Corrélation - Régression

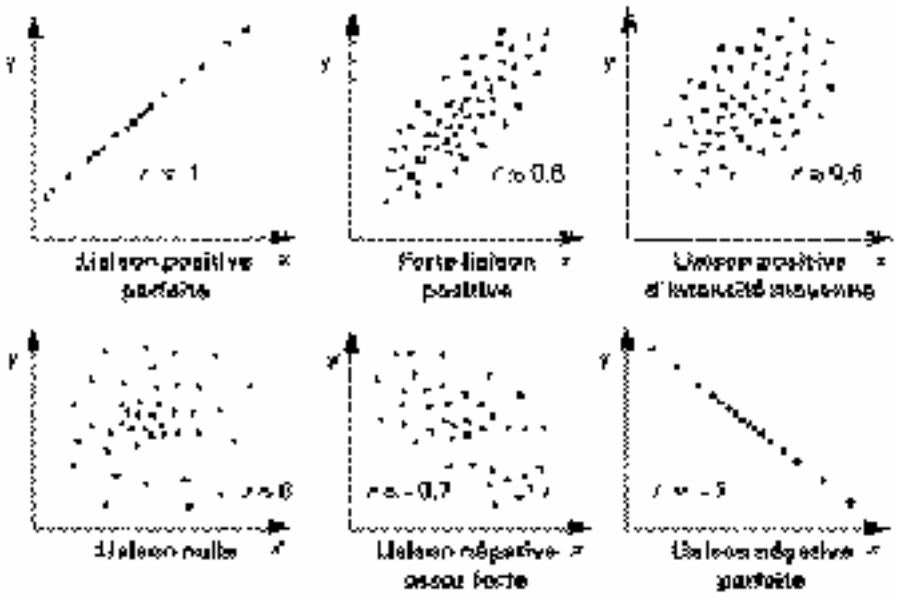
Corrélation

L'objectif de ces méthodes (Corrélation et Régression) est de tester l'existence de liens statistiques entre 2 variables QUANTITATIVES ou, ce qui est équivalent, de tester l'indépendance entre ces variables. La corrélation et la Régression nous permettront de tester l'existence de ce lien, son intensité et sa forme.

A1/ Principe et Calcul

Lorsque que l'on observe n couples (X,Y), r est le coefficient de corrélation mesuré dans un échantillon et ρ le coefficient de corrélation de la population dont est issue l'échantillon.

$$r = \frac{s_{XY}}{\sqrt{s_X^2 . s_Y^2}} = \frac{\sum^n (x-\bar{x})(y-\bar{y})}{\sqrt{\sum^n (x-\bar{x})^2 \sum^n (y-\bar{y})^2}}$$



A2/ Test de la Significativité de r (ρ)

Les hypothèses testées :

H_0 : $\rho = 0$ ($r = 0$) indépendances des variables X et Y, pas de liens statistiques entre ces variables
H_1 : $\rho \neq 0$ ($r \neq 0$) les variables X et Y sont dépendantes

Les conditions d'applications (cf. Régression)

La statistique du test :

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

t_r suit une loi de Student à n-2 ddl donc

si $|t_r| > t_{seuil}$ => Rejet de H_0 (avec $t_{seuil} = t_{1-\alpha/2,n-2}$)

A3/ Intervalles de Confiance de ρ

$$Z_r = \frac{1}{2} Ln\left(\frac{1+r}{1-r}\right)$$

Z_r suit une loi de Normale de moyenne E(Z_r) et de Variance Var(Z_r) avec :

$$E(Z_r) = \frac{1}{2} Ln\left(\frac{1+\rho}{1-\rho}\right) \quad \text{et} \quad Var(Z_r) = \frac{1}{n-3}$$

L'intervalle de confiance de Z_r à 1-α est $IC(1-\alpha) = \left[Z_r \pm U_{1-\alpha/2}\sqrt{\frac{1}{n-3}}\right]$, la transformation inverse permettra d'obtenir l'intervalle de confiance de ρ.

A4/ Comparaison de 2 Coefficients de Corrélation

$$Z_{r_1} = \frac{1}{2} Ln\left(\frac{1+r_1}{1-r_1}\right) \quad \text{et} \quad Z_{r_2} = \frac{1}{2} Ln\left(\frac{1+r_2}{1-r_2}\right) \quad \text{avec}$$

$$E(Z_{r_1}) = \frac{1}{2} Ln\left(\frac{1+\rho_1}{1-\rho_1}\right) \quad \text{et} \quad Var(Z_{r_1}) = \frac{1}{n_1-3}$$

$$E(Z_{r_2}) = \frac{1}{2} Ln\left(\frac{1+\rho_2}{1-\rho_2}\right) \quad \text{et} \quad Var(Z_{r_2}) = \frac{1}{n_2-3}$$

Les hypothèses testées :

$H_0 : \rho_1 = \rho_2 \ (Z_{r_1} = r_2)$
$H_1 : \rho_1 \neq \rho_2$

La statistique du test :

$$U = \frac{\left|Z_{r_1} - Z_{r_2}\right|}{\sqrt{1/(n_1-3) + 1/(n_2-3)}}$$

U suit une loi Normale donc

si $|U| > U_{1-\alpha/2}$ => Rejet de H_0

Régression

B1/ Principe et Calcul

Avec *n* couples (X,Y), la droite de régression est :

$$\hat{Y} = a.X + b$$

avec :

$$a = \frac{s_{XY}}{s_X^2} = r\frac{s_Y}{s_X} \quad \text{et} \quad b = \bar{y} - a.\bar{x}$$

B2/ Test de la Significativité de *a* (α_r)

Les hypothèses testées :

> H_0 : $\alpha_r = 0$ ($a = 0$) indépendances des variables X et Y, pas de liens statistiques entre ces variables
> H_1 : $\alpha_r \neq 0$ ($a \neq 0$) les variables X et Y sont dépendantes

Les conditions d'applications (cf. B3)

La statistique du test :

$$t_a = \frac{a}{\sqrt{Var(a)}} \quad \text{avec} \quad Var(a) = \frac{\sigma^2}{(n-1).s_X^2} = \frac{(s_Y/s_X)^2 - a^2}{n-2}$$

t_a suit une loi de Student à *n-2* ddl donc

$$\text{si } |t_a| > t_{seuil} \Rightarrow \text{ Rejet de } H_0 \quad (\text{avec } t_{seuil} = t_{1-\alpha/2,n-2})$$

On peut remarquer que :

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

B3/ Les Conditions d'Application

(i) Normalité des Y_i et Homogénéité des variances des Y_i (σ_Y^2). La distribution conditionnelle des Y_i par rapport à X ($Pr(Y_i|X_i)$) doit être Normale et de même variance.

(ii) Linéarité : les $E(Y_i|X_i)$ se situent sur une droite δ :

$$E(Y_i|X_i) = \alpha_r . X_i + \beta_r$$

(iii) Indépendances des Y_i :

$$\hat{Y}_i = \alpha_r . X_i + \beta_r + e_i$$

les e_i sont Normaux de moyenne $E(e_i) = 0$ et de variance σ_e^2 estimée par s_e^2 .

B4/ Variance Expliquée par la Droite de Régression

Les hypothèses testées :

> H_0 : La variance liée (expliquée par la droite de régression) est insuffi-
> -sante pour expliquer la variance des Y_i, la droite de régression ne diffère pas de la moyenne de Y_i ($\alpha_r = 0$)
> H_1 : La variance liée (expliquée par la droite de régression) explique une partie significative de la variance des Y_i, ($\alpha_r \neq 0$)

Les conditions d'applications (cf. B3)

Le tableau d'Analyse de Variance est le suivant :

Origine	**Σ des carrés** (a)	**ddl** (b)	**Variance** (a)/(b	**F**
Totale (A)	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	n-1		
Liée (B)	$\sum_{i=1}^{n}((a.x_i + b) - \bar{y})^2$	1	S_{YX}^2	$\frac{S_{YX}^2}{s_e^2}$
Résiduelle (C)	$\sum_{i=1}^{n}(y_i - (a.x_i + b))^2$	n-2	s_e^2	

On peut remarquer que : $F = \frac{S_{YX}^2}{s_e^2} = \frac{(n-2).r^2}{1-r^2}$

En effet, $s_{YX}^2 = r^2.s_Y^2$ et $s_e^2 = (1-r^2).s_Y^2/(n-2)$, le coefficient de détermination r^2, apparaît donc comme le pourcentage de la variance des Y_i, expliqué par la droite de régression.