Décomposition de la variance :

Considérons la regression linéaire multiple :

$$Y = X\beta + U \quad \dots (*)$$

On se pose la question est ce que les facteurs $X_1, X_2, \dots, X_k$ expliquent bien Y dans le modèle $(*)$.

La phrase "explique bien" signifie que le modèle $(*)$ doit donner des résultat sur Y très proches des valeurs des experiences si on répète de nouveau les expériences :

$$Y \simeq X\hat{\beta}_n$$

Posons :

$$\hat{Y}_i = (X\hat{\beta}_n)_i \text{ : Prévision de } Y_i.$$

On a :

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}_n)^2$$

$$= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_n)^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}_n)$$

or, puisque $\hat{Y}$ est la projection de Y sur $\mathcal{E}(X)$ alors :

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}_n) = \langle Y - \hat{Y}, \hat{Y} - \bar{Y}_n \rangle = 0$$

Par suite

$$\sum (Y_i - \bar{Y}_n)^2 = \|\hat{U}\|^2 + \sum (\hat{Y}_i - \bar{Y}_n)^2 \quad \text{où } \hat{U} = Y - \hat{Y}.$$

Définitions :

* La quantité $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ est appelée la variance totale.
* $\|\hat{U}\|^2$ est appelée le résidu de la regression linéaire.
* $S_{reg} = \sum (\hat{Y}_i - \bar{Y}_n)^2$ est appelée variance de la regression linéaire.

Rappels d'algèbre linéaire :

- Une matrice A, $(n \times n)$ est une matrice de projection orthogonale si $A^t = A$ et $A^2 = A$.

- Soit A une matrice de projection orthogonale. Alors : $\text{rang}(A) = \sum_{i=1}^{n} a_{ii} = r$ et il existe une matrice orthogonale B, $(n \times n)$ tels que :

$$A = B^t \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} B$$

où I_r est la matrice identité de $\mathbb{R}^r$.

Proposition :

Soit $X = (X_1, X_2, \dots, X_n)^t$ un vecteur gaussien avec $E(X) = 0$ et $C_X = I$.
Soit A une matrice de projection $(n \times n)$.
On pose : $q_A(x) = x^t A x$; $x \in \mathbb{R}^n$.
Alors :

$$Y = q_A(X) \hookrightarrow \chi^2_r$$

où $r = \text{rang}(A)$.

Preuve :

D'après le rappel, il existe une matrice orthogonale B telle que :

$$A = B^t \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} B$$

Posons $Z = BX = B(X_1, X_2, \dots, X_n)^t$
On a :

$$Y = q_A(X) = X^t A X = X^t B^t \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} B X$$

D'où :

$$Y = Z^t \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Z$$

D'autre part :

$$Z = BX \hookrightarrow N(0, BC_X B^t)$$

or : $BC_X B^t = BIB^t = BB^t = BB^{-1} = I.$

Ainsi :

$$Z \hookrightarrow N(0, I)$$

Donc

$$Y = q_A(X) = Z^t \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Z = \sum_{i=1}^{r} Z_i^2$$

Comme $Z \hookrightarrow N(0, I)$, alors $Z_i \hookrightarrow N(0,1)$ et les (Z_i) sont indépendantes.

D'où :

$$Y \hookrightarrow \chi^2_r.$$

Proposition : Soit $(X_1, X_2, \dots, X_n)$ un échantillon de loi $N(m, \sigma^2)$.

Soit $\overline{X}_n = \frac{1}{n} \sum X_i$ et $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$.

On a :

1) $\overline{X}_n \hookrightarrow N\left(m, \frac{\sigma^2}{n}\right)$.

2) $\frac{n-1}{\sigma^2} S_n^2 \hookrightarrow \chi^2_{n-1}$.

3) $\overline{X}_n$ et S_n^2 sont indépendantes.

4) $\sqrt{n} \, \frac{\overline{X}_n - m}{S_n} \hookrightarrow T$ (Loi de Student).

Preuve :

1) On a :

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\cdot n \cdot m = m$$

et

$$Var(\overline{X}_n) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\, n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

Comme $\overline{X}_n$ est une combinaison linéaire des X_i et $(X_1, X_2, \dots, X_n)$ est un vecteur gaussien, alors $\overline{X}_n$ est gaussien.

2) On a :

$$\begin{aligned}
\frac{n-1}{\sigma^2} S_n^2 &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2 \\
&= \sum_{i=1}^{n} \left(\frac{X_i - m}{\sigma} - \frac{\overline{X}_n - m}{\sigma}\right)^2 \\
&= \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2 \quad \text{où } Y_i = \frac{X_i - m}{\sigma} \\
&= \sum_{i=1}^{n} Y_i^2 - 2\overline{Y}\left(\sum_{i=1}^{n} Y_i\right) + n\overline{Y}_n^2 \\
&= \sum_{i=1}^{n} Y_i^2 - 2\overline{Y}_n \cdot n\overline{Y}_n + n\overline{Y}_n^2 \\
&= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}_n^2
\end{aligned}$$

Posons $z = (1, 1, \dots, 1)^t$ et $A = I_n - \frac{1}{n} z z^t$.

On peut écrire :

$$\begin{aligned}
Y^t A Y &= Y^t I_n Y - \frac{1}{n} Y^t z z^t Y \\
&= \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 \\
&= \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}(n\overline{Y}_n)^2 \\
&= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}_n^2 \\
&= \frac{n-1}{\sigma^2} S_n^2 .
\end{aligned}$$

D'autre part, il est facile de voir que $A^t = A$. De plus :

$$A^2 = \left(I - \frac{1}{n} z z^t\right)\left(I - \frac{1}{n} z z^t\right)$$

$$= I - \frac{2}{n} z z^t + \frac{1}{n^2}\left(z z^t\right)^2$$

$$= I - \frac{2}{n} z z^t + \frac{1}{n^2}\left(n\, z z^t\right)$$

$$= I - \frac{1}{n} z z^t$$

$$= A$$

Il en découle que A est une matrice de projection orthogonale.
En utilisant la proposition précédente et puisque :

$$\frac{n-1}{\sigma^2} S_n^2 = q_A(X)$$

alors $\dfrac{n-1}{\sigma^2} S_n^2 \hookrightarrow \chi^2_r$ avec :

$$r = \operatorname{rang}(A) = \operatorname{tr}(A) = n\left(1 - \frac{1}{n}\right) = n - 1.$$

3) Posons $G_i = \dfrac{X_i - \overline{X}_n}{\sigma}$.

On a : $S_n^2 = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} G_i^2$.

Le vecteur $(\overline{X}_n, G_1, G_2, \dots, G_n)$ est gaussien car toute combinaison linéaire de $\overline{X}_n, G_1, G_2, \dots, G_n$ est une combinaison linéaire de $X_1, X_2, \dots, X_n$.

D'autre part :

$$\operatorname{Cov}(\overline{X}_n, G_i) = \operatorname{Cov}\left(\overline{X}_n, \frac{X_i}{\sigma} - \frac{\overline{X}_n}{\sigma}\right)$$

$$= \operatorname{Cov}\left(\overline{X}_n, \frac{X_i}{\sigma}\right) - \operatorname{Cov}\left(\overline{X}_n, \frac{\overline{X}_n}{\sigma}\right)$$

On a :

$$\operatorname{Cov}\left(\overline{X}_n, \frac{X_i}{\sigma}\right) = \frac{1}{\sigma} \operatorname{Cov}\left(\frac{1}{n}\sum_{j=1}^{n} X_j, X_i\right)$$

$$= \frac{1}{n\sigma} \operatorname{cov}(X_1 + X_2 + \dots + X_n, X_i)$$

$$= \frac{1}{n\sigma} \left[\operatorname{cov}(X_1, X_i) + \dots + \operatorname{cov}(X_i, X_i) + \dots + \operatorname{cov}(X_n, X_i) \right]$$

$$= \frac{1}{n\sigma} \operatorname{var}(X_i)$$

$$= \frac{\sigma}{n}.$$

et :

$$\operatorname{cov}\left(\overline{X}_n, \frac{\overline{X}_n}{\sigma}\right) = \frac{1}{\sigma} \operatorname{cov}(\overline{X}_n, \overline{X}_n) = \frac{1}{\sigma} \operatorname{var}(\overline{X}_n) = \frac{\sigma^2}{n\sigma} = \frac{\sigma}{n}$$

Donc $\operatorname{cov}(\overline{X}_n, G_i) = 0$

Ainsi $\overline{X}_n$ et $(G_1, G_2, \dots, G_n)$ sont indépendantes

Finalement $\overline{X}_n$ est indépendante de $S_n^2 = h(G_1, G_2, \dots, G_n)$

4) On a :

$$\frac{\sqrt{n}\,(\overline{X}_n - m)}{S_n} = \sqrt{n-1}\; \frac{\frac{\sqrt{n}\,(\overline{X}_n - m)}{\sigma}}{\sqrt{(n-1)\frac{S_n^2}{\sigma^2}}}$$

Puisque $\sqrt{n}\,\frac{\overline{X}_n - m}{\sigma} \hookrightarrow N(0,1)$ et $\frac{(n-1)}{\sigma^2} S_n^2 \hookrightarrow \chi^2_{n-1}$,

alors $\frac{\sqrt{n}\,(\overline{X}_n - m)}{S_n} \hookrightarrow T_{n-1}$ où T_{n-1} est la loi de student avec $(n-1)$ degré de liberté.

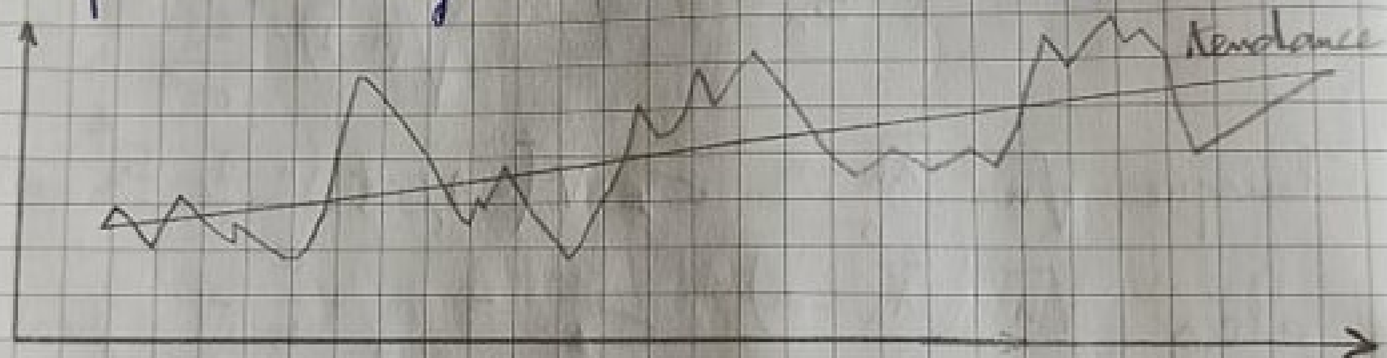
Chapitre 2 : Modèles avec saisonnalité

1) Introduction :

On considère une série chronologique $(Y_t)_{t \in T}$ avec $T = \mathbb{N}$ ou $\mathbb{Z}$ qui se présente comme résultat de différentes composantes fondamentales.

Le choix des éléments qui constituent les composantes fondamentales est très important pour expliquer les variations (les valeurs) de Y_t. Parmi ces composantes, on peut citer :

a) **La tendance T_t (trend)** : Elle représente l'évolution "globale" de la variable Y_t (phénomène Y_t), ou encore le comportement moyen de l'évolution de Y_t.



b) **La saisonnalité S_t** : Elle correspond à mettre en évidence un phénomène saisonnier. La saison naturelle est un cycle de temps qui se répète périodiquement. Ainsi la saison S_t représente les fluctuations périodiques (variations) de Y_t dûes aux saisons (ou encore des facteurs de la société : les fêtes, les week-end ...).

S_t est une fonction périodique de t et de période p.

c) **Le résidu U_t** : représente les fluctuations qui sont en général imprévisibles.

On retient le modèle suivant (modèle additif):

$$Y_t = T_t + S_t + U_t$$

On fait les hypothèses suivantes (conditions sur le modèle):

$$\sum_{j=1}^{p} S_j = 0 \quad \text{et} \quad \sum_{j=1}^{n} U_j$$

où n est le nombre d'observations.

Remarque: Il existe d'autre modèles comme le modèle multiplicatif $Y_t = T_t S_t + U_t$ ou encore $Y_t = T_t S_t U_t$.

2) Détermination des composantes T_t, S_t:

. Décomposition de Y_t: Il existe plusieurs modèles de décomposition avec un choix de T_t (déterministe).

1- Tendance linéaire: $T_t = b + at$

2- Tendance polynômiale: $T_t = a_0 + a_1 t + \dots + a_p t^p$.

3- Tendance exponentielle: $T_t = c + a e^{bt}$.

4- Tendance logarithmique: $T_t = b + a \log t$

5- Tendance hyperbolique: $T_t = \dfrac{1}{b + at}$

6- Tendance logistique: $T_t = \dfrac{1}{c + b e^{-at}}$

Les exemples de tendance précédents sont appliqués aux données là où on arrive à déceler la tendance. Si on arrive pas à mettre en évidence un choix de T_t, il existe des méthodes qui permettent de le faire.

Méthodes désaisonnalité (enlever les saisons): Parmi ces méthodes:

- Moyennes mobiles (méthode de lissage - Smoothing):

Définition: On appelle moyenne mobile des observation (Y_t) la quantité:

Exemples :

1) $b = 1$, $m = 3$:

$$M_3(Y_t) = \frac{1}{3}\left(Y_{t-1} + Y_t + Y_{t+1}\right)$$

2) $b = 1$, $m = 2$

$$M_2(Y_t) = \frac{1}{2}\left(\frac{1}{2}Y_{t-1} + Y_t + \frac{1}{2}Y_{t+1}\right)$$

On obtient donc une nouvelle série $\tilde{Y}_t = M_m(Y_t)$, $t = 1, 2, \ldots, T$.

On va montrer que dans cette nouvelle série $\tilde{Y}_t$, l'effet saisonnier "disparait".

En effet, à partir de $Y_t = T_t + S_t + U_t$, on applique la moyenne mobile d'ordre m, M_m :

$$M_m(Y_t) = M_m(T_t + S_t + U_t).$$

M_m est un opérateur linéaire, donc :

$$M_m(Y_t) = M_m(T_t) + M_m(S_t) + M_m(U_t).$$

Sous la condition $\sum_{j=1}^{p} S_j = 0$, on a :

- Si $p = m = 2b + 1$:

$$M_m(S_t) = \frac{1}{m}\left(S_{t-b} + S_{t-b+1} + \cdots + S_{t+b}\right)$$

or $t-b, t-b+1, \ldots, t+b$ sont $m = 2b+1$ entiers consécutifs et S_t est de période p, donc

$$S_{t-b} + S_{t-b+1} + \cdots + S_{t+b} = S_1 + S_2 + \cdots + S_p = 0.$$

D'où $M_m(S_t) = 0$.

- Si $p = m = 2b$:

$$M_m(S_t) = \frac{1}{m}\left(\frac{1}{2}S_{t-b} + S_{t-b+1} + \cdots + S_{t+b-1} + \frac{1}{2}S_{t+b}\right)$$

$$= \frac{1}{m}\left(S_{t-b} + S_{t-b+1} + \cdots + S_{t+b-1}\right) \text{ car } S_{t-b} = S_{t+b}$$

or $t-b, t-b+1, \ldots, t+b-1$ sont $m = 2b$ entiers consécutifs

Exemples :

1) $b=1, m=3$:

$$M_3(Y_t) = \frac{1}{3}\left(Y_{t-1}+Y_t+Y_{t+1}\right)$$

2) $b=1, m=2$

$$M_2(Y_t) = \frac{1}{2}\left(\frac{1}{2}Y_{t-1}+Y_t+\frac{1}{2}Y_{t+1}\right)$$

On obtient donc une nouvelle série $\tilde{Y}_t = M_m(Y_t)$, $t=1,2,\dots,T$.
On va montrer que dans cette nouvelle série $\tilde{Y}_t$, l'effet saisonnier "disparait".

En effet, à partir de $Y_t = T_t+S_t+U_t$, on applique la moyenne mobile d'ordre m, M_m :

$$M_m(Y_t) = M_m(T_t+S_t+U_t)$$

M_m est un opérateur linéaire, donc :

$$M_m(Y_t) = M_m(T_t)+M_m(S_t)+M_m(U_t).$$

Sous la condition $\sum_{j=1}^{p} S_j = 0$, on a :

- Si $p=m=2b+1$:

$$M_m(S_t) = \frac{1}{m}\left(S_{t-b}+S_{t-b+1}+\dots+S_{t+b}\right)$$

or $t-b, t-b+1, \dots, t+b$ sont $m=2b+1$ entiers consécutifs et S_t est de période p, donc

$$S_{t-b}+S_{t-b+1}+\dots+S_{t+b} = S_1+S_2+\dots+S_p = 0.$$

D'où $M_m(S_t) = 0$.

- Si $p=m=2b$:

$$M_m(S_t) = \frac{1}{m}\left(\frac{1}{2}S_{t-b}+S_{t-b+1}+\dots+S_{t+b-1}+\frac{1}{2}S_{t+b}\right)$$

$$= \frac{1}{m}\left(S_{t-b}+S_{t-b+1}+\dots+S_{t+b-1}\right) \text{ car } S_{t-b}=S_{t+b}$$

or $t-b, t-b+1, \dots, t+b-1$ sont $m=2b$ entiers consécutifs et S_t est de période p donc

$$S_{t-l} + S_{t-l+1} + \cdots + S_{t+l} = S_1 + S_2 + \cdots + S_p = 0$$

Ceci donne $M_m(S_t) = 0$.

On obtient donc pour $m = p$:

$$M_m(Y_t) = M_m(T_t) + M_m(U_t)$$

Posons $\tilde{Y}_t = M_m(Y_t)$; $\tilde{U}_t = M_m(U_t)$ et $\tilde{T}_t = M_m($

On a : $\tilde{Y}_t = \tilde{T}_t + \tilde{U}_t$.

$(\tilde{Y}_t)$ est une série chronologique sans effet "saison".
On a donc "désaisonalisé (Y_t)". Après, on doit ret
S_t par une autre opération.

Exemple :

t	Y_t	$M_2(Y_t)$	$M_3(Y_t)$
1	5	—	—
2	3	3,75	4
3	4	3,625	3,5
4	3,5	3,75	3,83
5	4	3,875	3,83
6	4	—	—

$$M_2(Y_t) = \frac{1}{2}\left(\frac{1}{2}Y_{t-1} + Y_t + \frac{1}{2}Y_{t+1}\right)$$

$$M_3(Y_t) = \frac{1}{3}\left(Y_{t-1} + Y_t + Y_{t+1}\right)$$