Rappels

Pour un échantillon de N éléments, soit x et y deux variables que l'on peut tracer, l'une en fonction de l'autre, sur un *diagramme de dispersion.* La droite de régression des moindres carrés s'écrit

$$Y_{est} = a_0 + a_1 X, \qquad \text{ou} \qquad X_{est} = a_2 + a_3 Y,$$

Une mesure de la dispersion de la droite de régression de Y sur X est

$$s_{Y,X} = \sqrt{\frac{\sum (Y - Y_{est})^2}{N}}.$$

Une mesure de la dispersion de la droite de régression de X sur Y est

$$s_{X,Y} = \sqrt{\frac{\sum (X - X_{est})^2}{N}}.$$

On a

$$s^2_{Y,X} = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}.$$

La variation totale est la somme de la variation expliquée et de la variation non expliquée par la droite des moindres carrés.

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{est})^2 + \sum (Y_{est} - \bar{Y})^2$$

Le coefficient de corrélation s'écrit

$$r = \pm \sqrt{\frac{\text{variation expliquée}}{\text{variation totale}}} = \pm \sqrt{\frac{\sum (Y_{est} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}}$$

ou

$$r = \sqrt{1 - \frac{s^2_{Y,X}}{s^2_Y}}$$

avec

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}}.$$

1 Les séries temporelles

y_t est une série chronologique aux différents temps $t = [1, 2, ..., n]$.
x_t est une série chronologique aux différents temps $t = [1, 2, ..., n]$.
$\bar{y}$ est la moyenne de y_t:

$$\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t.$$

$var(y)$ est la variance de y_t:

$$var(y) = \frac{1}{n-1} \sum_{t=1}^{n} (y_t - \bar{y})^2 = \frac{1}{n-1} \left(\sum_{t=1}^{n} y_t^2 - \frac{\left(\sum_{t=1}^{n} y_t \right)^2}{n} \right).$$

$covar(x, y)$ est la covariance de x_t et y_t:

$$covar(x, y) = \frac{1}{n-1} \left(\sum_{t=1}^{n} x_t y_t - \frac{\sum_{t=1}^{n} x_t \sum_{t=1}^{n} y_t}{n} \right).$$

r_{xy} est le coefficient de corrélation entre x_t et y_t:

$$r_{xy} = \frac{covar(x, y)}{\sqrt{var(x) var(y)}} = \frac{\sum_{t=1}^{n} x_t y_t - \frac{\sum_{t=1}^{n} x_t \sum_{t=1}^{n} y_t}{n}}{\sqrt{\left(\sum_{t=1}^{n} y_t^2 - \frac{\left(\sum_{t=1}^{n} y_t \right)^2}{n} \right) \left(\sum_{t=1}^{n} x_t^2 - \frac{\left(\sum_{t=1}^{n} x_t \right)^2}{n} \right)}},$$

ou encore

$$r_{xy} = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n} (x_t - \bar{x})^2 \sum_{t=1}^{n} (y_t - \bar{y})^2}}.$$

Corrélogrammes

La façon la plus simple d'analyser des variations autour de la tendance (la variabilité) est d'utiliser la fonction autocovariance. A partir de cette fonction de covariance on pourra calculer différents coefficients d'autocorrélations. Ces derniers nous permettrons de mettre en evidence des liens, des structures présents au sein de la série, que ces structures soient des composantes saisonnières ou des composantes cycliques.

Coefficients d'autocorrélations

Ces coefficients d'autocorrélations sont simplement des coefficients de corrélation entre les différentes observations de la série. En fait, ces coefficients de corrélation sont calculés entre les valeurs de la série et elles-même décalées d'un certain nombre de pas de temps k (i.e. lag).

Par conséquent on calculera $r_{yy}(k)$ avec

$$\begin{aligned} & r_{yy}(k) = \frac{\text{autocovariance}(k)}{\text{variance}}, \\ & y_t \text{ avec t=[1,2,...,n-k]}, \\ & y_{t-k} \text{ avec t=[k+1,k+2,...,n]}. \end{aligned}$$

$g(k)$ est la fonction d'autocovariance au lag k:

$$g(k) = \frac{1}{n-k-1} \sum_{j=k+1}^{n} (y_j - \bar{y})(y_{j-k} - \bar{y}).$$

Les séries étant supposées stationnaires, $\bar{y}$ peut être calculé sur la première partie de la série de 1 à $n-k$, sur la deuxième partie de la série de k à n ou sur l'ensemble de la série. On obtient

$$r_{yy}(k) = \frac{g(k)}{g(0)} = \frac{\sum\limits_{j=k+1}^{n} (y_j - \bar{y})(y_{j-k} - \bar{y})}{\sum\limits_{j=k+1}^{n} (y_j - \bar{y})^2},$$

ou encore

$$r_{yy}(k) = \frac{\sum\limits_{j=k+1}^{n} y_j y_{j-k} - \frac{\sum\limits_{j=k+1}^{n} y_j \sum\limits_{j=k+1}^{n} y_{j-k}}{n-k}}{\sum\limits_{j=k+1}^{n} y_j^2 - \frac{\left(\sum\limits_{j=k+1}^{n} y_j\right)^2}{n-k}}.$$

L'évolution de $r_{yy}(k)$ en fonction de k formera le corrélogramme (voir exemples).

Il est ensuite essentiel de savoir si les différentes valeurs de $r_{yy}(k)$ qui forment le corrélogramme sont dues au hasard ou si certaines de ces valeurs indiquent une dépendence au lag k (i.e. décalage k) ente les observations étudiées. On est donc amener à comparer les coeficients d'autocorrélation de la série étudiée à ceux d'une série qui serait aléatoire (e.g. un bruit blanc, une série générée par un processus gaussien, avec des valeurs indépendantes entre elles).

Pour une série aléatoire, approximativement un vingtième des valeurs de $r_{yy}(k)$ ne sont pas comprise entre $\pm 1.96/\sqrt{n}$. Cette valeur est la marge d'erreur associée avec l'estimation d'une moyenne. Pratiquement, on rejete donc le comportment aléatoire avec un seuil de 5% (voir exemples).

Le corrélogramme permet de mettre en évidence la périodicité des séries temporelles en présence d'un bruit qui pourrait masquer cette périodicité. Et ce que cette périodicité soit un saisonnalité ou une composante cyclique plus subtile.

Dans l'absolue, il est nécessaire de travailler sur des séries stationnaires et de supprimer la tendance de la série. Dans la pratique, un comportement périodique peut être mis en évidence même dans le cas où il existe une tendance. Les différentes relations mises en évidence dans la série pourront ensuite être incorporées dans un modèle qui décrira la série étudiée.

Corrélations partielles

Un deuxième corrélogramme est particulièrement important pour caractériser une série temporelle, celui calculé pour les autocorrélations partielles.

Comme dans le cas des corrélations "classiques" avec 3 variables significatives, on doit se demander si la corrélation ente 2 variables n'est pas due au fait que les 2 variables sont corrélées avec une troisème variable.

x_1 corrélée avec x_2

x_1 corrélée avec x_3

$\Longrightarrow$ x_2 corrélée avec x_3 du fait des 2 corréations précédentes.

L'utilisation des corrélations partielles consiste à calculer la corrélation entre x_2 et x_3 en supprimant l'effet de x_1.

On utilise cette démarche pour calculer les autocorrélations partielles au lag k en supprimant les autocorrélations au lag $k' < k$. On note $a_{i,j}$ l'autocorrélation au lag j en ayant supprimé les effets des corréations jusqu'à i. On a

$$a_{p,p} = \frac{r_{yy}(p) - \sum_{j=1}^{p-1} a_{j,p-1} r_{yy}(p-j)}{1 - \sum_{j=1}^{p-1} a_{j,p-1} r_{yy}(j)},$$

et

$$a_{j,p} = a_{j,p-1} - a_{p,p}a_{p-j,p-1}.$$

Les premiers termes de la série sont

$$a_{1,1} = r_{yy}(1).$$

$$a_{2,2} = \frac{r_{yy}(2) - a_{1,1}r_{yy}(1)}{1 - a_{1,1}r_{yy}(1)} = \frac{r_{yy}(2) - r_{yy}^2(1)}{1 - r_{yy}^2(1)}.$$

$$a_{1,2} = a_{1,1} - a_{2,2}a_{1,1}.$$

$$a_{3,3} = \frac{r_{yy}(3) - a_{1,2}r_{yy}(2) - a_{2,2}r_{yy}(1)}{1 - a_{1,2}r_{yy}(1) - a_{2,2}r_{yy}(2)}.$$

$$a_{1,3} = a_{1,2} - a_{3,3}a_{2,2}.$$

$$a_{2,3} = a_{2,2} - a_{3,3}a_{1,2}.$$

$$a_{4,4} = \frac{r_{yy}(4) - a_{1,3}r_{yy}(3) - a_{2,3}r_{yy}(2) - a_{1,2}r_{yy}(1)}{1 - a_{1,3}r_{yy}(1) - a_{2,3}r_{yy}(2) - a_{1,3}r_{yy}(3)}.$$

Autocorrélations croisées

Il est aussi intéressant de regarder le lien entre différentes séries (e.g. décés dues à la bronchite et l'indice de pollution atmosphérique). Dans un tel cas l'inter-dépendance entre x_t et y_t nous sera fourni par le biais d'autocorrélations entre les différentes séries. Pour une telle analyse, x_t et y_t doivent être stationnaires et les structures communes entre x_t et y_t doivent être enlevées des séries (sinon les corrélations partielles suffisent).

Blanchissement des séries,

residus = signal -tendance -composante cyclique

Les corrélations croisées entre les résidus des séries permet de regarder s'il n'existe pas des liens qui ne sont pas expliqués par des composantes tendancielles ou des composante cycliques similaires.

On a

$$g_{xy}(k) = covar(x_t, y_{t-k}),$$

et

$$g_{xy}(-k) = covar(x_t, y_{t+k}) = covar(y_{t+k}, x_t).$$

Du fait de la stationarité,

$$g_{xy}(-k) = covar(y_t, x_{t-k}) = g_{yx}(k).$$

$g_{xy}(k)$ n'est pas forcement identique à $g_{xy}(-k)$. En fait nous avons

$$\begin{aligned} g_{xy}(k) &= g_{yx}(-k), \\ g_{xy}(-k) &= g_{yx}(k). \end{aligned}$$

Par définition

$$g_{xy}(k) = \begin{cases} \dfrac{1}{n} \displaystyle\sum_{t=k+1}^{n} (x_t - \bar{x})(y_{t-k} - \bar{y}) & \text{si } k \geq 0, \\ \dfrac{1}{n} \displaystyle\sum_{t=1}^{n+k} (x_t - \bar{x})(y_{t-k} - \bar{y}) & \text{si } k < 0, \end{cases}$$

avec

$$r_{xy}(k) = \frac{g_{xy}(k)}{\sqrt{g_{xx}(0) g_{yy}(0)}} \text{ pour } -p \leq k \leq p,$$

et

$$g_{xx}(0) = var(x_t) = \frac{1}{n-1} \sum_{t=1}^{n} (x_t - \bar{x})^2,$$

$$g_{yy}(0) = var(y_t) = \frac{1}{n-1} \sum_{t=1}^{n} (y_t - \bar{y})^2.$$