(Suite du cours statistique MI)

L'ajustement linéaire par la Méthode des moindres carrés

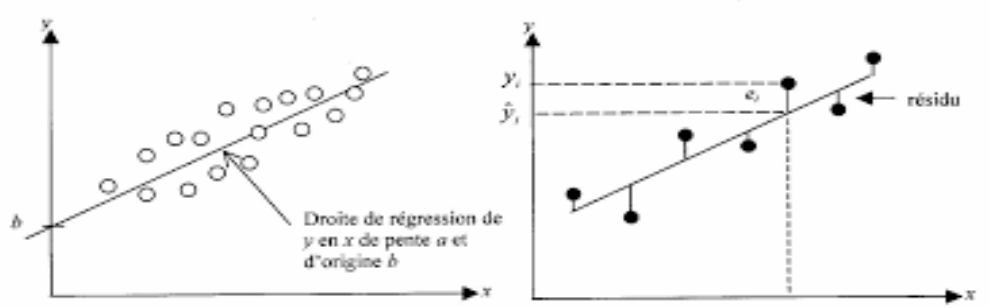
1) Introduction

Une situation courante est d'avoir à sa disposition deux ensembles de données de taille n, {y1, y2,...,yn} et {x1, x2,...,xn}, obtenus expérimentalement ou mesurés sur une population.
Le problème de l'ajustement linéaire ou de la régression consiste à rechercher une relation pouvant éventuellement exister entre les x et les y, par exemple de la forme Y = f(X).
Lorsque la relation recherchée est affine, c'est-à-dire de la forme y = ax + b, on parle de régression linéaire.
Même si une telle relation est effectivement présente, les données mesurées ne vérifient pas en général cette relation exactement. Pour tenir compte dans le modèle mathématique des erreurs observées, on considère les données {y1, y2,...,yn} comme autant de réalisations d'une variable aléatoire Y et parfois aussi les données {x1, x2,...,xn} comme autant de réalisations d'une variable aléatoire X. On dit que la variable Y est la variable dépendante ou variable expliquée et que la variable X est la variable explicative.

2) La droite d'ajustement linéaire

Les données {(xi, yi), i = 1,...,n} peuvent être représentées par un nuage de n points dans le plan (x, y), le diagramme de dispersion. Le centre de gravité de ce nuage peut se calculer facilement : il s'agit du point aux coordonnés $(\bar{X}, \bar{Y})$.
Rechercher une relation affine entre les variables X et Y revient à rechercher une droite qui s'ajuste le mieux possible à ce nuage de points. Parmi toutes les droites possibles, on retient celle qui rend ***minimale la somme des carrés des écarts*** des valeurs observées « yi » à la droite $\widehat{yi}$ =axi+b.
On pose $\varepsilon_i = y_i - \hat{y}_i$, qui n'est autre que la valeur qui représente cet écart (erreur), appelé aussi résidu.
Le principe des moindres carrés (MC) consiste à choisir les valeurs de « a »et de « b » qui minimisent la somme des erreurs commises sur chaque point.
Donc, on cherchera de minimiser cette fonction

$$E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (axi + b))^2$$



Calculons a et b

Remarquons d'abord que notre fonction E a pour inconnues (a,b) les xi et yi étant connus expérimentalement.

Les conditions initiales pour déterminer le minimum d'une fonction sont :

$$\frac{\partial E}{\partial a} = 0 \quad et \quad \frac{\partial E}{\partial b} = 0$$

$$\frac{\partial E}{\partial b} = 0 \Leftrightarrow \sum_{i=1}^{n} 2.(-1)(yi - axi - b) = 0$$
$$\Leftrightarrow \sum_{i=1}^{n}(yi - axi - b) = 0$$
$$\Leftrightarrow \sum_{i=1}^{n} yi - a.\sum_{i=1}^{n} xi - \sum_{i=1}^{n} b = 0$$
$$\Leftrightarrow n.\bar{y} - a.n.\bar{x} - n.b = 0$$
$$\Leftrightarrow \bar{y} - a.\bar{x} - b = 0$$
$$\Leftrightarrow b = \bar{y} - a.\bar{x}$$

Premier résultat : $\boxed{b = \bar{y} - a.\bar{x}}$

$$\frac{\partial E}{\partial a} = 0 \Leftrightarrow \sum_{i=1}^{n} 2.(-a.xi)(yi - axi - b) = 0$$
$$\Leftrightarrow \sum_{i=1}^{n} xi.(yi - axi - b) = 0$$
$$\Leftrightarrow \sum_{i=1}^{n} xi.yi - a.\sum_{i=1}^{n} xi^2 - \sum_{i=1}^{n} b.xi = 0$$
$$\Leftrightarrow n.\overline{xy} - a.n.\overline{x^2} - n.b\,\bar{x} = 0$$
$$\Leftrightarrow \overline{xy} - a.\overline{x^2} - b\,\bar{x} = 0$$
$$\Leftrightarrow \overline{xy} - a.\overline{x^2} - (\bar{y} - a.\bar{x})\bar{x} = 0 \quad car\ b = \bar{y} - a.\bar{x}$$
$$\Leftrightarrow \overline{xy} - a.\overline{x^2} - \bar{y}.\bar{x} + a.\bar{x}^2 = 0$$
$$\Leftrightarrow -a.\left(\overline{x^2} - \bar{x}^2\right) = -(\overline{xy} - \bar{y}.\bar{x}\,)$$
$$\Leftrightarrow a.V(X) = cov(X,Y)$$
$$\Leftrightarrow a = \frac{cov(X,Y)}{V(X)}$$

Deuxième résultat : $\boxed{a = \frac{cov(X,Y)}{V(X)}}$

Conclusion :

La droite de régression Y/x, (y par rapport à x) est : y=ax+b

Avec

$a = \frac{cov(X,Y)}{V(X)}$ et $b = \bar{y} - a.\bar{x}$

ie y= $\frac{cov(X,Y)}{V(x)}.x + \bar{y} - \frac{cov(X,Y)}{V(x)}.\bar{x}$

Remarques :

1) La minimisation de E implique d'autre conditions sur les dérivées du second ordre, ceci n'étant pas le but de notre cours, on ne s'attarde pas dessus mais elles sont vérifiés.
2) Dans le calcul de la droite des moindres carrés, les variables X et Y ne jouent pas des rôles interchangeables. La variable dépendante Y prend, comme son nom l'indique, des valeurs qui dépendent de celles de X.
   D'ailleurs si l'on échange les rôles de X et de Y, on aura une approximation linéaire de la forme $\widehat{xi}=a'yi + b'$, le critère des MC est alors

$$E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (x_i - \widehat{xi})^2 = \sum_{i=1}^{n} \left(x_i - (a'yi + b')\right)^2$$

et ce n'est plus le même et la droite que l'on obtient en général. Cette droite, tout comme la précédente, passe par le centre de gravité du nuage de point, mais c'est généralement leur seul point commun.

C'est le problème considéré qui indique s'il faut considérer Y ou plutôt X comme variable dépendante (et l'autre comme variable explicative).

Mais si l'on s'intéresse aux interactions entre deux variables X et Y dont ni l'une ni l'autre n'est clairement dépendante de l'autre, alors on pourra choisir de régresser Y en fonction de X ou bien le contraire. Mais on ne doit pas s'attendre à obtenir les mêmes résultats.

Ainsi la droite de régression de X/y (x par rapport à y) est : x=a'y+b'

Avec

$a' = \frac{cov(X,Y)}{V(Y)}$ et $b' = \bar{x} - a'.\bar{y}$

ie x= $\frac{cov(X,Y)}{V(y)} . y + \bar{x} - \frac{cov(X,Y)}{V(y)} . \bar{y}$

(Exercice : montrez que les deux droites de régression sont égales si et seulement si a.a'= 1.)

3) Le coefficient de corrélation linéaire

Pour évaluer la qualité de la régression linéaire et ainsi mesurer la qualité de l'approximation du nuage de points, on calcule son coefficient de corrélation linéaire défini par :

$$\rho_{xy} = r_{xy} = \frac{cov(x,y)}{\sigma_x . \sigma_y}$$

Avec $-1 \leq |\rho_{xy}| \leq +1$,

Il vaut +1 (resp. −1) si les points du nuage sont exactement alignés sur une droite de régression.

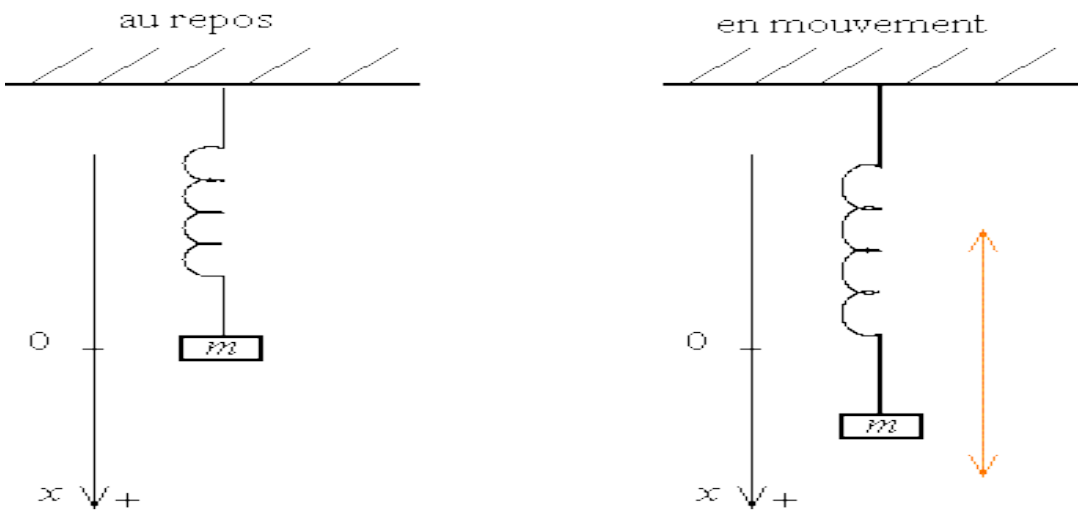
Ce coefficient est une mesure de la dispersion du nuage.

En pratique,

On estimera souvent la régression acceptable (et ainsi la prédiction acceptable) lorsque $|\rho_{xy}| \geq 0.85$

par analogie la régression sera inacceptable lorsque$|\rho_{xy}| < 0.85$.

Exemple



Dans un TP de physique à chaque masse mi (un poids) on obtiendra une extension xi, le tableau des données brutes est le suivant :

Mi (kg)	0	10	20	30	40
Xi (cm)	0	1.1	1.5	1.9	2.5

1) Donner la droite de régression de Y en fonction de X : y=ax+b. (avec Y la variable poids)
2) Donner la droite de régression de X en fonction de Y : x=a'y+b'. (avec Y la variable poids)
3) Cette droite est elle acceptable ? peut-on faire de la prédiction.
4) Si oui, quelle masse aurions nous si on veut une dilatation du ressort égale à xi= 3cm.

Solution

1.

						Somme
Y :mi	0	10	20	30	40	100
X :xi	0	0.5	1.1	1.5	1.9	5
mi.xi	0	5	22	45	76	148
X^2 :xi^2	0	0.25	1.21	2.25	3.61	7.32
Y^2 : mi^2	0	100	400	900	1600	3000

On obtient :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} xi = \frac{1}{5}(0 + 0.5 + \cdots + 1.9) = \frac{5}{5} = 1$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} mi = \frac{1}{5}(0 + 10 + \cdots + 40) = \frac{100}{5} = 20$$

$$\overline{XY} = \frac{1}{n}\sum_{i=1}^{n} xi.mi = \frac{1}{5}(0 + 5 + \cdots + 76) = \frac{148}{5} = 29.6$$

$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} xi^2 = \frac{1}{5}(0 + 0.25 + \cdots + 3.61) = \frac{7.32}{5} = 1.464$$

$$\overline{Y^2} = \frac{1}{n}\sum_{i=1}^{n} mi^2 = \frac{1}{5}(0 + 100 + \cdots + 1600) = \frac{3000}{5} = 600$$

Ainsi

$$V(X) = \overline{X^2} - (\bar{X})^2 = 1.464 - 1^2 = 0.464 \quad \Longrightarrow \quad \sigma_x = \sqrt{V(X)} = 0.681$$

$$V(Y) = \overline{Y^2} - (\bar{Y})^2 = 600 - 20^2 = 200 \quad \Longrightarrow \quad \sigma_y = \sqrt{V(Y)} = 14.142$$

$$cov(X,Y) = \overline{XY} - \bar{X}.\bar{Y} = 29.6 - 1 \times 20 = 9.6$$

En utilisant les formules obtenues avec la méthode des MC, la droite de régression qu'on obtiendra sera :

$$a = \frac{cov(X,Y)}{V(X)} = \frac{9.6}{0.464} = 20.689$$

$$b = \bar{y} - a.\bar{x} = 20 - 20.689 \times 1 = -0.689$$

Conclusion notre droite de régression (Δ) de y en fonction de x, sera la suivante :

(Δ) : y= (20.689)x-0.689

2. En utilisant les formules obtenues avec la méthode des MC, la droite de régression qu'on obtiendra sera :

$$a' = \frac{cov(X,Y)}{V(Y)} = \frac{9.6}{200} = 0.048$$

$$b' = \bar{x} - a'.\bar{y} = 1 - 0.048 \times 20 = 0.04$$

Conclusion notre droite de régression (Δ') de x en fonction de y sera la suivante :

(Δ') : x= (0.048)y+0.04

3. pour savoir si la droite de régression est acceptable, calculons le coefficient de corrélation :

$$\rho_{xy} = \frac{cov(x,y)}{\sigma_x.\sigma_y} = \frac{9.6}{0.681 \times 14.142} = 0.996$$

$$|\rho_{xy}| = 0.996 \geq 0.85 \quad \Longrightarrow la\ droite\ de\ régression\ (\Delta)\ est\ acceptée\ ainsi\ que\ les\ prédictions\ qui\ en\ découlerons$$

4.

Selon la réponse précédente, on peut faire la prédiction du poids si l'extension xi=3cm

y= (20.689)x-0.689 $\Longrightarrow$ yi = (20.689)xi $- 0.689$ = 20.689x3-0.689= 61.378kg

(Corrigé du TD2 sera sur le site)