Statistiques bivariées (HP)

I. Statistiques descriptives

I.1. Statistiques descriptives univariées

Commençons par un exemple afin de fixer les idées. Pour des raisons de santé publique, on s'intéresse à la concentration d'ozone O_3 dans l'air (en microgrammes par m^3) à Lyon. Pour cela on collecte les taux moyens d'ozone chaque jour. On obtient les données suivantes (données ATMO) pour le 1er de chaque mois en 2018 :

Mois	1	2	3	4	5	6	7	8	9	10	11	12
O_3 ($\mu g.m^{-3}$)	62,1	45,5	44,1	85,9	73	64,3	103,2	98,6	79,9	64,3	60,7	25,8

Pour étudier ces données avec **Scilab**, on peut choisir de les stocker dans une variable `y` avec la commande suivante :

```
y = [62.1, 45.5, 44.1, 85.9, 73, 64.3, 103.2, 98.6, 79.9, 64.3, 60.7, 25.8]
```

On notera dans la suite $y = (y_1, \dots, y_n)$ ces observations (ici $n = 12$).
Avant de traiter ces dernières, on commence par les décrire. Pour cela, on peut utiliser les mesures suivantes :

- la moyenne empirique : $\bar{y}_n = \dfrac{1}{n}\sum_{i=1}^{n} y_i$.

  Comme son nom l'indique, la moyenne empirique est la moyenne des observations. Elle s'obtient en **Scilab** avec la commande `mean`. Pour notre exemple, on entre l'instruction :

  ```
  mean(y)
  ```

  On obtient alors la sortie :

  ```
  ans  =
      67.283333
  ```

  Au cours de l'année 2018, la concentration en ozone à Lyon a donc été en moyenne de $67,3$ $\mu g.m^{-3}$.

Remarque
On peut établir une correspondance entre l'espérance et la moyenne empirique. On fait ainsi le lien entre un objet du « monde des variables aléatoires » et un objet du « monde des données » :

$$\mathbb{E}(Y) \quad \leftrightarrow \quad \frac{1}{n}\sum_{i=1}^{n} y_i$$

- la variance empirique : $s_y^2 = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \dfrac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y}_n)^2$.

Cette mesure permet de quantifier la dispersion des observations autour de la moyenne. Plus la variance est élevée, plus les observations comprennent des valeurs très éloignées de la moyenne. Pour notre exemple, on entre l'instruction :

```
variance(y)
```

On obtient alors la sortie :

```
ans  =
    511.82152
```

Cette valeur n'est pas facile à interpréter car son unité de mesure (le $(\mu g)^2.m^{-6}$) n'est pas la même que celles de nos données (ici le $\mu g.m^{-3}$). Intuitivement, c'est un peu comme si l'on comparait une surface (en m^2) à une longueur (en m) : cela n'a pas de sens. C'est pourquoi, pour l'interprétation, on lui privilégie la mesure suivante.

Remarque

× Notons que la définition de la variance empirique s_y^2 n'est pas sortie du chapeau. Elle est en effet à rapprocher de la définition de variance d'une variable aléatoire Y :

$$\mathbb{V}(Y) \;=\; \mathbb{E}\Big(\big(Y - \mathbb{E}(Y)\big)^2\Big) \;=\; \mathbb{E}\left(Y^2\right) - \big(\mathbb{E}(Y)\big)^2$$

× Pour s'en convaincre, reprenons la correspondance de la remarque précédente.

$$\begin{array}{lcl} & \mathbb{E}(Y) & \leftrightarrow \quad \dfrac{1}{n}\sum_{i=1}^{n} y_i \\ \text{donc} & Y-\mathbb{E}(Y) & \leftrightarrow \quad y_i - \dfrac{1}{n}\sum_{i=1}^{n} y_i \\ \text{d'où} & \big(Y-\mathbb{E}(Y)\big)^2 & \leftrightarrow \quad \left(y_i - \dfrac{1}{n}\sum_{i=1}^{n} y_i\right)^2 \\ \text{puis} & \mathbb{E}\Big(\big(Y-\mathbb{E}(Y)\big)^2\Big) & \leftrightarrow \quad \dfrac{1}{n}\sum_{i=1}^{n}\left(y_i - \dfrac{1}{n}\sum_{i=1}^{n} y_i\right)^2 \\ \text{ainsi} & \mathbb{V}(Y) & \leftrightarrow \quad \dfrac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y}_n)^2 = s_y^2 \end{array}$$

× De même, pour la deuxième expression de s_y^2 :

- d'une part :

$$\mathbb{E}(Y^2) \quad \leftrightarrow \quad \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

- d'autre part :

$$\begin{array}{lcl} & \mathbb{E}(Y) & \leftrightarrow \quad \dfrac{1}{n}\sum_{i=1}^{n} y_i \\ \text{donc} & \big(\mathbb{E}(Y)\big)^2 & \leftrightarrow \quad \left(\dfrac{1}{n}\sum_{i=1}^{n} y_i\right)^2 = (\bar{y}_n)^2 \end{array}$$

Ainsi :

$$\begin{array}{lcl} & \mathbb{E}(Y^2)-\big(\mathbb{E}(Y)\big)^2 & \leftrightarrow \quad \dfrac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y}_n)^2 \\ \text{d'où} & \mathbb{V}(Y) & \leftrightarrow \quad \dfrac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y}_n)^2 = s_y^2 \end{array}$$

- l'écart-type empirique : $s_y = \sqrt{s_y^2}$.
Cette mesure permet encore de quantifier la dispersion des observations autour de la moyenne, mais a l'avantage de s'exprimer dans la même unité de grandeur que les y_i, grâce à la composition par la fonction racine carrée. Pour notre exemple, on entre l'instruction :

```
1  stdev(y)
```

On obtient alors la sortie :

```
ans  =
    22.623473
```

Au cours de l'année 2018, la concentration d'ozone varie donc en moyenne de $22,6\ \mu g.m^{-3}$ autour de sa valeur moyenne ($\bar{y}_{12} \approx 67,3\ \mu g.m^{-3}$).

Remarque
Notons qu'il s'agit ici d'un écart à la moyenne particulier : l'écart-type est la racine carrée de l'écart **quadratique** à la moyenne (et non de l'écart à la moyenne en valeur absolue) :

$$s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y}_n)^2} \quad (\text{et non } \frac{1}{n}\sum_{i=1}^{n}|y_i-\bar{y}_n|)$$

Comme pour les quantités précédentes, cela est à rapprocher de la définition de l'écart-type d'une variable aléatoire :

$$\sigma(Y) = \sqrt{\mathbb{V}(Y)} = \sqrt{\mathbb{E}\Big(\big(Y-\mathbb{E}(Y)\big)^2\Big)} \quad (\text{et non } \mathbb{E}\big(|Y-\mathbb{E}(Y)|\big))$$

- la médiane.
Il s'agit de la valeur m telle que la moitié des observations sont inférieures à m et la moitié sont supérieures à m. Intuitivement, la médiane est le point milieu de l'ensemble des observations. Pour notre exemple, on entre l'instruction :

```
1  median(y)
```

On obtient alors la sortie :

```
ans  =
    64.3
```

Au cours de l'année 2018, la valeur médiane des concentrations en ozone est $m = 64,3\ \mu g.m^{-3}$. Cela signifie que l'on a eu la moitié de l'année des concentrations inférieures à cette valeur et l'autre moitié de l'année, des valeurs supérieures.

- d'autres quantiles.
  Le quantile d'ordre α est le réel q_α tel que une proportion α des observations est inférieure est q_α et une proportion $1-\alpha$ des observations est supérieure à q_α. On considère traditionnellement quelques quantiles particuliers :
  - × les quartiles. Les 3 quartiles coupent les données en 4 groupes de même cardinal. Ainsi, il y a 25% de données entre 2 quartiles consécutifs. Pour notre exemple, on entre l'instruction :

    ```
    1  quart(y)
    ```

    On obtient alors la sortie :

    ```
    ans  =
        53.1
        64.3
        82.9
    ```

    La commande `quart(y)` renvoie ainsi les 3 quartiles des observations contenues dans la variable `y`. Plus précisément :
    - 3 mois de l'année (un quart de l'année), la concentration d'ozone est inférieure à $53,1\ \mu g.m^{-3}$ (1^{er} quartile).
    - 3 mois de l'année, la concentration d'ozone est comprise entre $53,1\ \mu g.m^{-3}$ et $64,3\ \mu g.m^{-3}$ (1^{er} et $2^{\text{ème}}$ quartiles).
    - 3 mois de l'année, la concentration d'ozone est comprise entre $64,3\ \mu g.m^{-3}$ et $82,9\ \mu g.m^{-3}$ ($2^{\text{ème}}$ et $3^{\text{ème}}$ quartiles).
    - 3 mois de l'année, la concentration d'ozone est supérieure à $64,3\ \mu g.m^{-3}$.
  - × les déciles. Les 9 déciles coupent les données en 10 groupes de même cardinal. Ainsi, il y a 10% de données entre 2 déciles consécutifs.
  - × les centiles. Les 99 centiles coupent les données en 100 groupes de même cardinal. Ainsi, il y a 1% de données entre 2 centiles consécutifs.

  Remarque

  On notera au passage que la médiane est le quantile d'ordre $\frac{1}{2}$ (ou 50%).
- le minimum et le maximum.
  Il peut en effet être pertinent de s'intéresser à l'étendue des observations. Pour notre exemple, on entre l'instructions :

  ```
  1  min(y)
  ```

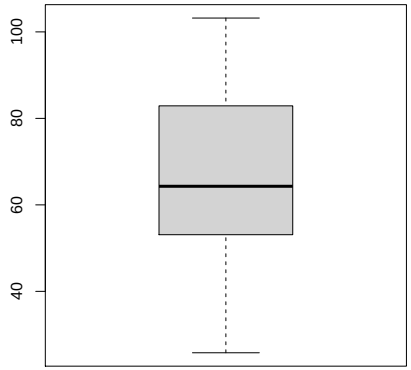
  On obtient alors la sortie :

  ```
  ans  =
      25.8
  ```

  La concentration minimale au cours de l'année 2018 est donc de $25,8\ \mu g.m^{-3}$. *(on pourrait faire de même avec la commande* `max(y)` *)*

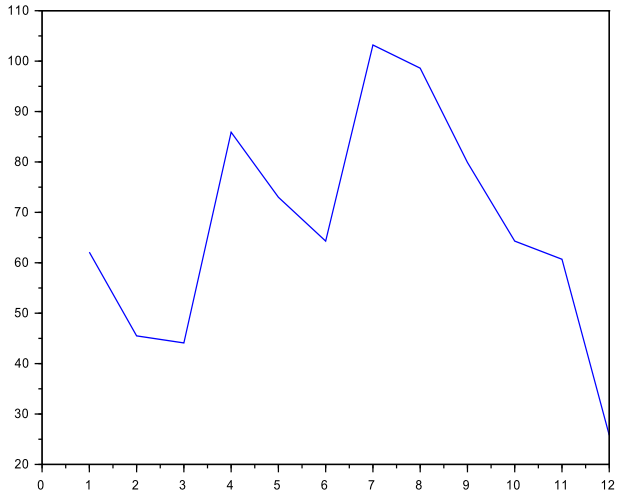
On peut représenter l'ensemble de ces caractéristiques dans un même graphe : une boîte à moustache. Celle-ci contiendra :

- × le minimum et le maximum des observations : les bornes de la boîte à moustaches complète,
- × les 1^{er} et $3^{\text{ème}}$ quartiles empiriques : les bornes du rectangle grisé,
- × la médiane : le trait horizontal tracé en gras.



Enfin, un bon réflexe est aussi de représenter ces données sur un graphe.

```
plot(y)
```



Il semble alors que la concentration d'ozone est plus élevée les mois d'été. On peut donc se demander s'il existe une dépendance entre la concentration d'ozone et la température.

I.2. Statistiques descriptives bivariées

On cherche maintenant à savoir s'il est possible d'expliquer le taux moyen d'ozone de la journée par la température t (en °C) à 16h. On obtient les données suivantes pour la température (Météo France) :

Mois	1	2	3	4	5	6	7	8	9	10	11	12
t (°C)	7	6,8	6,4	11,8	12,8	22,1	33,1	30,2	21,3	11,7	11,3	6,6
O_3 ($\mu g.m^{-3}$)	62,1	45,5	44,1	85,9	73	64,3	103,2	98,6	79,9	64,3	60,7	25,8

Comme dans la première partie, on stocke les données de température dans une variable `x`.

```
x = [7, 6.8, 6.4, 11.8, 12.8, 22.1, 33.1, 30.2, 21.3, 11.7, 11.3, 6.6]
```

On notera dans la suite $x = (x_1, \dots, x_n)$ les observations de température.
On pourrait évidemment faire de nouveau des statistiques univariées sur les données $(x_1, \dots, x_n)$ comme en partie précédente, mais on s'attachera ici à l'étude des 2 séries statistiques $(x_1, \dots, x_n)$ et $(y_1, \dots, y_n)$ de manière simultanée. On étudie alors les couples $\big((x_1, y_1), \dots, (x_n, y_n)\big)$. C'est ce n-uplet de couples que l'on appelle *observations* dans le cas de statistiques bivariées.
Comme précédemment, avant d'étudier ces observations, on commence par les décrire. Pour cela, on utilise essentiellement deux mesures :

- la covariance empirique : $\text{cov}(x,y) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) = \dfrac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \bar{x}_n\, \bar{y}_n$.
  Cette mesure permet de quantifier le degré de dépendance entre les 2 séries statistiques $(x_1, \dots, x_n)$ et $(y_1, \dots, y_n)$. Plus la covariance empirique est éloignée de 0, plus la dépendance entre les deux quantités étudiées est importante. Plus précisément :
  - × si $\text{cov}(x,y)$ est positive et éloignée de 0, alors les 2 quantités étudiées sont corrélées et la relation de dépendance entre les deux est croissante. Autrement dit, lorsque l'une des deux quantités augmente, alors l'autre augmente également. On dit que les 2 quantités sont *corrélées positivement*.
  - × si $\text{cov}(x,y)$ est négitive et éloignée de 0, alors les 2 quantités étudiées sont corrélées et la relation de dépendance entre les deux est décroissante. Autrement dit, lorsque l'une des deux quantités augmente, alors l'autre diminue. On dit que les 2 quantités sont *corrélées négativement*.

× si $\text{cov}(x,y)$ est proche de 0, alors les 2 quantités étudiées ne sont pas corrélées.

Pour notre exemple, on entre l'instruction :

```
corr(x, y, 1)
```

On obtient alors la sortie :

```
ans  =
    159.53236
```

On en déduit que la température et la concentration d'ozone sont corrélées positivement.

Remarque

× Notons une fois encore que la définition de covariance empirique est à rapprocher de la définition de la covariance de 2 v.a.r. X et Y :

$$\text{Cov}(X,Y) \;=\; \mathbb{E}\Big(\big(X-\mathbb{E}(X)\big)\big(Y-\mathbb{E}(Y)\big)\Big) \;=\; \mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)$$

× Pour la première expression de $\text{cov}(x,y)$:

- d'une part :

$$X-\mathbb{E}(X) \quad\leftrightarrow\quad x_i-\frac{1}{n}\sum_{i=1}^{n}x_i = x_i-\bar{x}_n$$

- d'autre part :

$$Y-\mathbb{E}(Y) \quad\leftrightarrow\quad y_i-\frac{1}{n}\sum_{i=1}^{n}y_i = y_i-\bar{y}_n$$

Ainsi :

$$\big(X-\mathbb{E}(X)\big)\big(Y-\mathbb{E}(Y)\big) \quad\leftrightarrow\quad (x_i-\bar{x}_n)(y_i-\bar{y}_n)$$

$$\text{donc}\quad \mathbb{E}\Big(\big(X-\mathbb{E}(X)\big)\big(Y-\mathbb{E}(Y)\big)\Big) \quad\leftrightarrow\quad \frac{1}{n}\sum_{i=1}^{n}\Big((x_i-\bar{x}_n)(y_i-\bar{y}_n)\Big)$$

$$\text{d'où}\quad \text{Cov}(X,Y) \quad\leftrightarrow\quad \text{cov}(x,y)$$

× Pour la deuxième expression de $\text{cov}(x,y)$:

- d'une part :

$$XY \quad\leftrightarrow\quad x_i\,y_i$$

$$\text{donc}\quad \mathbb{E}(XY) \quad\leftrightarrow\quad \frac{1}{n}\sum_{i=1}^{n}(x_i\,y_i)$$

- d'autre part :

$$\mathbb{E}(X)\,\mathbb{E}(Y) \quad\leftrightarrow\quad \bar{x}_n\,\bar{y}_n$$

Ainsi :

$$\mathbb{E}(XY)-\mathbb{E}(X)\,\mathbb{E}(Y) \quad\leftrightarrow\quad \frac{1}{n}\sum_{i=1}^{n}(x_i\,y_i)-\bar{x}_n\,\bar{y}_n$$

$$\text{donc}\quad \text{Cov}(X,Y) \quad\leftrightarrow\quad \text{cov}(x,y)$$

- le coefficient de corrélation linéaire empirique : $\rho_{x,y}=\dfrac{\text{cov}(x,y)}{s_x\,s_y}$.

Ce réel est toujours compris entre -1 et 1, et permet de quantifier le degré de dépendance **linéaire** entre les 2 séries statistiques $(x_1,\dots,x_n)$ et $(y_1,\dots,y_n)$. Plus précisément :

× si $\rho_{x,y}$ est proche de 1, alors les 2 quantités étudiées sont corrélées et la relation de dépendance entre les deux est une fonction **affine croissante**. Autrement dit, il existe une fonction f de la forme $f : x \mapsto ax+b$ telle que :

$$\forall i \in [\![1,n]\!],\; y_i = a\,x_i+b$$

où $(a,b)\in\mathbb{R}^2$ et $a \geqslant 0$.

× si $\rho_{x,y}$ est proche de -1, alors les 2 quantités étudiées sont corrélées et la relation de dépendance entre les deux est une fonction **affine décroissante**. Autrement dit, il existe une fonction f de la forme $f : x \mapsto ax+b$ telle que :

$$\forall i \in [\![1,n]\!],\; y_i = a\,x_i+b$$

où $(a,b)\in\mathbb{R}^2$ et $a \leqslant 0$.

× si $\rho_{x,y}$ est proche de 0, alors les 2 quantités étudiées ne sont pas corrélées.

Pour notre exemple, on entre l'instruction :

```
s1 = stdev(x) ;
s2 = stdev(y) ;
rho = corr(x,y,1)/(s1*s2)
```

On obtient alors la sortie :

```
ans  =
    0.7532372
```

Le coefficient de corrélation linéaire empirique est plutôt proche de 1, on peut donc penser qu'on peut approcher la relation entre concentration d'ozone et température par une relation linéaire.

Remarque
La définition du coefficient de corrélation linéaire empirique est à rapprocher de la définition du coefficient de corrélation linéaire de 2 v.a.r. X et Y :

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}(X)}\,\sqrt{\mathbb{V}(Y)}}$$

Commentaire

- La covariance empirique et le coefficient de corrélation linéaire empirique sont tous les deux des mesures du degré de dépendance entre deux quantités.
- Cependant, le coefficient de corrélation linéaire empirique nous donne une information supplémentaire quant à la forme de cette dépendance (dans les cas particuliers $\rho_{x,y} \approx 1$ et $\rho_{x,y} \approx -1$).

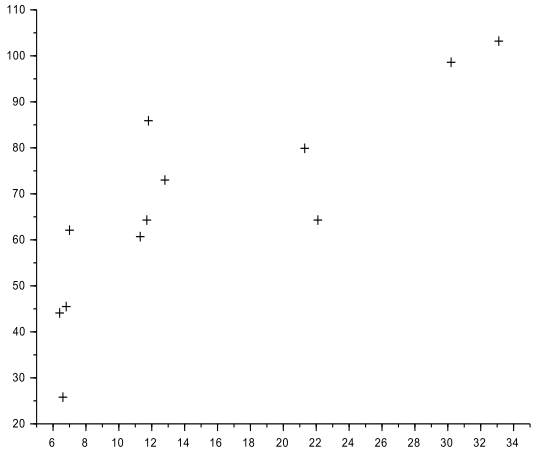
Enfin, il est toujours bon de représenter les observations $\big((x_1,y_1),\ldots,(x_n,y_n)\big)$ sur un graphe. On appelle le résultat le *nuage de points* des observations. Il nous permet de repérer certaines caractéristiques des observations. Par exemple :

- on peut relever une tendance sur le nuage :
  - des variations dans le même sens. On conjectura alors une dépendance croissante.
  - des variations en sens contraire. On conjectura alors une dépendance décroissante.
  - une allure de courbe particulière (dépendance affine, logarithmique, exponentielle).
  - des valeurs dispersées. On conjectura alors que les quantités étudiées ne sont pas corrélées.
- on peut repérer des valeurs aberrantes (des valeurs éloignées d'une tendance globale). Il faudra alors revenir sur la manière dont ont été collectées ces données : est-ce une erreur de mesure ? une erreur de manipulation ? une erreur de recopie ? ou tout simplement, il n'y a pas d'erreur et cette valeur est alors à prendre en compte pour une analyse plus poussée.

Dans notre exemple, on entre l'instruction :

```
plot2d(x, y, style = -1)
```

L'argument `style = -1` permet de représenter les points sous forme de croix.



À partir de ce nuage, il apparaît clairement que la relation de dépendance entre la concentration d'ozone et la température est une relation croissante (ce qui est confirmé par les valeurs de $\text{cov}(x,y)$ et $\rho_{x,y}$). Il semble de plus qu'on ait le choix entre deux types de conjectures (au moins) :

- cette relation de dépendance est linéaire et le point $(6.6, 25.8)$ (donnée du 1^{er} décembre) est une valeur aberrante.
- cette relation de dépendance est de la forme $y = \sqrt{a\,x + b}$ (et dans ce cas, le point $(6.6, 25.8)$ n'est pas une valeur aberrante).

II. Régression

II.1. Généralités

- Après cette étape de description, pour pouvoir aller plus loin dans l'étude, on souhaite analyser la relations entre les x_i (température) et les y_i (ozone). Plus précisément, on cherche à trouver une fonction f telle que :

$$\forall i \in [\![1,n]\!],\ y_i \approx f(x_i)$$

- *Effectuer une régression* sur les observations $\big((x_1,y_1),\dots,(x_n,y_n)\big)$ c'est chercher à trouver le meilleure fonction f (dans un sens à définir) pour les données observées. On dit qu'on *ajuste* le modèle aux observations.
  D'un point de vue pratique, le but de cette régression est double :
  - × expliquer la variable y en fonction de la variable x (ici expliquer la concentration d'ozone en fonction de la température),
  - × prédire les valeurs de y_i pour de nouvelles valeurs x_i.
- Plus formellement, on suppose qu'il existe une fonction f telle que :

$$\forall i \in [\![1,n]\!], \quad y_i \ = \ f(x_i) + e_i$$

  où $(e_1,\dots,e_n)$ est une réalisation d'un n-uplet de v.a.r. $(\varepsilon_1,\dots,\varepsilon_n)$. Les variables ε_1, …, ε_n sont des **variables aléatoires** appelée *erreurs* ou *bruits*. Un modèle statistique vérifiant une équation de ce type est appelé *modèle de régression*.

Remarque

- × Notons que les v.a.r. ε_i portent **seules** l'aléa dans cette modélisation : les x_i et y_i ne sont pas supposés être des réalisations de v.a.r. X_i et Y_i respectivement.
- × La présence des v.a.r. ε_i dans le modèle vient du fait que les réalisations (x_i,y_i) ne sont jamais parfaitement positionnées sur la courbe représentative de la fonction f. En pratique, cela peut provenir des imperfections des mesures.

- On impose traditionnellement 3 hypothèses sur les v.a.r. ε_1, …, ε_n :
  (i) *v.a.r. centrées* : $\forall i \in [\![1,n]\!]$, $\mathbb{E}(\varepsilon_i) = 0$
  (ii) *v.a.r. non corrélées* : $\forall (i,j) \in [\![1,n]\!]^2$, avec $i \neq j$, $\text{Cov}(\varepsilon_i,\varepsilon_j) = 0$
  (iii) *homoscédasticité* : $\forall i \in [\![1,n]\!]$, $\mathbb{V}(\varepsilon_i) = \sigma^2$
  (les v.a.r. ε_1, …, ε_n ont même variance)
- Assez régulièrement, on suppose même les assertions suivantes (plus restrictives que les précédentes) :
  (i') *indépendance* : les v.a.r. ε_1, …, ε_n sont indépendantes
  (ii') $\forall i \in [\![1,n]\!]$, $\varepsilon_i \hookrightarrow \mathcal{N}\left(0,\sigma^2\right)$
  On dit dans ce cas que ε est un *bruit blanc*.

On vérifie aisément que les hypothèses ***(i)'*** et ***(ii')*** impliquent les hypothèses ***(i)***, ***(ii)*** et ***(iii)***.

Remarque
On peut se convaincre que les hypothèses ***(i)***, ***(ii)*** et ***(iii)*** ont du sens.

× Montrons d'abord qu'on peut toujours se ramener à un modèle de régression vérifiant ***(i)***, c'est-à-dire tel que :

$$\forall i \in [\![1,n]\!],\ y_i = f(x_i) + e_i \quad \text{avec} \quad \mathbb{E}(\varepsilon_i) = 0$$

Supposons : $\forall i \in [\![1,n]\!]$, $y_i = g(x_i) + e'_i$ où $\mathbb{E}(\varepsilon'_i) = m \neq 0$.
On note alors :

$$f : x \mapsto g(x) + m \quad \text{et} \quad \varepsilon = \varepsilon' - m$$

Dans ce cas :

- d'une part, pour tout $i \in [\![1,n]\!]$:

$$\begin{aligned} y_i &= g(x_i) + e'_i \\ &= \big(f(x_i) - \cancel{m}\big) + (e_i + \cancel{m}) \qquad \textit{(par définition de } f \textit{ et } \varepsilon\textit{)} \\ &= f(x_i) + e_i \end{aligned}$$

- d'autre part, par linéarité de l'espérance, pour tout $i \in [\![1,n]\!]$:

$$\mathbb{E}(\varepsilon_i) \;=\; \mathbb{E}(\varepsilon'_i - m) \;=\; \mathbb{E}(\varepsilon'_i) - m \;=\; \cancel{m} - \cancel{m} \;=\; 0$$

(notons que la v.a.r. ε_i admet bien une espérance en tant que transformée affine de la v.a.r. ε'_i qui en admet une)

× Quant aux hypothèses ***(ii)*** et ***(iii)***, elles sont rendues naturelles par ce que modélisent les v.a.r. ε_i : des erreurs. Ces erreurs (ces bruits) n'ont pas de raison a priori de dépendre les uns des autres (hypothèse de non corrélation), ni d'être d'amplitudes distinctes (hypothèse d'homoscédasticité).

- La fonction f est appelée *fonction de régression* et c'est elle que l'on souhaite déterminer (ou approcher).
  Pour cela, on ne va pas chercher cette fonction f parmi toutes les fonctions possibles : il y en a beaucoup trop. On commencera donc par convenir, à l'avance, d'un ensemble de fonctions $\mathcal{F}$ (choisi en fonction de l'expérience, des statistiques descriptives effectuées auparavant, etc) et on supposera que la vraie fonction f inconnue appartient bien à cet ensemble.
  On distingue deux grandes familles d'ensembles de fonctions $\mathcal{F}$:
  × les ensembles définis à l'aide d'un nombre fini de paramètres. On dit alors que le modèle statistique est un *modèle paramétrique*. Par exemple :
    - modèle de régression linéaire : $\mathcal{F} = \{f : x \mapsto ax + b \mid (a,b) \in \mathbb{R}^2\}$
    - $\mathcal{F} = \{f : x \mapsto a\,\mathrm{e}^{bx} + c \mid (a,b,c) \in \mathbb{R}^3\}$
  × les ensembles définis à l'aide d'un nombre infini de paramètres. On dit alors que le modèle statistique est un *modèle non paramétrique*. Par exemple :
    - modèle des forêts aléatoires : $\mathcal{F}$ est l'ensemble des fonctions constantes par morceaux.
    - modèle d'estimation par noyau :

$$\mathcal{F} = \{f : x \mapsto \sum_{k\in\mathbb{Z}} \beta_k\, K(x - \alpha_k) \mid \forall k \in \mathbb{Z},\ (\alpha_k, \beta_k) \in \mathbb{R}^2\}$$

      où K est une fonction positive d'intégrale 1 appelée noyau (**K**ernel en allemand).

- Revenons maintenant à nos observations. Pour analyser la relations entre les x_i (température) et les y_i (ozone), on cherche donc une fonction $f \in \mathcal{F}$ telle que :

$$\forall i \in [\![1,n]\!],\ y_i \approx f(x_i)$$

  Dans cette écriture, il conviendra de préciser le sens de $\approx$. Pour cela, il faut se donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données. Intuitivement, on cherche à minimiser la « distance » (en un sens à définir) entre les valeurs fournies par le modèle (les $f(x_i)$) et les données (les y_i).

- Mathématiquement, le problème revient alors à trouver une fonction $f \in \mathcal{F}$ pour laquelle la quantité :

$$\sum_{i=1}^{n} L\big(y_i - f(x_i)\big)$$

  est minimale.
  La fonction L est appelée *fonction de coût* ou *fonction de perte* (**L**oss en anglais). Deux fonctions sont classiquement utilisées comme fonction de coût :
  × le coût absolu : $L_1 : u \mapsto |u|$.
  × le coût quadratique : $L_2 : u \mapsto u^2$.
  Pour des raisons de régularité, on a tendance à privilégier le coût quadratique.

Commentaire

Ces deux fonctions de coût peuvent être reliées à des notions de distance :
× la fonction $(u,v) \mapsto |v-u| = L_1(v-u)$ est exactement la distance sur $\mathbb{R}$,
× la fonction $(u,v) \mapsto (v-u)^2 = L_2(v-u)$ est le carré de la distance sur $\mathbb{R}$.
On retrouve donc bien notre intuition : la fonction f la mieux ajustée aux données est celle qui minimise la *distance* entre les prédictions du modèle et les données.

II.2. Régression linéaire

II.2.a) Définition et utilisation

Définition (Modèle de régression linéaire)
Un *modèle de régression linéaire* est un modèle de régression pour lequel la classe de fonctions $\mathcal{F}$ choisie est l'ensemble des fonctions affines de $\mathbb{R}$ dans $\mathbb{R}$:

$$\mathcal{F} \;=\; \{f : x \mapsto ax+b \mid (a,b) \in \mathbb{R}^2\}$$

Le modèle de régression s'écrit alors :

$$\forall i \in [\![1,n]\!],\; y_i \;=\; a\,x_i + b + e_i$$

Cas d'utilisation :
En pratique, on pense à utiliser un modèle de régression linéaire lorsque le coefficient de corrélation linéaire empirique $\rho_{x,y}$ est proche de 1. La justification mathématique qui se cache derrière cette pratique est la propriété de cours suivante.

Théorème 1.

Soit (X,Y) un couple de v.a.r. discrètes.

On suppose que X et Y admettent des variances non nulles.

1) *Inégalité de Cauchy-Schwarz :*

$$\boxed{|\operatorname{Cov}(X,Y)| \;\leqslant\; \sigma(X)\,\sigma(Y)}$$

On en déduit (reformulation) : $\boxed{|\rho(X,Y)| \leqslant 1}$

2) •
$$\boxed{\rho(X,Y) \;=\; 1 \quad\Leftrightarrow\quad \begin{array}{l}\textit{une des v.a.r. est une fonction affine}\\ \textit{strictement croissante de l'autre v.a.r.}\\ \textit{presque sûrement}\end{array}}$$

- *Autrement dit, $\rho(X,Y)=1$ ssi il existe $a>0$ et $b\in\mathbb{R}$ tels que, p.s. :*

$$X = a\,Y + b \qquad \texttt{OU} \qquad Y = a\,X + b$$

3) •
$$\boxed{\rho(X,Y) \;=\; -1 \quad\Leftrightarrow\quad \begin{array}{l}\textit{une des v.a.r. est une fonction affine}\\ \textit{strictement décroissante de l'autre}\\ \textit{v.a.r. presque sûrement}\end{array}}$$

- *Autrement dit, $\rho(X,Y)=-1$ ssi il existe $a>0$ et $b\in\mathbb{R}$ tels que, p.s. :*

$$X = -a\,Y + b \qquad \texttt{OU} \qquad Y = -a\,X + b$$

Démonstration.

1) • Considérons la fonction $h : t \mapsto \mathbb{V}(Y-tX)$.
Remarquons tout d'abord que cette fonction est bien définie. En effet, la v.a.r. $Y-tX$ admet une variance en tant que somme de v.a.r. qui admettent une variance.

- Par ailleurs :

$$\begin{aligned}
h(t) \;&=\; \mathbb{V}(Y-tX) \;=\; \operatorname{Cov}(Y-tX, Y-tX)\\
&=\; \operatorname{Cov}(Y, Y-tX) - \operatorname{Cov}(tX, Y-tX) && \textit{(par linéarité à gauche)}\\
&=\; \operatorname{Cov}(Y,Y) - \operatorname{Cov}(Y,tX) + \operatorname{Cov}(tX,tX) - \operatorname{Cov}(tX,Y) && \textit{(par linéarité à droite)}\\
&=\; \operatorname{Cov}(Y,Y) - t\operatorname{Cov}(Y,X) + t^2\operatorname{Cov}(X,X) - t\operatorname{Cov}(X,Y)\\
&=\; \operatorname{Cov}(Y,Y) - 2t\operatorname{Cov}(X,Y) + t^2\operatorname{Cov}(X,X)\\
&=\; \mathbb{V}(Y) - 2\operatorname{Cov}(X,Y)\,t + \mathbb{V}(X)\,t^2
\end{aligned}$$

Ainsi, la fonction h est polynomiale de degré 2.

- Or, pour tout $t \in \mathbb{R}$:
$$h(t) = \mathbb{V}(Y - tX) \geqslant 0$$
Cette fonction polynomiale étant de signe constant, on en déduit que le discriminant du polynôme P associé est de signe négatif. Or :
$$\Delta \;=\; (-2\mathrm{Cov}(X,Y))^2 - 4\mathbb{V}(X)\mathbb{V}(Y) \;=\; 4(\mathrm{Cov}(X,Y))^2 - 4\mathbb{V}(X)\mathbb{V}(Y)$$
Et enfin :
$$\begin{aligned}
\Delta \leqslant 0 \quad &\Leftrightarrow \quad \cancel{4}\,(\mathrm{Cov}(X,Y))^2 \leqslant \cancel{4}\,\mathbb{V}(X)\mathbb{V}(Y) \\
&\Leftrightarrow \quad \sqrt{(\mathrm{Cov}(X,Y))^2} \leqslant \sqrt{\mathbb{V}(X)\mathbb{V}(Y)} \qquad \textit{(par stricte croissance de } \sqrt{\cdot}\textit{)} \\
&\Leftrightarrow \quad |\,\mathrm{Cov}(X,Y)\,| \leqslant \sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)} \\
&\Leftrightarrow \quad |\,\mathrm{Cov}(X,Y)\,| \leqslant \sigma(X)\,\sigma(Y) \\
&\Leftrightarrow \quad \frac{|\,\mathrm{Cov}(X,Y)\,|}{\sigma(X)\,\sigma(Y)} \leqslant 1 \\
&\Leftrightarrow \quad |\,\rho(X,Y)\,| \leqslant 1
\end{aligned}$$

2) et ***3)*** D'après ce qui précède :
$$\begin{aligned}
&\rho(X,Y) \in \{-1,1\} \\
\Leftrightarrow\ &|\rho(X,Y)| = 1 \\
\Leftrightarrow\ &\text{Le polynôme } P \text{ admet une unique racine } a \in \mathbb{R} \\
\Leftrightarrow\ &\exists\,!\, a \in \mathbb{R},\ \mathbb{V}(Y - a\,X) = 0 \\
\Leftrightarrow\ &\text{Il existe un unique } a \in \mathbb{R} \text{ tel que la v.a.r. } Y - a\,X \text{ est presque sûrement constante} \\
\Leftrightarrow\ &\exists\,!\, a \in \mathbb{R},\ \exists b \in \mathbb{R},\ Y - a\,X = b \text{ presque sûrement} \\
\Leftrightarrow\ &\exists\,!\, a \in \mathbb{R},\ \exists b \in \mathbb{R},\ Y = a\,X + b \text{ presque sûrement}
\end{aligned}$$

Par la formule des racines des polynômes de second degré, on obtient que l'unique racine a de P s'écrit sous la forme :
$$\begin{aligned}
a \;&=\; \frac{\cancel{2}\,\mathrm{Cov}(X,Y)}{\cancel{2}\,\mathbb{V}(X)} \\
&=\; \frac{\mathrm{Cov}(X,Y)}{\mathbb{V}(X)}\,\frac{\sigma(X)\,\sigma(Y)}{\sigma(X)\,\sigma(Y)} \\
&=\; \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\,\sigma(Y)}\,\frac{\sigma(X)\,\sigma(Y)}{\mathbb{V}(X)} \\
&=\; \rho(X,Y)\,\frac{\sigma(Y)}{\sigma(X)} \qquad\qquad \textit{(car } \mathbb{V}(X) = \big(\sigma(X)\big)^2\textit{)}
\end{aligned}$$
On en déduit finalement, d'après ce qui précède :

$$\boxed{\rho(X,Y) = 1 \quad \Leftrightarrow \quad \text{il existe } b \in \mathbb{R} \text{ tel que } Y = \frac{\sigma(Y)}{\sigma(X)}\,X + b \text{ p.s.}}$$

(dans ce cas, la v.a.r. Y est une transformée affine strictement croissante de X presque sûrement)

$$\boxed{\rho(X,Y) = -1 \quad \Leftrightarrow \quad \text{il existe } \beta \in \mathbb{R} \text{ tel que } Y = -\frac{\sigma(Y)}{\sigma(X)}\,X + b \text{ p.s.}}$$

(dans ce cas, la v.a.r. Y est une transformée affine strictement décroissante de X presque sûrement)

□

Remarque
Dans les cas où $\rho(X,Y) \in \{-1,1\}$, c'est-à-dire les cas où la v.a.r. Y est une transformée affine de X, on remarque que la démonstration ci-dessus nous fournit la valeur du coefficient directeur de cette transformée affine :
$$a \;=\; \frac{\mathrm{Cov}(X,Y)}{\mathbb{V}(X)}$$

Dans le cas d'un modèle de régression linéaire :

$$\forall i \in [\![1,n]\!],\ y_i = a x_i + b + e_i$$

On peut donc penser que valeur de a à choisir pour ajuster le modèle aux données sera :

$$a^* \ = \ \frac{\text{cov}(x,y)}{s_x^2}$$

Nous verrons en partie suivante que c'est effectivement le cas.

II.2.b) Méthodes des moindres carrés (HEC 2008)

On se place désormais dans le cadre d'un modèle de régression linéaire.

- D'après la partie **II.1**, on sait que l'on cherche alors une fonction $f \in \mathcal{F}$, où $\mathcal{F} = \{f : x \mapsto ax+b \mid (a,b) \in \mathbb{R}^2\}$ telle que la quantité $\sum_{i=1}^{n} L\big(y_i - f(x_i)\big)$ est minimale.
- Cela équivaut à chercher un couple $(a,b) \in \mathbb{R}^2$ tel que la quantité $\sum_{i=1}^{n} L\big(y_i - (a\,x_i + b)\big)$ est minimale.
- De plus, comme expliqué toujours en partie **II.1**, on va choisir la fonction de coût quadratique L_2 comme fonction de perte pour des raisons de régularité. Finalement, le problème revient à minimiser la fonction :

$$\begin{array}{rccc} h & : & \mathbb{R}^2 & \to & \mathbb{R} \\ & & (a,b) & \mapsto & \displaystyle\sum_{i=1}^{n} (y_i - a\,x_i - b)^2 \end{array}$$

  La méthode consistant à minimiser la fonction h pour effectuer une régression linéaire est appelée *méthode des moindres carrés.*

Théorème 2.
La fonction h admet un unique minimum local (a^*, b^*) sur $\mathbb{R}^2$ où :

$$a^* = \frac{cov(x,y)}{s_x^2} \quad et \quad b^* = \bar{y}_n - a^*\,\bar{x}_n = \bar{y}_n - \frac{cov(x,y)}{s_x^2}\,\bar{x}_n$$

Le couple (a^*, b^*) est appelée estimation des moindres carrés.

Démonstration.

- La fonction h est de classe $\mathcal{C}^2$ sur $\mathbb{R}^2$ en tant que fonction polynomiale.
- Elle admet donc en particulier des dérivées partielles d'ordre 1 (et 2) sur $\mathbb{R}^2$. Soit $(a,b) \in \mathbb{R}^2$.

$$\begin{aligned} \partial_1(h)(a,b) &= \sum_{i=1}^{n} \Big(2(-x_i)\,(y_i - a\,x_i - b)\Big) && \textit{(par linéarité de la dérivée)} \\ &= -2\,\left(\sum_{i=1}^{n} x_i(y_i - a\,x_i - b)\right) \\ &= -2\left(\sum_{i=1}^{n} x_i\,y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i\right) \end{aligned}$$

On cherche alors à exprimer les 3 sommes ci-dessus à l'aide des quantités $\bar{x}_n$, $\bar{y}_n$, s_x^2 et cov(x,y) définies en partie **I.** pour simplifier l'expression de $\partial_1(h)$.

× Tout d'abord :

$$\frac{1}{n}\sum_{i=1}^{n} x_i\,y_i - \bar{x}_n\,\bar{y}_n = \text{cov}(x,y)$$

$$\text{donc} \quad \frac{1}{n}\sum_{i=1}^{n} x_i\,y_i = \text{cov}(x,y) + \bar{x}_n\,\bar{y}_n$$

$$\text{d'où} \quad \sum_{i=1}^{n} x_i\,y_i = n\,(\text{cov}(x,y) + \bar{x}_n\,\bar{y}_n)$$

× De plus :

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2 = s_x^2$$

$$\text{donc} \quad \frac{1}{n}\sum_{i=1}^{n} x_i^2 = s_x^2 + \bar{x}_n^2$$

$$\text{d'où} \quad \sum_{i=1}^{n} x_i^2 = n\left(s_x^2 + \bar{x}_n^2\right)$$

× Enfin : $\sum_{i=1}^{n} x_i = n\,\bar{x}_n$.

On en déduit :

$$\begin{aligned}\partial_1(h)(a,b) &= -2\Big(n\big(\mathrm{cov}(x,y)+\bar{x}_n\,\bar{y}_n\big)-a\,n\big(s_x^2+\bar{x}_n^2\big)-b\,n\,\bar{x}_n\Big)\\ &= -2n\Big(\mathrm{cov}(x,y)+\bar{x}_n\,\bar{y}_n-\big(s_x^2+\bar{x}_n^2\big)\,a-\bar{x}_n\,b\Big)\end{aligned}$$

De même :

$$\begin{aligned}\partial_2(h)(a,b) &= \sum_{i=1}^{n}\Big(2\,(-1)\,(y_i-a\,x_i-b)\Big)\\ &= -2\sum_{i=1}^{n}(y_i-a\,x_i-b)\\ &= -2\left(\sum_{i=1}^{n}y_i-a\sum_{i=1}^{n}x_i-b\,n\right)\\ &= -2\big(n\,\bar{y}_n-a\,n\,\bar{x}_n-b\,n\big)\\ &= -2n\,\big(\bar{y}_n-a\,\bar{x}_n-b\big)\end{aligned}$$

- On déterminer alors les points critiques de h.
  Soit $(a,b)\in\mathbb{R}^2$.

$$\begin{aligned}&(a,b)\text{ est un point critique de }h\\ \Leftrightarrow\ &\nabla(h)(a,b)=0_{\mathscr{M}_{2,1}(\mathbb{R})}\\ \Leftrightarrow\ &\begin{cases}\partial_1(h)(a,b) &= 0\\ \partial_2(h)(a,b) &= 0\end{cases}\\ \Leftrightarrow\ &\begin{cases}-2\,n\,\Big(\mathrm{cov}(x,y)+\bar{x}_n\,\bar{y}_n-\big(s_x^2+\bar{x}_n^2\big)\,a-\bar{x}_n\,b\Big) &= 0\\ -2\,n\,(\bar{y}_n-a\,\bar{x}_n-b) &= 0\end{cases}\\ \Leftrightarrow\ &\begin{cases}\mathrm{cov}(x,y)+\bar{x}_n\,\bar{y}_n-\big(s_x^2+\bar{x}_n^2\big)\,a-\bar{x}_n\,b &= 0\\ \bar{y}_n-a\,\bar{x}_n-b &= 0\end{cases}\end{aligned}$$

Ainsi :

$$\begin{aligned}&(a,b)\text{ est un point critique de }f\\ \Leftrightarrow\ &\begin{cases}\mathrm{cov}(x,y)+\bar{x}_n\,\bar{y}_n-\big(s_x^2+\bar{x}_n^2\big)\,a-\bar{x}_n\,b &= 0\\ \bar{y}_n-a\,\bar{x}_n-b &= 0\end{cases}\\ \Leftrightarrow\ &\begin{cases}\mathrm{cov}(x,y)+\bar{x}_n\,\bar{y}_n &= \big(s_x^2+\bar{x}_n^2\big)\,a+\bar{x}_n\,b\\ \bar{y}_n-a\,\bar{x}_n &= b\end{cases}\\ \Leftrightarrow\ &\begin{cases}\mathrm{cov}(x,y)+\bar{x}_n\,\bar{y}_n &= \big(s_x^2+\bar{x}_n^2\big)\,a+\bar{x}_n\,\big(\bar{y}_n-a\,\bar{x}_n\big)\\ b &= \bar{y}_n-a\,\bar{x}_n\end{cases}\\ \Leftrightarrow\ &\begin{cases}\mathrm{cov}(x,y)+\cancel{\bar{x}_n\,\bar{y}_n} &= \big(s_x^2+\cancel{\bar{x}_n^2}-\cancel{\bar{x}_n^2}\big)\,a+\cancel{\bar{x}_n\,\bar{y}_n}\\ b &= \bar{y}_n-a\,\bar{x}_n\end{cases}\\ \Leftrightarrow\ &\begin{cases}a &= \dfrac{\mathrm{cov}(x,y)}{s_x^2}\\ b &= \bar{y}_n-a\,\bar{x}_n\end{cases}\\ \Leftrightarrow\ &\begin{cases}a &= \dfrac{\mathrm{cov}(x,y)}{s_x^2}\\ b &= \bar{y}_n-\dfrac{\mathrm{cov}(x,y)}{s_x^2}\,\bar{x}_n\end{cases}\end{aligned}$$

On en déduit que h admet un unique point critique :

$$(a^*,b^*)=\left(\frac{\mathrm{cov}(x,y)}{s_x^2},\ \bar{y}_n-\frac{\mathrm{cov}(x,y)}{s_x^2}\,\bar{x}_n\right)$$

Commentaire

- La difficulté de la recherche de points critiques réside dans le fait qu'il n'existe pas de méthode générale pour résoudre l'équation $\nabla(h)(a,b) = 0_{\mathscr{M}_{2,1}(\mathbb{R})}$. On est donc confronté à une question bien plus complexe qu'une résolution de système d'équations linéaires (que l'on résout aisément à l'aide de la méthode du pivot de Gauss).
- Lors de la recherche de points critiques, on doit faire appel à des méthodes ad hoc. Ici, on fait apparaître une équation du type :

$$b = \psi(a)$$

En injectant cette égalité dans la seconde équation, on obtient une nouvelle équation qui ne dépend plus que d'une variable et qu'il est donc plus simple de résoudre. C'est la stratégie qu'on a adoptée précédemment.

- On cherche alors le signe des valeurs propres de la matrice $H = \nabla^2(h)(a^*,b^*)$ pour connaître la nature du point critique (a^*,b^*).
  - × Commençons par déterminer les dérivées partielles secondes de h sur $\mathbb{R}^2$. Soit $(a,b) \in \mathbb{R}^2$.
    - Tout d'abord :
    $$\partial^2_{1,1}(h)(a,b) \;=\; 2\,n\,\left(s_x^2 + \bar{x}_n^2\right)$$
    - Ensuite :
    $$\partial^2_{1,2}(h)(a,b) \;=\; \partial^2_{2,1}(h)(a,b) \;=\; 2\,n\,\bar{x}_n$$
    La première égalité est obtenue en vertu du théorème de Schwarz puisque la fonction h est de classe $\mathcal{C}^2$ sur l'**ouvert** $\mathbb{R}^2$.
    - Enfin :
    $$\partial^2_{2,2}(h)(a,b) \;=\; 2\,n$$

Commentaire

- Il faut penser à utiliser le théorème de Schwarz dès que la fonction à deux variables considérée est de classe $\mathcal{C}^2$ sur un ouvert $U \subset \mathbb{R}^2$.
- Ici, le calcul de $\partial^2_{1,2}(h)(a,b)$ et $\partial^2_{2,1}(h)(a,b)$ est aisé. Il faut alors concevoir le résultat du théorème de Schwarz comme une mesure de vérification : en dérivant par rapport à la 1ère variable puis par rapport à la 2ème, on doit obtenir le même résultat que dans l'ordre inverse.

  - × La matrice hessienne de h en (a,b) est donc :
  $$\nabla^2(h)(a,b) \;=\; \begin{pmatrix} \partial^2_{1,1}(h)(a,b) & \partial^2_{1,2}(h)(a,b) \\ \partial^2_{2,1}(h)(a,b) & \partial^2_{2,2}(h)(a,b) \end{pmatrix} \;=\; \begin{pmatrix} 2\,n\,(s_x^2 + \bar{x}_n^2) & 2\,n\,\bar{x}_n \\ 2\,n\,\bar{x}_n & 2\,n \end{pmatrix}$$
  Ainsi : $H = \nabla^2(h)(a^*,b^*) = \begin{pmatrix} 2\,n\,(s_x^2 + \bar{x}_n^2) & 2\,n\,\bar{x}_n \\ 2\,n\,\bar{x}_n & 2\,n \end{pmatrix}$.
  - × Soit $\lambda \in \mathbb{R}$.
  $$\begin{aligned} \det(H - \lambda\, I_2) \;&=\; \det\left(\begin{pmatrix} 2\,n\,(s_x^2 + \bar{x}_n^2) - \lambda & 2\,n\,\bar{x}_n \\ 2\,n\,\bar{x}_n & 2\,n - \lambda \end{pmatrix}\right) \\ &=\; \left(2\,n\,(s_x^2 + \bar{x}_n^2) - \lambda\right)\,(2\,n - \lambda) - (2\,n\,\bar{x}_n)^2 \\ &=\; \lambda^2 - \left(2\,n\,(s_x^2 + \bar{x}_n^2) + 2\,n\right)\lambda + 4\,n^2\,(s_x^2 + \bar{x}_n^2) - \cancel{4\,n^2\,\bar{x}_n^2} \\ &=\; \lambda^2 - 2\,n\,(s_x^2 + \bar{x}_n^2 + 1)\,\lambda + 4\,n^2\,s_x^2 \end{aligned}$$
  On en déduit :
  $$\begin{aligned} \lambda \text{ est valeur propre de } H \;&\Leftrightarrow\; H - \lambda\, I_2 \text{ non inversible} \\ &\Leftrightarrow\; \det(H - \lambda\, I_2) = 0 \\ &\Leftrightarrow\; \lambda^2 - 2\,n\,(s_x^2 + \bar{x}_n^2 + 1)\,\lambda + 4\,n^2\,s_x^2 = 0 \\ &\Leftrightarrow\; \lambda \text{ est racine de } Q \end{aligned}$$

où Q est le polynôme de degré 2 défini par :

$$Q(X) \;=\; X^2 - 2\,n\,(s_x^2 + \bar{x}_n^2 + 1)\,X + 4\,n^2\,s_x^2$$

Commentaire

À ce stade de l'étude de la nature d'un point critique, le calcul de $\det(H - \lambda\, I_2)$ nous fournit toujours un polynôme de degré 2 en λ. Notons le Q. Deux cas se présentent alors :

× l'expression de Q est « simple » (c'est par exemple le cas lorsque les coefficients de Q sont numériques). Dans ce cas :

1) on détermine explicitement les racines de Q par factorisation ou calcul de discriminant.

2) les racines de Q sont les valeurs propres de H d'après les équivalences ci-dessus.

3) on en déduit le signe des valeurs propres de H et ainsi la nature du point critique étudié.

× l'expression de Q est « compliquée » (c'est par exemple le cas lorsque l'expression de Q dépend de plusieurs paramètres, comme ici). Dans ce cas, **on ne cherchera pas** à déterminer les racines de Q explicitement. On procèdera de la manière suivante :

1) on justifie l'existence de valeurs propres λ_1 et λ_2 de H (la matrice H est symétrique).

2) les valeurs propres de H sont racines de Q d'après les équivalences ci-dessus. On en déduit la factorisation de Q suivante :

$$Q(X) \;=\; (X - \lambda_1)\,(X - \lambda_2)$$

3) on identifie les coefficients des deux expressions de Q pour en déduire des relations sur λ_1 et λ_2 (elles sont appelées *relations coefficients / racines*).

4) on détermine le signe de λ_1 et λ_2 (valeurs propres de H) grâce à ces relations, et on obtient ainsi la nature du point critique étudié.

× La matrice H est une matrice symétrique (réelle). Elle est donc diagonalisable. On note λ_1 et λ_2 ses valeurs propres **éventuellement égales**.

× D'après les équivalences précédentes, on en déduit que λ_1 et λ_2 sont racines de Q. Ainsi :

$$\begin{aligned} Q(X) &= (X - \lambda_1)\,(X - \lambda_2) \\ &= X^2 - (\lambda_1 + \lambda_2)\,X + \lambda_1\,\lambda_2 \end{aligned}$$

D'où, par définition de Q :

$$X^2 - 2\,n\,(s_x^2 + \bar{x}_n^2 + 1)\,X + 4\,n^2\,s_x^2 \;=\; X^2 - (\lambda_1 + \lambda_2)\,X + \lambda_1\,\lambda_2$$

Par identification des coefficients de ces polynômes de degré 2, on en déduit le système d'équations suivant :

$$\begin{cases} \lambda_1 + \lambda_2 \;=\; 2\,n\,(\sigma_x^2 + \bar{x}^2 + 1) & (*) \\ \lambda_1\,\lambda_2 \;=\; 4\,n^2\,\sigma_x^2 & (**) \end{cases}$$

- L'équation $(**)$ implique : $\lambda_1\,\lambda_2 > 0$.
  On en déduit que λ_1 et λ_2 ont même signe. Le point (a^*, b^*) est donc un extremum local de h sur $\mathbb{R}^2$.
- L'équation $(*)$ implique : $\lambda_1 + \lambda_2 > 0$.
  Or λ_1 et λ_2 ont même signe. D'où : $\lambda_1 > 0$ et $\lambda_2 > 0$.

On en conclut que la fonction h admet un mimimum local en (a^*, b^*).
Comme (a^*, b^*) était l'unique point critique de h sur $\mathbb{R}^2$, on en déduit que la fonction h admet un unique minimum local : (a^*, b^*). □

Théorème 3.
Par méthode des moindres carrés, la meilleure droite de régression pour le modèle de régression linéaire :

$$\forall i \in [\![1, n]\!], \quad y_i \;=\; a\,x_i + b + e_i$$

est la droite d'équation : $y = a^*\,x + b^*$ où :

$$a^* = \frac{cov(x, y)}{s_x^2} \quad et \quad b^* = \bar{y}_n - a^*\,\bar{x}_n = \bar{y}_n - \frac{cov(x, y)}{s_x^2}\;\bar{x}_n$$

Remarque

- On peut noter que, comme $b^* = \bar{y}_n - a^* \bar{x}_n$, la droite de régression des moindres carrés passe toujours par le centre de gravité du nuage de point : le point $(\bar{x}_n, \bar{y}_n)$.
- La méthode des moindres carrés fait l'objet de l'exercice du sujet `HEC 2008`. On y retrouve donc toute la démarche de cette section.
- Si l'on effectue une régression linéaire sur nos données d'ozone et de température, on obtient les estimations des moindres carrés à l'aide des commandes suivantes :

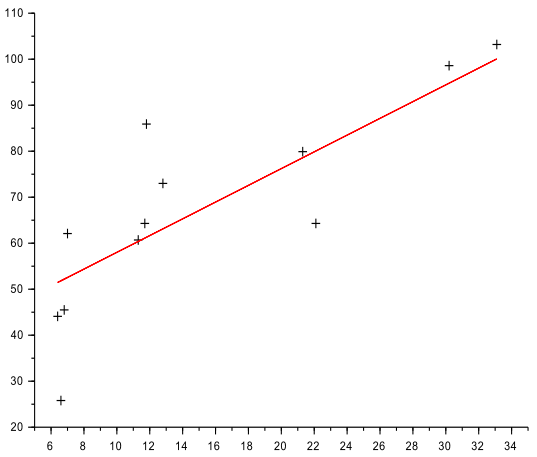
```
xbar = mean(x)
ybar = mean(y)
s_x2 = variance(x)
a = corr(x, y, 1) / s_x2
b = ybar - a * xbar
```

```
a  =
    1.8202594
b  =
    39.812585
```

Représentons enfin graphiquement la droite de régression sur le nuage de points.

```
plot2d(x, y, style = -1)
plot(x, a * x + b, 'r')
```

L'argument `'r'` permet de représenter la droite de régression en rouge.



II.2.c) Maximum de vraisemblance (`HEC 2016`)

On considère toujours dans cette partie un modèle de régression linéaire :

$$\forall i \in [\![1, n]\!],\ y_i = a x_i + b + e_i$$

où $(e_1, \ldots, e_n)$ est une réalisation de $(\varepsilon_1, \ldots, \varepsilon_n)$. On suppose en plus que l'on se place sous des hypothèses d'erreurs gaussiennes, c'est-à-dire on suppose que le modèle vérifie les hypothèses ***(i')*** et ***(ii')*** citées en **II.1** et que l'on rappelle :

(i') *indépendance* : les v.a.r. ε_1, …, ε_n sont indépendantes

(ii') $\forall i \in [\![1, n]\!]$, $\varepsilon_i \hookrightarrow \mathcal{N}\left(0, \sigma^2\right)$

Autrement dit, le n-uplet $(\varepsilon_1, \ldots, \varepsilon_n)$ est un n-échantillon d'une v.a.r. ε de loi $\mathcal{N}\left(0, \sigma^2\right)$.

- On dispose donc d'observations $(e_1, \ldots, e_n) = (y_1 - a x_1 - b, \ldots, y_n - a x_n - b)$.
- Ces observations proviennent d'un n-échantillon $(\varepsilon_1, \ldots, \varepsilon_n)$ d'une v.a.r. ε dont la loi dépend des paramètres (a, b), a priori inconnu et qu'on cherche à déterminer. Pour ce faire, une idée naturelle consiste à considérer que la valeur du couple (a, b) qui a permis de générer les observations est celle qui avait la plus grande probabilité de les générer. C'est cette idée qui guide la méthode que nous allons utiliser maintenant : la méthode dite du *maximum de vraisemblance*.

Remarque
Pour mieux comprendre cette méthode, plaçons nous un moment dans le cas où ε serait une v.a.r. discrète (la méthode est plus simple à appréhender dans ce cas). L'idée est de choisir comme estimation de (a,b) le couple $(\hat{a},\hat{b})$ tel que la **vraisemblance** d'avoir obtenu les observations de notre jeu de données soit maximisée. Autrement dit, le couple $(\hat{a},\hat{b})$ tel que la probabilité :

$$\begin{aligned}\mathcal{L}(a,b) &= \mathbb{P}_{a,b}([\varepsilon_1 = e_1] \cap \ldots \cap [\varepsilon_n = e_n]) \\ &= \mathbb{P}_{a,b}([\varepsilon_1 = e_1]) \times \ldots \times \mathbb{P}_{a,b}([\varepsilon_n = e_n]) && \textit{(par indépendance de } \varepsilon_1, \ldots, \varepsilon_n\textit{)} \\ &= \prod_{i=1}^{n} \mathbb{P}_{a,b}([\varepsilon_i = e_i])\end{aligned}$$

soit maximale.

- Par analogie avec le cas discret, la vraisemblance pour notre modèle de régression linéaire est donc la fonction $\mathcal{L}$ ($\mathcal{L}$ikelihood) définie par :

$$\begin{array}{rccc}\mathcal{L} & : & \mathbb{R}^2 & \to & \mathbb{R} \\ & & (a,b) & \mapsto & \displaystyle\prod_{i=1}^{n} f_{\varepsilon_i}(e_i)\end{array}$$

  où, pour tout $i \in [\![1,n]\!]$, f_{ε_i} désigne une densité de la v.a.r. ε_i.
- Simplifions l'expression de $\mathcal{L}$.
  Soit $(a,b) \in \mathbb{R}^2$.

$$\begin{aligned}\mathcal{L}(a,b) &= \prod_{i=1}^{n} f_{\varepsilon_i}(e_i) \\ &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\,\pi\,\sigma^2}}\ \exp\left(-\frac{e_i^2}{2\,\sigma^2}\right)\right) && \textit{(car : } \forall i \in [\![1,n]\!], \varepsilon_i \hookrightarrow \mathcal{N}\left(0,\sigma^2\right)\textit{)} \\ &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\,\pi\,\sigma^2}}\ \exp\left(-\frac{(y_i - a\,x_i - b)^2}{2\,\sigma^2}\right)\right)\end{aligned}$$

  La dernière égalité est vérifiée car on s'est placé dans le cadre du modèle de régression linéaire :

$$\forall i \in [\![1,n]\!],\ y_i = ax_i + b + e_i$$

- On rappelle qu'on cherche ici à maximiser la fonction $\mathcal{L}$. Cependant son expression sous forme de produit n'est pas facile à manipuler. C'est pourquoi on va plutôt chercher à maximiser la fonction $\ell = \ln \circ \mathcal{L}$ appelée *log-vraisemblance*. En effet, grâce aux propriétés de la fonction ln, l'expression de la fonction ℓ sera alors sous forme de somme (bien plus facile à dériver notamment).
  × Démontrons tout d'abord que trouver un maximum pour la fonction $\mathcal{L}$ est bien équivalent à trouver un maximum pour la fonction ℓ.

$$\begin{aligned}&(\hat{a},\hat{b}) \text{ est un maximum de } \mathcal{L} \text{ sur } \mathbb{R}^2 \\ \Leftrightarrow\ & \forall (a,b) \in \mathbb{R}^2,\ \mathcal{L}(a,b) \leqslant \mathcal{L}(\hat{a},\hat{b}) \\ \Leftrightarrow\ & \forall (a,b) \in \mathbb{R}^2,\ \ln\Big(\mathcal{L}(a,b)\Big) \leqslant \ln\Big(\mathcal{L}(\hat{a},\hat{b})\Big) && \textit{(par stricte croissance de } \ln \textit{ sur }]0,+\infty[\textit{)} \\ \Leftrightarrow\ & \forall (a,b) \in \mathbb{R}^2,\ \ell(a,b) \leqslant \ell(\hat{a},\hat{b}) \\ \Leftrightarrow\ & (\hat{a},\hat{b}) \text{ est un maximum de } \ell \text{ sur } \mathbb{R}^2\end{aligned}$$

  × Simplifions maintenant l'expression de la fonction ℓ.
  Soit $(a,b) \in \mathbb{R}^2$.

$$\begin{aligned}\ell(a,b) &= \ln\Big(\mathcal{L}(a,b)\Big) \\ &= \ln\left(\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\,\pi\,\sigma^2}}\ \exp\left(-\frac{(y_i - a\,x_i - b)^2}{2\,\sigma^2}\right)\right)\right) \\ &= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\,\pi\,\sigma^2}}\ \exp\left(-\frac{(y_i - a\,x_i - b)^2}{2\,\sigma^2}\right)\right) \\ &= \sum_{i=1}^{n}\left(-\frac{1}{2}\ln\left(2\,\pi\,\sigma^2\right) - \frac{(y_i - a\,x_i - b)^2}{2\,\sigma^2}\right) \\ &= -\frac{n}{2}\ \ln\left(2\,\pi\,\sigma^2\right) - \frac{1}{2\,\sigma^2}\ \sum_{i=1}^{n}(y_i - ax_i - b)^2 \\ &= -\frac{n}{2}\ \ln\left(2\,\pi\,\sigma^2\right) - \frac{1}{2\,\sigma^2}\ h(a,b)\end{aligned}$$

  où h est la fonction définie en section **II.2.b)**.

× Démontrons enfin que maximiser la fonction ℓ est équivalent à minimiser la fonction h.

$$
\begin{aligned}
& (\hat{a},\hat{b}) \text{ est un maximum de } \ell \text{ sur } \mathbb{R}^2 \\
\Leftrightarrow\ & \forall (a,b)\in\mathbb{R}^2,\ \ell(a,b)\leqslant \ell(\hat{a},\hat{b}) \\
\Leftrightarrow\ & \forall (a,b)\in\mathbb{R}^2,\ -\frac{n}{2}\ \ln\left(2\,\pi\,\sigma^2\right)-\frac{1}{2\,\sigma^2}\ h(a,b)\leqslant -\frac{n}{2}\ \ln\left(2\,\pi\,\sigma^2\right)-\frac{1}{2\,\sigma^2}\ h(\hat{a},\hat{b}) \\
\Leftrightarrow\ & \forall (a,b)\in\mathbb{R}^2,\ -\frac{1}{2\,\sigma^2}\ h(a,b)\leqslant -\frac{1}{2\,\sigma^2}\ h(\hat{a},\hat{b}) \\
\Leftrightarrow\ & \forall (a,b)\in\mathbb{R}^2,\ h(a,b)\geqslant h(\hat{a},\hat{b}) \qquad\qquad (car\ -2\,\sigma^2<0) \\
\Leftrightarrow\ & (\hat{a},\hat{b}) \text{ est un minimum de } h \text{ sur } \mathbb{R}^2
\end{aligned}
$$

× Or, on a déjà démontré en **II.2.b)** que le couple (a^*,b^*) est l'unique minimum local de h. On en déduit que l'estimation du maximum de vraisemblance est encore (a^*,b^*).

Théorème 4.
*Sous les hypothèses **(i')** et **(ii')**, par méthode du maximum de vraisemblance, la meilleure droite de régression pour le modèle de régression linéaire :*

$$\forall i\in[\![1,n]\!],\quad y_i\ =\ a\,x_i+b+e_i$$

est la droite d'équation : $y=a^*\,x+b^*$ où :

$$a^*=\frac{cov(x,y)}{s_x^2}\quad et\quad b^*=\bar{y}_n-a^*\,\bar{x}_n=\bar{y}_n-\frac{cov(x,y)}{s_x^2}\ \bar{x}_n$$

Remarque
La régression linéaire par méthode du maximum de vraisemblance fait l'objet du problème du sujet `HEC 2016` (Partie III). On y retrouve donc toute la démarche de cette section. On y démontre même que (a^*,b^*) est un maximum global de $\mathcal{L}$ sur $\mathbb{R}^2$ (très calculatoire).

II.2.d) Variantes de la régression linéaire

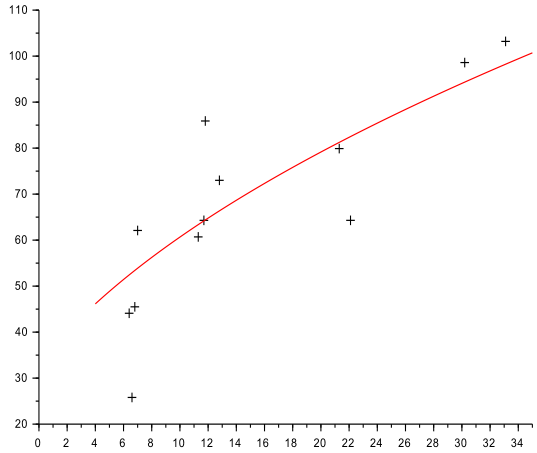
- On peut tout à fait appliquer le principe de régression linéaire à des transformées de la variable y. Cela permet d'envisager des relations de dépendance entre x et y plus diverses.
- Plus précisément, plutôt que de considérer un modèle de régression linéaire classique :
$$\forall i\in[\![1,n]\!],\quad y_i\ =\ a\,x_i+b+e_i$$
On pourra considérer le modèle suivant :
$$\forall i\in[\![1,n]\!],\quad \psi(y_i)\ =\ a\,x_i+b+e_i$$
où ψ est une fonction bijective sur un intervalle I contenant toutes les observations $(y_1,\dots,y_n)$.
On effectue alors une régression linéaire non plus entre les variable x et y, mais entre les variables x et $z=\psi(y)$. Le procédé d'estimation de a et b reste exactement le même.
- Revenons sur l'exemple de l'ozone. D'après le nuage de points tracé en partie **I.2.**, il nous avait semblé que l'on pouvait aussi supposer que la relation de dépendance entre x et y était de la forme :
$$\forall i\in[\![1,n]\!],\quad y_i\approx\sqrt{a\,x_i+b}$$
On considère alors le modèle de régression linéaire :
$$\forall i\in[\![1,n]\!],\quad z_i=y_i^2=a\,x_i+b+e_i$$
- Par la méthode des moindres carrés, on obtient les estimations de a et b suivantes :
$$a^*=\frac{\mathrm{cov}(x,z)}{s_x^2}\quad\text{et}\quad b^*=\bar{z}_n-a^*\,\bar{x}_n=\bar{z}_n-\frac{\mathrm{cov}(x,z)}{s_x^2}\ \bar{x}_n$$
- À l'aide des commandes **Scilab** suivantes, on obtient les estimations des coefficients de la droite de régression entre les variables x et $z=y^2$.

```
z = y.^2
xbar = mean(x)
zbar = mean(z)
s_x2 = variance(x)
a = corr(x, z, 1) / s_x2
b = zbar - a * xbar
```

```
a  =
    258.49757
b  =
    1095.0576
```

- On finit en représentant la courbe de régression sur le nuage de points.

```
1 plot2d(x, y, style = -1)
2 plot(x, sqrt(a * x + b), 'r')
```



Remarque
Notons que dans cet exemple, le gain d'ajustement entre le modèle de régression linéaire et le modèle de régression par racine carrée n'est pas flagrant. On conservera donc le modèle de régression linéaire pour sa facilité d'utilisation.