Outils mathématiques pour la gestion (S2)
Partie 2 : Régression et séries chronologiques

Jérôme Casse

IUT de Sceaux GEA2

1er mars 2024

Programme du cours

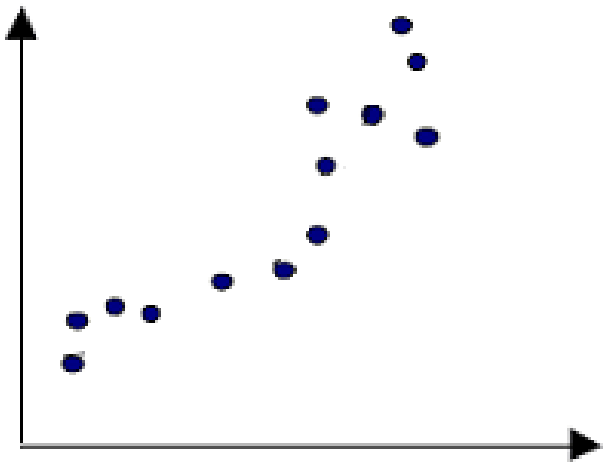
1. Régression

2. Séries chronologiques
   - Choix du modèle
   - Tendance I
   - Saisonalité
   - Tendance II

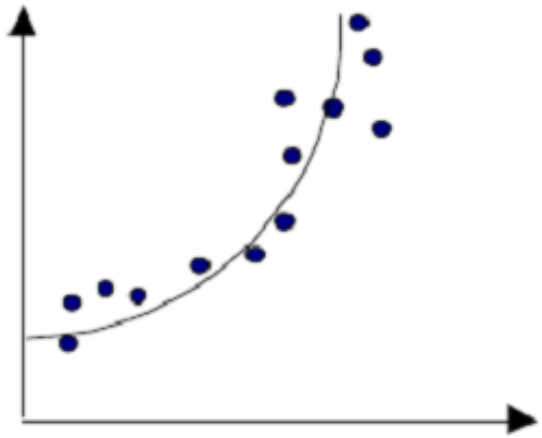
Chapitre 1 : Régression

Motivations

- Comprendre un nuage de points.
- Trouver une fonction pour le modéliser.



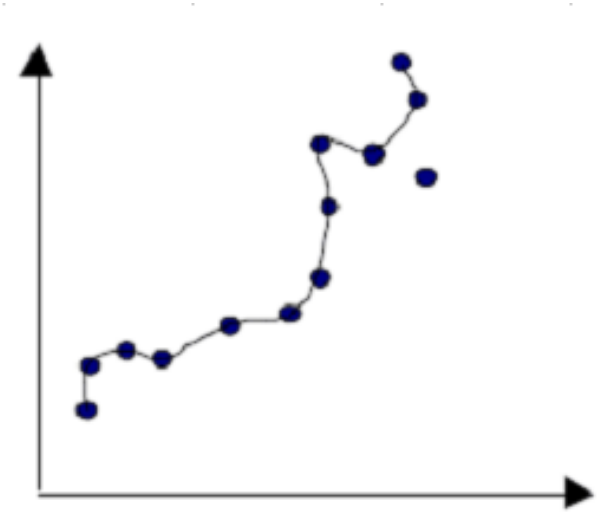
Nuage de point



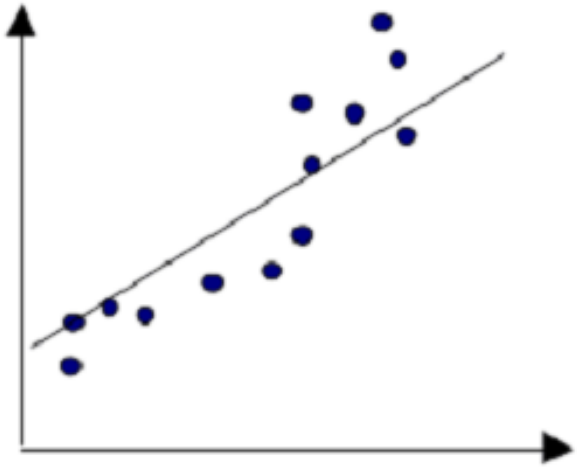
La fonction de régression

Quelle fonction ?

- Ni trop complexe.
- Ni trop simple.



Modèle trop complexe



Modèle trop simple

Éventail de fonctions

Dans le cadre du cours, notre éventail de fonctions est

- $f(x) = a\,x + b$ (affine, aussi appelée les droites).
  2 paramètres à trouver : a et b.

Éventail de fonctions

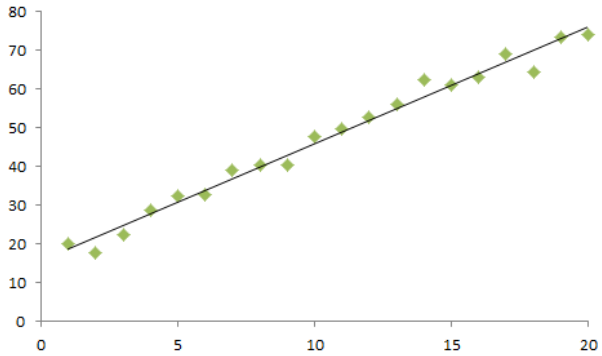
Dans le cadre du cours, notre éventail de fonctions est

- $f(x) = a\,x + b$ (affine, aussi appelée les droites).
  2 paramètres à trouver : a et b.

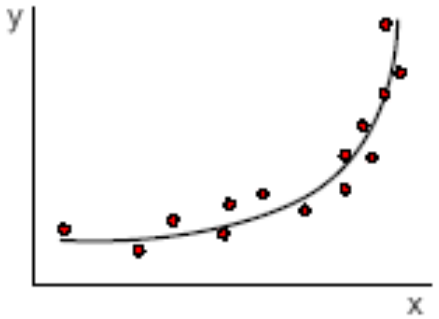
Si on sait faire les droites, on peut faire d'autres fonctions :

- $f(x) = c\,(x-k)^{\alpha} + b$ (monomiale) avec $\alpha \in \mathbb{R}$.
  4 paramètres à évaluer : c, k, α et b.
- $f(x) = c\,e^{\alpha x} + b$ (exponentielle)avec $\alpha \in \mathbb{R}$.
  3 paramètres à évaluer : c, α et b.
- $f(x) = c\,\ln(x-k) + b$ (logarithmique)
  3 paramètres à évaluer : c, k et b.
- Des "par morceaux" de ces fonctions.

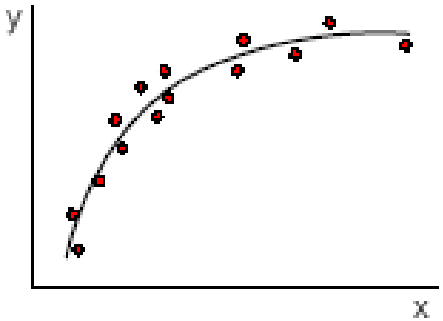
Exemples



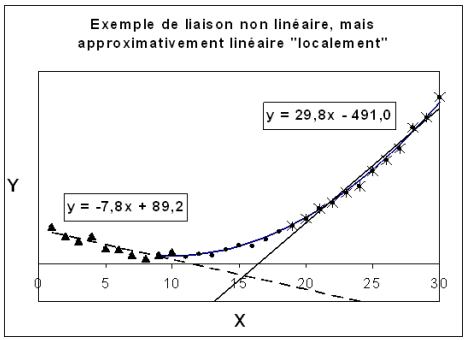
Régression affine



Régression exponentielle



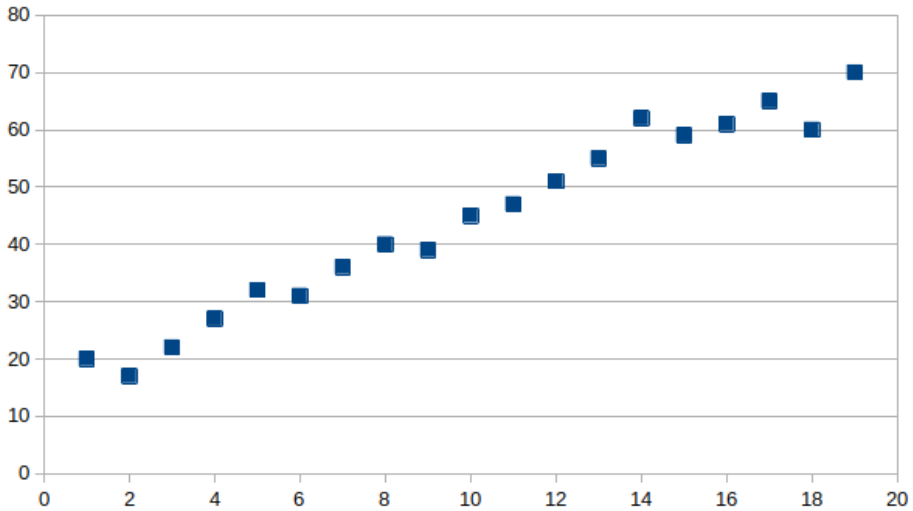
Régression logarithmique



Régression par morceaux

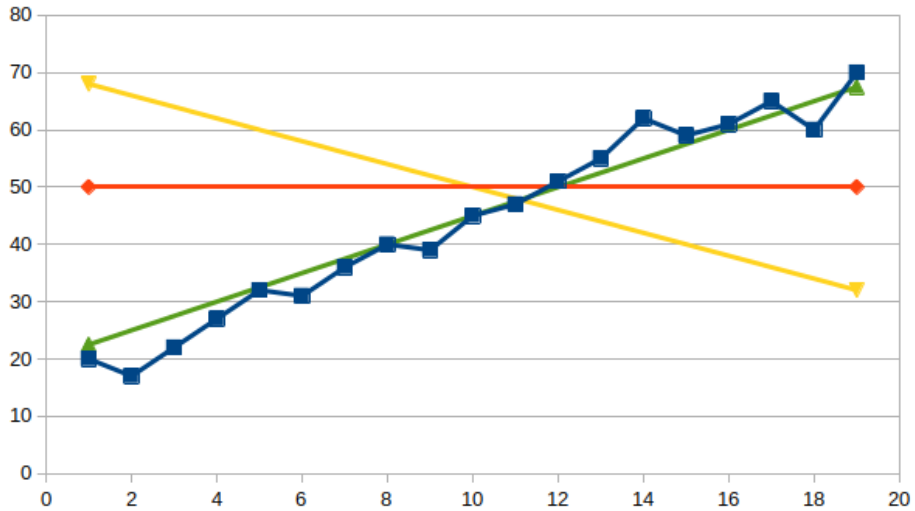
C'est quoi la régression affine ?

1. On a n données $(x_i, y_i)_{i=1...n}$.
   Exemple : "prix des voitures en fonction du revenu du ménage".
   Ménage 1 à $x = 100000$ par an : Mercedes + Clio -> 45000 = y.
   Ménage 2 à $x = 15000$ par an : Twingo d'occasion -> 3000 = y.



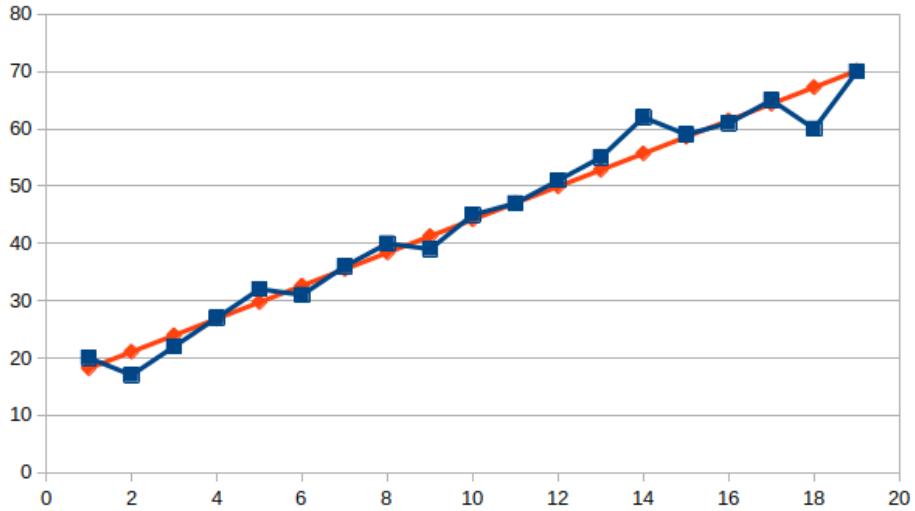
C'est quoi la régression affine ?

1. On a n données $(x_i, y_i)_{i=1\ldots n}$.
2. On a l'ensemble infini des droites affines : $y = ax + b$.



C'est quoi la régression affine ?

1. On a n données $(x_i, y_i)_{i=1\ldots n}$.
2. On a l'ensemble infini des droites affines : $y = ax + b$.
3. But : trouver la droite la plus proche des données.

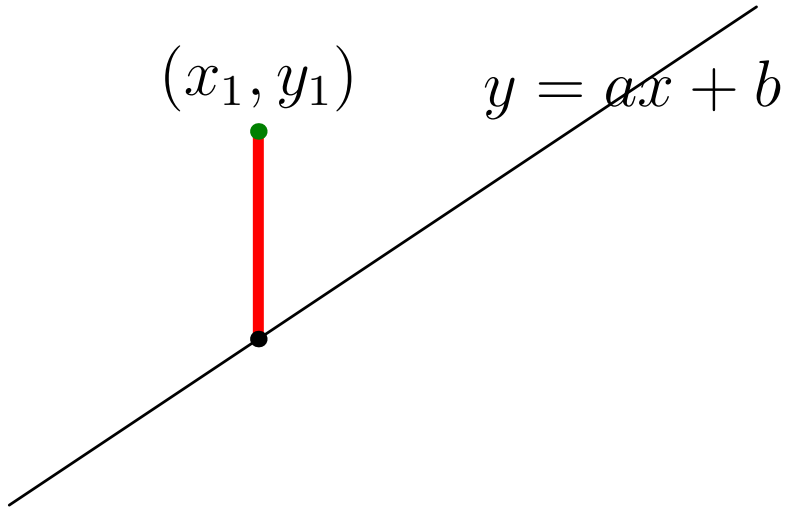


On peut ainsi dire :
En moyenne, un ménage de 50000 par an possède 20000 euros de véhicules.

Comment on calcule cette droite ?

Il faut mathématiser "la droite la plus proche des données".

1. La distance d'un point (x_1, y_1) à la droite $y = ax + b$.

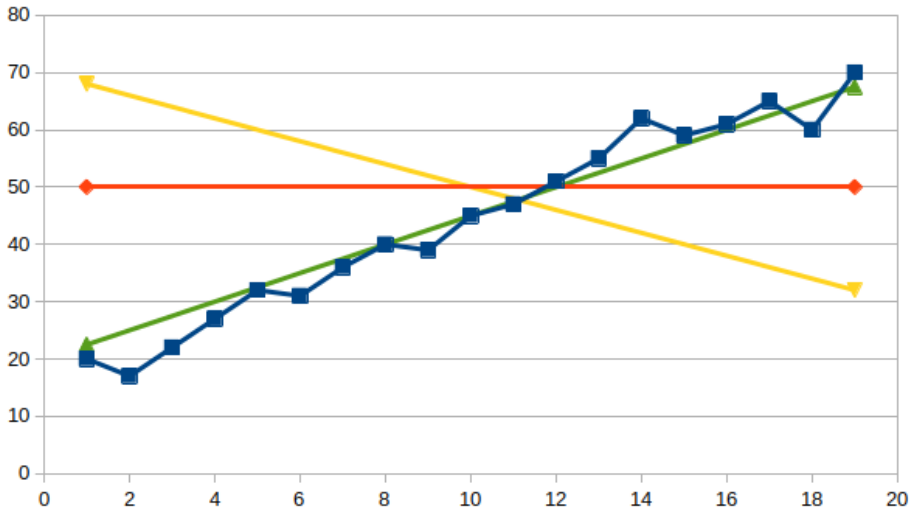


Cette distance vaut $|y_1 - (ax_1 + b)|$.

Comment on calcule cette droite ?

Il faut mathématiser "la droite la plus proche des données".

1. La distance d'un point (x_1, y_1) à la droite $y = ax + b$.
   Cette distance vaut $|y_1 - (ax_1 + b)|$.
2. La distance d'un nuage de n points $(x_i, y_i)_{i=1\dots n}$ à la droite $y = ax + b$.
   La somme des distances au carré :
   $|y_1 - (ax_1 + b)|^2 + |y_2 - (ax_2 + b)|^2 + \cdots + |y_n - (ax_n + b)|^2$.

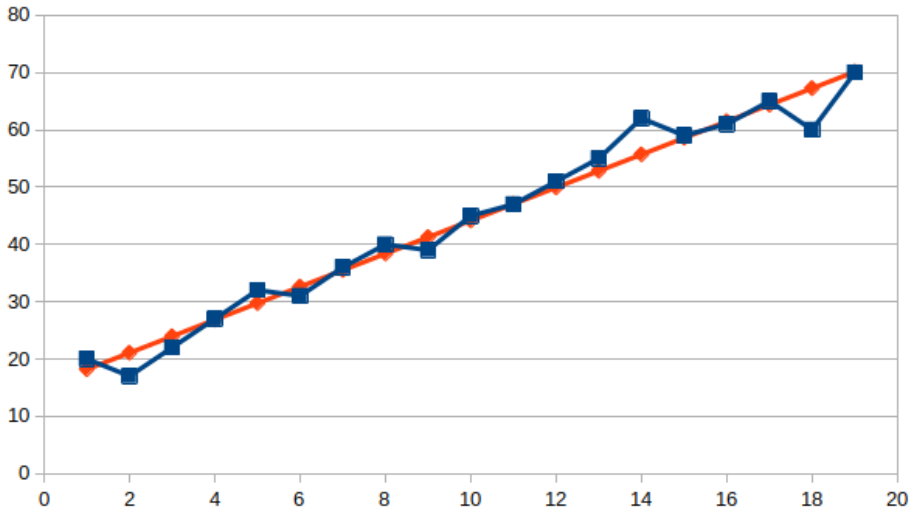


- Distance à rouge : 5535
- Distance à jaune : 14395
- Distance à vert : 237.5

Comment on calcule cette droite ?

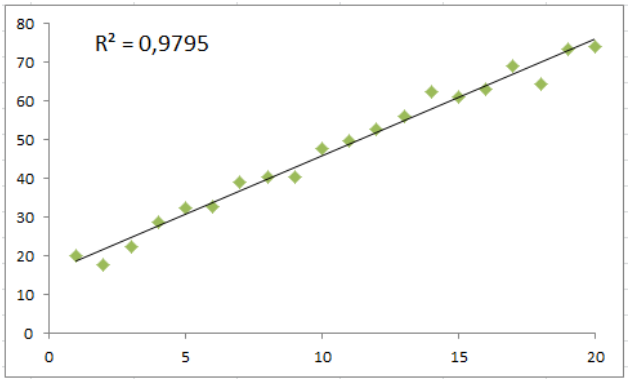
Il faut mathématiser "la droite la plus proche des données".

1. La distance d'un point (x_1, y_1) à la droite $y = ax + b$.
   Cette distance vaut $|y_1 - (ax_1 + b)|$.
2. La distance d'un nuage de n points $(x_i, y_i)_{i=1\dots n}$ à la droite $y = ax + b$.
   La somme des distances au carré :
   $|y_1 - (ax_1 + b)|^2 + |y_2 - (ax_2 + b)|^2 + \cdots + |y_n - (ax_n + b)|^2$.
3. On cherche la droite $y = ax + b$ telle que cette distance soit minimale.

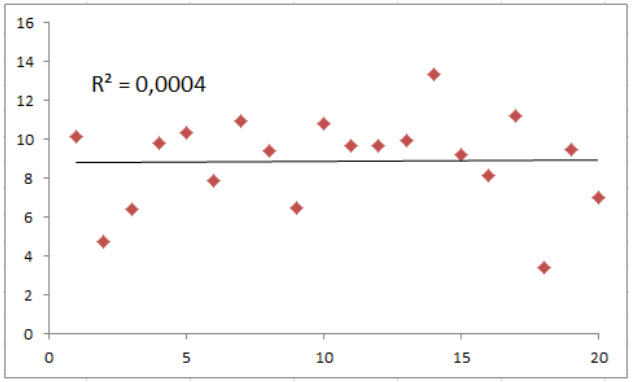


- Distance à rouge : 139.1

À quel point les données sont proche de la droite ?



Coefficient de détermination R proche de 1
Relation affine



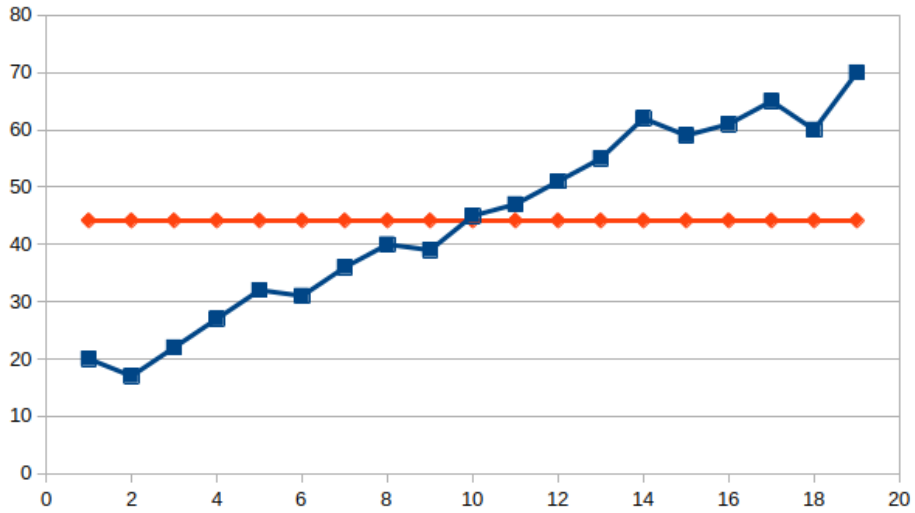
Coefficient de détermination R proche de 0
Pas de relation affine

Rappel du S1 et réinterprétation

- Données $(x_i, y_i)_{i=1\dots n}$.
- Moyenne m des y : $m = \frac{y_1+\dots+y_n}{n}$.
- Variance des y : $\text{Var}(y) = \frac{\sum_{i=1}^{n}(y_i-m)^2}{n}$.

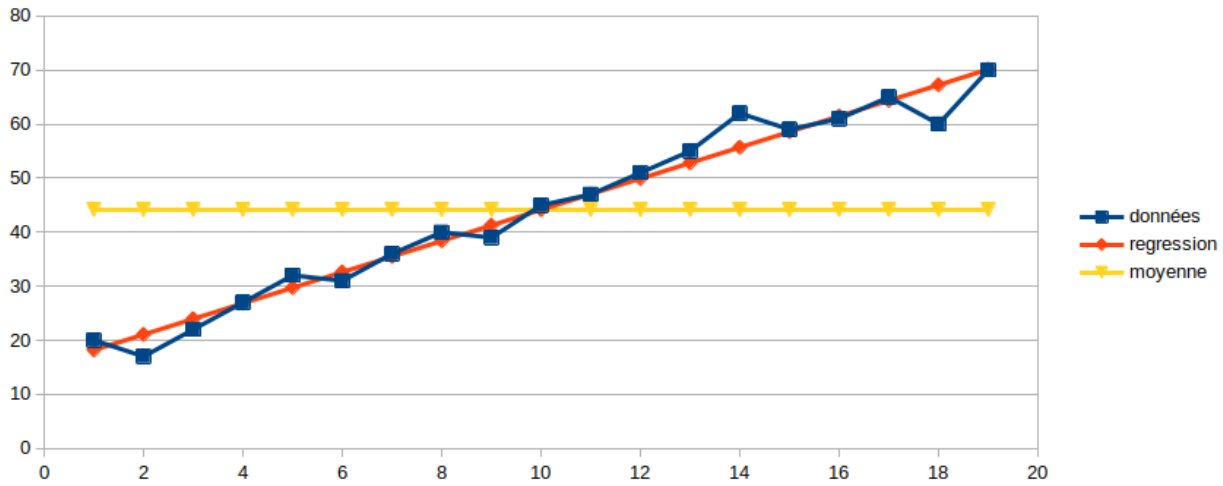
Rappel du S1 et réinterprétation

- Données $(x_i, y_i)_{i=1\dots n}$.
- Moyenne m des y : $m = \frac{y_1+\dots+y_n}{n}$.
- Variance des y : $\text{Var}(y) = \frac{\sum_{i=1}^{n}(y_i-m)^2}{n}$.



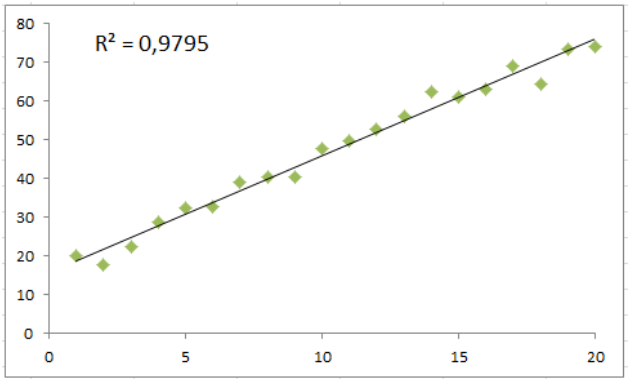
- Parmi les droites horizontales $y = b$, la meilleure est $y = m$.
- La distance de $y = m$ aux données est $n\text{Var}(y)$.

Coefficient de détermination

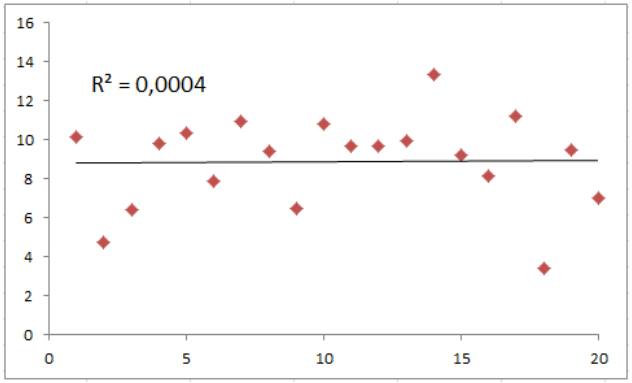


- La distance entre les données et la droite de régression :
  SCR, la variance résiduelle.
- La distance entre les données et la moyenne :
  SCT, la variance totale.
- "La distance entre la moyenne et la droite de régression" :
  SCE = SCT-SCR, la variance expliquée (par la connaissance de x).
- Coefficient de détermination : $R^2 = \frac{\text{SCE}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}}$.

Exemple



R proche de 1
Relation linéaire



R proche de 0
Pas de relation linéaire

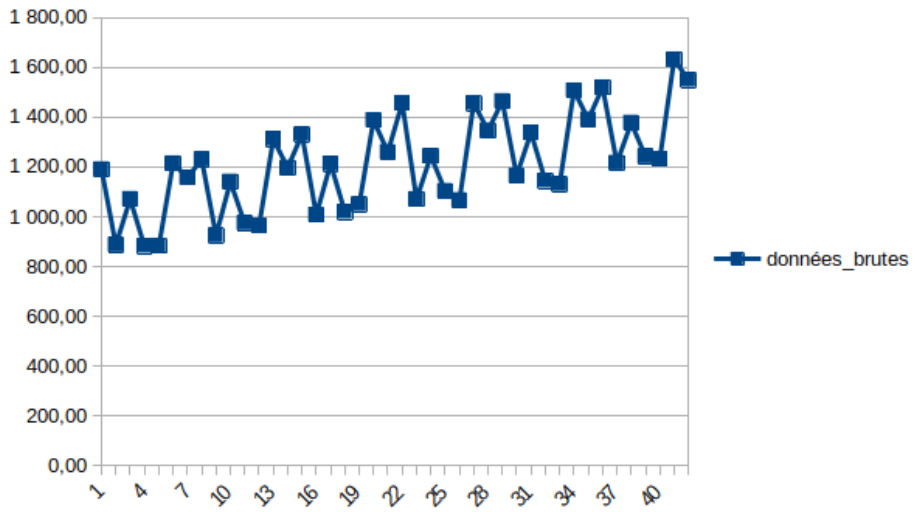
Commentaires

- Excel va faire cela pour nous.
- Pourquoi la somme des carrées des distances $\sum_i |y_i - (ax_i + b)|^2$ et pas la somme des distance $\sum_i |y_i - (ax_i + b)|$?
  - C'est la méthode des moindres carrés, canonique.
  - Faisable avec $\sum_i |y_i - (ax_i + b)|^\alpha$ si $\alpha > 0$.
  - Mais, avec $\alpha = 2$, les maths sont plus "simples".
  - Lien avec la variance, etc.
- Un exemple très simple à la main sera donné dans un TD ?
- Pour les fonctions non affine, on montrera en TP comment se ramener au cas affine.

Chapitre 2 : Séries chronologiques.

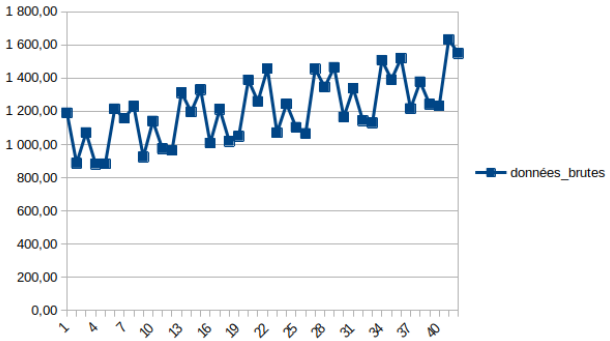
Motivations

- Comprendre une variable (aléatoire) qui dépend du temps.
- Analyser la tendance longue et sa saisonnalité.
- Prédire le futur.



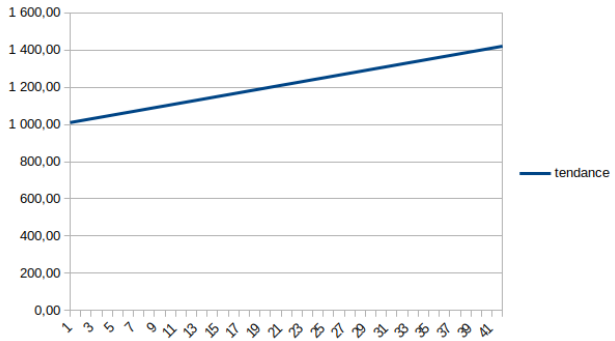
Données en fonction du temps.

Finalité (WoW)



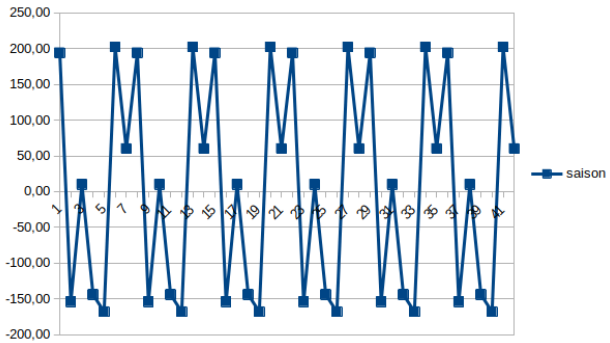
Données brutes

=



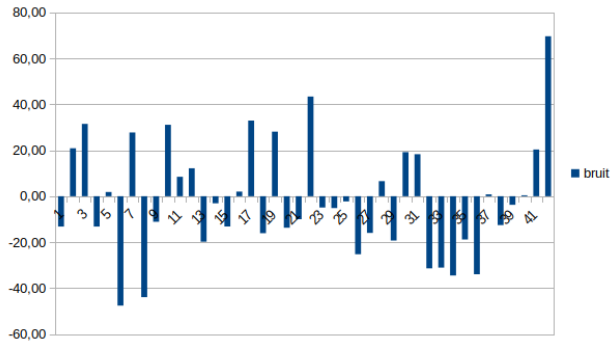
Tendance longue

+



Saisonalité

+



Bruit

Interprétation

Exemple : nombre de personnes sur une plage mois par mois.

- Tendance : attractivité du lieu.
- Saisonalité : quantifier la différence entre les mois.
  - Phénomènes sur 12 mois : $S(\text{janvier})$, $S(\text{février})$, ..., $S(\text{décembre})$.
  - On s'attend à $S(\text{février}) < S(\text{août})$.
  - On va le voir et le quantifier.
- Bruit : choses très éphémères ou aléatoires.
  - Jour de pluie.
  - Les février bissextiles (29 jours) et non bissextiles (28 jours).

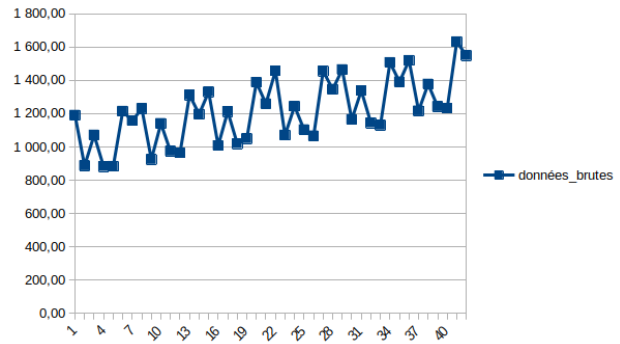
Comment fait-on ?

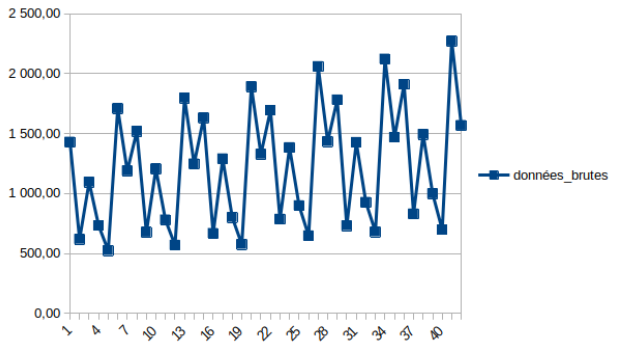
1. Choix d'une modélisation parmi 2 (dans ce cours).
2. Pré-étude de la tendance en temps long.
3. Étude de la saisonalité.
4. Étude de la tendance en temps long.
5. Étude du bruit (pour le S5).

2.A : Choix du modèle

Les deux modèles possibles (dans ce cours)

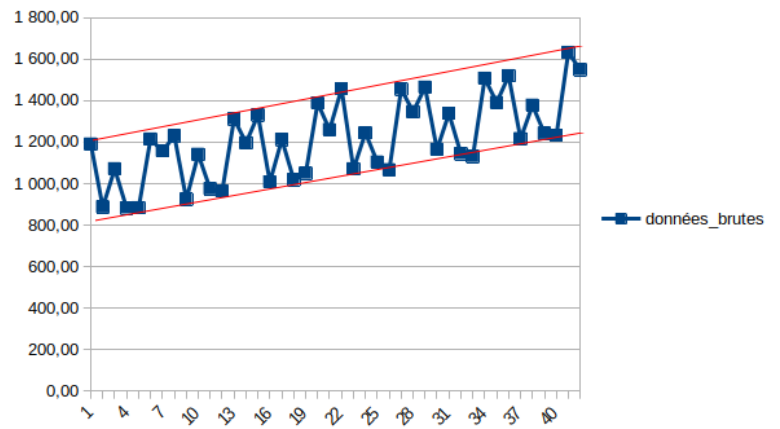
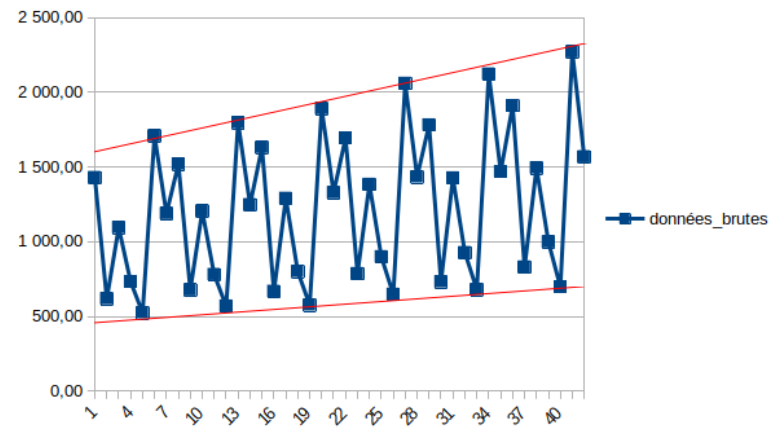
Modèle additif	Modèle mixte
$X(t) = T(t) + S(t) + \epsilon(t)$	$X(t) = T(t) \times S(t) + \epsilon(t)$
$T(t)$ tendance (en temps long)	
$S(t)$ saisonalité, périodique de période τ	
$\frac{1}{\tau}\sum_{i=1}^{\tau} S(i) = 0$	$\frac{1}{\tau}\sum_{i=1}^{\tau} S(i) = 1$
$\epsilon(t)$ bruit aléatoire (bruit blanc gaussien)	





Comment choisir entre les deux ?

- Droite qui passe par les maxima.
- Droite qui passe par les minima.
- Sont-elles à peu près parallèles ?

OUI	NON
	
Modèle additif	Modèle mixte

2.B : Pré-étude de la tendance

Moyenne mobile arithmétique

Exemple : CA par trimestre

Date	1^{er} 22	2^{e} 22	3^{e} 22	4^{e} 22	1^{er} 23	2^{e} 23	3^{e} 23	4^{e} 23
Données	10	25	55	30	12	25	58	29
MM								

- La moyenne mobile, c'est la moyenne autour du trimestre sur un an.
- Pour le 3^{e} 22, c'est
  - la seconde moitié du 1^{er} 22,
  - le 2^{e} 22,
  - le 3^{e} 22,
  - le 4^{e} 22,
  - la première moitié du 1^{er} 23.
- La moyenne mobile : $M(3^{e}22) = \frac{1}{4}\left(\frac{10}{2} + 25 + 55 + 30 + \frac{12}{2}\right) = 30.25$.

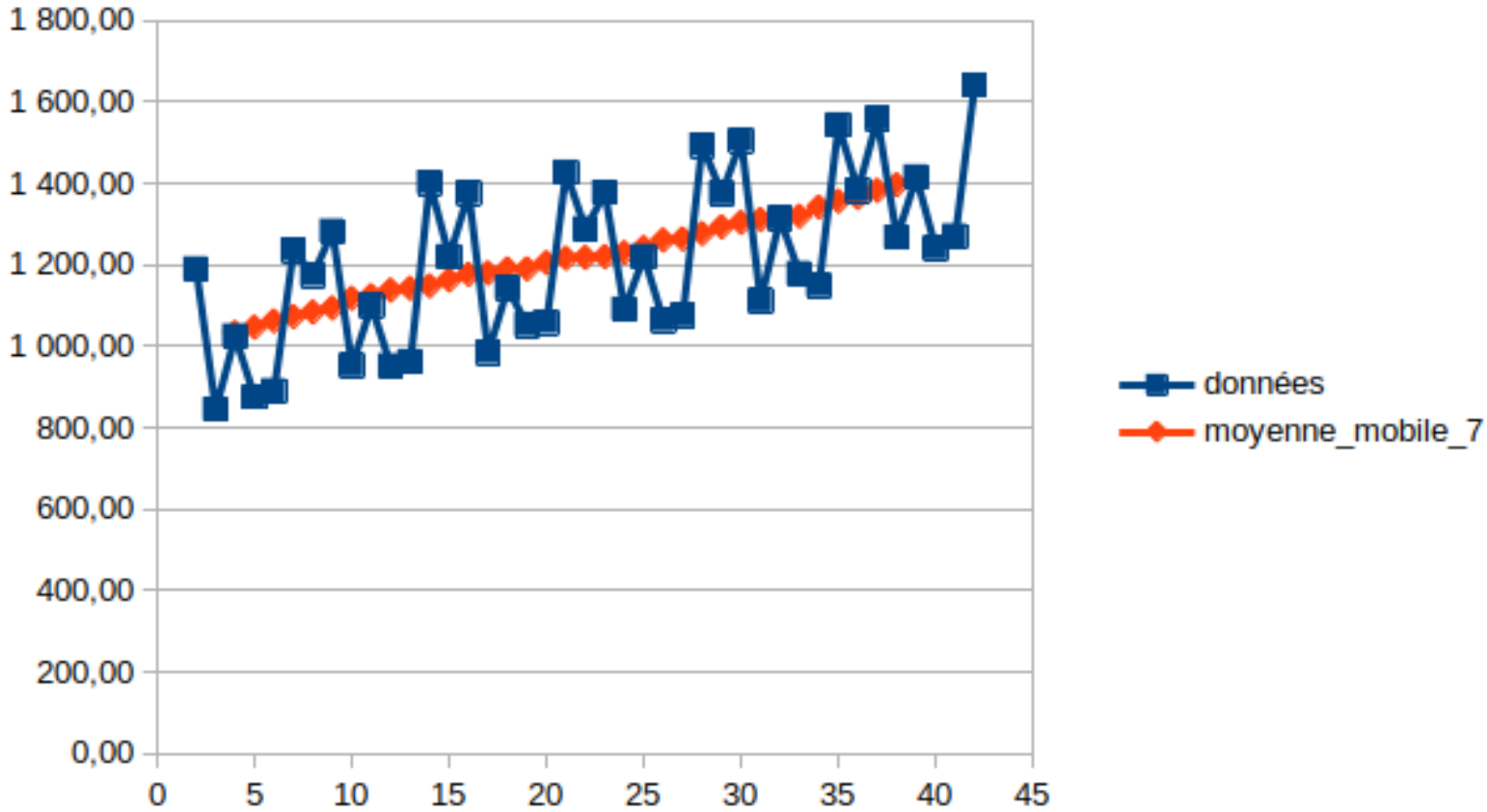
Moyenne mobile arithmétique

Exemple : CA par trimestre

Date	1er 22	2e 22	3e 22	4e 22	1er 23	2e 23	3e 23	4e 23
Données	11	25	55	30	12	25	58	29
MM			30.25	30.5	30.875	31.125		

- La moyenne mobile, c'est la moyenne autour du trimestre sur un an.
- Pas calculable pour les 2 premiers trimestres et les 2 derniers.
  - Pour le 2e 22, il faut connaître (la moitié du) 4e 21.
- Fluctuations moindres que les données.
- Ne dépend plus de la saisonalité.

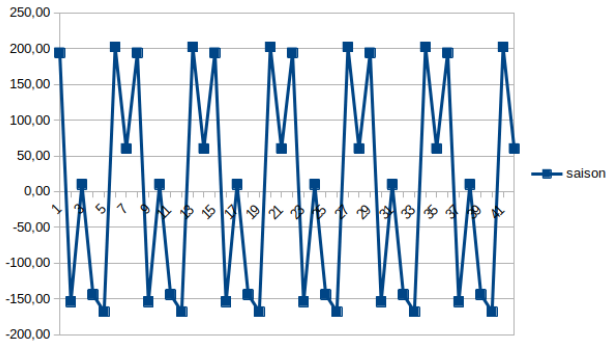
Exemple


1 800,00
1 600,00
1 400,00
1 200,00
1 000,00
800,00
600,00
400,00
200,00
0,00
0
5
10
15
20
25
30
35
40
45
données
moyenne_mobile_7

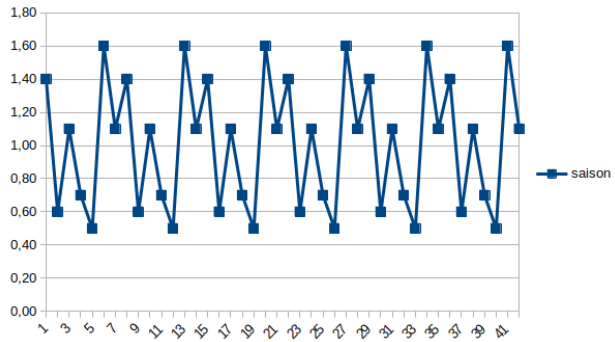
2.C : Saisonalité

La saisonalité

- Périodique : période τ
  1 semaine = 7 jours ; 1 année = 12 mois ; 1 année = 4 trimestres.
- Exemple : $S(\text{lundi})$, $S(\text{mardi})$, $S(\text{mercredi})$, $S(\text{jeudi})$, $S(\text{vendredi})$, $S(\text{samedi})$, $S(\text{dimanche})$



Modèle additif



Modèle mixte

- "On vend en moyenne pour 200 euros de plus le lundi que la moyenne de la semaine" (modèle additif).
- "On vend en moyenne 40% de plus le lundi et 40% de moins le mardi que la moyenne de la semaine" (modèle mixte).

Trouver la saisonalité : cas additif

1. Regarder $X(t) - M(t)$: données - moyenne mobile.
2. Faire la moyenne (ou la médiane) de $X(t) - M(t)$ sur tous les 1^{er} trimestre (pour notre exemple).
   Noter $S^*_{1^{er}}$ (pre-saisonalité).
3. Idem pour les autres trimestres (pour notre exemple).
4. Faire la somme $S^* = \frac{1}{4}\left(S^*_{1^{er}} + \cdots + S^*_{4^{e}}\right)$.
   Rappel : on veut $S^* = 0$.
5. Si $S^* \neq 0$, $S_i = S^*_i - S^*$.
   En mot : on soustrait à chaque coefficient, sa part du surplus ou sousplus.

Exemple

1. Différence $X(t) - M(t)$:

Date	1^{er} 22	2^{e} 22	3^{e} 22	4^{e} 22	1^{er} 23	2^{e} 23	3^{e} 23	4^{e} 23
X(t)	10	25	55	30	12	25	58	29
M(t)			30.2	30.5	30.9	31.1		
X(t)-M(t)			24.7	-0.5	-18.9	-6.1		

2. Moyenne ou médiane (ici : une seule donnée), donc :
   $S^*_{1^{er}} = -18.9$, $S^*_{2^{e}} = -6.1$, $S^*_{3^{e}} = 24.7$ et $S^*_{4^{e}} = -0.5$.
3. $S^* = \frac{1}{4}(-18.9 - 6.1 + 24.8 - 0.5) = -0.2$.
4. La saisonalité est :
   $S^*_{1^{er}} = -18.7$, $S^*_{2^{e}} = -5.9$, $S^*_{3^{e}} = 24.9$ et $S^*_{4^{e}} = -0.3$.

Trouver la saisonalité : cas mixte

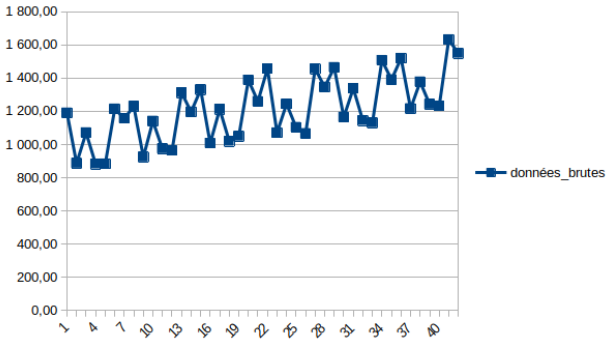
1. Regarder X_t / T_t.
2. Faire la moyenne (ou la médiane) de X_t / T_t sur tous les "lundis" (pour notre exemple).
   Noter S^*_{lundi}.
3. Idem pour les autres jours (pour notre exemple).
4. Faire la somme $S^* = \frac{1}{7}\left(S^*_{\text{lundi}} + \cdots + S^*_{\text{dimanche}}\right)$.
   Rappel : on veut que cette somme vaille 1.
5. Si $S^* \neq 1$, poser $S_i = S_i^* / S^*$.
   En mot : on divise chaque coefficient pour que la somme vaille 1.

2.D : Tendance (en temps long)

1. Calcul de $C(t) = X(t) - S(t)$ (modèle additif)
   ou $C(t) = X(t)/S(t)$ (modèle mixte) :
   correction des données par la saisonalité.
2. Régression sur ces données corrigées $(t, C(t))$ donne $T(t)$.
3. Le bruit : $\epsilon(t) = X(t) - S(t) - T(t)$,
   ou $\epsilon(t) = X(t) - T(t)S(t)$.

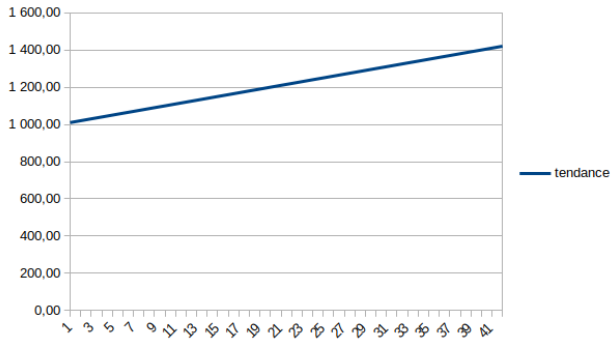
Date	1er 22	2e 22	3e 22	4e 22	1er 23	2e 23	3e 23	4e 23
X(t)	10	25	55	30	12	25	58	29
M(t)			30.2	30.5	30.9	31.1		
S(t)	-18.7	-5.9	24.9	-0.3	-18.7	-5.9	24.9	-0.3
C(t)	28.7	30.9	30.1	30.3	30.7	30.9	33.1	29.3

Fin

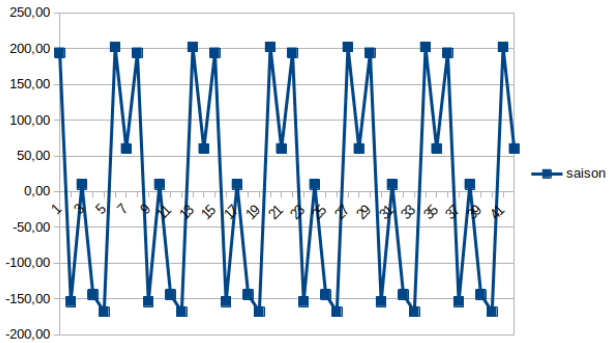


Données brutes

=

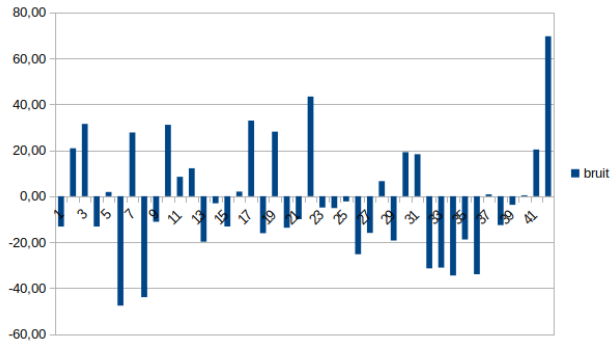


Tendance longue



\+ Saisonalité

\+



Bruit