Chapitre 3 _ Suite _

II _ Séries Statistiques à deux variables.

→ La Régression Linéaire Simple

1. Introduction :

Dans cette partie, nous allons nous interesser à des séries statistiques à deux variables x et y. On se demande s'il existe une relation de type linéaire entre ces deux variables, sachant que d'autres types de relations peuvent exister. L'objectif serait donc de trouver cette relation linéaire, mesurer le degre de liaison entre x et y et si cette liaison est assez forte, procéder à la prévision de y sachant x ou le contraire.

2. Les séries statistiques à deux variables.

La série statistique est décrite par deux variables, x et y qui prennent les valeurs x_i et y_j. Ces données sont regroupées dans un tableau à double entrée (tableau de contingence) ds lequel apparaissent les effectifs n_{ij} et les distributions marginales n_i et n_j de x et y.

2.1. Moyennes

Elles se calculent pour chacune des deux variables séparément de l'autre.

$$\bar{X} = \frac{1}{N} \sum_i n_i x_i \quad , \quad \bar{Y} = \frac{1}{N} \sum_j n_j y_j$$

2.2. Variances et ecarts-types

$$\sigma_x^2 = \frac{1}{N} \sum_i n_i x_i^2 - \bar{X}^2 \qquad \sigma_x = \sqrt{\sigma_x^2}$$

$$\sigma_y^2 = \frac{1}{N} \sum_j n_j y_j^2 - \bar{Y}^2 \qquad \sigma_y = \sqrt{\sigma_y^2}$$

2.3. Covariance entre variables

Elle mesure le sens de variation entre les variables

$$Cov(x,y) = \frac{1}{N}\sum_i \sum_j n_{ij}(x_i - \bar{x})(y_j - \bar{y})$$

ou

$$Cov(x,y) = \frac{1}{N}\sum_i \sum_j n_{ij}\, x_i\, y_j - \bar{x}\bar{y}$$

NB : La covariance peut être négative. Le signe moins (–) traduit une relation inverse entre les variables.

Par définition, $Cov(x,x) = S_x^2$ et $Cov(y,y) = S_y^2$, et si les variables x et y sont indépendantes alors la covariance est nulle.

3. Détermination de la liaison entre x et y.

Sur un graphique, on représente les points (x_i, y_j), on obtient un nuage statistique où tous les points ont le même poids. En fonction de la forme du nuage, ont peut décider d'ajuster une droite ou de déterminer une relation linéaire entre les variables, tel que :

$$y = ax + b \quad \text{ou} \quad x = cy + d$$

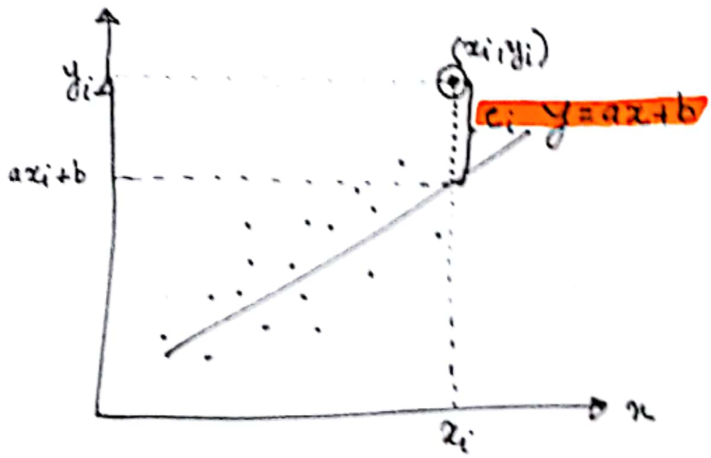
Il reste que le sens de la relation dépend de l'existence d'un rapport logique dentre les variables.

Plusieurs moyens sont disponibles pour déterminer la relation supposée linéaire entre x et y, ici nous focalisons l'attention sur la méthode des moindres carrés.

– La méthode des moindres carrés :

On considère une série statistique (x_i, y_j) de N observations ; l'idée est d'ajuster une droite qui minimise les écarts entre les observations réelles et celles qui sont sur la droite.

Pour déterminer la droite d'ajustement $y = ax + b$, on minimise la somme des carrés des écarts entre les observations réelles et leurs projections verticales sur la droite.



Soit la droite d'équation $y = ax + b$

- les observations réelles x_i, y_i
- les observations théoriques x_i, $ax_i + b$

On minimise la somme des carrés des

$$e_i = y_i - ax_i - b$$

qui revient à annuler les dérivées partielles de cette expression par rapport à a et b

On trouve les résultats suivants :

$$a = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$$

$$b = \bar{y} - a\bar{x}$$

NB : de la même manière si $x = cy + d$

$$c = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(y)} \quad , \quad d = \bar{x} - c\bar{y}$$

4. Ajustement se ramenant à un ajustement linéaire

4.1. Ajustement de la forme $y = b a^x$

en prenant le log décimal

$$\log y = \log b + x \log a$$

$$Y = B + Ax$$

$$Y = Ax + B$$

$Y = \log y$

$A = \log a$

$B = \log b$

$$A = \frac{\mathrm{Cov}(x,Y)}{\mathrm{Var}(x)} \qquad a = 10^A$$

$$B = \bar{Y} - A\bar{x} \qquad b = 10^B$$

4.2. Ajustement se présentant sous la forme $y = b x^a$

En posant :

$$Y = \log y$$
$$B = \log b$$
$$X = \log x$$

$y = b x^a$ devient $Y = a X + B$

a et B sont déterminés en travaillant sur les séries ($X = \log x$, $Y = \log y$)

On obtient :

$$a = \frac{Cov(X, Y)}{Var(X)}$$

$$B = \bar{Y} - a \bar{X} \qquad b = 10^B$$

NB : une telle relation s'obtient sur du papier log-log et met en evidence une relation de type puissance.

5. La Prévision.

Une fois la relation déterminé, entre x et y on peut faire des prévisions. Connaissant les valeurs de la variable explicative x on peut déterminer la valeur de la variable expliqué y.

6. La Corrélation

On appelle coefficient de corrélation linéaire, entre deux variables x et y, le réel r défini comme suit en utilisant la covariance entre x et y :

$$r = \frac{Cov(x, y)}{\delta_x \, \delta_y}$$

ou

$$r = \frac{\sum X_i Y_i}{\sqrt{\sum X_i^2 \sum Y_i^2}}$$

avec $X = x - \bar{x}$ (écart par rapport à $\bar{x}$)
$Y = y - \bar{y}$ (écart " " $\bar{y}$)

(C'est la méthode des écarts à la moyenne)

r peut être positif ou négatif selon le signe de la covariance $Cov(x,y)$

- $r = 0$: il n'y a pas de relation entre x et y.
- $|r| = 1$: la relation linéaire entre x et y est parfaite.
- $0.87 < |r| < 1$: on admet qu'il y a une forte corrélation entre x et y et on accepte l'ajustement linéaire
- r^2 est le coefficient de détermination et exprime le pourcentage de la variance expliqué par la droite de regression.

NB : avant tout calcul, il faut verifier la qualité de la relation mesurée par r.

une méthode rapide de le faire est celle des écarts à la moyenne :

1. Calculer $\bar{x}$ et $\bar{y}$
2. Calculer $X_i = x_i - \bar{x}$ et $Y_i = y_i - \bar{y}$, $\sum_i X_i^2$ et $\sum_i Y_i^2$
3. " $\sum_i X_i Y_i$
4. Le coefficient de corrélation r est :

$$r = \frac{\sum_i X_i Y_i}{\sqrt{\sum X_i^2 \sum Y_i^2}}$$

7. La régression multiple.

Dans la regression simple, il s'agit de trouver la relation entre deux variables x et y. Pour un type linéaire, il s'agit de trouver les paramètres de la regression linéaire a et b tel que :

$y = ax + b$ (x : variable explicative y : var. expliqué)

pour estimer a et b, on résoud un système de 2 equations et 2 inconnues en minimisant la somme des carés des ecarts entre les valeurs observées et celles théoriques.

même chose si on fait la regression de x en y, c'est-à-dire

$x = cy + d$, l'estimation de c et d obeit au même principe.

(x : variable expliqué, y : var. explicative.