

Statistiques bivariées (correction)

Exercice 1 (D'après HEC ECE 2016). On s'intéresse à la fonction de production d'une entreprise qui produit un certain bien à une époque donnée. On note respectivement X et Y les quantités de travail et de capital requises pour produire une certaine quantité de ce bien, et l'on suppose que $X > 0$ et $Y > 0$.

On suppose que la production totale de l'entreprise est une variable aléatoire Q telle que $Q = BX^aY^{1-a}e^R$, où $a \in]0; 1[$, $B > 0$ et R est une variable aléatoire suivant une loi normale centrée. Enfin, on pose :

$$b = \ln(B), \quad U = \ln(X) - \ln(Y), \quad \text{et} \quad T = \ln(Q) - \ln(Y).$$

On sélectionne $n \in \mathbb{N}^*$ entreprises qui produisent le bien considéré à l'époque donnée. On mesure pour chaque entreprise $i \in \llbracket 1; n \rrbracket$ la quantité de travail x_i et la quantité de capital y_i utilisées ainsi que la quantité produite q_i . On suppose que, pour tout $i \in \llbracket 1; n \rrbracket$, $x_i > 0$, $y_i > 0$ et $q_i > 0$.

Pour tout $i \in \llbracket 1; n \rrbracket$, x_i , y_i et q_i sont des réalisations de variables aléatoires X_i , Y_i , Q_i ayant respectivement les mêmes lois que X , Y , et Q . On a $Q_i = BX_i^aY_i^{1-a}\exp(R_i)$ et $q_i = Bx_i^ay_i^{1-a}\exp(r_i)$. Ici r_1, \dots, r_n sont des réalisations de R_1, R_2, \dots, R_n qui sont des variables aléatoires supposées indépendantes et de même loi que R . On pose, pour tout $i \in \llbracket 1; n \rrbracket$:

$$U_i = \ln X_i - \ln Y_i, \quad T_i = \ln Q_i - \ln Y_i \quad \text{et} \quad t_i = \ln q_i - \ln y_i$$

Ainsi, pour chaque entreprise $i \in \llbracket 1; n \rrbracket$, t_i est une réalisation de la variable aléatoire T_i .

On a relevé pour $n = 16$ entreprises qui produisent le bien considéré à l'époque donnée, les deux séries statistiques $(u_i)_{1 \leq i \leq n}$ et $(t_i)_{1 \leq i \leq n}$ implémentées dans Python par les listes

```
u=[1.06,0.44,2.25,3.88,0.61,1.97,3.43,2.10,1.50,1.68,2.72,1.35,2.94,2.78,3.43,3.58]
```

```
t=[2.58,2.25,2.90,3.36,2.41,2.79,3.32,2.81,2.62,2.70,3.17,2.65,3.07,3.13,3.07,3.34]
```

- 1) Vérifier que $T = aU + b + R$ et, pour tout $i \in \llbracket 1; n \rrbracket$, $t_i = au_i + b + r_i$.
- 2) Représenter sur un même graphique :
 - le nuage des points $(u_1, t_1), (u_2, t_2), \dots, (u_n, t_n)$.
 - la droite de régression de T en U .
 - la droite de régression de U en T .
- 3) Interpréter le point d'intersection des deux droites de régression.
- 4) Estimer graphiquement les moyennes empiriques \bar{u} et \bar{t} .

Correction :

1) On a $\ln(Q) = \ln(BX^aY^{1-a}e^R) = \ln(B) + a\ln(X) + (1-a)\ln(Y) + R$. Donc

$$\ln(Q) - \ln(Y) = a(\ln(X) - \ln(Y)) + \ln(B) + R$$


et donc $T = aU + b + R$.

Pour tout $i \in \llbracket 1; n \rrbracket$, $\ln(q_i) = \ln(Bx_i^ay_i^{1-a}e^{r_i}) = \ln(B) + a\ln(x_i) + (1-a)\ln(y_i) + r_i$. Donc

$$\ln(q_i) - \ln(y_i) = a(\ln(x_i) - \ln(y_i)) + \ln(B) + r_i$$

et donc $t_i = au_i + b + r_i$.

2) Si on note $u = (u_1, \dots, u_n)$ et $t = (t_1, \dots, t_n)$, alors

- l'équation de la droite de régression de T en U est $y = a_1x + b_1$ où $a_1 = \frac{\text{Cov}(u, t)}{\sigma_u^2}$ et $b_1 = \bar{t} - a_1\bar{u}$.
- l'équation de la droite de régression de T en U est $x = a_2y + b_2$ où $a_2 = \frac{\text{Cov}(u, t)}{\sigma_t^2}$ et $b_2 = \bar{u} - a_2\bar{t}$. Donc il s'agit de la droite d'équation $y = \frac{x - b_2}{a_2}$.  Attention à cette subtilité!!!

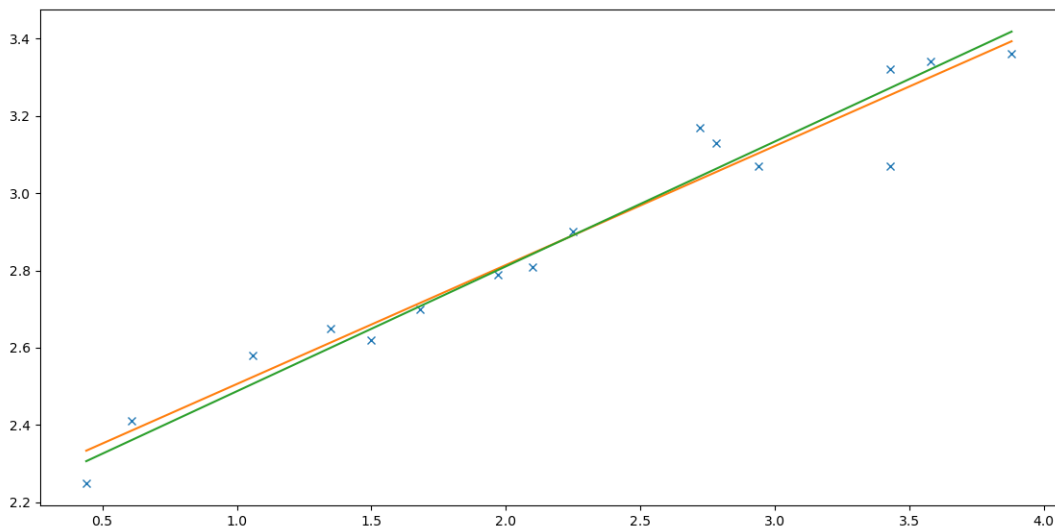
Voici le code Python :

```

1 #Les importations
2 import numpy as np
3 import matplotlib.pyplot as plt
4
5 #Tracé du nuage de points
6 u=[1.06,0.44,2.25,3.88,0.61,1.97,3.43,2.10,1.50,1.68,
7     2.72,1.35,2.94,2.78,3.43,3.58]
8 t=[2.58,2.25,2.90,3.36,2.41,2.79,3.32,2.81,2.62,2.70,
9     3.17,2.65,3.07,3.13,3.07,3.34]
10 plt.plot(u,t,'x')
11
12 #Calcul des moyennes et variances
13 moy_u=np.mean(u); moy_t=np.mean(t)
14 var_u=np.var(u); var_t=np.var(t)
15
16 #Calcul de la covariance
17 S=0
18 for k in range(len(u)):
19     S=S+u[k]*t[k]
20 cov=S/len(u)-moy_u*moy_t
21
22 #Calcul de a1,b1,a2,b2
23 a1=cov/var_u; b1=moy_t-a1*moy_u
24 a2=cov/var_t; b2=moy_u-a2*moy_t
25
26 #Tracé des courbes
27 m=np.min(u); M=np.max(u)
28 plt.plot([m,M],[a1*m+b1,a1*M+b1])
29 plt.plot([m,M],[(m-b2)/a2,(M-b2)/a2])
30
31 plt.show()

```

On obtient :



3) Un point (x, y) est intersection des deux droites si et seulement si $y = a_1x + b_1$ et $x = a_2y + b_2$ si et seulement si $y = a_1x + b_1 = a_1(x - \bar{u}) + \bar{t}$ et $x = a_2y + b_2 = a_2(y - \bar{t}) + \bar{u}$ si et seulement si $y - \bar{t} = a_1(x - \bar{u})$ et $x - \bar{u} = a_2(y - \bar{t})$ si et seulement si $y - \bar{t} = a_1a_2(y - \bar{t})$ et $x - \bar{u} = a_2(y - \bar{t})$.

- Si $a_1 a_2 = 1$, alors il y a une infinité de solutions. Cela correspond au cas où les deux droites sont confondues. C'est le cas où le coefficient de corrélation linéaire est égal à 1 (on constate que $a_1 a_2$ est le carré du coefficient de corrélation).
 - Si $a_1 a_2 \neq 1$, alors (x, y) est intersection si et seulement si $y - \bar{t} = 0$ et $x - \bar{u} = 0$ si et seulement si $(x, y) = (\bar{u}, \bar{t})$. C'est le point moyen.
- 4) Ici, les deux droites ont un unique point d'intersection. Ainsi \bar{u} et \bar{t} sont respectivement l'abscisse et l'ordonnée de ce point d'intersection.

Exercice 2 (Développement de bactéries). On étudie dans cet exercice le nombre de bactéries présentes dans un bouillon de culture au fur et à mesure du temps.

On dispose de la matrice de données suivante :

```
donnees=np.array([[0,1,2,3,4,5,6],[32,47,65,92,132,190,275]])
```

Cette matrice contient deux lignes, la seconde indiquant le nombre de bactéries (par unité de volume) présentes dans le bouillon de culture au cours de l'expérience : pour tout $i \in \llbracket 0;6 \rrbracket$, $x_i = \text{donnees}[0,i] = i$ et $y_i = \text{donnees}[1,i]$ est le nombre de bactéries par unité de volume présentes dans le bouillon de culture après i heures.

- 1) Pour ce relevé statistique, quelle est la variable explicative X ? Quelle est la variable à expliquer Y ?
- 2) Représenter le nuage de points des données.
- 3) Avec Python calculer les moyennes \bar{x} et \bar{y} , les variances σ_x^2 et σ_y^2 des séries statistiques X et Y , ainsi la covariance et le coefficient de corrélation linéaire r de la série statistique.
- 4) Superposer au nuage de points la droite de régression de y en x .
- 5) Le nuage de points de la série statistique peut laisser penser que Y est une fonction « exponentielle de X » : il semble qu'il existe α et λ deux réels strictement positifs tels que $Y \approx \lambda \cdot e^{\alpha X}$. Remarquons que

$$Y \approx \lambda \cdot e^{\alpha X} \quad \Longleftrightarrow \quad \ln(Y) \approx \alpha X + \ln \lambda.$$

Ainsi, il semble que la « nouvelle » variable à expliquer $\ln(Y)$ est une fonction « presque » affine de X . Nous allons alors étudier la série statistique bivariée de variable explicative X et de variable à expliquer $Z = \ln(Y)$, et nous allons déterminer la droite de régression linéaire de cette série.

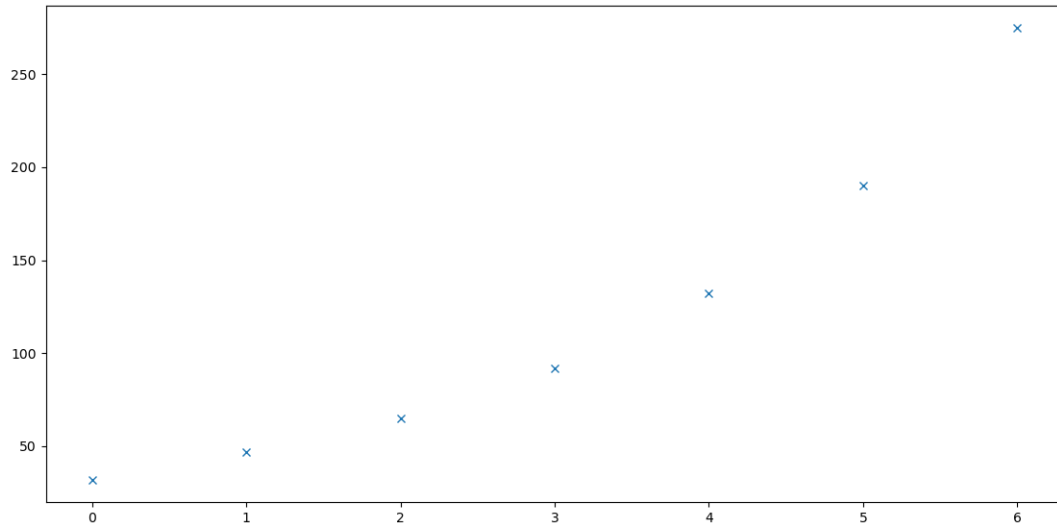
- a) Définir en Python un vecteur Z contenant $(z_0, \dots, z_6) = (\ln(y_0), \dots, \ln(y_6))$.
- b) Déterminer l'équation de la droite de régression linéaire de Z en X , notée $z = \alpha x + \beta$.
- c) Représenter alors le nuage de points de la série statistique $((x_0, z_0), (x_1, z_1), \dots, (x_6, z_6))$ et lui superpose la droite en question. Commenter.
- d) Superposer à la courbe de la question 1, la courbe d'équation $y = e^{\beta} e^{\alpha x}$. Commenter.

Correction :

- 1) Ici X est le nombre de secondes écoulées et Y le nombre de bactéries présentes après X secondes.

```
2)
1 #Les importations
2 import numpy as np
3 import matplotlib.pyplot as plt
4
5 #On extrait les données
6 donnees=np.array([[0,1,2,3,4,5,6],[32,47,65,92,132,190,275]])
7 x=donnees[0,:]
8 y=donnees[1,:]
9
10 #On trace le nuage de points
11 plt.plot(x,y,'x')
12 plt.show()
```

On obtient :



```

3) 1 #Calcul des moyennes et variances
    2 moy_x=np.mean(x); moy_y=np.mean(y)
    3 var_x=np.var(x); var_y=np.var(y)
    4
    5 #Calcul de la covariance
    6 S=0
    7 for k in range(len(x)):
    8     S=S+x[k]*y[k]
    9 cov=S/len(x)-moy_x*moy_y
    10
    11 #Calcul du coefficient de corrélation
    12 cor=cov/np.sqrt(var_x*var_y)

```

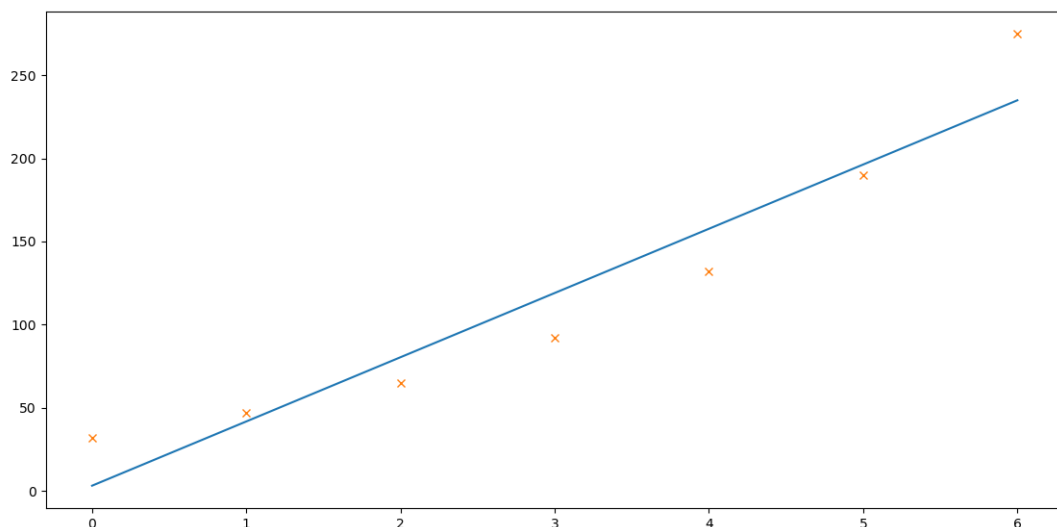
On obtient un coefficient de corrélation égal environ à 0,954. Cela pourrait laisser penser à une relation linéaire.

```

4) 1 #Calcul des coefficients a et b de la droite de régression
    2 a=cov/var_x
    3 b=moy_y-a*moy_x
    4 m=np.min(x); M=np.max(x)
    5 plt.plot([m,M],[a*m+b,a*M+b])
    6 plt.plot(x,y,'x')
    7 plt.show()

```

On obtient :



Bien que le coefficient de corrélation soit très proche de 1, la relation ne semble pas linéaire.

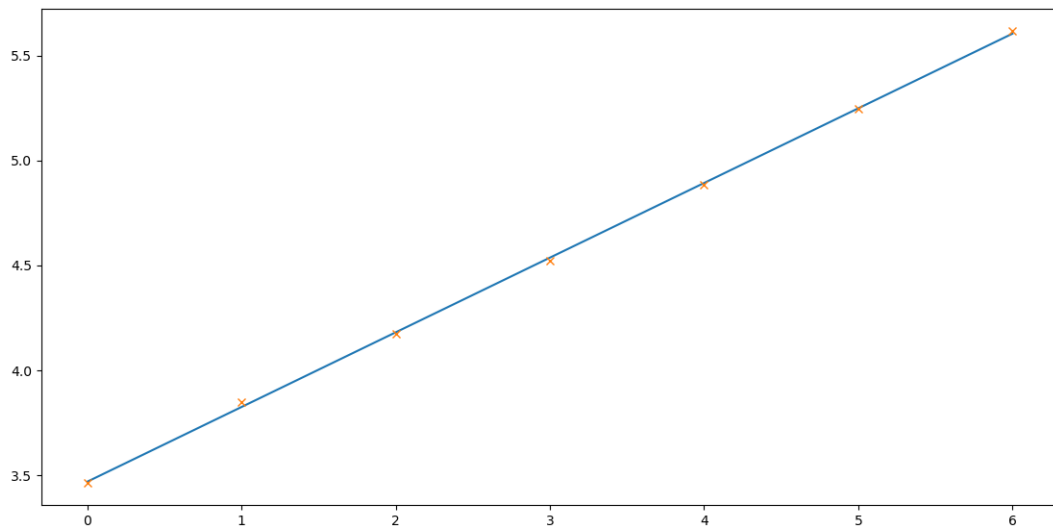
5)

```

1 #On définit z
2 z=np.log(y)
3
4 #Calcul de la nouvelle covariance
5 moy_z=np.mean(z)
6 S=0
7 for k in range(len(x)):
8     S=S+x[k]*z[k]
9 cov_xz=S/len(x)-moy_x*moy_z
10
11 #On définit alpha et beta
12 alpha=cov_xz/var_x
13 beta=moy_z-alpha*moy_x
14
15 #On trace le nuage de point des (x_i,z_i) et la droite
16 plt.plot([m,M],[alpha*m+beta,alpha*M+beta])
17 plt.plot(x,z,'x')
18 plt.show()

```

On obtient :

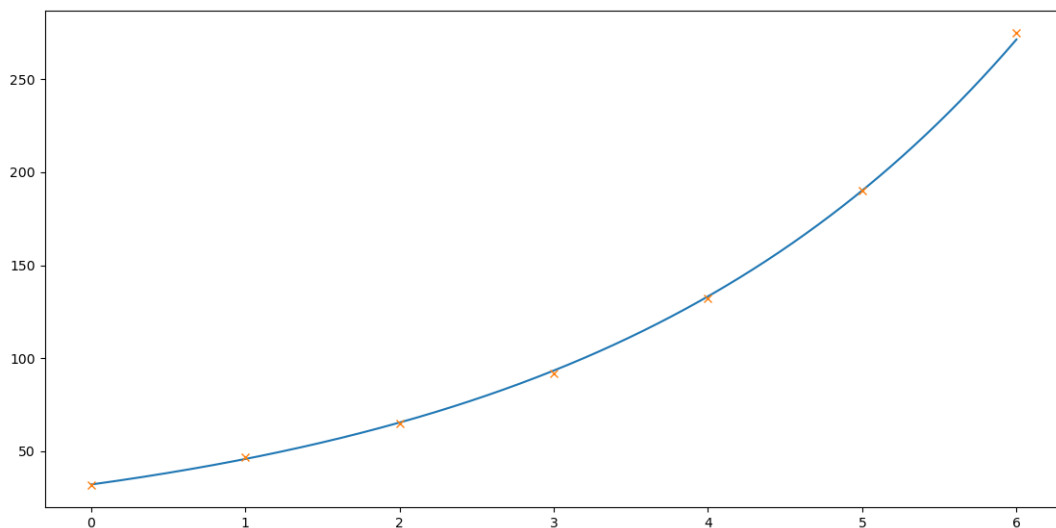


```

1 #On trace le nuage de point des (x_i,y_i) et la courbe
2 abs=np.linspace(m,M,1000)
3 ord=[np.exp(beta+alpha*x) for x in abs]
4 plt.plot(abs,ord)
5 plt.plot(x,y,'x')
6 plt.show()

```

On obtient :



Exercice 3 (Corrélation ne veut pas dire causalité).

Une bonne corrélation entre deux séries de données ne signifie pas pour autant qu'il existe un lien de cause à effet entre les deux. Voici deux exemples :

- 1) Ce tableau donne la consommation de margarine (en livre par personne) aux USA, et la taux de divorce dans l'état du Maine (en divorce pour 1000 personnes).

Année	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Taux de divorce	8,2	7	6,5	5,3	5,2	4	4,6	4,5	4,2	3,7
Consommation de margarine	5	4,7	4,6	4,4	4,3	4,1	4,2	4,2	4,2	4,1

Calculer le coefficient de corrélation. Commenter.

- 2) Ce tableau donne la consommation de mozzarella (en livre par personne) et le nombre de doctorats en génie civil décernés aux USA.

Année	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Consommation de mozzarella	9,3	9,7	9,7	9,7	9,9	10,2	10,5	11	10,6	10,6
Nombre de doctorats	480	501	540	552	547	622	655	701	712	708

Calculer le coefficient de corrélation. Commenter.

Correction :

```
1) 1 import numpy as np
    2 X=[8.2,7,6.5,5.3,5.2,4,4.6,4.5,4.2,3.7]
    3 Y=[5,4.7,4.6,4.4,4.3,4.1,4.2,4.2,4.2,4.1]
    4 M=np.cov(X,Y)
    5 print(M[0,1]/(M[0,0]*M[1,1])**1/2)
```

On obtient 0.992. C'est très proche de 1 : il y a une forte corrélation.

```
2) 1 import numpy as np
    2 X=[9.3,9.7,9.7,9.7,9.9,10.2,10.5,11,10.6,10.6]
    3 Y=[480,501,540,552,547,622,655,701,712,708]
    4 M=np.cov(X,Y)
    5 print(M[0,1]/(M[0,0]*M[1,1])**1/2)
```

On obtient 0.958. C'est très proche de 1 : il y a une forte corrélation.

Exercice 4 (Autre démonstration de la droite de régression). Soient x_1, \dots, x_n des réels qui ne sont pas tous égaux. Soient y_1, \dots, y_n des réels. On note

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, \quad \bar{y} = \frac{1}{n} \sum_{k=1}^n y_k, \quad \overline{x^2} = \frac{1}{n} \sum_{k=1}^n x_k^2, \quad \overline{xy} = \frac{1}{n} \sum_{k=1}^n x_k y_k,$$

$$\sigma_x^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \overline{x^2} - \bar{x}^2, \quad \text{et} \quad \text{Cov}(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) = \overline{xy} - \bar{x}\bar{y}.$$

- 1) Montrer que $\sigma_x^2 > 0$.

- 2) Montrer que la fonction

$$F : (a, b) \in \mathbb{R}^2 \mapsto \sum_{i=1}^n (y_i - ax_i - b)^2$$

est de classe C^1 sur \mathbb{R}^2 et admet pour unique point critique $(\hat{a}, \hat{b}) = \left(\frac{\text{Cov}(x, y)}{\sigma_x^2}, \bar{y} - \frac{\text{Cov}(x, y)}{\sigma_x^2} \bar{x} \right)$.

- 3) Montrer que, pour tout $(a, b) \in \mathbb{R}^2$,

$$F(a + \hat{a}, b + \hat{b}) - F(\hat{a}, \hat{b}) = a^2 \sigma_x^2 + (a\bar{x} + b)^2.$$

- 4) En déduire que F admet un unique minimum en (\hat{a}, \hat{b}) .

Correction :

- 1) On a $\sigma_x^2 = 0$ si et seulement si, pour tout $i \in \llbracket 1; n \rrbracket$, $x_i = \bar{x}$ si et seulement si tous les x_i , $1 \leq i \leq n$ sont tous égaux. Ici ce n'est pas le cas par hypothèse donc $\sigma_x^2 \neq 0$ et donc $\sigma_x^2 > 0$.
- 2) Pour tout $(a, b) \in \mathbb{R}^2$,

$$\begin{aligned} F(a, b) &= \sum_{i=1}^n (y_i^2 + a^2 x_i^2 + b^2 - 2ax_i y_i - 2by_i + 2abx_i) \\ &= n(\bar{y}^2 + a^2 \bar{x}^2 + b^2 - 2a\bar{x}\bar{y} - 2b\bar{y} + 2ab\bar{x}). \end{aligned}$$

La fonction F est polynomiale donc elle est de classe C^1 sur \mathbb{R}^2 . Pour tout $(a, b) \in \mathbb{R}^2$,

$$\nabla F(a, b) = \left(n(2a\bar{x}^2 - 2\bar{x}\bar{y} + 2b\bar{x}), n(2b - 2\bar{y} + 2a\bar{x}) \right).$$

On a

$$\begin{aligned} (a, b) \text{ est critique} &\iff \begin{cases} n(2a\bar{x}^2 - 2\bar{x}\bar{y} + 2b\bar{x}) = 0 \\ n(2b - 2\bar{y} + 2a\bar{x}) = 0 \end{cases} \\ &\iff \begin{cases} a\bar{x}^2 - \bar{x}\bar{y} + b\bar{x} = 0 \\ b = \bar{y} - a\bar{x} \end{cases} \\ &\iff \begin{cases} b = \bar{y} - a\bar{x} \\ a\bar{x}^2 - \bar{x}\bar{y} + (\bar{y} - a\bar{x})\bar{x} = 0 \end{cases} \\ &\iff \begin{cases} b = \bar{y} - a\bar{x} \\ a(\bar{x}^2 - \bar{x}^2) - \bar{x}\bar{y} + \bar{x}\bar{y} = 0 \end{cases} \\ &\iff \begin{cases} b = \bar{y} - a\bar{x} \\ a\sigma_x^2 = \text{Cov}(x, y) \end{cases} \\ &\iff \begin{cases} a = \frac{\text{Cov}(x, y)}{\sigma_x^2} \\ b = \bar{y} - \frac{\text{Cov}(x, y)}{\sigma_x^2} \bar{x} \end{cases} \end{aligned}$$

Ainsi (\hat{a}, \hat{b}) est l'unique point critique.

- 3) Soit $(a, b) \in \mathbb{R}^2$. On a

$$\begin{aligned} F(a + \hat{a}, b + \hat{b}) - F(\hat{a}, \hat{b}) &= n(\bar{y}^2 + (a + \hat{a})^2 \bar{x}^2 + (b + \hat{b})^2 - 2(a + \hat{a})\bar{x}\bar{y} - 2(b + \hat{b})\bar{y} + 2(a + \hat{a})(b + \hat{b})\bar{x}) \\ &\quad - n(\bar{y}^2 + \hat{a}^2 \bar{x}^2 + \hat{b}^2 - 2\hat{a}\bar{x}\bar{y} - 2\hat{b}\bar{y} + 2\hat{a}\hat{b}\bar{x}) \\ &= n \left((a^2 + 2a\hat{a})\bar{x}^2 + (b^2 + 2b\hat{b}) - 2a\bar{x}\bar{y} - 2b\bar{y} + 2(ab + a\hat{b} + \hat{a}b)\bar{x} \right) \\ &= n \left(a^2 \bar{x}^2 + 2a(\hat{a}\bar{x}^2 - \bar{x}\bar{y} + \hat{b}\bar{x}) + b^2 + 2b(\hat{b} - \bar{y} + \hat{a}\bar{x}) + 2ab\bar{x} \right) \end{aligned}$$

On a $\hat{b} = \bar{y} - \hat{a}\bar{x}$ donc le coefficient devant b ci-dessus est nul. Ensuite

$$\hat{a}\bar{x}^2 - \bar{x}\bar{y} + \hat{b}\bar{x} = \hat{a}\bar{x}^2 - \bar{x}\bar{y} + \bar{x}\bar{y} - \hat{a}\bar{x}^2 = \hat{a}\sigma_x^2 - \text{Cov}(x, y) = 0$$

donc le coefficient devant b ci-dessus est nul aussi. On en déduit que

$$F(a + \hat{a}, b + \hat{b}) - F(\hat{a}, \hat{b}) = a^2 \bar{x}^2 + b^2 + 2ab\bar{x} = a^2 \sigma_x^2 + a^2 \bar{x}^2 + b^2 + 2ab\bar{x} = a^2 \sigma_x^2 + (a\bar{x} + b)^2.$$

- 4) On a donc, pour tout $(a, b) \in \mathbb{R}^2$,

$$F(a + \hat{a}, b + \hat{b}) - F(\hat{a}, \hat{b}) = a^2 \bar{x}^2 + b^2 + 2ab\bar{x} = a^2 \sigma_x^2 + a^2 \bar{x}^2 + b^2 + 2ab\bar{x} = a^2 \sigma_x^2 + (a\bar{x} + b)^2 \geq 0.$$

Ainsi F admet un minimum global en (\hat{a}, \hat{b}) . Pour l'unicité deux méthodes :

Première méthode. S'il y en avait un autre point en lequel F est minimale sur \mathbb{R}^2 , ce serait un point critique. Or (\hat{a}, \hat{b}) est le seul point critique.

Deuxième méthode. La fonction F admet un minimum en un point (a, b) si et seulement si $F(a, b) = F(\hat{a}, \hat{b})$ si et seulement si $F(h + \hat{a}, k + \hat{b}) = F(\hat{a}, \hat{b})$ avec $h = a - \hat{a}$ et $k = b - \hat{b}$. Le calcul de la question précédente entraîne que c'est le cas si et seulement si $h^2 \sigma_x^2 + (h\bar{x} + k)^2 = 0$ si et seulement si $h = 0$ (car $\sigma_x^2 \neq 0$) et $h\bar{x} + k = 0$ si et seulement si $h = k = 0$ si et seulement si $a = \hat{a}$ et $b = \hat{b}$. D'où l'unicité.