Chapitre IV: Ajustement linéaire - Régression.

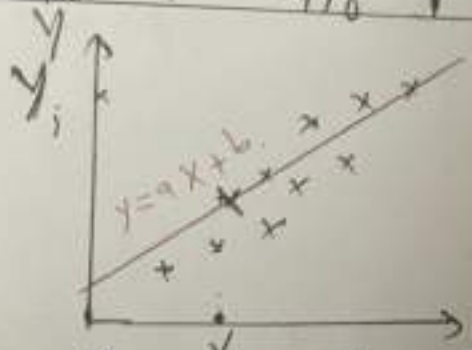
① Ajustement à l'aide d'une droite: Supposons que nous ayons une population de N individus et que nous désirons l'étudier sur deux caractère simultanement. Notre but étant de déterminer et d'analyser la liaison entre ces deux caractères. Par exemple nous disposons d'un échantillon de 10 personnes sur chacune desquelles nous avons mesuré la taille X et le poids Y, nous pouvons classer les données dans un tableau comme suit:

Individus	Taille X(m)	Poids Y(kg)
1	1,55	58,2
2	1,60	58,1
3	1,62	61,3
4	1,64	65,3
5	1,65	69,5
6	1,70	69,7
7	1,72	70,3
8	1,73	75,4
9	1,76	74,2
10	1,78	82,0

Si nous présentons graphiquement ces données, on obtient un nuage de points qui peut être représenté par une droite d'équation:

$$y = ax + b.$$

Il est demandé de déterminer a et b pour établir définitivement la relation entre X et Y.



$$\begin{cases} a = \dfrac{cov(x,y)}{Var(x)} \\ \text{et} \\ b = \bar{y} - a\bar{x}. \end{cases}$$

Exemple de calcul: $\bar{x} = 1,675 \qquad \bar{y} = 68,4.$

$$Var(x) = \left[\frac{1}{10}\sum x_i^2\right] - \bar{x}^2 = 0,005.$$

$$cov(x,y) = \left[\frac{1}{10}\sum_{i=1}^{10} x_i y_i\right] - \bar{x}\bar{y} = 0,4913.$$

et donc: $\begin{cases} a = \dfrac{cov(x,y)}{Var(x)} = 98,86. \\ b = \bar{y} - a\bar{x} = -97,19 \end{cases}$

La droite d'ajustement est alors: $Y = 98,86X - 97,19.$

② Variance résiduelle : La méthode qui nous a permis de définir les paramètre de la droite de régression (a et b), s'appelle la méthode des moindres carrés. Cette méthode nous donne la meilleure droite au sens où elle est la plus proche des points, cela signifie qui minimise la quantité :

$$\sum_{i=1}^{N} [Y_i - (aX_i + b)]^2,$$

ce qui revient au même, celle qui minimise la quantité :

$$Var_r(Y) = \frac{1}{N} \sum_{i=1}^{N} [Y_i - (aX_i + b)]^2.$$

Cette quantité s'appelle la variance résiduelle.
Le calcul de la variance résiduelle dans le cas de la droite se fait par la formule :

$$Var_r(Y) = Var(Y)(1 - \rho^2).$$

En effet : $Var(Y) = \frac{1}{N} \sum_{i=1}^{N} [Y_i - aX_i - b]^2.$

$$= \frac{1}{N} \sum_{i=1}^{N} [Y_i - aX_i - \bar{Y} + a\bar{X}]^2.$$

$$= \frac{1}{N} \sum_{i=1}^{N} [(Y_i - \bar{Y}) - a(X_i - \bar{X})]^2.$$

En développant l'expression entre crochets on obtient :

$$Var_r(Y) = \frac{1}{N} \sum_{i=1}^{N} [(Y_i - \bar{Y})^2 - 2a \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})(X_i - \bar{X}) + a^2 \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

$$= Var(Y) - 2a\, cov(X,Y) + a^2 Var(X).$$

$$= Var(Y) - \frac{2\, cov^2(X,Y)}{Var(X)} + \frac{cov(X,Y)}{Var(X)} \quad \left(\text{car } a = \frac{cov(X,Y)}{Var(X)} \right)$$

$$Var_r(Y) = Var(Y) - \frac{cov^2(X,Y)}{Var(X)}$$

$$= Var(Y) - \frac{cov^2(X,Y) \cdot \sigma_Y^2}{\sigma_X^2 \sigma_Y^2} = Var(Y)\left[1 - \rho^2\right].$$