CORRÉLATION - RÉGRESSION

G ACHAZ

Covariance

On s'intéresse ici à une population d'objets aléatoires caractérisés par 2 variables aléatoires, X et Y. On peut par exemple penser à des individus caractérisés par leur taille et leur poids (ce sera le cas dans l'exemple traité en cours).

Chacune de ces variables aléatoires possède une distribution *marginale* caractérisée par deux valeurs quantitatives (nommées moments) de la distribution : une espérance ($\mathbb{E}[X]$, $\mathbb{E}[Y]$) et une variance ($\mathbb{V}\text{ar}[X]$, $\mathbb{V}\text{ar}[Y]$). Les distributions marginales de X ou Y ne "considèrent pas" les valeurs de l'autre variable aléatoire (plus précisément, elles sont intégrées sur l'espace de l'autre VA).

On rappelle que pour un échantillon on définit un estimateur de la moyenne de X, $\bar{X}$ et de sa variance s_X^2 comme :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

Il existe une distribution jointe des deux variables aléatoires que l'on peut représenter aisément dans un espace bidimensionnel, comme par exemple dans le repère (O,x,y).

[FIG - avec x,y et les deux moyennes. 3 cas de covariance pos, neg et nulle]

Une valeur qui caractérise la distribution jointe des deux variables aléatoire est la covariance ($\mathbb{C}\text{ov}[X, Y]$) entre ces deux VA, qui dans le cas d'un échantillon de n objet est estimée comme :

$$\text{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

On note immédiatement que $\text{cov}(x, y) = s_X^2$, c'est à dire que la covariance est une généralisation de la variance à deux variables aléatoires. On notera aussi que la covariance de l'échantillon

peut se lire graphiquement en découpant l'espace en 4 quarts, définit par les moyenne empirique $\bar{X}$ et $\bar{Y}$. Les points situés (x_i, y_i) dont les deux valeurs sont supérieures aux deux moyennes contribuent positivement à la covariance : $(x_n - \bar{X}) > 0, (y_n - \bar{Y}) > 0$. De même pour les points dont les deux VA sont situées en dessous des moyennes $(x_n - \bar{X}) < 0, (y_n - \bar{Y}) < 0$. Ainsi, grossièrement si les points sont principalement répartis dans ces deux quarts, la covariance est positive. S'ils sont situés plutôt dans les deux autres quarts la covariance est négative. S'ils sont équi-répartis dans les 4 quarts, la covariance est proche de 0.

La covariance caractérise donc par son signe la relation entre les deux variables aléatoires. On peut montrer qu'elle est maximale lorsque les deux valeurs ont une relation linéaire, c-a-d qu'elles se représentent sur une droite. Dans ce cas, la valeur absolue de la covariance est $\sqrt{s_X^2 s_Y^2}$.

Corrélation

La corrélation entre deux VA aléatoires X et Y est caractérisée par le coefficient de corrélation, noté $r_{x,y}$ ou encore $\mathrm{cor}(x, y)$ définit comme :

$$\begin{aligned} r_{x,y} &= \mathrm{cov}(x, y)/\sqrt{s_X^2 s_Y^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{X})^2 \sum_{k=1}^{n}(y_k - \bar{Y})^2}} \end{aligned}$$

En clair, le coefficient de corrélation est une covariance normalisée par sa valeur maximale (en absolu). Les valeurs possibles de ce coefficient de corrélation sont donc $r_{x,y} \in [-1, 1]$. Il vaut 0 lorsque la covariance est nulle. Il vaut 1 lorsqu'elle est positive et maximale : les couples x_i, y_i sont sur une droite parfaite de pente positive. Il vaut -1 lorsqu'elle négative et maximale (une droite de pente négative).

Si les deux variables aléatoires sont indépendantes ($\mathbb{C}\mathrm{ov}[X, Y] = 0$), il est cependant possible qu'un échantillon de ces variables aléatoires aient un covariance estimée non-nulle et donc un $r_{x,y} \neq 0$. Par exemple, dans le cas d'un échantillon de taille 2 $\{(x_1, y_1), (x_2, y_2)\}$, il est toujours de faire passer une droite parfaite de pente non nulle (sauf exception) entre ces deux points. Pourtant les deux VA pourraient être complètement indépendante. Il existe donc un test statistique permettant de tester l'indépendance entre les deux VA à partir de $r_{x,y}$.

Si les deux variables aléatoires sont indépendantes, leur coefficient de corrélation est d'espérance nulle (en moyenne, il vaut zéro) mais peut s'en écarter. On peut montrer que, si $\mathbb{C}\mathrm{ov}[X, Y] = 0$, alors pour un échantillon de taille n, la VA T_r (qui est une transformation de ce coefficient de corrélation) suit assymptotiquement une loi de Student à $n - 2$ degrés de libertés. Pour une valeur de $r_{x,y}$, noté simplement r ci-dessous, on calcule t_r comme :

$$t_r = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$

On regarde ensuite où se situe t_r dans la distribution de Student pour rejetter l'indépendance et/ou calculer une p-valeur.

Régression

Bien que la covariance et la corrélation permettent de caractériser quantitativement et statistiquement la relation entre deux VA, on pourrait souhaiter produire un modèle qui, à partir de d'une valeur x_i pourrait permettre de prédire sa valeur y associée, notée $\hat{y}_i$. Une des méthodes possibles pour déterminer un bon modèle est la méthode des moindres carrés. Dans cette méthode, on calcule la somme des carrés des écarts entre les valeurs prédites $\hat{Y}$ et les valeur observées Y, ici notée S :

$$S = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

On cherche un modèle dont les prédictions sont au plus proches des données observées, c-a-d tel que cette somme soit minimale. Un des modèles les plus simples est celui d'une relation affine (linéaire à un centrage près) entre les deux variables $Y = aX + b$. Dans ce cas, S devient :

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

La somme est indicée par (a,b) pour indiquée que sa valeur changera en fonction de la pente (a) et de l'intercept (b) choisie. Dans la méthode des moindres carrés, on choisira ces deux valeurs tels que $S_{a,b}$ soit minimale.

Dans ce cas, on peut montrer (en calculant les valeurs de a et b où les dérivées partielles de cette somme s'annulent) que ces deux valeurs $(\hat{a}, \hat{b})$ sont données par :

$$\begin{aligned} \hat{a} &= \frac{\mathrm{cov(x,y)}}{s_X^2} \\ \hat{b} &= \bar{Y} - \hat{a}\bar{X} \end{aligned}$$

La pente est dont du même signe que la covariance (ouf) et la droite passe par le point d'intersection des deux moyennes, comme illustrer sur la figure ci-dessous :

[FIG - avec x,y une regression illustrée avec les points remarquables]