Comment Stata calcule des coefficients de régression ?

Pier-André Bouchard St-Amant

École nationale d'administration publique

19 mars 2018

Objectifs

Comprendre l'origine des coefficients d'une régression linéaire, soit comprendre :

- l'intuition derrière une formule de covariance ;
- ce qu'est la minimisation des termes d'erreurs ;
- le système d'équation servant à la résolution des coefficients de régression ;
- ce qu'est le biais de variable omises.

Un exemple concret : température et baigneurs

TABLE – Échantillon recueilli par l'équipe de planification

Observation i	Température x_{i1}	Personnes y_i
1	24	20
2	26	50
3	30	40

Un exemple concret : température et baigneurs

On cherche à estimer une relation linéaire « mystère » liant la température au nombre de personnes à la piscine :

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i. \tag{1}$$

1. Les termes β_0 et β_1 sont les paramètres inconnus de la relation linéaire qu'on cherche à estimer. β_0 est l'ordonnée à l'origine et β_1 est la pente. C'est l'équivalent du b et du a dans la relation $y = ax + b$.
2. Le terme e_i est une nouvelle variable qui représente **l'erreur d'estimation**. De toute évidence, une propriété désirable du modèle qu'on cherche à établir est que les erreurs de prédiction soient *minimales*.
3. Question centrale à la mécanique de régression : quels sont les nombres β_0^*, β_1^* qui représentent le mieux les données observées ?

Un exemple concret : température et baigneurs

TABLE – Échantillon, valeur prédite et erreur d'estimation

Obs. i	Temp. y_i	Pers. x_{i1}	Valeur prédite $-30 + \frac{5}{2}x_{i1}$	Erreur d'est. e_i^*
1	24	20	30	-10
2	26	50	35	15
3	30	40	45	-5

Paraboles multivariées

Une **forme quadratique** $f(\beta_0, \beta_1)$ à deux variables a la forme suivante :

$$f(\beta_0, \beta_1) = c_1\beta_0^2 + c_2\beta_0 + c_3\beta_0\beta_1 + c_4\beta_1 + c_5\beta_1^2 + c_6, \qquad (2)$$

1. les c_i sont des nombres connus, équivalents à a, b et c dans la version univariée.
2. elle dépend de deux variables, soit β_0 et β_1.
3. la plus haute puissance de chaque produit de variable est deux (2) ;

Paraboles multivariées (2 variables)

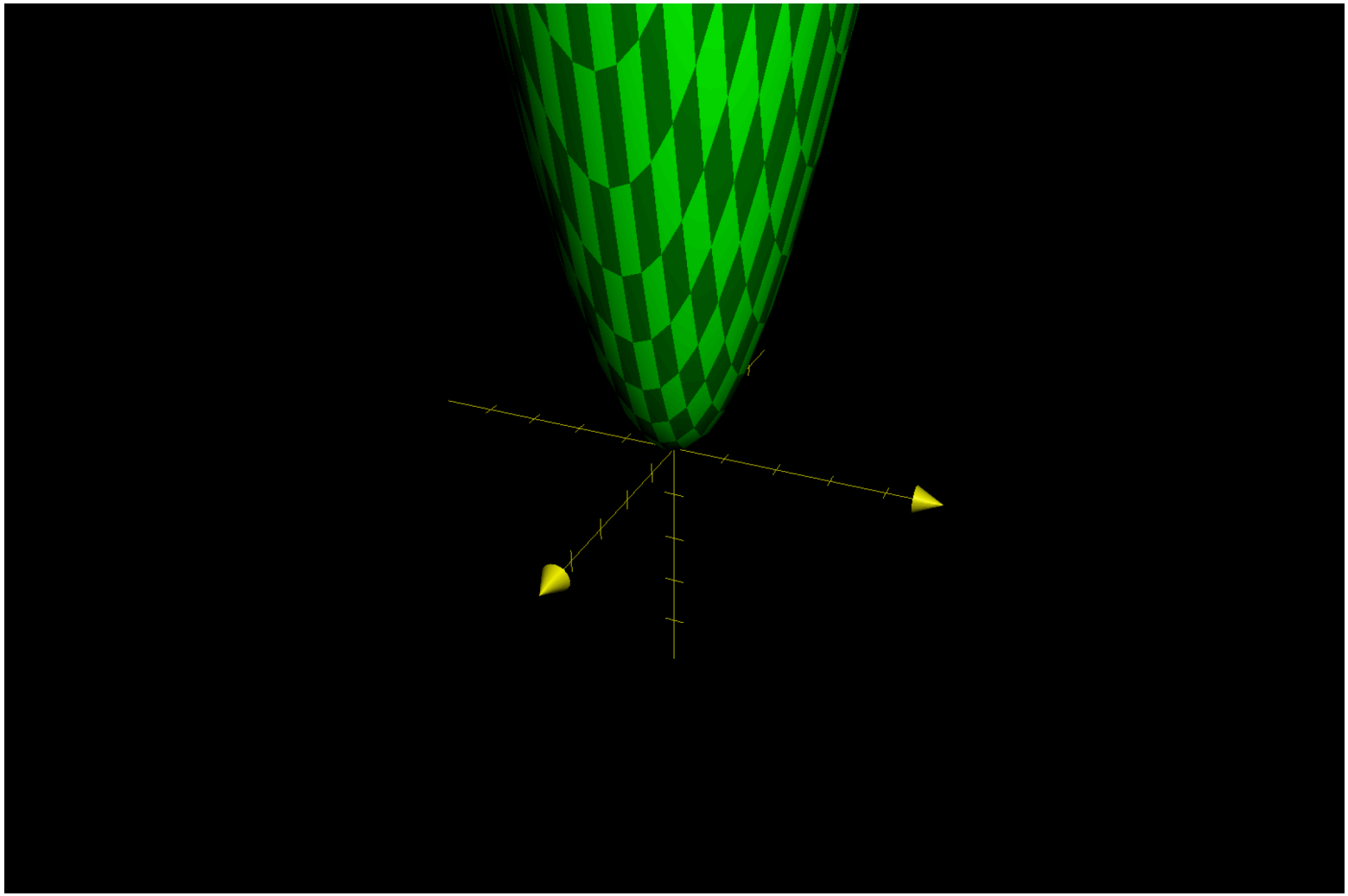


FIGURE – Représentation de $\beta_0^2 + \beta_1^2$ dans un espace tri-dimensionnel

Paraboles multivariées (3 variables)

1. Une équation quadratique à trois variables a quant-à-elle la forme suivante :

$$\begin{aligned} f(\beta_0, \beta_1, \beta_2) = c_1\beta_0^2 + c_2\beta_0 + c_3\beta_1 + c_4\beta_2 + \dots \\ \cdots + c_5\beta_0\beta_1 + c_6\beta_0\beta_2 + c_7\beta_1\beta_2 + c_8\beta_1^2 + + \dots \\ \cdots + c_9\beta_2^2 + c_{10}. \end{aligned} \tag{3}$$

2. On cherche le « creux » de la parabole multidimensionnelle.
3. Dans une forme quadratique univarée, on se souvient que cette valeur minimale est donnée à la valeur $\beta_0^* = -\frac{b}{2a}$

Trouver le sommet d'une forme quadratique

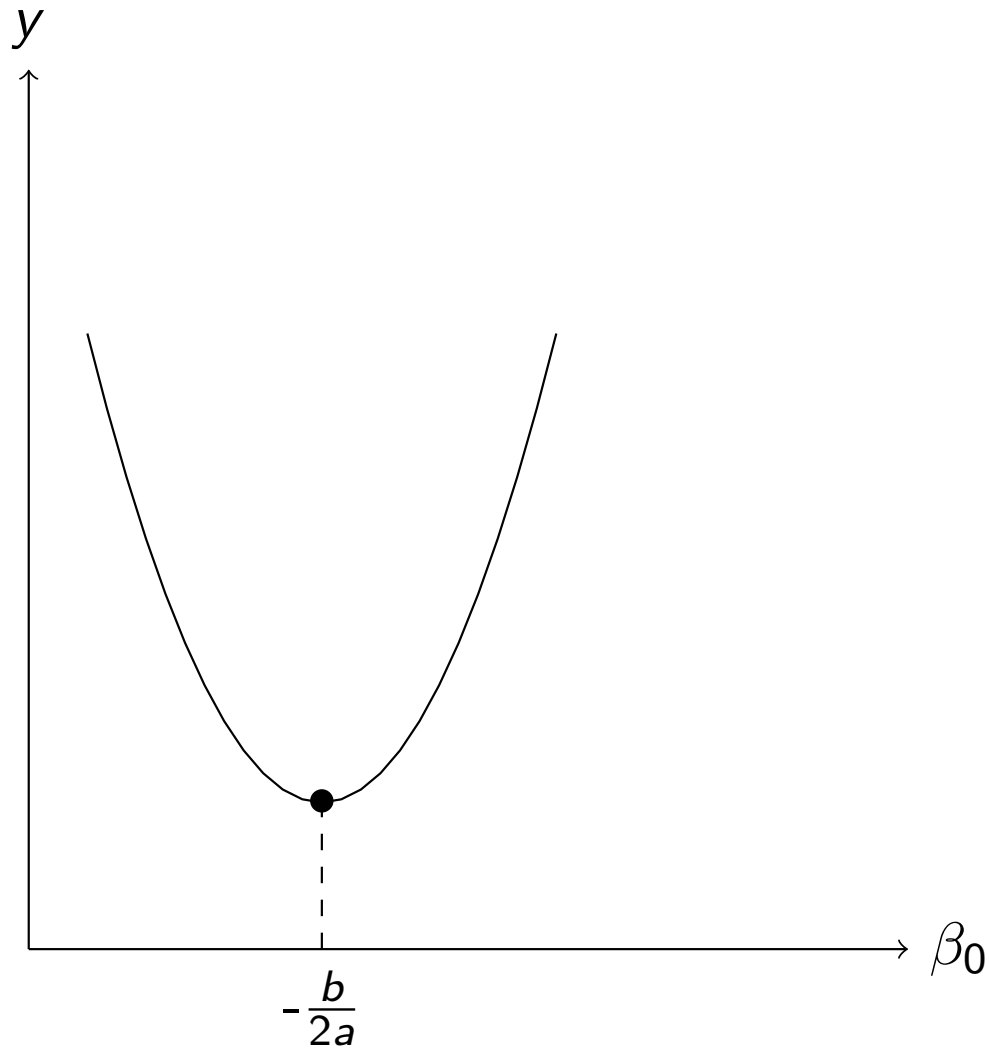


FIGURE – Une parabole univariée $y = a\beta_0^2 + b\beta_0 + c$ prends sa valeur minimale à $\beta_0^* = -\frac{b}{2a}$.

Trouver le sommet d'une forme quadratique

1. Pour une forme quadratique à plusieurs variables, le minimum de la parabole sera aussi au point spécifié par les coordonnées « $-\frac{b}{2a}$ ».
2. Il faut cependant spécifier cette coordonnée pour chacune des dimensions de la parabole $(\beta_0, \beta_1, \dots)$.
3. Pour trouver les coordonnées selon chaque dimension, l'idée consiste à organiser la forme quadratique en fonction d'une seule variable, comme si c'était une équation univariée. Il faut ensuite identifier les termes équivalents à b et a pour appliquer la formule $-\frac{b}{2a}$ selon cette dimension.
4. On procède ainsi pour chaque dimension.

Exemple à deux variables

$$f(\beta_0, \beta_1) = \underbrace{c_1}_{a} \beta_0^2 + \underbrace{(c_2 + c_3\beta_1)}_{b} \beta_0 + \underbrace{\left(c_4\beta_1 + c_5\beta_1^2 + c_6\right)}_{c}. \quad (4)$$

Le sommet en β_0 sera donné par l'équivalent de la formule $-\frac{b}{2a}$, soit :

$$\beta_0 = -\frac{c_2 + c_3\beta_1}{2c_1} \quad (5)$$

En faisant une démarche similaire pour la variable β_1, on trouve que le sommet sera donné à la coordonnée :

$$\beta_1^* = -\frac{c_4 + c_3\beta_0^*}{2c_5} \quad (6)$$

C'est un système à deux équations et deux inconnus en β_0, β_1. La résolution nous donne le minimum de la fonction.

Exemple à trois variables

En ce qui concerne une forme quadratique à trois variables, on peut reprendre la même démarche et la réorganiser en fonction de chacune des variables. Les coordonnées pour le sommet sont données par l'application de la formule $-\frac{b}{2a}$ à chacune des équations :

$$\beta_0 = -\frac{c_2 + c_5\beta_1 + c_6\beta_2}{2c_1}, \tag{7}$$

$$\beta_1 = -\frac{c_3 + c_5\beta_0 + c_7\beta_2}{2c_8}, \tag{8}$$

$$\beta_2 = -\frac{c_4 + c_6\beta_0 + c_7\beta_3}{2c_9}, \tag{9}$$

ce qui constitue un système de trois équations à trois inconnues. On peut le résoudre avec les techniques usuelles d'algèbre.

Plusieurs variables

On peut généraliser la démarche identifiée ci-dessus pour trouver le sommet d'une forme quadratique à $k+1$ variables en appliquant une démarche similaire :

1. Regrouper les termes en fonction d'une seule variable et identifier la formule $-\frac{b}{2a}$ associée à cette variable. Répéter pour chaque variable.
2. Cela génèrera un système de $k+1$ équations à $k+1$ inconnues qu'il faut ensuite résoudre avec les techniques usuelles d'algèbre.
3. Ce système d'équations aura pour solution les coordonnées de la valeur minimale que peut prendre la forme quadratique.

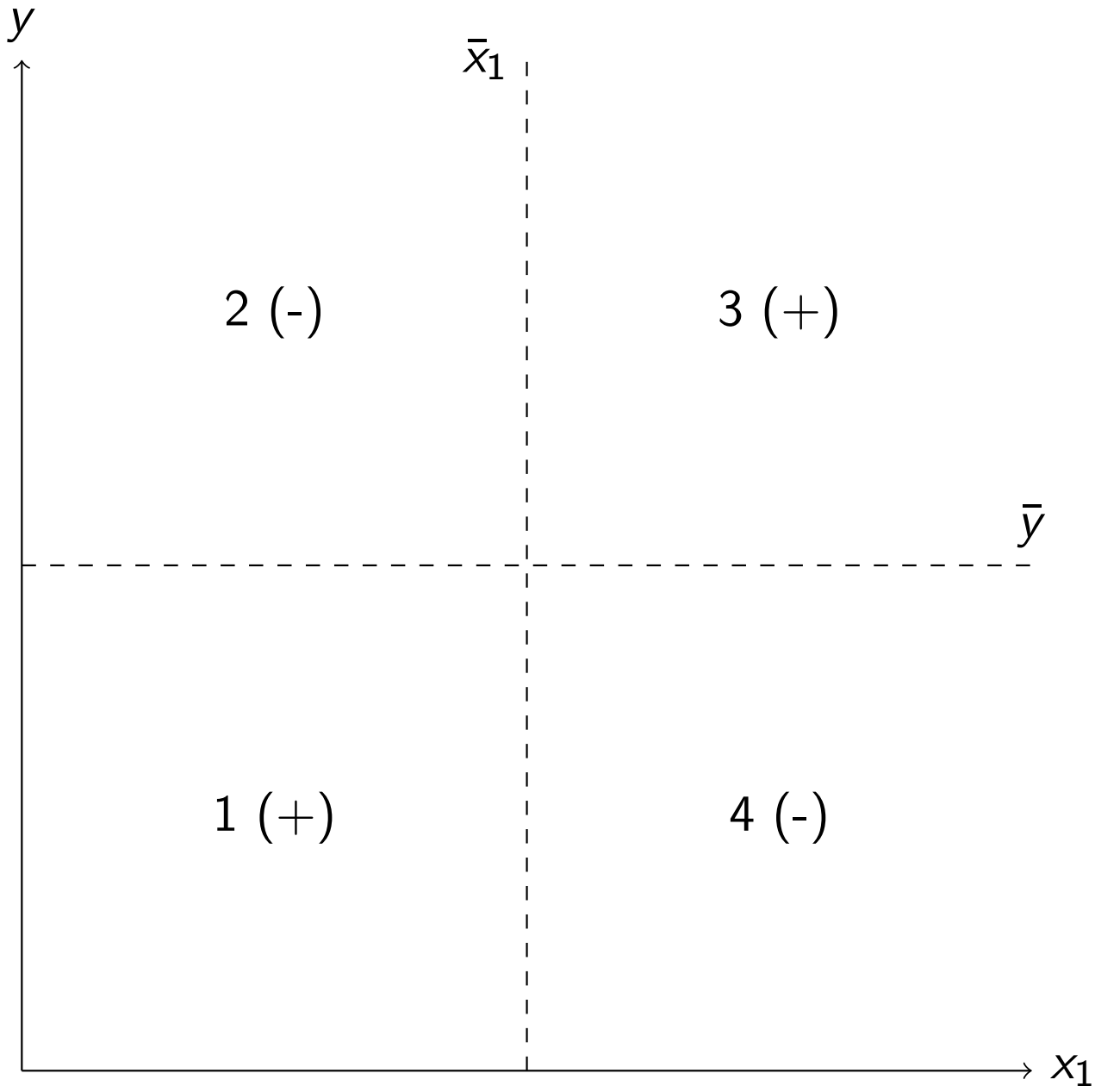
La covariance

1. covariance : « co » signifie « ensemble » et variance signifie « bouger » ou « changement ».
2. La covariance mesure à quel point deux variables statistiques « bougent ensemble ».
3. Le signe révèle la direction du changement entre les variables.
4. La magnitude reflète l'ampleur des changements.
5. Pour deux variables statistiques x_1 et y, la covariance est donnée par la formule :

$$\text{cov}(x_1, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(y_i - \bar{y}), \qquad (10)$$

6. Si $x_1 = y$: la covariance devient la variance.

Comment interpréter cette formule ?



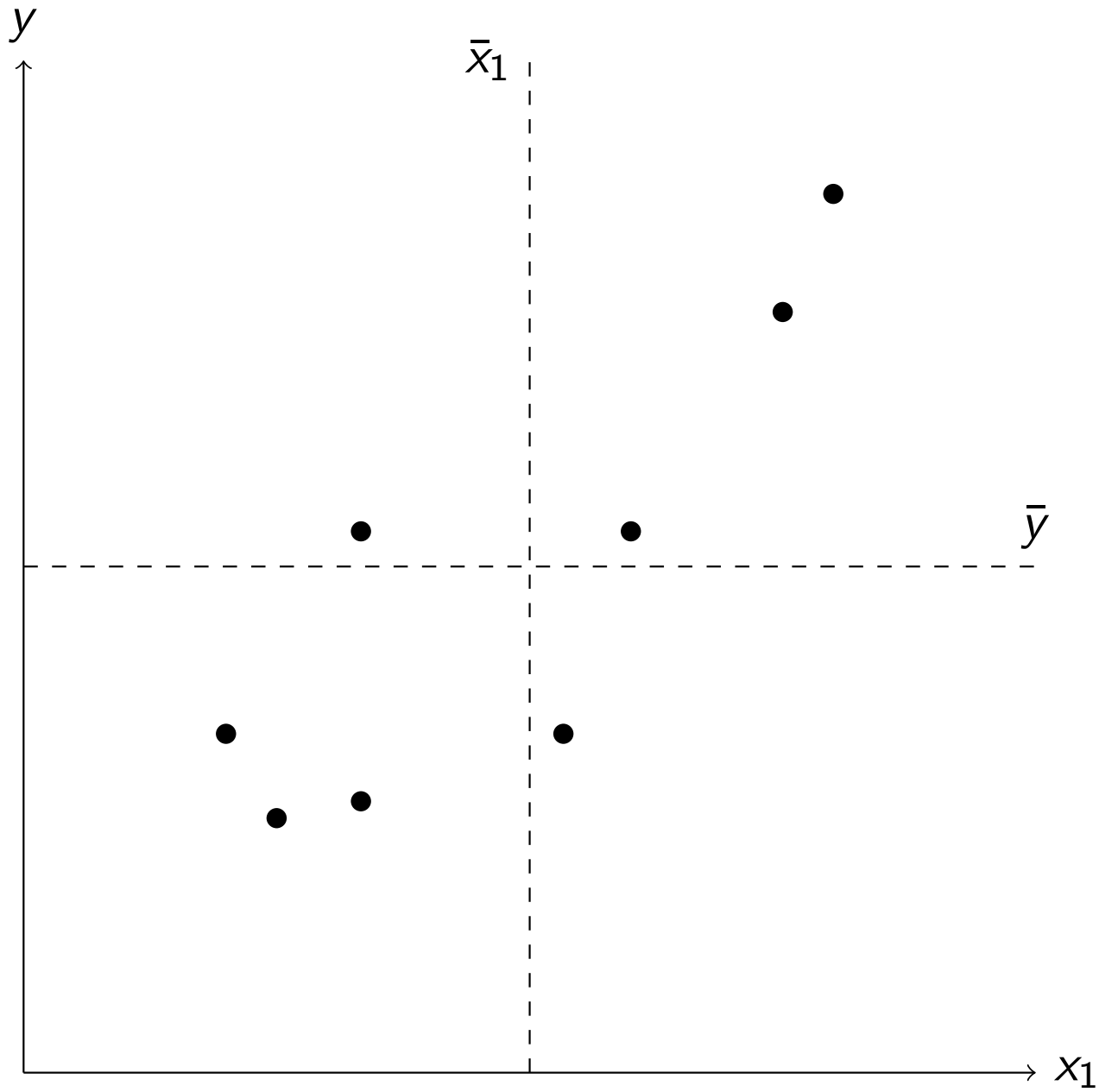


FIGURE – La covariance de ces points est positive.

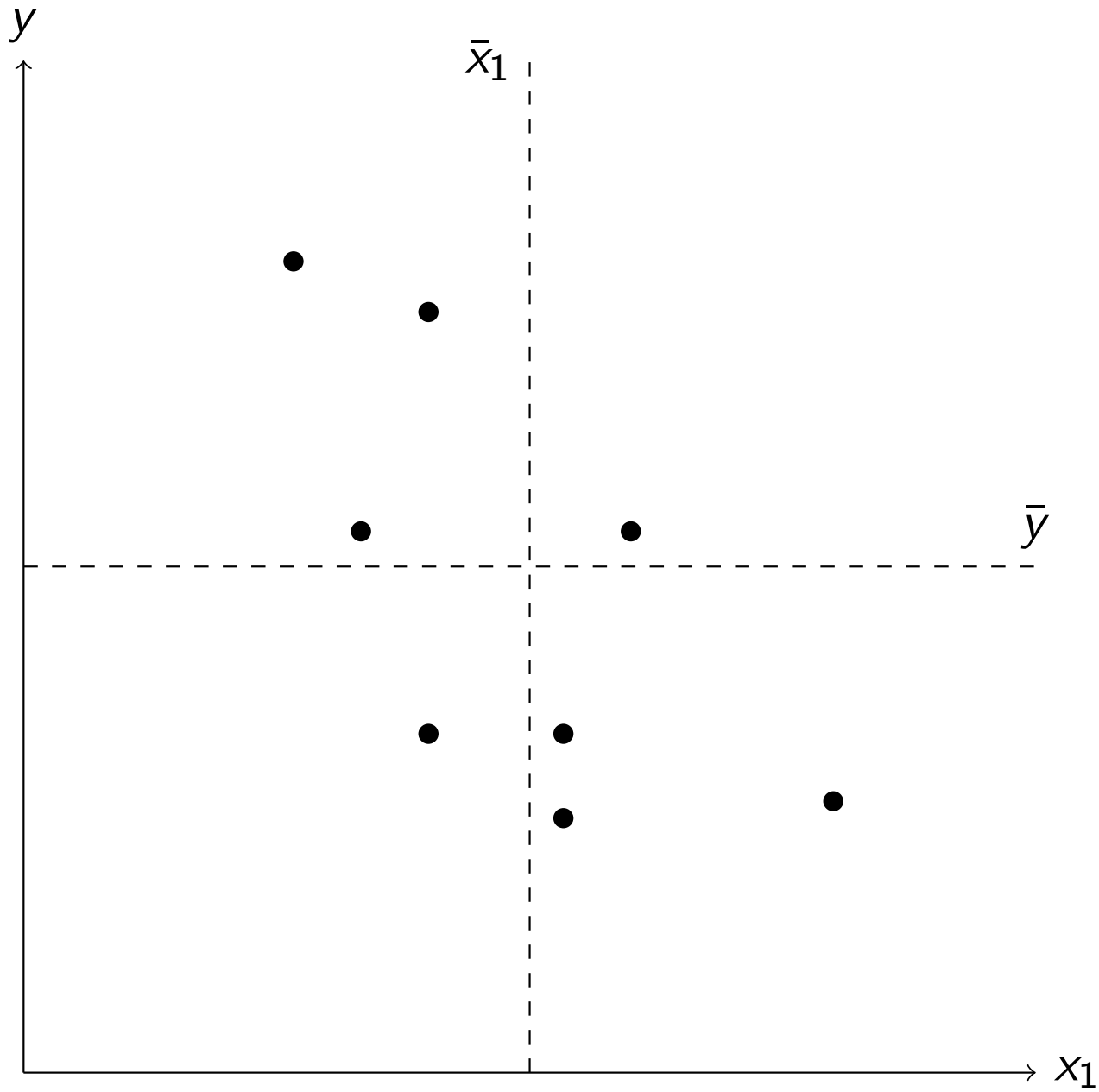


FIGURE – La covariance de ces points est négative.

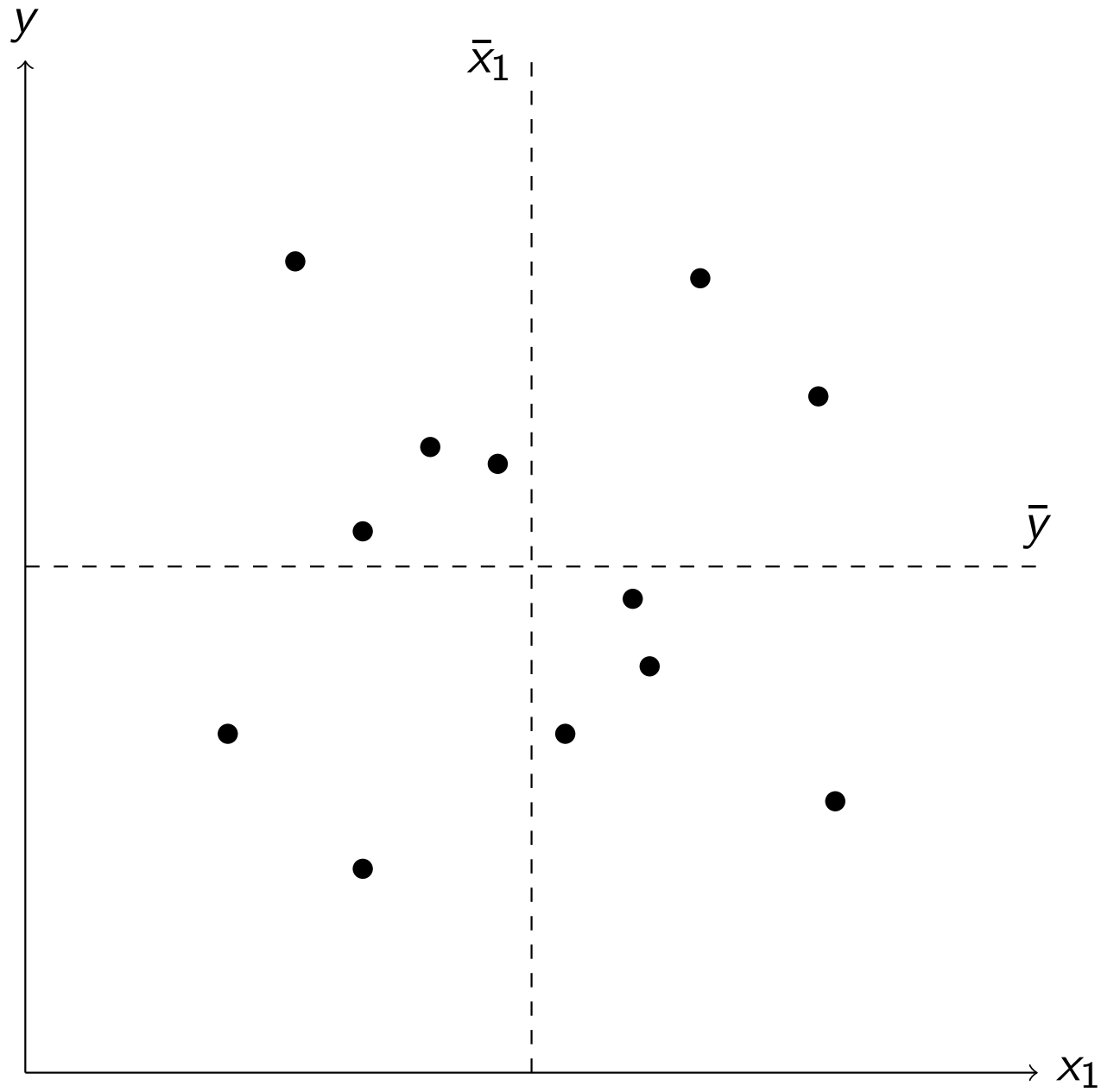


FIGURE – La covariance de ces points est proche de zéro.

Quelques propriétés de la covariance

$$\begin{aligned} \operatorname{cov}(y, x_1) &= \operatorname{cov}(x_1, y), && \text{(symmétrie)} \\ \operatorname{cov}(ax + b, x) &= a\operatorname{var}(x), && \text{(transformation linéaire)} \\ \operatorname{cov}(b, x) &= 0. && \text{(covariance avec une constante)} \end{aligned}$$

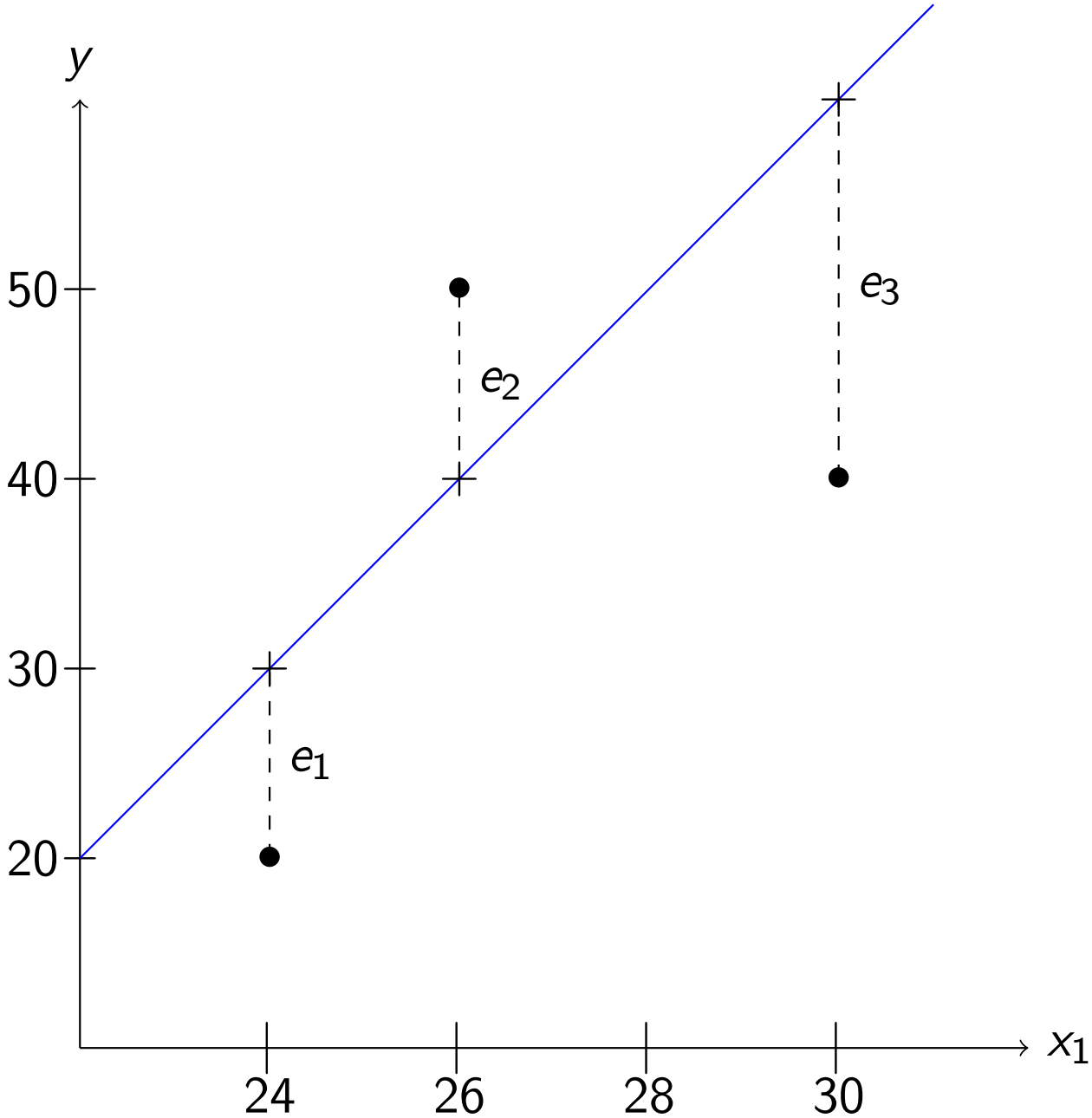
Régression à une variable

Rappel : on cherche β_0, β_1 pour :

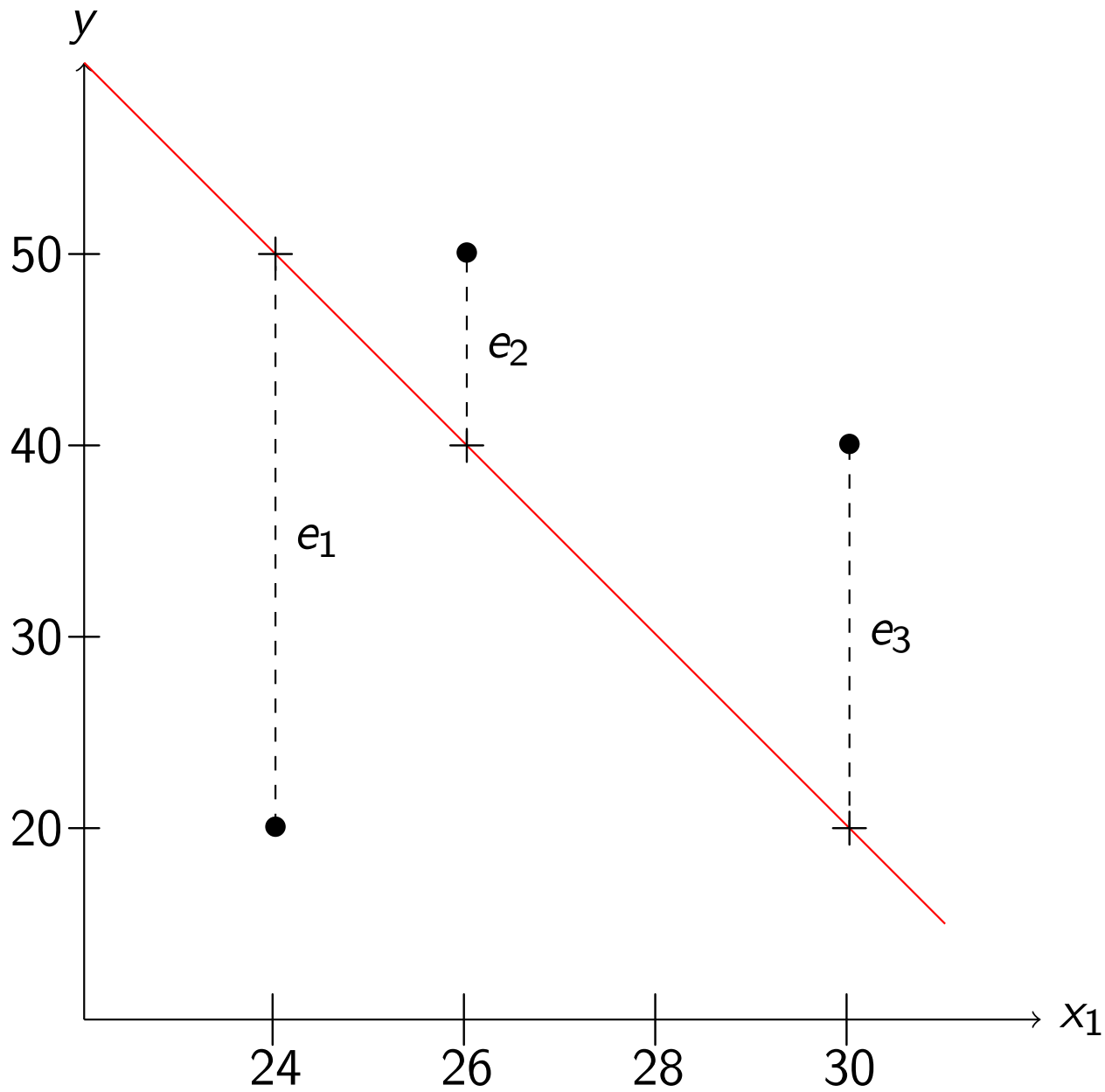
$$y_i = \beta_0 + \beta_1 x_{i1} + e_i \tag{11}$$

1. Quels sont les coefficients β_0, β_1 qui nous permettent de bien représenter les données observées ?
2. On aimerait que l'ensemble des erreurs de prédiction, soit la somme des écarts entre chaque variable observée et sa valeur prédite associée, soit la plus petite possible.

Solution possible : $\beta_1 = 1, \beta_0 = 2$



La solution possible : $\beta_1 = -1, \beta_0 = 10$



Minimisation des carrés

1. Pour mesurer adéquatement l'erreur de prévision, on emploie le *carré* des erreurs de prédiction : la distance est toujours positive.
2. Pour un échantillon de taille n, une droite de prédiction génèrera n erreurs de prédiction $e_1, e_2, \ldots, e_n$. On cherche à minimiser :
3. $$\min SSR = \min \sum_{i=1}^{n} e_i^2. \tag{12}$$
4. SSR : *sum of square residuals*
5. On peut donc remplacer le terme e_i^2 par cette expression dans la fonction SSR :

$$SSR = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2 . \tag{13}$$

Minimisation des carrés

1. L'expression pour SSR est une forme quadratique. Avec un peu d'algèbre (voir les notes complètes) :

$$SSR = \beta_1^2 \left(\sum_{i=1}^{n} x_{i1}^2 \right) + \sum_{i=1}^{n} y_i^2 - \beta_0 \left(2 \sum_{i=1}^{n} y_i \right) \cdots$$
$$\cdots - \beta_1 \left(2 \sum_{i=1}^{n} x_{i1} y_i \right) + \beta_0 \beta_1 \left(2 \sum_{i=1}^{n} x_{i1} \right) + n\beta_0^2 \tag{14}$$

2. Le minimum de la forme quadratique est donné en trouvant le sommet $(-\frac{b}{2a})$ selon chaque dimension :

$$\beta_1^* = \frac{\sum_{i=1}^{n} x_{i1} y_i - \beta_0^* \sum_{i=1}^{n} x_{i1}}{\sum_{i=1}^{n} x_{i1}^2}, \tag{15}$$

$$\beta_0^* = \frac{2 \sum_{i=1}^{n} y_i - \beta_1^* 2 \sum_{i=1}^{n} x_{i1}}{2n}. \tag{16}$$

Solution

Avec un peu d'algèbre (voir les notes complètes), on trouve que les deux coefficients β_0^* et β_1^* qui minimisent le carré des erreurs de prédictions peuvent s'écrire comme :

$$\beta_1^* = \frac{\text{cov}(x_1, y)}{\text{var}(x_1)}, \tag{17}$$

$$\beta_0^* = \bar{y} - \beta_1^* \bar{x}_1. \tag{18}$$

Les deux sont valides peu importe la taille de l'échantillon (n).

Interprétation de la solution

1. Le signe du coefficient β_1^* (pente), dépend de la covariance entre x_1 et y. Le signe de la covariance dépend de la disposition des points observés par rapport à la moyenne.
2. La magnitude de la pente dépend quant-à-elle de deux facteurs, soit l'ampleur de la covariance entre x_1 et y et l'ampleur de la variance de x_1.
3. Le coefficient β_0 représente la valeur moyenne des observations qui n'est pas explicable par la moyenne en x_1.

Rappel : température et baigneurs

TABLE – Échantillon recueilli par l'équipe de planification

Observation i	Température x_{i1}	Personnes y_i
1	24	20
2	26	50
3	30	40

Application

1. Tiré du tableau précédent (et calculs) :

$$\bar{x}_1 = \frac{80}{3} \qquad \bar{y} = \frac{110}{3}$$
$$\text{cov}(x_1, y) = \frac{70}{3} \qquad \text{var}(x_1) = \frac{84}{9}.$$

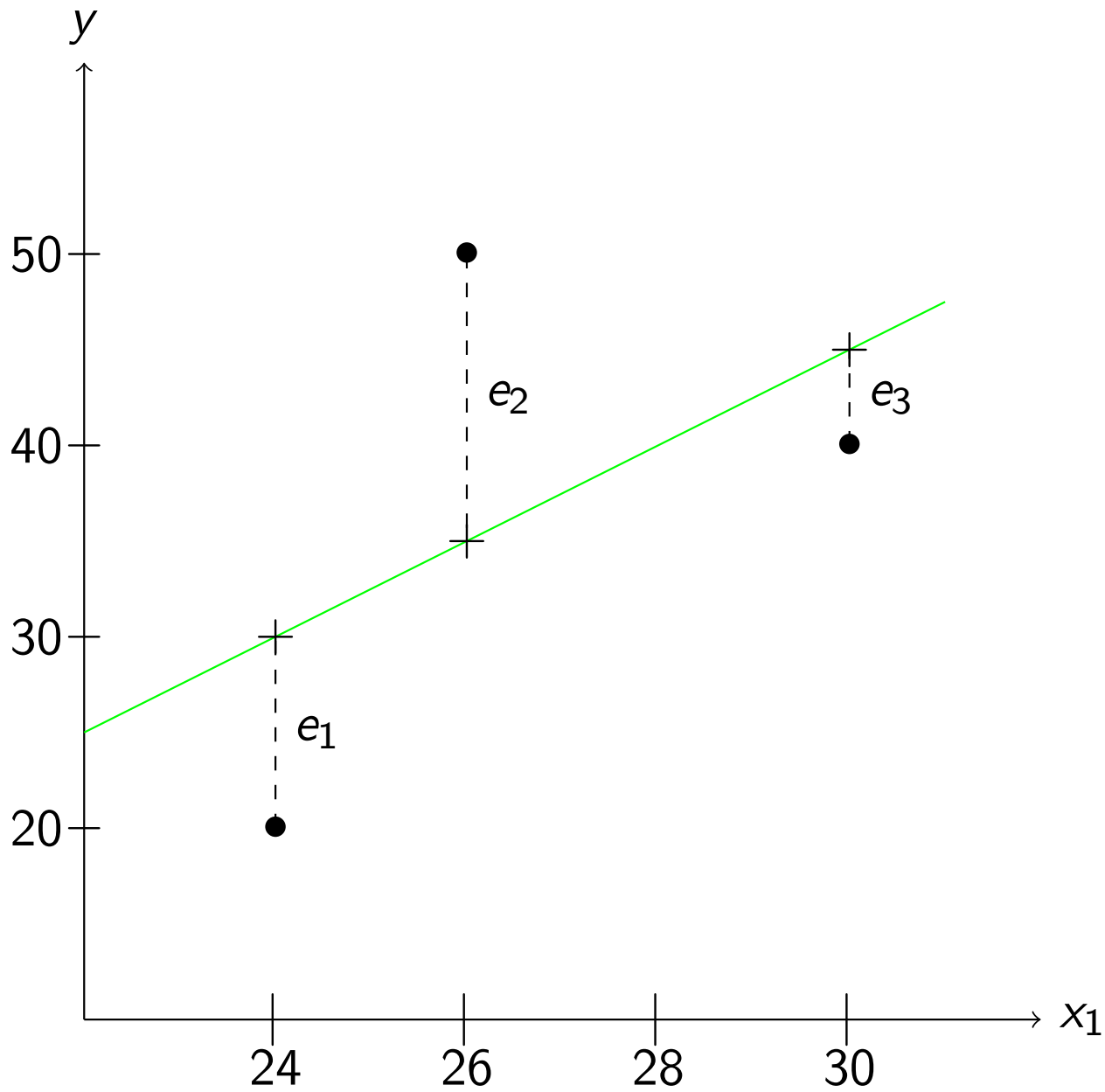
2. On peut calculer les coefficients estimés qui minimisent le carré des termes d'erreur :

$$\beta_1^* = \frac{\text{cov}(x_1, y)}{\text{var}(x_1)} = \frac{5}{2} = 2.5. \tag{19}$$

$$\beta_0^* = \bar{y} - \beta_1^* \bar{x}_1 = \frac{110}{3} - 2.5\frac{80}{3} = -30. \tag{20}$$

3. Avec cette équation, les erreurs de prédiction sont respectivement $e_1^* = -10, e_2^* = 15, e_3^* = -5$.

La solution optimale : $\beta_0^* = -30, \beta_1^* = 2.5$



Régression à trois variables

1. L'idée à plusieurs variables est la même : il faut développer le SSR en fonction du modèle de régression et trouver le minimum de la parabole multivariée
2. Avec un peu d'algèbre, on peut obtenir l'ensemble du système exprimé sous forme d'équations compréhensibles :

$$\beta_0^* = \bar{y} - \beta_1^* \bar{x}_1 - \beta_2^* \bar{x}_2, \tag{21}$$

$$\beta_1^* = \frac{\mathrm{cov}(y, x_1) - \beta_2^* \mathrm{cov}(x_1, x_2)}{\mathrm{var}(x_1)}, \tag{22}$$

$$\beta_2^* = \frac{\mathrm{cov}(y, x_2) - \beta_1^* \mathrm{cov}(x_1, x_2)}{\mathrm{var}(x_2)}. \tag{23}$$

Régression à trois variables (interprétation)

1. β_0^* : la valeur résiduelle de $\bar{y}$ une fois qu'on a soustrait le produit des coefficients avec la valeur moyenne de chaque variable.
2. β_1^* : outre la variance au dénominateur, l'ampleur du coefficient est déterminé par la covariance **qui peut seulement provenir** de la variable indépendante associée à ce coefficient.
3. β_2^* : l'interprétation est similaire.

Interprétation intuitive

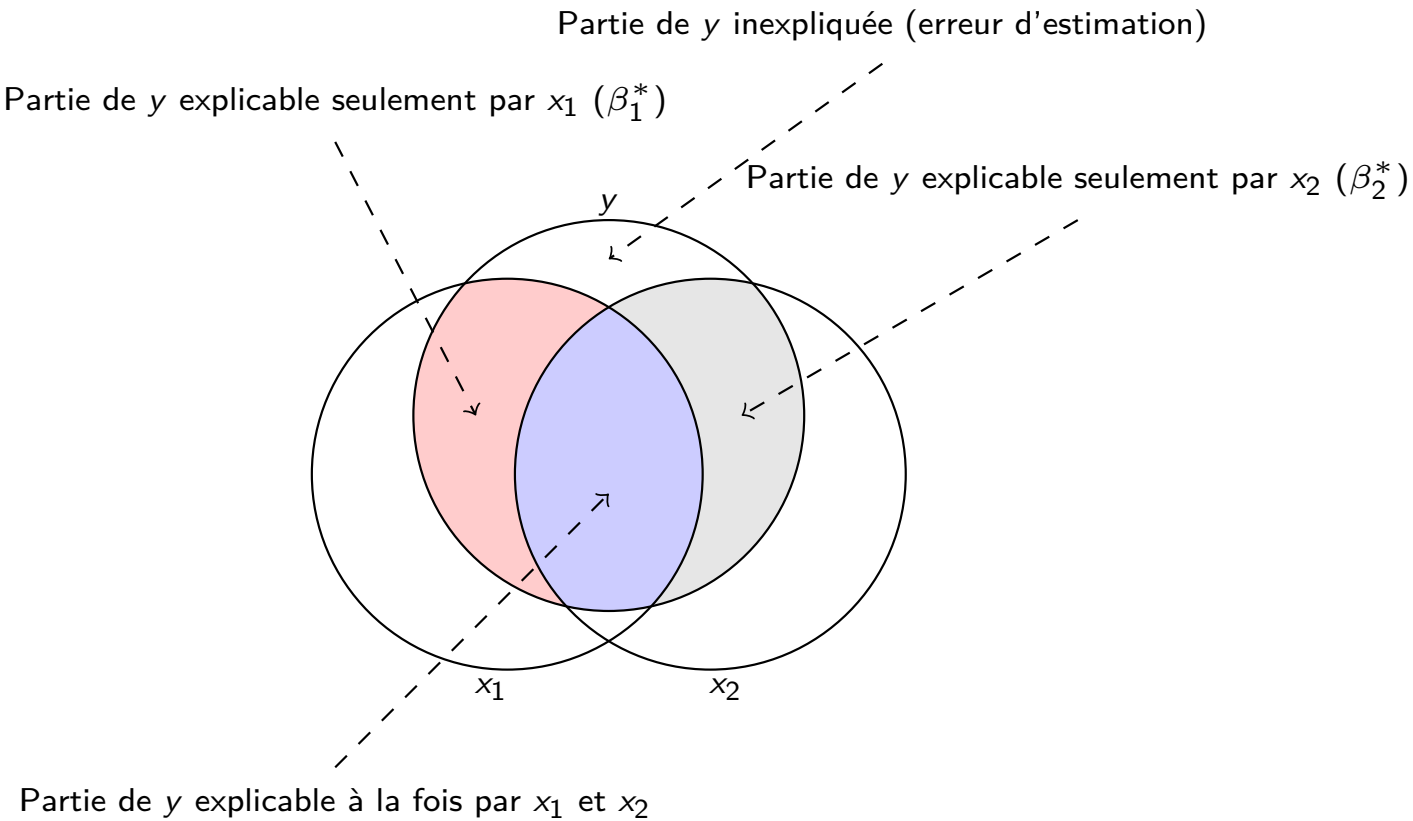


FIGURE – Explication de la formule d'un coefficient β_i^* en termes de comouvements.

Biais de variable(s) omise(s)

1. L'ajout d'une variable additionnelle x_2 change le signe... et même la valeur du coefficient β_1^*.
2. En conséquence, omettre une variable d'une régression peut introduire un **biais** aux coefficients estimés.
3. Omettre une variable du modèle peut donc affecter sévèrement l'interprétation qu'on se fait entre les variables d'un modèle.

Généralisation des formules d'estimation à $k+1$ variables

On peut s'intéresser à estimer l'effet de k variables indépendantes $x_{i1}, x_{i2}, \ldots x_{ik}$ sur une variable dépendante : y_i via un modèle linéaire :

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i, \tag{24}$$

$$= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \cdots + \beta_k x_{ik} + e_i. \tag{25}$$

Dans ce cas, les coefficients estimés auront la forme :

$$\beta_0^* = \bar{y} - \sum_{j=1}^{k} \beta_j^* \bar{x}_j, \tag{26}$$

$$\beta_l^* = \frac{\mathrm{cov}(y, x_l) - \sum_{j \neq l} \beta_j^* \mathrm{cov}(x_l, x_j)}{\mathrm{var}(x_l)} \; l = 1 \ldots k. \tag{27}$$