Séries chronologiques

Agnès LAGNOUX

lagnoux@univ-tlse2.fr

ISM-AG 1 - MI0B702T

PLAN DU COURS

1. Introduction aux séries chronologiques (S.C.)
   - Vocabulaire, exemples et objectifs
   - Description schématique de l'étude complète d'une S.C.
2. Modélisation déterministe
   - Modèles additif, multiplicatif, mixte
3. Moyennes mobiles
4. Décomposition d'une S.C.

CHAPITRE I
Introduction aux séries chronologiques

Introduction aux séries chronologiques

On s'intéresse à l'évolution au cours du temps d'un phénomène.
On souhaite DÉCRIRE, EXPLIQUER et PRÉVOIR ce phénomène au cours du temps.

On dispose d'observations à des dates différentes, c-à-d d'une suite de valeurs numériques indicées par le temps.

Exemples : - évolution du nombre de voyageurs utilisant le train,
- accroissement relatif mensuel de l'indice des prix,
- occurence d'un phénomène naturel (comme le nombre de taches solaires).

Cette suite d'observations d'une famille de variables aléatoires réelles notées $(X_t)_{t\in\Theta}$ est appelée série chronologique ou série temporelle.

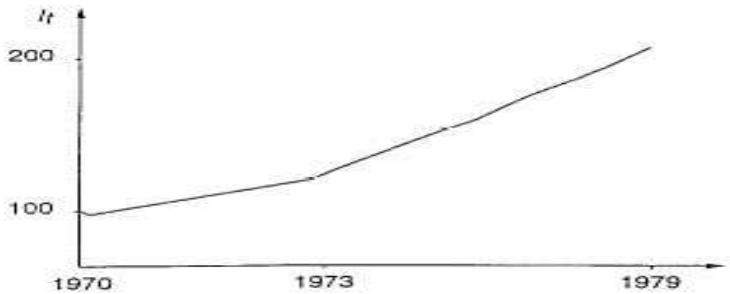
L'ensemble Θ est appelé espace des temps et peut être

- **discret** : $\Theta \subset \mathbb{Z}$. Les dates d'observation sont le plus souvent équidistantes (relevés mensuels, trimestriels...) et on dispose d'un nombre fini T d'observations. Si h est l'intervalle de temps séparant deux observations et t_0 l'instant de la première observation, on a le schéma suivant

$$\begin{array}{ccccc} t_0 & t_0+h & \dots & t_0+(T-1)h \\ \downarrow & \downarrow & \dots & \downarrow \\ X_{t_0} & X_{t_0+h} & \dots & X_{t_0+(T-1)h} \\ \downarrow & \downarrow & \dots & \downarrow \\ X_1 & X_2 & \dots & X_T \end{array}$$

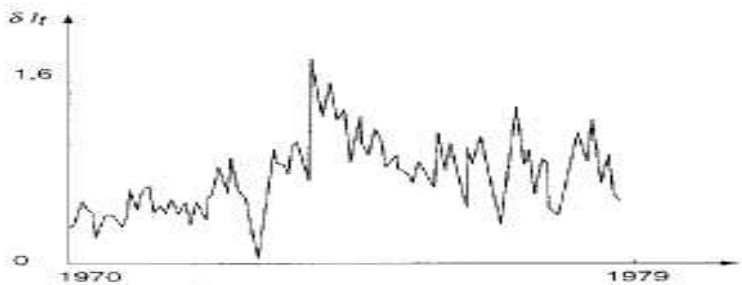
- **continu** : $\Theta \subset \mathbb{R}$ et on dispose (au moins potentiellement) d'une infinité d'observations issues d'un processus $(X_t)_{t\in\Theta}$. Les méthodes présentées dans ce cadre sont différentes de celles pour les séries chronologiques à temps discret et présentées dans la suite.

Des exemples (I)

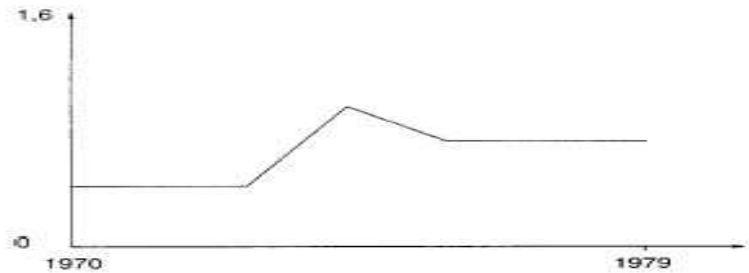


(a) Indice mensuel des prix à la consommation I_t.

(b) Trafic voyageur de la SNCF en 2ième classe

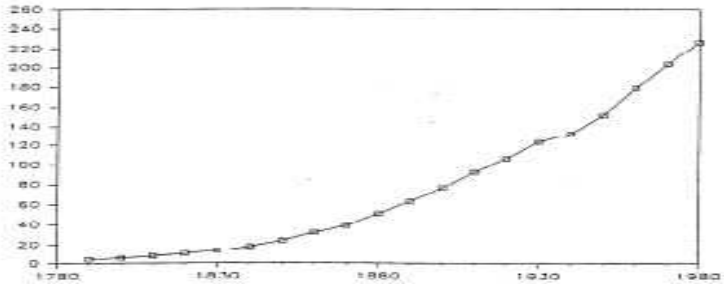


(c) Accroissement relatif mensuel de l'indice des prix

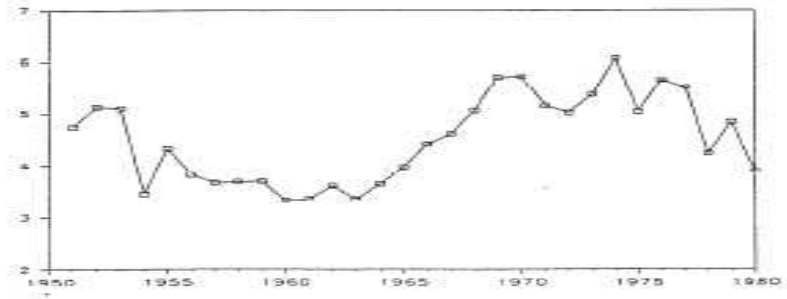


(d) Évolution à moyen terme de l'accroissement relatif mensuel de l'indice des prix

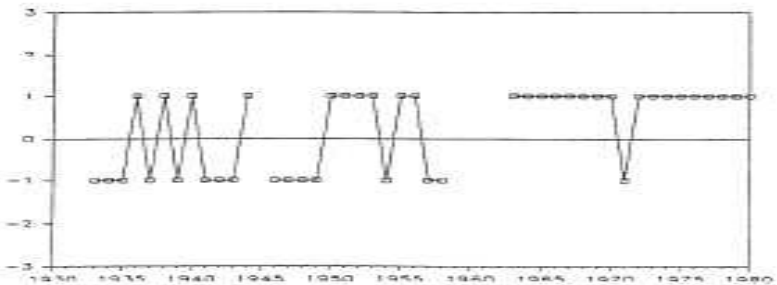
Des exemples (II)



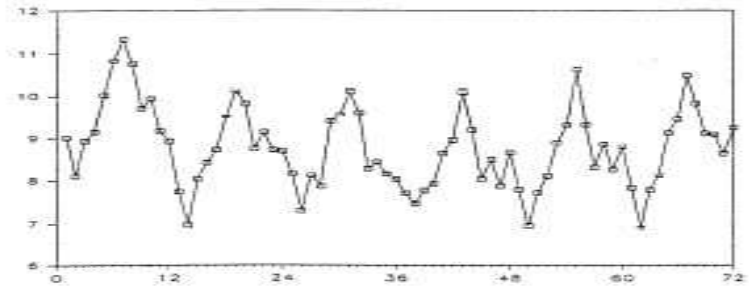
(a) Population des Etats-Unis, 1790–1980



(b) Nombre de grêves aux Etats-Unis, 1951-1980



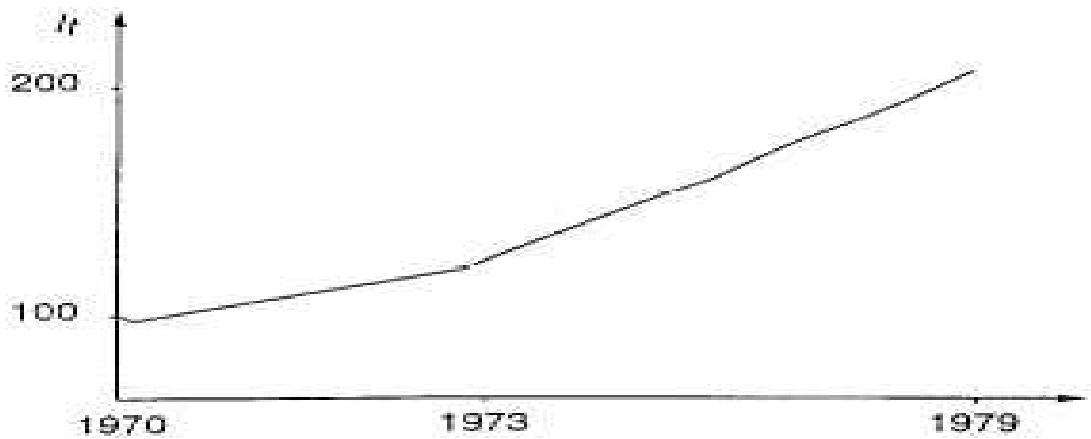
(c) "All Star Baseball Games", 1933-1980



(d) Nombre mensuel de décès accidentels aux USA, 1973-1978

Description d'une SC

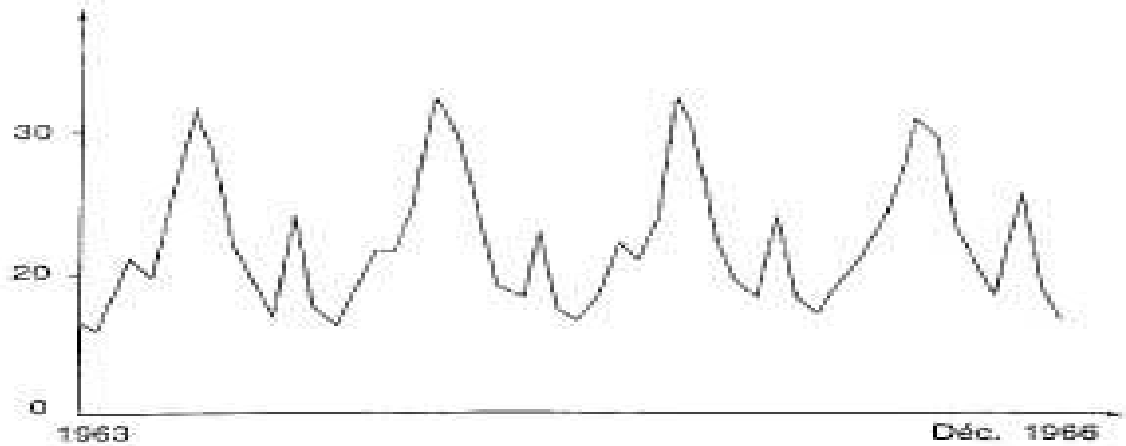
L'observation graphique de la série est souvent une aide et permet de se faire une idée de ses différentes composantes.



(a) Indice mensuel des prix à la consommation I_t.

Description d'une SC

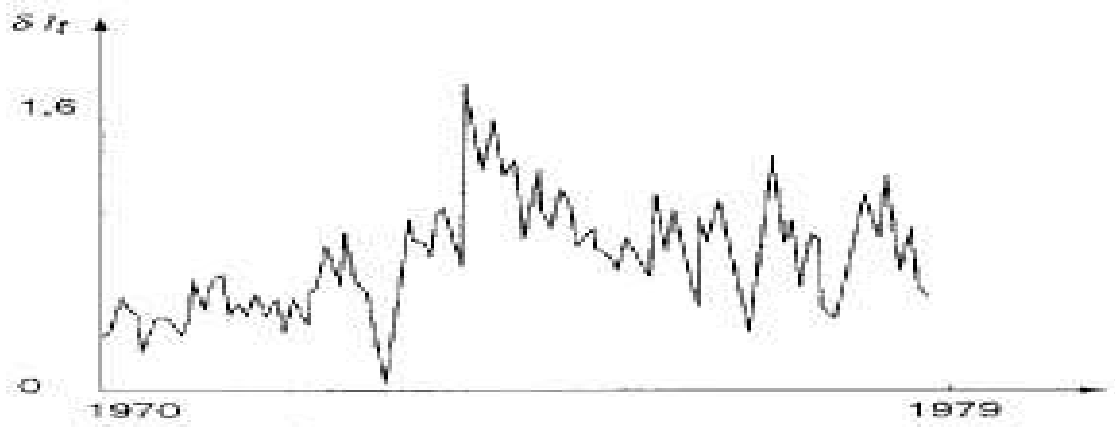
L'observation graphique de la série est souvent une aide et permet de se faire une idée de ses différentes composantes.



(b) Trafic voyageur de la SNCF en 2ième classe

Description d'une SC

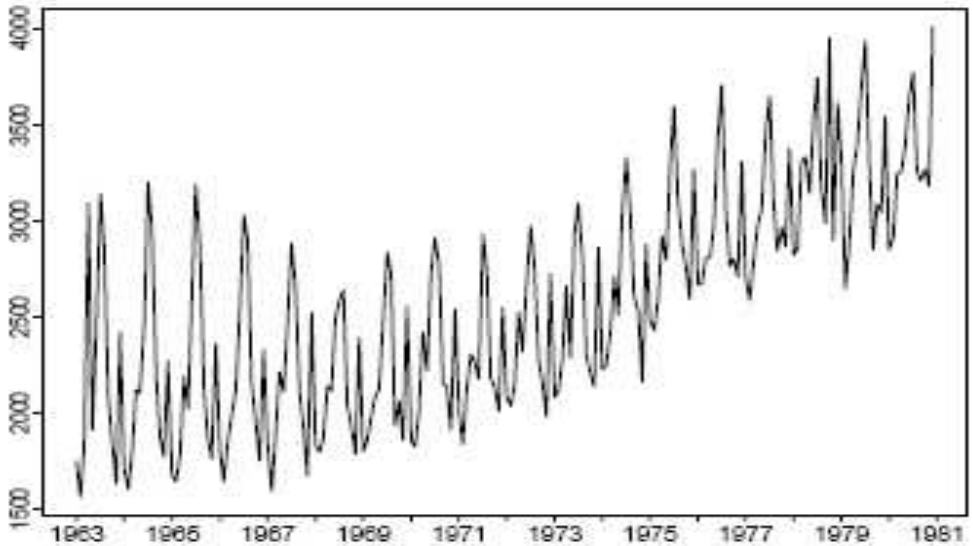
L'observation graphique de la série est souvent une aide et permet de se faire une idée de ses différentes composantes.



(c) Accroissement relatif mensuel de l'indice des prix

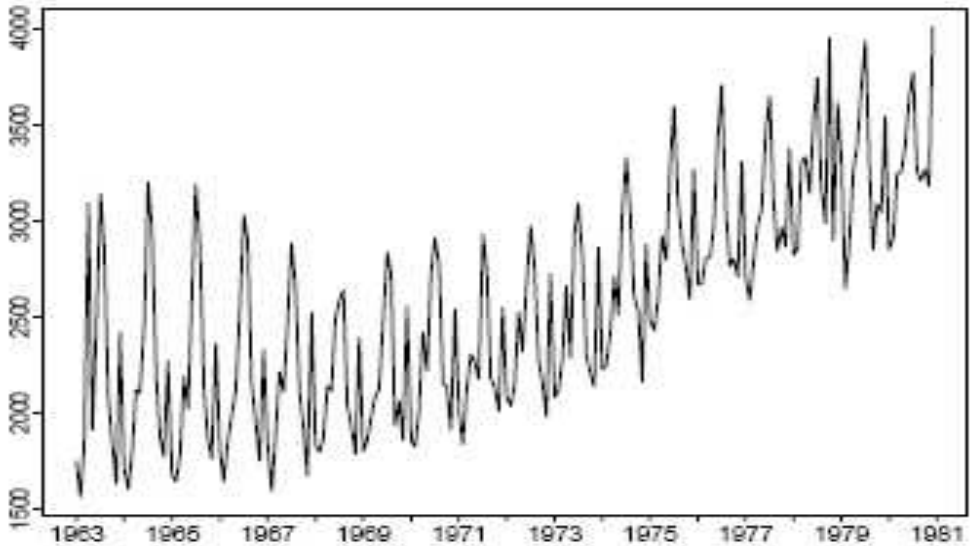
Description d'une SC

Un autre exemple : évolution du trafic voyageur SNCF de 1960 à 1980



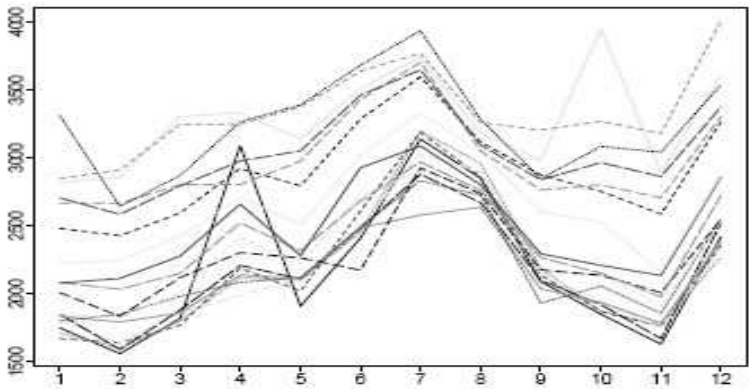
Description d'une SC

Un autre exemple : évolution du trafic voyageur SNCF de 1960 à 1980



Nous remarquons

Description d'une SC



Description d'une SC

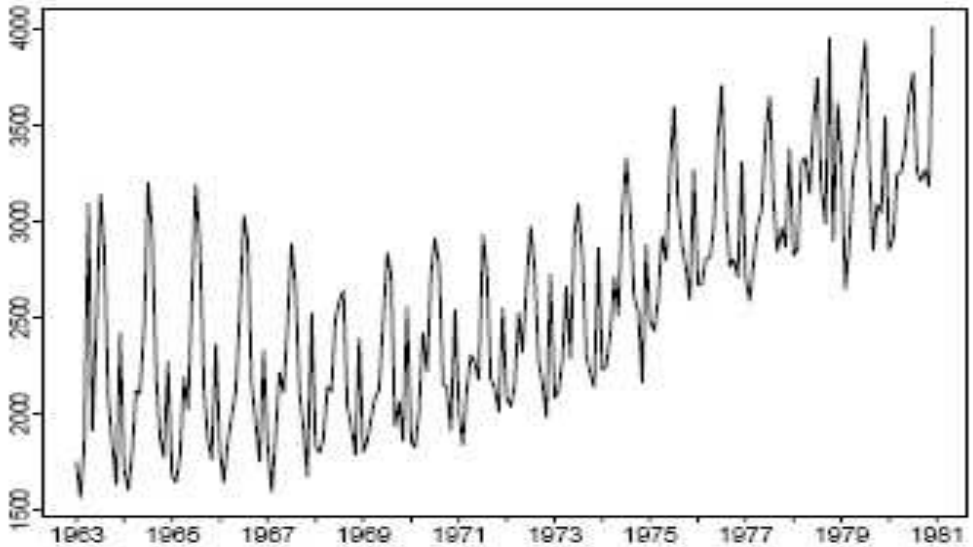


Figure – Évolution du trafic voyageur SNCF de 1960 à 1980

Description d'une SC

On considère qu'une série chronologique (X_t) est la résultante de différentes composantes fondamentales :

1. la tendance (ou trend) (Z_t) représente l'évolution à long terme de la série et traduit son comportement "moyen".
2. la composante saisonnière (ou saisonnalité) (S_t) correspond à un phénomène qui se répète à intervalles de temps réguliers (périodiques).
3. la composante résiduelle (ou bruit ou résidu) (ϵ_t) correspond à des fluctuations irrégulières, de faible intensité mais de nature aléatoire.
4. des phénomènes accidentels (grèves, météo exceptionnelle, crash financier) peuvent notamment intervenir.
5. un phénomène cyclique : c'est souvent le cas en climatologie et en économie (ex : récession et expansion...). Il s'agit d'un phénomène se répétant mais sur des durées qui ne sont pas fixes ($\neq$ saisonnalité) et généralement longues.

Objectifs

1. La description : analyser, décrire un phénomène au cours du temps et en tirer les conséquences : par exemple pour des prises de décision (marketing, ...).
2. Le contrôle : contrôle pour la gestion de stocks, contrôle d'un processus chimique...
3. La détection de rupture : il arrive souvent qu'une série chronologique soit affectée par la survenue d'événements accidentels (grèves, changement de législation, catastrophe climatique). Ces interventions vont parfois modifier brutalement la tendance de la série se traduisant par des données aberrantes.
4. La prévision : ayant observé X_1, X_2, ... , X_T, on veut prédire les valeurs futures X_{T+1}, X_{T+2}, ... Nous utiliserons essentiellement des modèles de décomposition pour faire de la prévision. Les méthodes de ce cours visent à faire des prévisions à court terme et à proposer des modélisations pour les $\neq$ composantes de la SC (tendance, saisonnalité et résidu).

Description schématique de l'étude complète d'une SC

1. correction des données
2. observation de la série
3. modélisation (avec un nombre fini de paramètres)
4. analyse de la série à partir de ses composantes
5. diagnostic du modèle
6. prédiction (= prévision)

Description schématique de l'étude complète d'une SC

Etape 1 - Correction des données = prétraitement des données

- évaluation des données manquantes, remplacement des données accidentelles
- découpage en sous-séries
- standardisation afin de se ramener à des intervalles de lg fixe
- transformation des données (ex. : en économie, on utilise les transf. de Box-Cox $Y_t = \frac{1}{\lambda}\left[(X_t)^{\lambda} - 1\right], \quad \lambda \in \mathbb{R}*.$)

Description schématique de l'étude complète d'une SC

Etape 2 - Observation de la série

- Une règle générale en Statistique Descriptive consiste à commencer par regarder les données avant d'effectuer le moindre calcul : une fois la série corrigée et prétraitée, on trace son graphique c-à-d la courbe de coordonnées (t, X_t).
- L'observation de ce graphique est souvent une aide à la modélisation de la série chronologique et permet de se faire une idée des différentes composantes de la série (tendance, saisonnalité, résidu).
- Lorsqu'on veut mettre en évidence le phénomène de saisonnalité à l'aide d'un graphique, on découpe la série en sous-séries de longueur de période du saisonnier et on les représente sur un même graphique.

Description schématique de l'étude complète d'une SC

Etape 3 - Modélisation (I)

- Un modèle est une image simplifiée de la réalité qui vise à traduire les mécanismes de fonctionnement du phénomène étudié et permet de mieux les comprendre.
- Un modèle peut être meilleur qu'un autre pour décrire la réalité. Plusieurs questions se posent alors :
  - comment mesurer cette qualité ?
  - comment diagnostiquer un modèle ?
- Nous présentons dans la suite une petite liste qui sert à résumer et classifier les différents modèles envisagés dans ce cours.

Description schématique de l'étude complète d'une SC

Etape 3 - Modélisation (II)

Les modèles déterministes : l'observation de la série en t est donnée par

$$X_t = f(t, \epsilon_t)$$

où les ϵ_t sont décorrélées.

Cas les plus usités :

- le modèle additif : $X_t = Z_t + S_t + \epsilon_t$.
- le modèle multiplicatif : $X_t = Z_t(1 + S_t)(1 + \epsilon_t)$.
- le modèle mixte : $X_t = Z_t(1 + S_t) + \epsilon_t$.

Description schématique de l'étude complète d'une SC

Etape 3 - Modélisation (III)

Les modèles stochastiques : du même type que les modèles déteministes à ceci près que les ϵ_t sont corrélées et sont une fonction des valeurs passées ($\pm$ lointaines suivant le modèle) et d'une erreur η_t :

$$\epsilon_t = g(\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \eta_t).$$

La modélisation porte ici sur la forme du processus (ϵ_t).

Cas les plus usités : g linéaire (ex. modèles autorégressifs $\epsilon_t = a_1\epsilon_{t-1} + a_2\epsilon_{t-2} + \eta_t$.)

Les deux types de modèles ci-dessus induisent des techniques de prévision bien particulières.
Dans ce cours, nous n'étudierons que les modèles déterministes.
Les modèles stochastiques seront vus au second semestre en renforcement.

Description schématique de l'étude complète d'une SC

Etape 4 - Analyse de la série à partir de ses composantes

Schématiquement,

- on s'interesse tout d'abord à la tendance et à la saisonnalité.
- on les isole : ces deux opérations s'appellent la détendancialisation et la désaisonnalisation de la série.
- on les modélise.
- on les élimine de la série.
- on obtient la série des résidus qui est
  - décorrélée dans les modèles déterministes et il n'y a plus rien à faire.
  - stationnaire (on l'espère du moins) dans les modèles stochastiques et on l'étudie à part.

Description schématique de l'étude complète d'une SC

Etape 5 - Diagnostic du modèle

Une fois le modèle construit et ses paramètres estimés, on vérifie que le modèle proposé est bon c'est-à-dire l'ajustement au modèle :

- en étudiant les résidus
- en faisant des tests
- ...

Description schématique de l'étude complète d'une SC

Etape 6 - Prévision, prédiction

- Une fois ces différentes étapes réalisées, nous sommes en mesure de faire de la prédiction. Ayant observé X_1, X_2, ... , X_T, on veut prédire les valeurs futures X_{T+1}, X_{T+2}, ...
- Nous utiliserons les modèles de décomposition pour faire de la prévision : on prédira la tendance et la saisonnalité séparément grâce aux modélisations.
- On appelle alors prévision à l'horizon h la valeur X_{T+h} qui fournit une évaluation de la valeur de la série à la date $T+h$.

CHAPITRE II
La modélisation déterministe

Le modèle additif

Nous considérons dans cette section une série $X = (X_t)_t$ admettant une décomposition additive

$$X_t = Z_t + S_t + \epsilon_t, \quad t = 1 \dots T,$$

où Z_t est la composante tendancielle, S_t la composante saisonnière et ϵ_t représente l'erreur ou l'écart au modèle.

- La **tendance** Z_t exprime un mouvement à moyen terme de la série.
- La **composante saisonnière** S_t exprime un phénomène qui se reproduit de manière analogue sur chaque intervalle de temps successif. En plus de la notion de périodicité, on suppose par ailleurs que l'effet du saisonnier est en moyenne nul sur une période, ce qui signifie que

$$\sum_{i=1}^{P} c_i = 0.$$

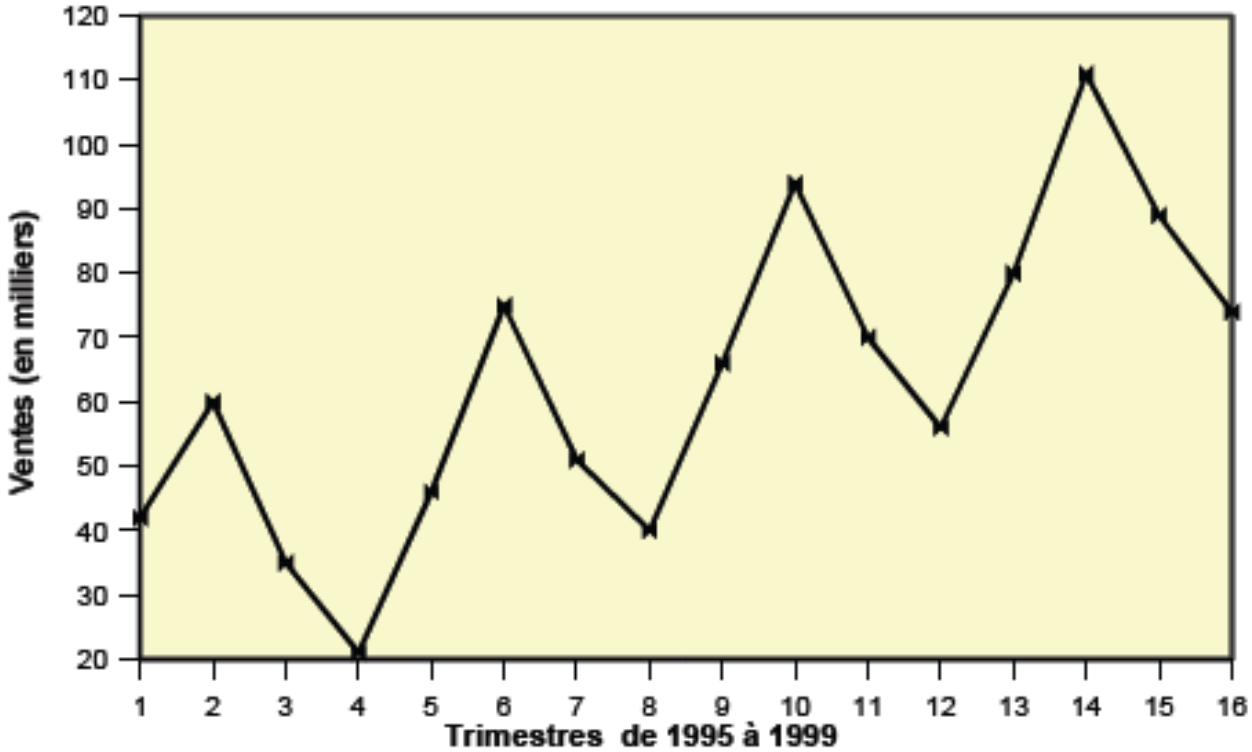
- Les **erreurs** ε_t sont des variables aléatoires.

Le modèle additif

☞ **A retenir.** Dans ce modèle, l'amplitude de la série reste constante au cours du temps. Ceci se traduit graphiquement par des fluctuations autour de la tendance Z_t constantes au bruit près.

Le modèle additif - Un exemple

Que peut-on dire des composantes présentes sur cet exemple ?



- tendance générale ?
- période ?
- que signifient des valeurs $c_2 = +20$ et $c_4 = -10$?
- que signifie une fluctuation irrégulière $\epsilon_{14} = -2$?

Le modèle multiplicatif

Nous considérons dans cette section une série $X = (X_t)_t$ admettant une décomposition muliplicative

$$X_t = Z_t(1 + S_t)(1 + \epsilon_t), \quad t = 1 \dots T,$$

où Z_t est la composante tendancielle, S_t la composante saisonnière et ϵ_t représente l'erreur ou l'écart au modèle.
Là encore, la composante saisonnière vérifie

$$\sum_{i=1}^{P} c_i = 0.$$

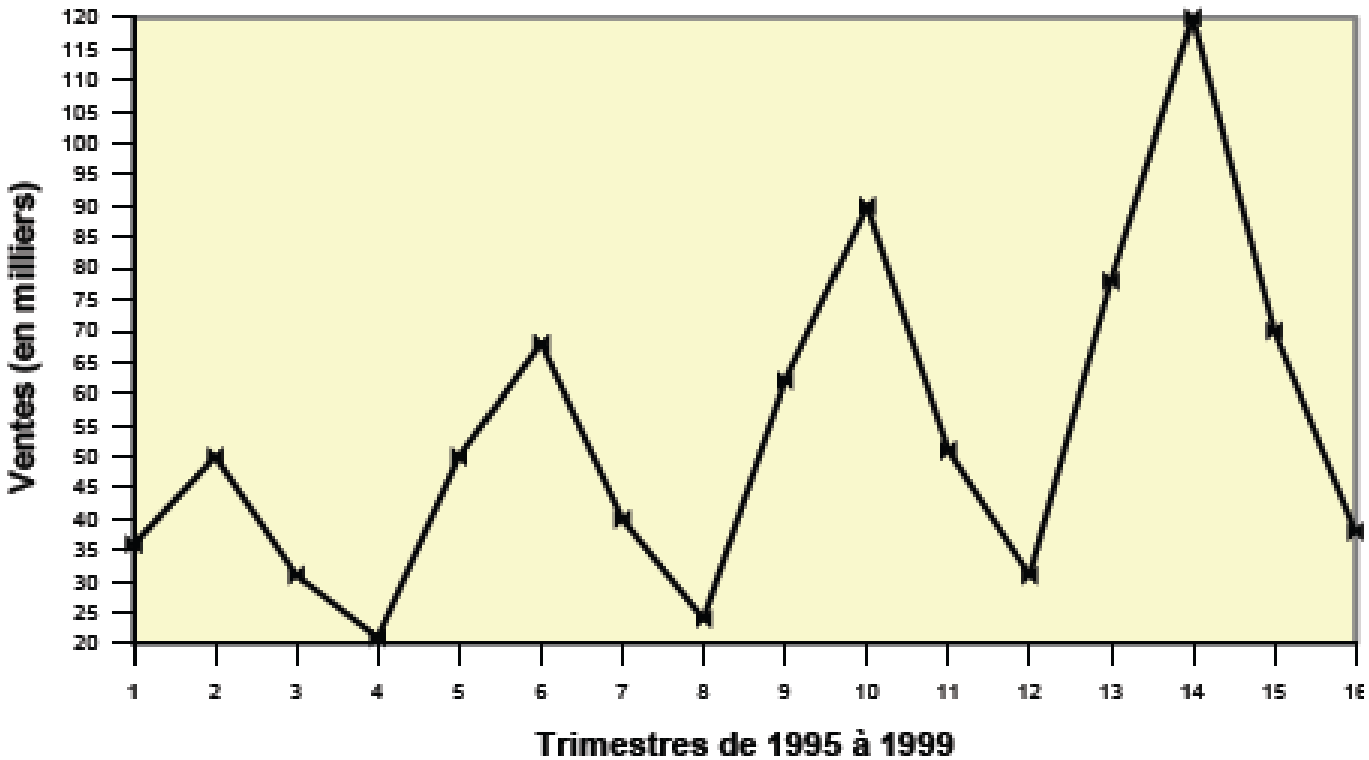
Le modèle multiplicatif

☞ **A retenir.** L'amplitude de la série n'est plus constante au cours du temps : elle varie au cours du temps proportionnellement à la tendance Z_t au bruit près. Dans ce modèle, on considère que les amplitudes des fluctuations dépendent du niveau.

Le modèle multiplicatif est généralement utilisé pour des données de type économique.

Le modèle multiplicatif - Un exemple

Reprenons l'exemple des ventes trimestrielles.



- que signifient des valeurs $c_2 = +0.8$ et $c_4 = -0.5$?
- que signifie une valeur $\epsilon_9 = +0.2$?

Les modèles mixtes

Il s'agit là de modèles où addition et multiplication sont utilisées. On peut supposer par exemple que la composante saisonnière agit de façon multiplicative alors que les fluctuations irrégulières sont additives :

$$X_t = Z_t \widetilde{S}_t + \epsilon_t, \quad t = 1 \dots T,$$

avec l'hypothèse ici que $\sum_{i=1}^{P} c_i = P$.

Choix du modèle

Avant toute modélisation et étude aprofondie du modèle, on tente d'abord de déterminer si on est en présence d'une série dans laquelle pour une observation X donnée

- la variation saisonnière S s'ajoute simplement à la tendance Z ; c'est le modèle additif.
- la variation saisonnière S multiplie la tendance Z ; c'est le modèle multiplicatif.

Afin de faire cette distinction, on va se baser sur deux méthodes graphiques (bande et profil) et utiliser une méthode analytique.

Choix du modèle - Un exemple

Nouvelles immatriculations de voitures particulières, commerciales et utilitaires neuves selon le mois :

	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août
1996	2006	3224	3789	4153	3100	2527	3015	1504
1997	2247	3862	3586	4047	2838	2727	2730	1648
1998	2433	3723	4325	4493	3399	3083	3247	1928
1999	3127	4437	5478	4384	3552	3678	3611	2260
2000	3016	4671	5218	4746	4814	3545	3341	2439

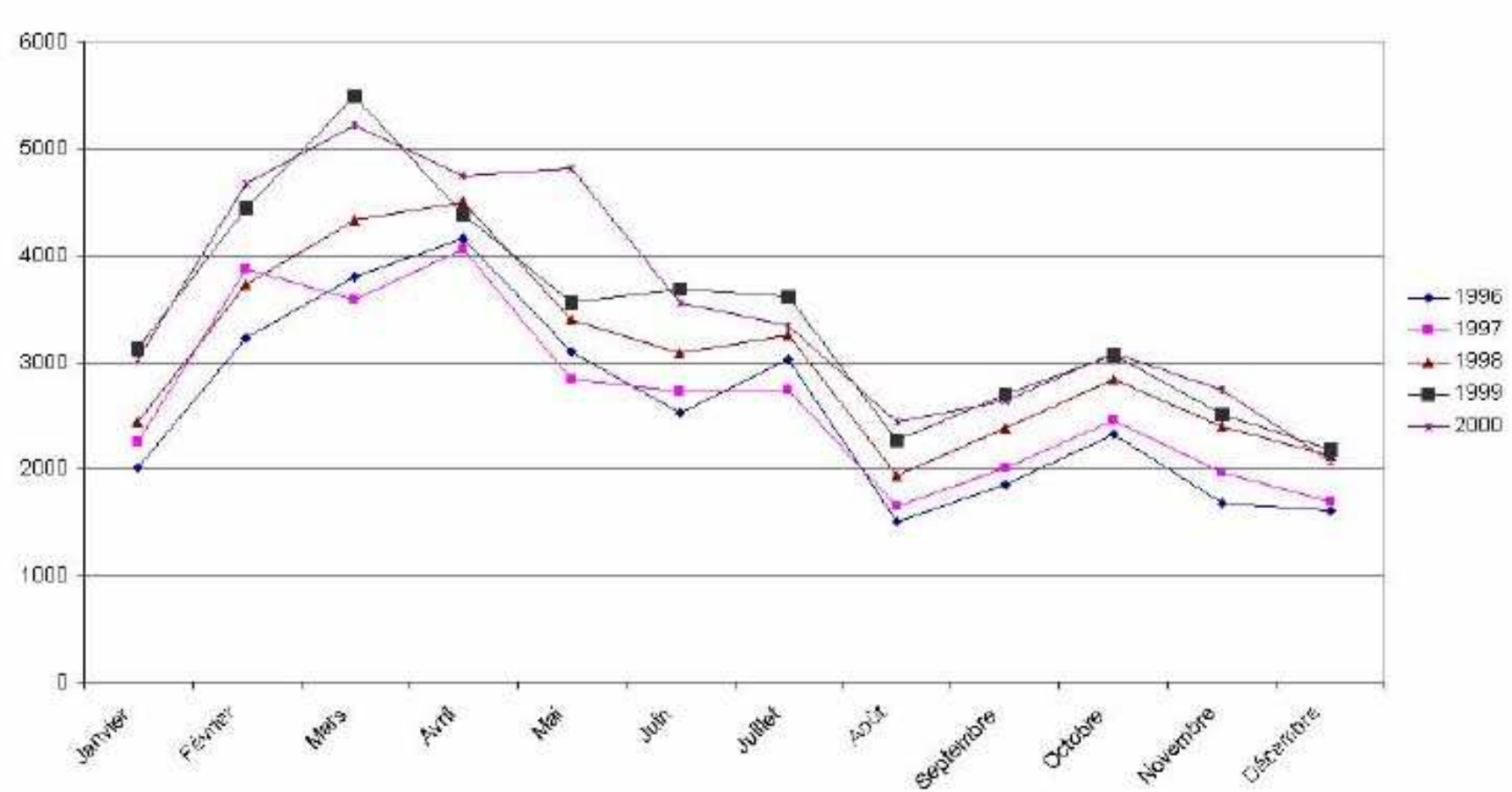
	Sept.	Oct.	Nov.	Déc.
1996	1847	2314	1673	1602
1997	2007	2450	1966	1695
1998	2377	2831	2388	2126
1999	2699	3071	2510	2182
2000	2637	3085	2737	2055

Choix du modèle - Méthode du profil

Pour faire la détermination entre modèle additif et modèle multiplicatif graphiquement, on peut par exemple superposer les saisons représentées par des courbes de profil sur un même graphique.

Si ces courbes sont parallèles, le modèle est additif, autrement le modèle est multiplicatif.

Choix du modèle - Méthode du profil sur l'exemple

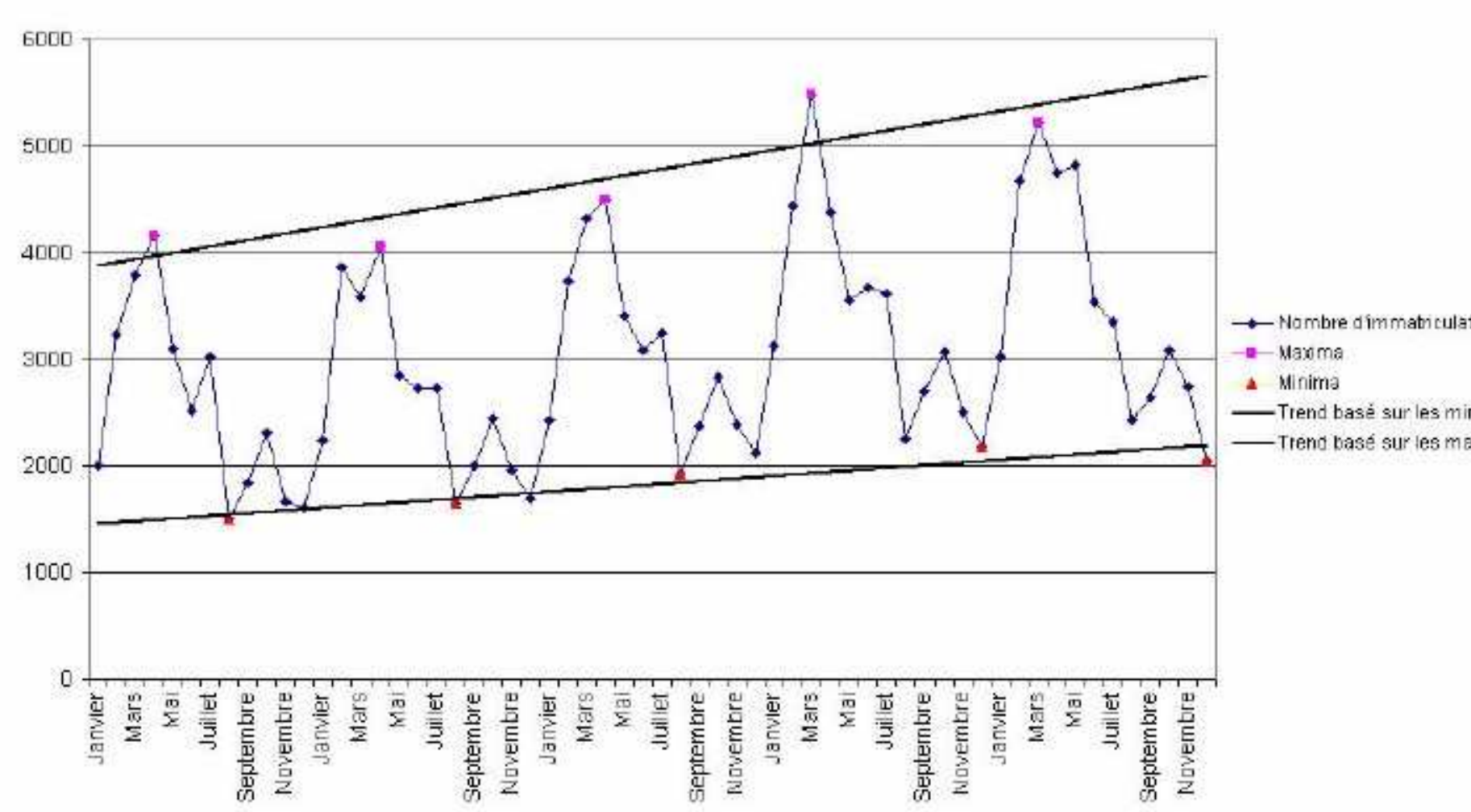


Choix du modèle - Méthode de la bande

On fait un graphique représentant la série chronologique, puis on trace une droite passant respectivement par les minima et par les maxima de chaque saison.

Si ces deux droites sont parallèles, nous sommes en présence d'un modèle additif. Dans le cas contraire, c'est un modèle multiplicatif.

Choix du modèle - Méthode de la bande sur l'exemple



Nous ne sommes donc pas en mesure de conclure

Choix du modèle - Méthode analytique

On calcule les moyennes $\overline{x}$ et les écarts-types σ pour chacune des périodes considérées puis la droite des moindres carrés $\sigma = a\overline{x} + b$. Pour des rappels sur la droite des moindres carrés voir le chapitre suivant.

Si a est nul, c'est le modèle additif, sinon c'est le modèle multiplicatif.

Choix du modèle - Conclusion

☞ **A retenir.** Il faut bien tester avec les trois méthodes pour décider du modèle !

Rappels sur la moyenne, la variance et la covariance empiriques

Dans le cadre de N observations discrètes, on dispose d'un échantillon $x_1, \ldots, x_N$ d'une variable X.

La moyenne empirique est donnée par

$$\overline{X}_N = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

La variance empirique est donnée par

$$\sigma_N^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{X}_N)^2 = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - (\overline{X}_N)^2.$$

Si on dispose d'un échantillon $y_1, \ldots, y_N$ d'une seconde variable Y, la covariance empirique entre X et Y est donnée par

$$\frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{X})(y_i - \overline{Y}) = \frac{1}{N} \sum_{i=1}^{N} x_i y_i - \overline{X}\ \overline{Y}.$$

Rappels sur la régression linéaire (I)

Lorsqu'une liaison linéaire forte entre deux variables X et Y semble raisonnable au vu du nuage de points, on a une relation du type :

$$Y \simeq aX + b,$$

où les coefficients a et b sont inconnus.

$\Leftarrow$ On souhaite déterminer les valeurs de a et b.

Si les points du nuage sont parfaitement alignés (sur une même droite), il serait facile de donner des valeurs à a et b : il suffirait en effet de prendre pour a la pente de la droite sur laquelle se trouvent les points du nuage et pour b la valeur en $x = 0$ (la solution se trouve en résolvant un système de deux équations à deux inconnues à partir de deux points du nuage).

Rappels sur la régression linéaire (II)

Le problème est que les points du nuage sont rarement (parfaitement) alignés : ils sont proches d'une droite. Le problème est donc d'estimer ces coefficients grâce aux valeurs observées sur l'échantillon.

Nous cherchons maintenant la droite qui passe au plus près des points du nuage. Pour cela, il faut donc mesurer l'éloignement des points du nuage par rapport à une droite D d'équation $y = ax + b$ puis minimiser un critère d'erreur donné.

On peut envisager de minimiser

- la somme des erreurs en valeur absolue : $\min_{a,b} \sum_{i=1}^{n} |y_i - ax_i - b|$.
- la somme des erreurs au carré : $\min_{a,b} \sum_{i=1}^{n} (y_i - ax_i - b)^2$.

Rappels sur la régression linéaire par moindres carrés (III)

On démontre en minimisant la fonction de deux variables

$$g(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

que le couple solution $(\widehat{a}, \widehat{b})$ est donné par

$$\widehat{a} = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

$$\widehat{b} = \frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{a}\frac{1}{n}\sum_{i=1}^{n} x_i = \overline{Y} - \widehat{a}\overline{X}$$

La droite d'équation $y = \widehat{a}x + \widehat{b}$ est appelée droite de régression de Y en X et est notée : $\Delta_{Y/X}$.

Rappels sur la régression linéaire par moindres carrés (IV)

Propriétés :

1) Cette droite passe par le point moyen $M(\overline{X};\overline{Y})$. Puisqu'il suffit de deux points pour tracer une droite, on pourra, pour tracer $\Delta_{Y/X}$, placer les points $B(0;b)$ et $M(\overline{X};\overline{Y})$.

2) Le coefficient directeur $\widehat{a}$ de $\Delta_{Y/X}$, $\mathrm{Cov}(X,Y)$ et $r(X,Y)$ sont de même signe :

- lorsqu'ils sont positifs, on parle de corrélation positive (y augmente quand x augmente).
- lorsqu'ils sont négatifs, on parle de corrélation négative (y diminue quand x augmente).

Rappels sur la régression linéaire par moindres carrés (V)

Afin de confirmer qu'il est raisonnable d'approximer le nuage de points par une droite, on calcule le coefficient de corrélation linéaire encore appelé coefficient de Bravais-Pearson :

$$r(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Rappels sur la régression linéaire par moindres carrés (VI)

Propriétés :

a) Le coefficient de corrélation linéaire est symétrique :

$$r(X,Y) = r(Y,X).$$

b) L'inégalité de Cauchy-Schwarz donne :

$$-1 \leq r(X,Y) \leq 1.$$

c) Transformation affine des données. Soient a, b, c et d quatre nombres réels quelconques ($a \neq 0$ et $c \neq 0$). Posons $Z = aX + b$ et $T = cY + d$. On a alors :

$$r(Z,T) = \begin{cases} r(X,Y), \text{ si } a \text{ et } c \text{ sont de même signe,} \\ -r(X,Y), \text{ si } a \text{ et } c \text{ sont de signes opposés.} \end{cases}.$$

En particulier, pour $a = c = 0$, on voit que le coefficient de corrélation linéaire est invariant par translations et pour $b = d = 0$, il est invariant au signe près par homothéties.

Rappels sur la régression linéaire par moindres carrés (VII)

- Si $r(X,Y) = 0$.
  Dans ce cas l'éloignement des points du nuage avec la droite de régression de Y en X est maximal. On dira alors que X et Y sont linéairement indépendants.

- Si $0 < r(X,Y) < 1$.
  Dans ce cas la droite de régression de Y en X est croissante ; on parle alors de corrélation linéaire croissante entre X et Y. Lorsque $r(X,Y)$ est proche de 1, les points du nuage sont donc presque alignés, on a donc une forte corrélation linéaire croissante (ou positive) entre X et Y.

  Arbitrairement, on considèrera la corrélation linéaire croissante faible lorsque $0 < r(X,Y) < 0,3$, moyenne lorsque $0,3 \leq r(X,Y) \leq 0,7$ et forte lorsque $r > 0,7$.

Rappels sur la régression linéaire par moindres carrés (VIII)

- Si $r(X,Y) = 1$, les points du nuage sont alors parfaitement alignés, on peut donc parler de corrélation linéaire croissante totale : pour un individu, sa donnée suivant X détermine entièrement sa donnée suivant Y.

- Si $-1 < r(X,Y) < 0$.
  Dans ce cas la droite de régression de Y en X est décroissante ; on parle alors de corrélation linéaire décroissante entre X et Y. Lorsque $r(X,Y)$ est proche de -1, les points du nuage sont donc presque alignés, on a donc une forte corrélation linéaire décroissante (ou négative) entre X et Y.

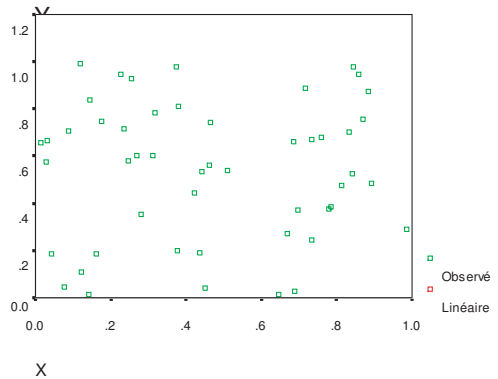
  Arbitrairement, on considèrera la corrélation linéaire décroissante faible lorsque $-0,3 < r(X,Y) < 0$, moyenne lorsque $-0,7 \leq r(X,Y) \leq -0,3$ et forte lorsque $r < -0,7$.

Rappels sur la régression linéaire par moindres carrés (IX)

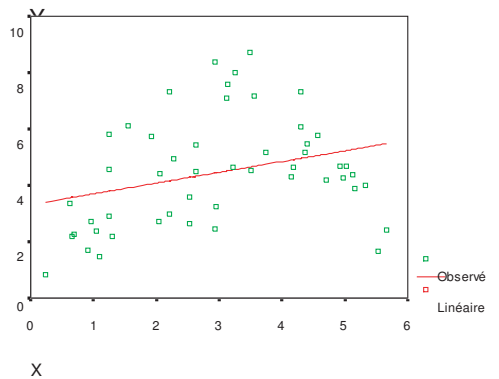
- Si $r(X, Y) = -1$, les points du nuage sont alors parfaitement alignés, on peut donc parler de corrélation linéaire décroissante totale : pour un individu, sa donnée suivant X détermine entièrement sa donnée suivant Y.

Rappels sur la régression linéaire par moindres carrés (X)

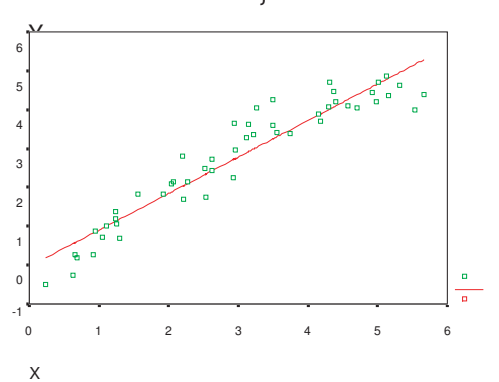
r=0



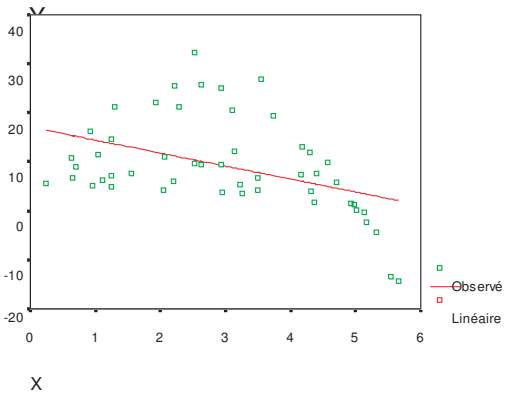
r=0,30



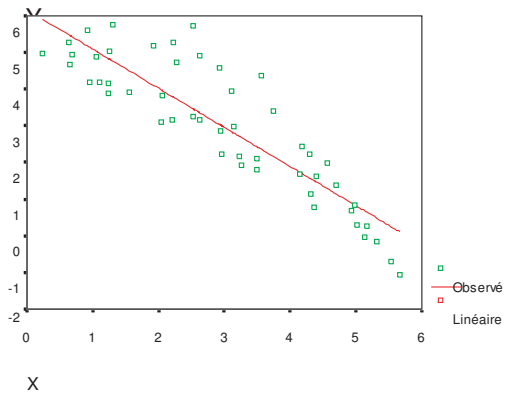
r=0,95



r=-0,43



r=-0,89



Choix du modèle - Méthode analytique sur l'exemple

	Moyenne $\overline{X}$	Ecart-type σ
1996		
1997		
1998		
1999		
2000		

Détail des calculs

$$\overline{\overline{X}} = \qquad \overline{\sigma} =$$

$$\mathrm{Var}(\overline{X}) = \qquad \mathrm{Var}(\sigma) =$$

$$\mathrm{Cov}(\overline{X}, \sigma) =$$

En calculant la droite des moindres carrés, on obtient $a =$ et $b =$, ce qui permet de trancher pour le modèle