LA CORRELATION ET LA REGRESSION MULTIPLES

La réalité se réduit rarement à une relation entre deux variables y et x. Elle est plutôt plus complexe et fait appel à plus d'une variable, d'où le recours à des méthodes plus élaborées comme l'analyse multivariée qui fait intervenir plusieurs variables simultanément. A ce titre, il y a plusieurs outils d'analyse selon l'objectif recherché comme la régression multiple, l'analyse factorielle.

La corrélation et la régression multiples permettent l'analyse de la relation entre une variable expliquée ou dépendante (y) et plusieurs variables explicatives ou indépendantes xi (x1, x2,...xn). C'est la généralisation du modèle de la corrélation et de la régression simple à plusieurs variables xi, la même méthode va être utilisée ici à la différence qu'on a désormais affaire à plusieurs variables indépendantes x1, x2…xn.

I - LE MODELE

1 - Le modèle général

Le modèle général s'écrit : $y = f(xi) \pm \varepsilon$ avec xi les variables indépendantes et ε l'élément résiduel. La relation peut être linéaire ou courbe et nous avons vu lors de l'examen de la régression simple comment linéariser une bonne partie des courbes. L'équation s'écrit, sous sa forme linéaire, comme suit :

$y = a_1x_1 + a_2x_2 + ...a_n x_n + b \pm \varepsilon$ ou $y = \Sigma aixi + b \pm \varepsilon$

2 - Le choix du modèle

Pour le choix du modèle, on suit les mêmes règles et étapes examinées ci-dessus dans la régression simple: les considérations théoriques, le nuage de point et la corrélation maximale. Le nuage de points ne permet, pas dans ce cas, directement le choix du modèle dans la meure où on a plusieurs variables à la fois. D'autre part, le nuage de points dans un espace à trois dimensions (ou plus) est peu lisible. Les nuages de points effectués séparément entre y et chacune des variables explicatives xi permettent de voir le type de relations.

3 - La linéarisation

Lorsqu'on a une relation multiplicative ou courbe, on passe par la linéarisation de l'équation en utilisant la transformation logarithmique ou toute autre transformation de nature à linéariser la relation.

Le modèle puissance à deux variables s'écrit par exemple selon l'équation $y = b.x_1^{a1}x_2^{a2}$. En utilisant la transformation logarithmique, on obtient: $\log y = a_1.\log x_1 + a_2.\log x_2 + \log b$, soit $y' = a_1x_{1'} + a_2.x_{2'} + b'$ avec y: log y, x': $\log x_1$, $x_{2'}$: $\log x_2$ et b': logb.

II - LA RESOLUTION

Comme pour la corrélation simple, il s'agit de voir dans quelle mesure il y a une relation entre la variable y et les variables indépendantes xi? Si cette relation est vérifiée significative, et importante à la fois, on passe à la régression pour le calcul des paramètres ai,

1- La corrélation multiple

Le coefficient de corrélation multiple est le rapport entre la covariance des valeurs observées (y) et calculées (y') d'un côté et le produit de leur écart-type.

Ryx1, x2...xn = Cov (y, y')/σy σy'

Cov (y, y') = Σai cov y xi ai : coefficient de régression

En remplaçant les différents paramètres (ai) par leur valeur, les variances et les covariances par les coefficients de corrélation, on obtient :

$R^2 yx_1, x_2...xn = (\Sigma ai.Cov\ y\ xi)/\sigma^2 y$

$Ryx1, x2...xn = ((\Sigma ai.Cov\ y\ xi)/\sigma^2 y)^{1/2}$.

Cas de 2 variables

Pour deux variables par exemple (x1, x2) on a, selon les données disponibles (corrélations, variances) la formule de la corrélation multiple R:

$Ry\ x_1, x_2 = (\sigma^2\ x_1.Cov(yx_2) + \sigma^2 x_2.Cov(yx_1) - 2Cov(yx_1).Cov(yx_2).Cov(x_1x_2))/\sigma^2 y.$
$Ry\ x_1, x_2 = ((a_1.Cov\ Yx_1 + a_2.Cov\ Yx_2)/\sigma^2 y)^{1/2}.$
$Ry\ x_1, x_2 = ((r^2 y\ x_1 + r^2 yx_2 - 2\ ryx_1\ .ryx_2.\ rx_1x_2)\ /\ (1 - r^2\ x_1x_2\ ))\ ½$

Si les variables xi sont indépendantes on obtient ($r_{x1x2} = 0$) : $R^2 yx_1, x_2...xn = \Sigma r^2 yxi$

Les coefficients a_j sont extraits des équations :

$$Cov_{p,1} = a_1 Var_1 + a_2 Cov_{1,2} + ... + a_{p-1} Cov_{1,p-1}$$

$$Cov_{p,2} = a_1 Cov_{2,1} + a_2 Var_2 + ... + a_{p-1} Cov_{2,p-1}$$

...

$$Cov_{p,p-1} = a_1 Covp-1,_1 + a_2 Cov_{1,2} + ... + a_{p-1} Var_{p-1}$$

Les *p-1* coefficients sont ensuite obtenus par résolution du système. Avec deux variables explicatives X_1 et X_2 et une variable à expliquer Y on a par exemple :

$$a_1 = \frac{(Var_{X_2} * Cov_{YX_1}) - (Cov_{YX_2} * Cov_{X_1X_2})}{(Var_{X_1} * Var_{X_2}) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * (r_{YX_1} - (r_{YX_2} * r_{X_1X_2}))}{\sigma_{X_1} * (1 - r^2_{X_1X_2})}$$

$$a_2 = \frac{(Var_{X_1} * Cov_{Y_{21}}) - (Cov_{YX_1} * Cov_{X_1X_2})}{(Var_{X_1} * Var_{X_2}) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * (r_{YX_2} - (r_{YX_1} * r_{X_1X_2}))}{\sigma_{X_{21}} * (1 - r^2_{X_1X_2})}$$

Le coefficient de corrélation multiple est alors donnée par :

$$R_{Y,X_1X_2} = \sqrt{\frac{\left(r_{YX_1}^2 + r_{YX2}^2 - 2\left(r_{YX_1} * r_{YX_2} * r_{X_1X_2}\right)\right)}{1 - r_{X_1X_2}^2}} = r_{YY'}$$

III - LA CORRELATION PARTIELLE

La corrélation partielle est la corrélation entre deux variables lorsque les autres variables sont (réellement ou supposées) constantes. Elle est notée $Ry.x_1,x_2$: corrélation partielle entre y et x_1 lorsque x_2 est constante ou supposée fixe. Elle exprime la corrélation entre la variable y et x_1 dans les zones qui ont la même valeur de x_2. C'est le cas par exemple de la corrélation entre le niveau de dépense et la scolarisation dans les régions qui ont le même taux d'urbanisation : Rds,u. La corrélation partielle permet ainsi d'éliminer l'effet de certaines variables (elles sont ou supposées constantes) afin de ne pas brouiller la relation entre les variables concernées par l'étude ce qui correspond dans le domaine des sciences de la matière à fixer les conditions de l'expérimentation (la température, la pression ou l'humidité….).

Contrairement à la corrélation simple ou multiple où on a un seul coefficient, on a plusieurs corrélations partielles en fonction du nombre de variables. L'ordre est exprimé par le nombre de variables secondaires supposées constantes. Chaque ordre s'exprime aussi par celui immédiatement inférieur. Un coefficient de corrélation partielle d'ordre n, s'exprime de (n -1) manières différentes.

eX : si la relation entre les ventes

territoires de vente	ventes (X1)	population (X2)	revenu (X3)
1	3	2	6
2	6	5	4
3	4	3	5
4	6	7	2
5	4	2	6
6	8	6	3
7	6	4	4
8	9	9	3
9	9	8	2
10	5	4	5