Regression Multiple

- Régression multiple
- Coefficient de corrélation partielle
- Test de signification des coefficients de corrélation partielle

Régression multiple

Dans le cadre d'une régression simple, la corrélation se fait entre 2 variables seulement mais en pratique cela ne suffit pas car la majorité des variables dépendent de plus d'un facteur ou d'une variable.

Si on désigne par y le rendement agricole d'une culture, il est évident que ce rendement dépend de plus d'un facteur dont on peut citer : les engrais (X_1), eau (X_2), climat (X_3), …….X_n , donc on peut résumer ceci pas l'équation suivante:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots\ldots\ldots + \beta_n X_n$$

où X_1, X_2, X_3, ……, X_n sont des variables explicatives qui doivent être indépendantes, Y la variable dépendante á expliquer et β_0, β_1, β_2, …….β_n sont des constantes de régression.

Si on se limite au cas d'une régression multiple de 3 variables avec Z comme variable dépendante et X et Y comme variables explicatives (c-a-d, $Z=\beta_0+\beta_1 X+\beta_2 Y$), le principe consiste à estimer le modèle de la régression multiple par la détermination de β_0, β_1, et β_2, sur la base des données empériques des variables étudiées par la méthode des **moindres carrés,** et d'apprécier l'intensité des corrélation de Y et X1, ou de Y et X2 ou de X1 et X2 par exemple en faisant abstraction des autres variables par le biais des **cœfficients de corrélation partielle** ainsi que **le coefficient de corrélation multiple.**

Un coefficient de corrélation partielle de Z et Y est le coefficient de corrélation des résidus de la régression linéaire simple de Z en fonction X et de la régression linéaire simple de Y en fonction de X. Il est toujours compris entre -1 et +1 et il est désigné par $\mathbf{r_{yZ,\,x}}$ et calculé á partir des cœfficients de corrélation simple grâce á la relation suivante : $r_{yZ,X} = \dfrac{\left(r_{yZ} - r_{yX}.r_{XZ}\right)}{\sqrt{\left(1 - r_{yX}^2\right)\left(1 - r_{ZX}^2\right)}}$.

Le coefficient de corrélation multiple de Z en fonction de X et de Y exprime la corrélation des valeurs observées de la valeurs dépendante Z avec ses valeurs estimés Z^* obtenues par le modèle de régression multiple linéaire de Z en fonction de X et Y. Il est désigné par $R_{z,xy}$. Le $R^2_{z,xy}$ qui est nommé coefficient de détermination multiple

variant entre 0 et 1, peut être lui aussi calculé grâce aux cœfficients de corrélation simple par la relation suivante: $R^2_{z,xy} = \left(\frac{r^2_{XZ} + r^2_{yZ} - 2r_{Xy} r_{Xz} r_{yZ}}{1 - r^2_{Xy}} \right)$

L'exemple suivant cité par Schwartz illustre bien la corrélation multiple:

Exemple: á partir de 200 dossiers (n = 200) dans une maternité, on a prélevé:

X : âge de la mère

Y : poids du bébé à la naissance.

Z : rang de la naissance du bébé.

On a trouvé les corrélations suivantes $r_{xy} = 0{,}24$, $r_{yz} = 0{,}28$, $r_{xz} = 0{,}60$

Sous forme de matrice de corrélation on a:

	X	Y	Z
X	1	r_{xy}	r_{xz}
Y	r_{yx}	1	r_{yz}
Z	r_{zx}	r_{zy}	1

Ces cœfficients qui ont été estimés deux á deux comme dans le cas de la régression simples ne suffisent pas pour expliquer la dépendance mutuelle et multiple entre les variables prises en même temps dans le cadre d'une expérimentation.

On remarque que le poids à la naissance est lié positivement, d'une part à l'âge de la mère et d'autre part au rang de la naissance, mais ces 2 variables elles mêmes sont très liées par un coefficient de corrélation $r_{xz} = 0{,}60$ qui est très significative. Il est donc intéressant de connaître le rôle relatif des variables X et Y: pour des naissances de même rang le poids est il encore lié à l'âge de la mère ? et pour des mères d'âge donné, le poids est-il lié au rang de naissance ?

On peut donc calculer le coefficient de corrélation entre **le poids à la naissance** et **l'âge de la mère** pour différents rangs de naissance, et le coefficient de corrélation entre le poids á la naissance et le rang de naissance pour différentes tranches d'ages de la mère. On doit donc fixer Z pour trouver la relation entre Y, X, et fixer X, pour trouver la relation entre Y ou Z.

Sans recourir á tout cela, **on peut à partir des 3 coefficient de corrélation** r_{xy}, r_{xz}, r_{yz} estimer les **coefficient de corrélation partielles** par les formules suivantes:

1- coefficient de corrélation entre Y et Z pour X constant :

$$r_{yZ,X} = \frac{\left(r_{yZ} - r_{yX}.r_{XZ}\right)}{\sqrt{\left(1 - r^2_{yX}\right)\left(1 - r^2_{ZX}\right)}}$$

2- coefficient de corrélation entre Z et X pour Y constant :

$$r_{ZX,y} = \frac{\left(r_{ZX} - r_{Zy}.r_{Xy}\right)}{\sqrt{\left(1-r_{Zy}^2\right)\left(1-r_{Xy}^2\right)}}$$

3- coefficient de corrélation entre X et y pour Z constant :

$$r_{Xy,Z} = \frac{\left(r_{Xy} - r_{XZ}.r_{yZ}\right)}{\sqrt{\left(1-r_{XZ}^2\right)\left(1-r_{yZ}^2\right)}}$$

Exemple: des données précédentes on peut calculer les cœfficients de corrélation partielles suivants :

$$r_{Xy,Z} = \frac{0,24 - (0,60.0,28)}{\sqrt{(1-0,60^2)(1-0,28^2)}} = 0,09$$

$$r_{ZX,y} = \frac{0,60 - (0,28.0,24)}{\sqrt{(1-0,28^2)(1-0,24^2)}}$$

$$r_{yZ,X} = \frac{0,28 - (0,24.0,60)}{\sqrt{(1-0,24^2)(1-0,60^2)}} = 0,18 \qquad r_{xy} = 0,24,\ r_{yz} = 0,28,\ r_{xz} = 0,60$$

Le coefficient de corrélation multiple est :

$$R_{z,xy}^2 = \left(\frac{0,6^2 + 0,28^2 - 2*0,24*0,6*0,28}{1-0,24^2}\right) = 0,38 \Rightarrow R_{z,xy} = 0,62$$

Test de signification des coefficients de corrélation partielle :

Le test est semblable à celui des coefficients de corrélation entre 2 variables, le ddl étant toute fois pour 3 variables ddl = n – 3 on peut utiliser soit la table de r, soit la table de Student.

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-3}$$

Exemple précédent:

$H_0 : r = 0$

Pour $r_{xy,z} = 0,09 \qquad \Rightarrow \qquad t = \frac{0,09}{\sqrt{1-0,09^2}}\sqrt{200-3}$

$$t = 1,26$$

à ddl = 197 et α =5% $\Rightarrow$ 1,26 < 1,96 le test n'est pas significative au seuil de 5% $\Rightarrow$ H_0 est acceptée, cela veut dire, pour une série de naissance de rang donné il n'y a pas de relation significative entre le poids et l'âge.

Pour $r_{yz,x} = 0,18 \Rightarrow t = \frac{0,18}{\sqrt{1-0,18^2}}\sqrt{200-3} = 2,56$

à ddl = 197 et α =5% $\Rightarrow$ 2,56 > 1,96 le test est significative $\Rightarrow H_0$ est rejetée, donc pour des séries de naissance correspondantes à un même âge donné de la mère, le poids à la naissance est **lié significativement** au rang de la naissance.

Il semble bien á partir de ces résultats, que des 2 variables qui paraissent intervenir, une seule «le rang de naissance, Z) soit intéressante, l'âge de la mère n'étant lié au poids que par l'intermédiaire du rang de naissance.

Remarque :

Pour chaque coefficient de corrélation on doit calculer un intervalle de confiance comme précédemment. Cependant le problème majeur rencontré quant on veut procéder aux méthodes de la régression multiple est celui du choix des variables explicatives car il est très important de savoir quelle variable doit être expliquée par le modèle de régression. Si non, on procède á faire toutes les combinaisons possibles avec toutes les variables mise en jeu et on obtient ainsi pour 3 variables par exemple 2^3-1=7 équations correspondant á 7 modèles de régression simple et multiple différents. On ne retient dans ce cas que les équations de régression qui ont une variance résiduelle minimum.

Il faut remarquer que pour 4 variables on aura 2^4-1=15 équations de régression, et pour 10 variables on aura 1023 équations et ainsi de suite, c'est pour cela il est conseillé dans la pratique de se limiter á 3 ou 4 variables dans l'emploi de la régression multiple afin de maîtriser l'analyse et l'interprétation des corrélations qui en découlent.