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2 = 0,$$

ce qui implique $y_i = a + bx_i$ pour tout $i \in \{1, \ldots, n\}$: tous les points du nuage sont alignés sur la droite de régression ; l'ajustement est parfait. Plus $|r|$ s'éloigne de 1 vers 0, plus l'ajustement est douteux.

Le graphique suivant illustre le lien existant entre la pertinence de l'ajustement d'un nuage de points par une droite, caractérisée par la corrélation linéaire entre Y et X, et la valeur associée de r :



Source : `https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`

Critères

Partant de r^2, on adopte les critères numériques suivants :

- si $0,75 \leq r^2 \leq 1$, alors il existe une bonne corrélation linéaire entre Y et X,
- si $0,25 \leq r^2 < 0,75$, alors il existe une faible corrélation linéaire entre Y et X,
- si $0 \leq r^2 < 0,25$, alors il existe une mauvaise corrélation linéaire entre Y et X.

Utilisation de la calculatrice

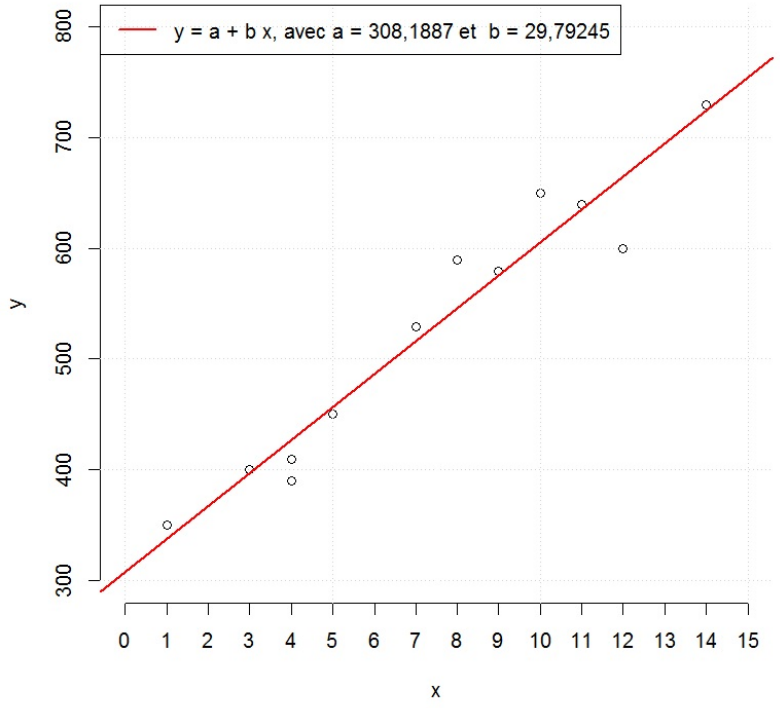
Avec l'option adéquate, la plupart des calculatrices graphiques donnent les valeurs de r et r^2 en même temps que celles de a et b.

Exemples

Retour sur l'exemple 1. Une étude a été menée auprès de 12 étudiants afin d'expliquer le score à un examen de mathématiques à partir du temps consacré à la préparation de cet examen. Pour chaque étudiant, on dispose du temps de révision en heures (variable X) et du score obtenu sur 800 points (variable Y). Les résultats sont :

x_i	4	9	10	14	4	7	12	1	3	8	11	5
y_i	390	580	650	730	410	530	600	350	400	590	640	450

La méthode des moindres carrés propose la droite (de régression) suivante :



On a utilisé les valeurs a et b données par l'option "régression linéaire" de la calculatrice. On peut toutefois retrouver ces valeurs avec les formules. Après calculs, on obtient :

$\overline{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i^2$	$\overline{y}$	$\frac{1}{n}\sum_{i=1}^{n} x_i y_i$	$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$	$\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} \times \overline{y}$
$7,333333$	$68,5$	$526,6667$	$4300,833$	$14,72223$	$438,6107$

Ainsi, la droite de régression est la droite d'équation $y = a + bx$, avec

$$b = \frac{\sigma_{x,y}}{\sigma_x^2} = \frac{438,6107}{14,72223} = 29,7924$$

et

$$a = \overline{y} - b\overline{x} = 526,6667 - 29,7924 \times 7,333333 = 308,189.$$

Avec l'équation de la droite de régression, on peut faire des prévisions.

Par exemple, une valeur estimée du score d'un étudiant ayant consacré 16 heures de préparation à l'examen est :

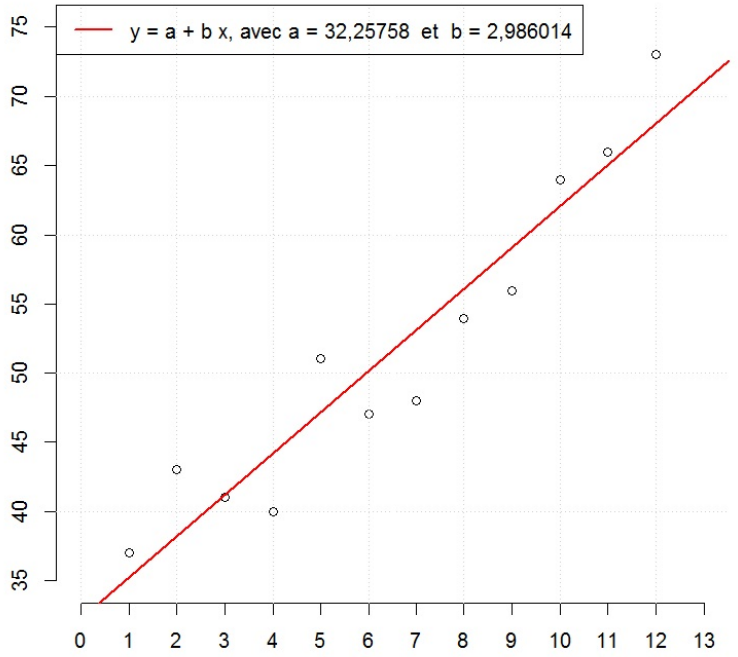
$$y = a + bx = 308,189 + 29,7924 \times 16 = 784,8674.$$

Ainsi, on prévoit un score de 785 pour un tel étudiant.

Retour sur l'exemple 2. On étudie l'évolution du nombre d'inscriptions à un jeu en ligne au cours du temps. Pour chaque mois de l'année 2016, on dispose du rang du mois (variable X ; janvier est rang 1, février est le rang 2...) et du nombre d'inscriptions en milliers (variable Y). Les résultats sont :

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	37	43	41	40	51	47	48	54	56	64	66	73

La méthode des moindres carrés propose la droite (de régression) suivante :



On a utilisé les valeurs a et b données par l'option "régression linéaire" de la calculatrice. On peut toutefois retrouver ces valeurs avec les formules.

Après calculs, on obtient :

$\overline{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i^2$	$\overline{y}$	$\frac{1}{n}\sum_{i=1}^{n} x_i y_i$	$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$	$\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} \times \overline{y}$
$6,5$	$54,16667$	$51,66667$	$371,4167$	$11,91667$	$35,58334$

Ainsi, la droite de régression est la droite d'équation $y = a + bx$, avec

$$b = \frac{\sigma_{x,y}}{\sigma_x^2} = \frac{35,58334}{11,91667} = 2,986014$$

et

$$a = \overline{y} - b\overline{x} = 51,66667 - 2,986014 \times 6,5 = 32,25758.$$

Avec cette équation, on peut alors faire des prévisions.

Par exemple, au rang 13 correspondant au mois de janvier 2017, une valeur estimée du nombre d'inscriptions au jeu en milliers est :

$$y = a + bx = 32,25758 + 2,986014 \times 13 = 71,07576.$$

Ainsi, on prévoit 71076 inscriptions en janvier 2017.

Pour compléter, intéressons nous au coefficient de corrélation linéaire. La valeur de r^2 renvoyée par la calculatrice est

$$r^2 = 0,9025686$$

(On aurait aussi pu utiliser la formule : $r = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$ et élever le résultat au carré). Comme $0,75 \leq r^2 \leq 1$, il existe une bonne corrélation linéaire entre Y et X.

5 Pour s'entraîner

Énoncé 1. On souhaite expliquer le nombre de cellules végétales au millimètre carré (variable Y) à partir du temps d'exposition au soleil en jours (variable X). Pour 7 expériences indépendantes, les résultats sont :

x_i	2	4	8	10	24	40	52
y_i	6	11	15	20	39	62	85

Énoncé 2. Pour 12 sites internet de commerce électronique, on compte sur une semaine le nombre de connexions (variable X) et le nombre de commandes (variable Y). Les résultats sont :

x_i	90	100	115	110	70	125	105	90	110	95	80	75
y_i	42	50	62	56	28	75	63	55	53	49	30	27

Énoncé 3. On teste 7 candidats à un jeu électronique et on relève pour chacun d'entre eux le nombre de tentatives (variable X) et le score maximum (variable Y). Les résultats sont :

x_i	12	20	15	23	8	30	24
y_i	26	47	24	54	5	81	61

Énoncé 4. Une entreprise souhaite expliquer le coût total de production mensuel en kilos euro (variable Y) en fonction de la production en tonnes (variable X). Les résultats sont :

x_i	1	2	4	6	8	10
y_i	32,5	38,5	44,6	48,4	51,1	53,3

Énoncé 5. On souhaite expliquer le chemin de freinage en mètres d'un véhicule (distance parcourue entre le début du freinage et l'arrêt total) (variable Y) à partir de sa vitesse en kilomètres heure (variable X). Pour 12 expériences indépendantes, les résultats sont :

x_i	40	50	60	70	80	90	100	110	120	130	140	150
y_i	9	11	20	27	39	45	58	78	79	93	108	124

Énoncé 6. On étudie l'évolution du prix d'un produit au cours du temps. Pour les 4 dernières années, on dispose du rang de l'année (variable X) et du prix en euros du produit (variable Y). Les résultats sont :

x_i	1	2	3	4
y_i	10,6	11,5	12,2	13,1

Énoncé 7. On étudie l'évolution du nombre de visiteurs sur un site internet au cours du temps. Pour les 10 premiers mois du site, on dispose du rang du mois (variable X) et du nombre de visiteurs (variable Y). Les résultats sont :

x_i	1	2	3	4	5	6	7	8	9	10
y_i	325	328	331	333	337	341	342	345	346	345

Énoncé 8. On étudie l'évolution de la dépense des ménages en produits informatiques au cours du temps. Pour les années comprises entre 1990 à 1999, on dispose du rang de l'année (variable X) et de la dépense en millions d'euros (variable Y). Les résultats sont :

x_i	1	2	3	4	5	6	7	8	9	10
y_i	398	451	423	501	673	956	1077	1285	1427	1490

Énoncé 9. On étudie l'évolution du nombre d'ordinateurs fonctionnels dans un grand lycée au cours du temps. Pour les 8 dernières rentrées scolaires, on dispose du rang de l'année (variable X) et le nombre d'ordinateurs fonctionnels (variable Y). Les résultats sont :

x_i	1	2	3	4	5	6	7	8
y_i	140	160	180	220	260	320	380	450

Énoncé 10. Avant la commercialisation d'un produit, une entreprise effectue une étude de marché afin de déterminer la quantité demandée en milliers (variable Y) en fonction du prix de vente en euros (variable X). Pour 6 prix de vente différents, les résultats sont :

x_i	15	20	25	30	35	40
y_i	44,4	27,0	16,3	10,0	6,2	3,5

6 Quelques compléments

Ajustement non affine/non-linéaire

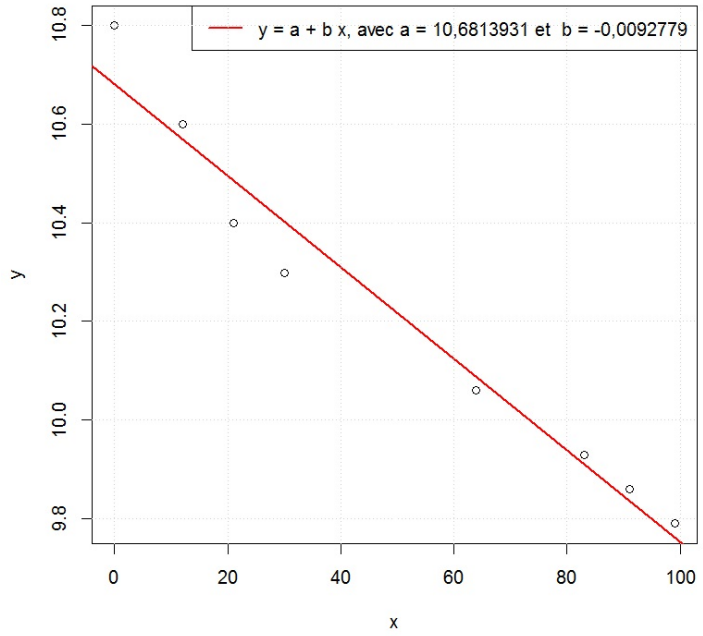
Première approche. Parfois,

- un nuage de points peut être visuellement mieux ajusté par une courbe que par une droite,
- même quand les points sont presque alignés, un ajustement affine ne permet pas toujours de faire des prévisions réalistes.

Pour illustrer ce dernier point, prenons un exemple. On étudie l'évolution des records de l'épreuve d'athlétisme du 100 mètres masculin. On dispose du rang de l'année (variable X ; 0 pour 1900, 30 pour 1930...) et de la performance record en secondes (variable Y). Les résultats sont :

x_i	0	12	21	30	64	83	91	99
y_i	10,80	10,60	10,40	10,30	10,06	9,93	9,86	9,79

Les points du nuage sont presque alignés et la méthode des moindres carrés propose la droite (de régression) suivante :



La droite de régression est d'équation : $y = a + bx$, avec $a = 10,6813931$ et $b = -0,0092779$. En outre, on a $r^2 = 0,9562$, confirmant ainsi une bonne corrélation linéaire entre Y et X. Cependant, cet ajustement ne permet pas des prévisions à long terme. Par exemple, si on considère l'année 3050 (de rang 1150), une estimation du record de l'épreuve d'athlétisme du 100 mètres masculin est :

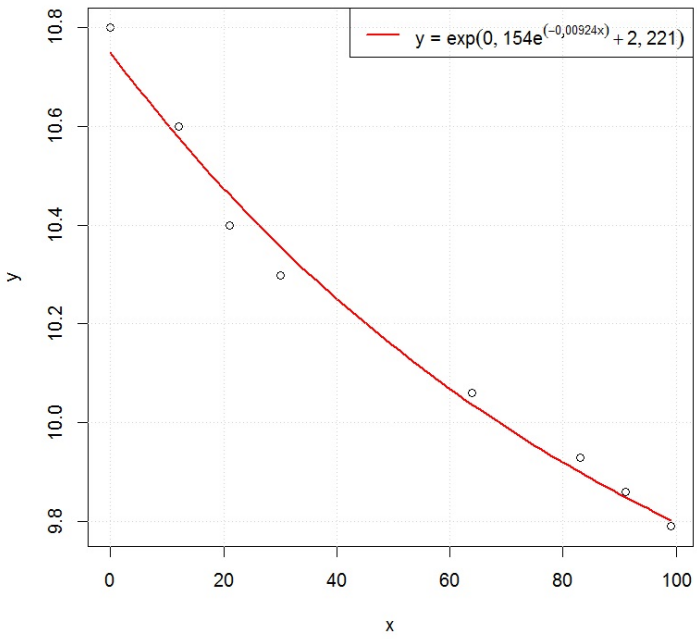
$$y = 10,6813931 - 0,0092779 \times 1150 = 0,0118081.$$

Or $0,012$ secondes est absurde. Ainsi, un ajustement du nuage de points par une courbe décroissante qui converge avec le temps vers une certaine valeur limite semble plus approprié. D'ailleurs, si on regarde attentivement la silhouette du nuage de points, une légère courbure allant dans ce sens peut être remarquée. Des études statistiques montrent que la courbe d'équation :

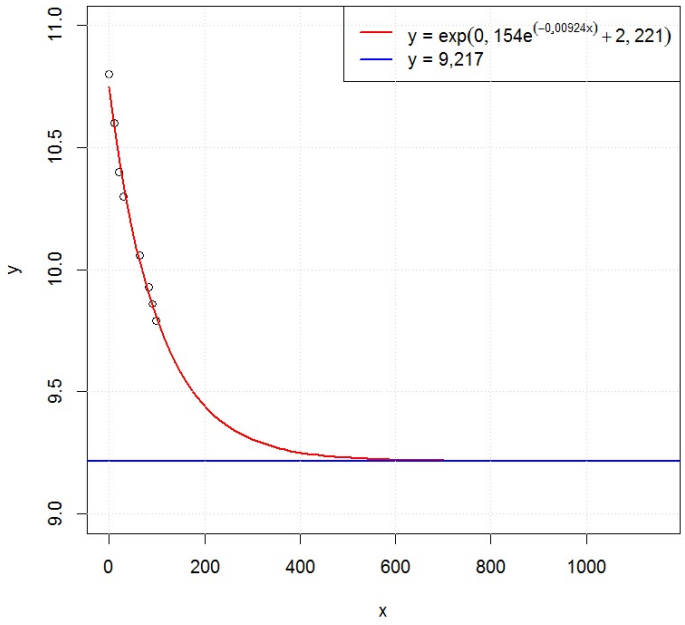
$$y = e^{(0,154 \times e^{-0,00924x} + 2,221)}$$

est acceptable, mettant le record du 100 mètres masculin égal à $9,217$ secondes à très long terme.

L'ajustement du nuage de points par cette courbe est représenté dans le graphique suivant :



On peut aussi représenter des prédictions à plus long terme, avec une stabilisation à $y = 9,217$:



Méthodes. Lorsque l'ajustement du nuage de points par une courbe est judicieux, on peut encore utiliser les méthodes vues précédemment en transformant intelligemment les variables X et Y. Pour ce faire, on choisit deux fonctions $f : \mathbb{R} \to \mathbb{R}$ et $g : \mathbb{R} \to \mathbb{R}$ de sorte à ce que la silhouette du nuage constitué des points M_1^*, M_2^*,..., M_n^* de coordonnées respectives $(f(x_1); g(y_1))$, $(f(x_2); g(y_2))$, ..., $(f(x_n); g(y_n))$ soit très étirée dans une direction. Dès lors, une relation non-linéaire entre Y et X est envisageable : on suppose l'existence de deux coefficients réels inconnus α et β tels que

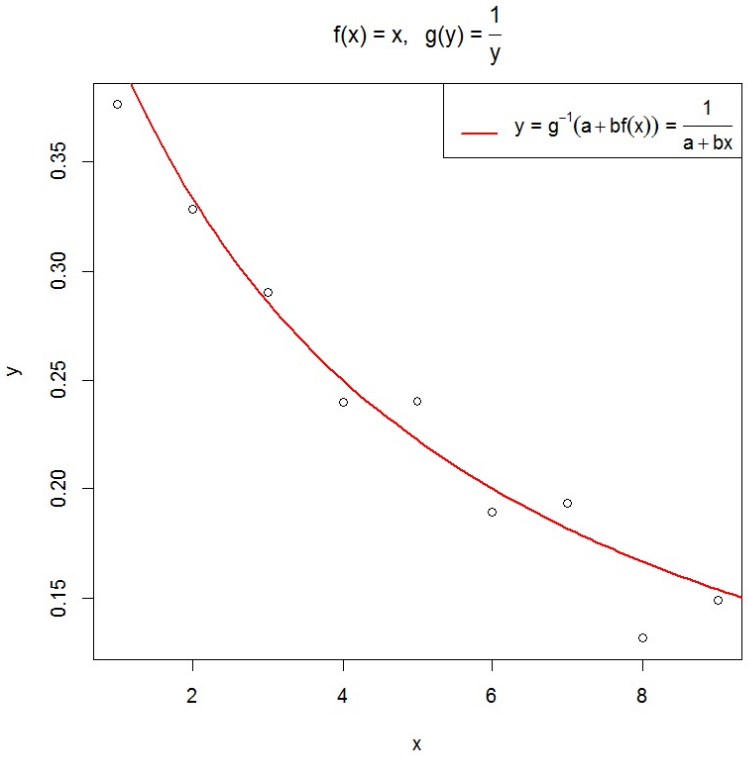
$$g(Y) = \alpha + \beta f(X).$$

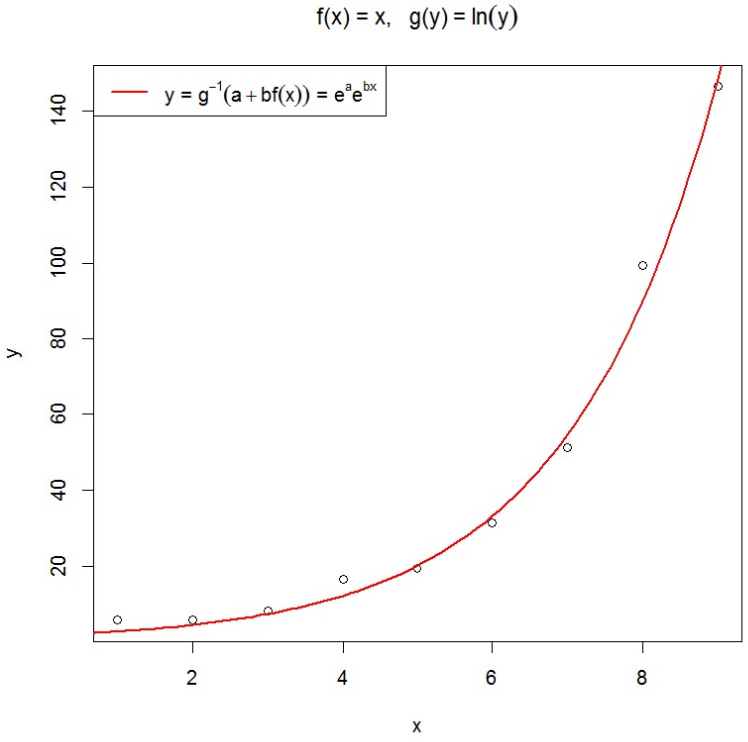
Ainsi, pour toute valeur x de X, une valeur estimée y de Y vérifie :

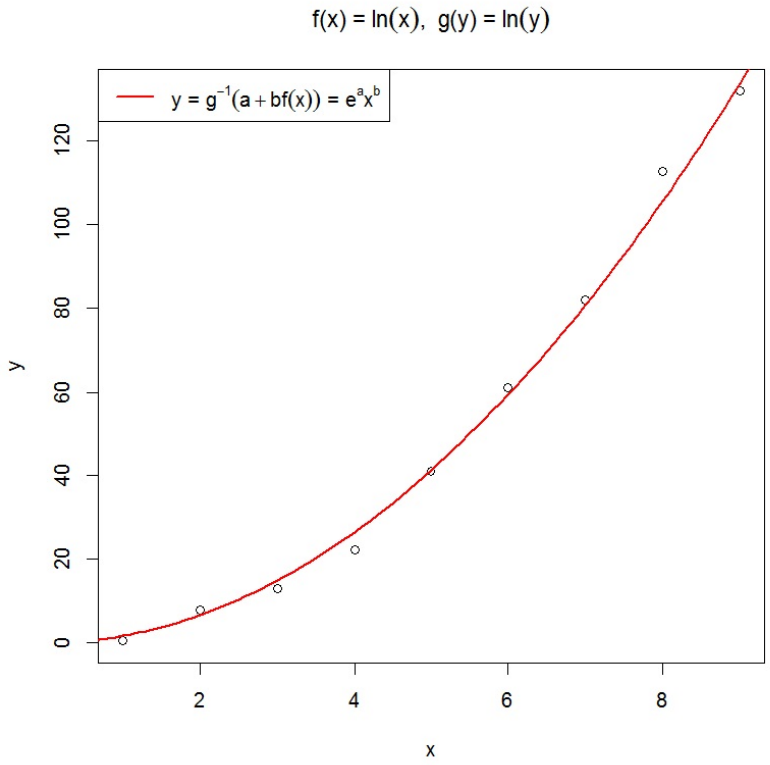
$$g(y) = a + bf(x) \qquad \Leftrightarrow \qquad y = g^{-1}(a + bf(x)),$$

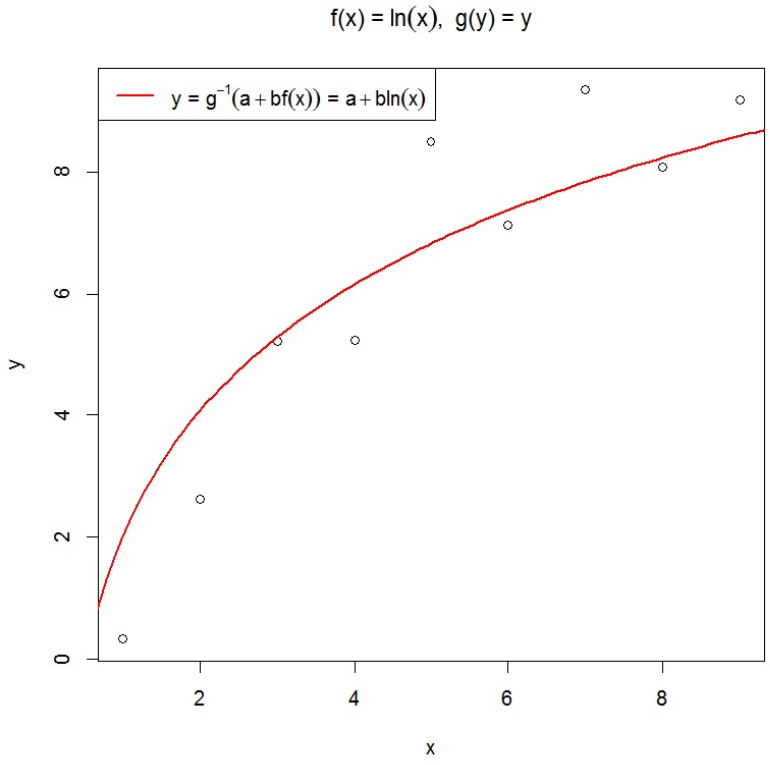
où a désigne une valeur estimée de α et b désigne une valeur estimée de β. On peut les calculer avec l'une des méthodes présentées précédemment et les données transformées : $(f(x_1); g(y_1))$, ..., $(f(x_n); g(y_n))$.

Des exemples d'ajustements non affines, avec différentes fonctions f et g, sont présentées ci-après.






f(x) = ln(x), g(y) = ln(y)
y = g^-1(a + bf(x)) = e^a x^b
y
x
0
20
40
60
80
100
120
2
4
6
8


f(x) = ln(x), g(y) = y
y = g^-1(a + bf(x)) = a + bln(x)
y
x
0
2
4
6
8
2
4
6
8

Autres méthodes d'ajustement

Il existe d'autres méthodes d'ajustement déterminant une droite qui minimise une mesure totale des écarts entre les points du nuage et celle-ci. Quelques-unes sont présentées ci-dessous.

Rappels : Méthode des moindres carrés : La méthode des moindres carrés propose d'ajuster le nuage de points par la droite de régression d'équation $y = a + bx$, avec a et b qui rendent minimale la somme des carrés :

$$\sum_{i=1}^{n} d_i^2,$$

avec d_i la distance entre le point du nuage M_i et le point de même abscisse sur la droite :

$$d_i = |y_i - (a + bx_i)|.$$

On peut alors montrer que

$$b = \frac{\sigma_{x,y}}{\sigma_x^2}, \qquad a = \overline{y} - b\overline{x}.$$

Méthode des moindres distances : La méthode des moindres distances propose d'ajuster le nuage de points par la droite d'équation $y = a + bx$, avec a et b qui rendent minimale la somme des carrés :

$$\sum_{i=1}^{n} d_i^2,$$

avec d_i la distance entre le point du nuage M_i et le point sur la droite mesurée de façon perpendiculaire :

$$d_i = \frac{1}{1+b^2}|y_i - (a + bx_i)|.$$

On peut alors montrer que

$$b = -c \pm \sqrt{1+c^2}, \qquad c = \frac{\sigma_x^2 - \sigma_y^2}{2\sigma_{x,y}}, \qquad a = \overline{y} - b\overline{x}.$$

Ainsi, deux solutions de signes opposés sont possibles pour b. Le bon est celui qui a le même signe que le coefficient de corrélation linéaire r.

Cette droite est appelée droite des moindres distances.

Méthode des moindres rectangles : La méthode des moindres rectangles propose d'ajuster le nuage de points par la droite des moindres rectangles d'équation $y = a + bx$, avec a et b qui rendent minimale la somme :

$$\sum_{i=1}^{n} s_i,$$

avec s_i la surface du rectangle défini de la manière suivante : ses côtés sont parallèles aux axes, le point M_i constitue un premier sommet et les deux sommets qui lui sont adjacents sont situés sur la droite :

$$s_i = |y_i - (a + bx_i)| \times \left| x_i - \frac{1}{b}(y_i - a) \right|.$$

On peut alors montrer que

$$b = \pm \frac{\sigma_y}{\sigma_x}, \qquad a = \overline{y} - b\overline{x}.$$

Ainsi, deux solutions de signes opposés sont possibles pour b. Le bon est celui qui a le même signe que le coefficient de corrélation linéaire r.

Remarque : Les droites présentées ci-dessus ont un point commun : elles passent toutes par le point moyen G du nuage de points.

Question : Pourquoi la méthode des moindres carrés est la plus populaires ?

Réponses :

- Les formules de a et b sont aisément calculables.
- Sous certaines hypothèses portant sur le terme d'erreur (que l'on a passé sous silence ici), on a des garanties théoriques attestant de la précision de notre ajustement.
- La méthode des moindres carrés s'étend au cas où l'on souhaite prévoir les valeurs d'une variable numérique Y à partir des valeurs de plusieurs variables numériques $X_1, X_2, \ldots, X_p$. Les formules qui en découlent sont aisément manipulables.

Preuve "par curiosité" : Équation de la droite de régression

Rappels : La méthode des moindres carrés propose d'ajuster le nuage de points par la droite de régression d'équation $y = a + bx$, avec a et b qui rendent minimale la somme des carrés : $\sum_{i=1}^{n}(y_i - (a + bx_i))^2$. On peut alors montrer que

$$b = \frac{\sigma_{x,y}}{\sigma_x^2}, \qquad a = \overline{y} - b\overline{x}.$$

Preuve : On considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(u,v) = \sum_{i=1}^{n}(y_i - (u + vx_i))^2$. Les conditions nécessaires pour que a et b rendent minimale $f(u,v)$ sont $\frac{\partial}{\partial u}f(a,b) = 0$ et $\frac{\partial}{\partial v}f(a,b) = 0$. Ainsi, on a

$$\begin{cases} \frac{\partial}{\partial u}f(a,b) = 0, \\ \\ \frac{\partial}{\partial v}f(a,b) = 0, \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} 2(-1)(y_i - (a + bx_i)) = 0, \\ \\ \sum_{i=1}^{n} 2(-x_i)(y_i - (a + bx_i)) = 0, \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} y_i - an - b\sum_{i=1}^{n} x_i = 0, \\ \\ \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0, \end{cases}$$

$$\Rightarrow \begin{cases} a = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i\right), \\ \\ \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0, \end{cases} \Rightarrow \begin{cases} a = \overline{y} - b\overline{x}, \\ \\ b = \dfrac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} \times \overline{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} = \dfrac{\sigma_{x,y}}{\sigma_x^2}. \end{cases}$$

Pour montrer que (a,b) est un minimum pour $f(u,v)$, on utilise les notations de Monge :

$$r = \frac{\partial^2}{\partial u^2}f(a,b) = 2n, \qquad s = \frac{\partial^2}{\partial u \partial v}f(a,b) = 2\sum_{i=1}^{n} x_i, \qquad t = \frac{\partial^2}{\partial v^2}f(a,b) = 2\sum_{i=1}^{n} x_i^2.$$

Ainsi, on a $r = 2n > 0$ et, par l'inégalité de Cauchy-Schwarz,

$$s^2 - rt = 4\left(\sum_{i=1}^{n} x_i\right)^2 - 4n\sum_{i=1}^{n} x_i^2 < 0.$$

On en conclut que (a,b) est un minimum (local) pour $f(u,v)$.