Chapitre 3

Modèles linéaires

3.1 Notions de base

Observation

Lors d'une expérimentation, on peut être amené à differents mements à mesurer differents paramètres liés au phénomène ou processus étudié. Ainsi des medecins scolaires notent pour chaque élève d'une école, leur taille et leur poids. À chaque élève est associée une observation. L'ensemble des élèves est parfois appelé aussi une population et chaque élève est un individu.

Exemple 3.1.1 *On a mesuré les hauteurs en centimètres des plantes de blé d'un même champ dont on a tiré un echantillon chaque semaine. Les observations sont données dans le tableau suivant :*

âge en semaine x_i	1	2	3	4	5	6	7	8
hauteur en cm y_i	5	12	15	22	35	35	41	48

Les variables observées simultanément sont donc :

X : âge en semaines.

Y : hauteur en centimètres.

Relation fonctionnelle

Étant données deux variables statistiques X et Y, on dit qu'il existe une relation fonctionnelle de X vers Y si à chaque valeur de la variable X est associée une valeur et une seule de la variable Y.

Modèle

S'il existe une relation fonctionnelle entre les variable X et Y et si f est la fonction donnant pour toute valeur x de X la valeur correspondante y de Y, $(Y = f(x))$ on dit que f est un modèle du phénomène étudié.

Valeur observée

Étant donnée une variable statistique X on appelle valeur obsevée de cette variable, toute valeur de cette variable relevée au cours d'une observation.

Valeur théorique

S'il existe une relation fonctionnelle entre les variable X et Y de X vers Y et si f un modèle de cette relation $(Y = f(x))$. À toute valeurs x_i de X est associée une valeur $y_i = f(x_i)$. y_i est la valeur théorique ou valeur expliquée par le modèle.

Variable statistique centrée réduite

Soit X une variable statistique et $\overline{X}$ sa moyenne, σ_X son écart-type.

On appelle varible statistique centrée la varible $X - \overline{X}$.

On appelle varible statistique centrée réduite la varible $\dfrac{X - \overline{X}}{\sigma_X}$.

Varible explicative et variable expliquée

S'il existe une relation fonctionnelle entre les variable X et Y de X vers Y et si f un modèle de cette relation $(Y = f(X))$, on dit que X est la variable explicative et Y est la variable expliquée.

Nuage de points

Soit $\mathbb{R}$ un repère cartésien du plan.

Si au cours de n observations $\omega_1, \omega_2, \ldots, \omega_n$ on relève les valeurs $X(\omega_i)$ et $Y(\omega_i)$ prises par deux variables statistiques X et Y, on appelle nuage de points l'ensemble des points du plan dont les cordonnées dans le repère $\mathbb{R}$ sont les couples $(X(\omega_i), Y(\omega_i))$.

Point moyen

On appelle point moyen d'un nuage de points, le point dont les cordonnées sont les moyennes arithmetiques des cordonnées des points du nuage.

3.2 L'ajustement linéaires

Principe d'ajustement

Lorsqu'on procède à un ajustement, on recherche à partir des observations une fonction notée f telle que les deux variables X et Y soient liées par la relation $Y = f(X)$.

Intérêt de l'ajustement

Connaissant l'équation d'ajustement de tendance générale, il est alors aisé de prévoire les résultas des autres observations. Par exemple le résultat de notre experience à une date ulterieur si notre experience dépond du temps.

3.2.1 régression linéaire simple

Dans ce cas le nuage statistique est approché par une droite, la fonction f est donc donnée par $y = f(x) = ax + b$. Le nombre a est appelé pente de la droite.

La méthode des points moyens (méthode de Mayer)

Après le classement ordonné croissant selon les x_i, on partage le nuage en deux sous nuages égaux (ou égaux à une unité près si le nombre d'obsevations est impair). On considèr pour chaque sous nuage un point appelé point moyen.

Les cordonnées du premier point sont $(\bar{x_1}, \bar{y_1})$.

$\bar{x_1}$: Moyenne arithmétique des abscisses x_i du premier sous nuage.

$\bar{y_1}$: Moyenne arithmétique des ordonnées y_i du premier sous nuage.

On note ce point $G_1(\bar{x_1}, \bar{y_1})$.

On procède en suite de même pour le reste des valeurs du second sous nuage, on obtient un second point moyen noté $G_2(\bar{x_2}, \bar{y_2})$. La droite de Mayer est l'unique droite qui passe par ces deux points, son équation est de la forme $y = ax + b$.

Or $G_1(\bar{x_1}, \bar{y_1})$ appartient à la droite, ces cordonnées vérifient donc l'équation et en conséquence : $\bar{y_1} = a\bar{x_1} + b$.

Il en est de même pour $G_2(\bar{x_2}, \bar{y_2})$ et donc $\bar{y_2} = a\bar{x_2} + b$.

Pour déterminer a et b, on résoud le système de deux équations à deux inconues donnés par :

$$\begin{cases} \bar{y_1} = a\bar{x_1} + b \\ \bar{y_2} = a\bar{x_2} + b \end{cases}$$

En retranchant membre à membre la première équation de la deuxième, on trouve

$$\bar{y_2} - \bar{y_1} = a(\bar{x_2} - \bar{x_1}),$$

alors

$$a = \frac{\bar{y_2} - \bar{y_1}}{\bar{x_2} - \bar{x_1}}.$$

Pour trouver b, il suffit de roporter la valeur de a dans l'une des deux équations précédentes.

Exemple 3.2.1

Le premier sous nuage de la distribution des couples (x_i, y_i) *est l'ensemble des points* $(1,5)$, $(2,12)$, $(3,15)$, $(4,22)$.

On a

$$\bar{x_1} = \frac{1+2+3+4}{4} = 2.5, \quad \bar{y_1} = \frac{5+12+15+22}{4} = 13.5$$

d'où

$$G_1(\bar{x_1}, \bar{y_1}) = G_1(2.5, 13.5).$$

De même, le deuxième sous nuage de la distribution des couples (x_i, y_i) *est l'ensemble des points* $(5,34)$, $(6,35)$, $(7,41)$, $(8,48)$.

On a

$$\bar{x_2} = \frac{5+6+7+8}{4} = 6.5, \quad \bar{y_2} = \frac{34+35+41+48}{4} = 39.5$$

d'où $G_2(\bar{x_2}, \bar{y_2}) = G_2(6.5, 39.5)$.

Pour déterminer a et b, on résoud le système

$$\begin{cases} 13.5 = 2.5a + b \\ 39.5 = 6.5a + b \end{cases}$$

d'où $a = 6.5$, *en portant cette valeur dans la première équation on obtient* $b = -2.75$, *alors* $Y = 6.5x - 2.75$.

La méthode des moindres carrés

Les points du nuage sont numérotés $M_1, M_2, \ldots, M_n$.

Lorsque les points du nuage semblent à peut-près alignés, on essaie de trouver l'équation d'une droite (Δ) *approchant mieux le nuage.*

Le point $G(\bar{x}, \bar{y})$ *s'appelle point moyen du nuage.*

Soit (Δ) *la droite d'équation* $Y = ax + b$.

Chercher l'équation $Y = ax + b$ *de la droite d'ajustement* (Δ) *de y en x par la méthode des moindes carrés revient à chercher a et b, t.q*

$$\begin{cases} a = \dfrac{COV(X,Y)}{V(X)} \\ b = \overline{Y} - a\overline{X} \end{cases}$$

D'autre part $r = \dfrac{COV(X,Y)}{\sigma_X \sigma_Y}$ *est le coefficient de corrélation de X et Y.*

Relation entre le coefficient de corrélation et l'ajustement mathématique

1. *Le coefficient de corrélation mesure la qualité d'ajustement affine.*
2. $-1 \leq r \leq 1$.
3. *Si* $r = 0$ *alors il n y a pas de corrélation entre X et Y (X et Y sont independantes.)*
4. *Si* $0 < r < 1$ *alors il y a une corrélation positive faible, moyenne ou forte entre X et Y.*
5. *Si* $-1 < r < 0$ *alors il y a une corrélation négative faible, moyenne ou forte entre X et Y.*
6. $|r| = 1$ *si et seulement si* $Y = ax + b$.

Remarque 3.2.1

Une relation statistique, détectée par le coefficient de corréltion ou par un grapphique, ne montre jamais de relation causale entre deux variables. La causalité ne peut être déduite que d'une analyse non statistique des données.

Conséquence importante

La droite d'ajustement toujours passe par le point moyen $(\bar{x}, \bar{y})$.

Exemple 3.2.2

On calcule dans un premier temps les résultas intermédiaires en utilisant le tableau suivant :

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	5	5	1	25
2	12	24	4	144
3	15	45	9	225
4	22	88	16	484
5	34	170	25	1156
6	35	210	36	1225
7	41	287	49	1681
8	48	384	64	2304
$\bar{x} = 4.5$	$\bar{y} = 26.5$	$\sum x_i y_i = 1213$	$\sum x_i^2 = 204$	$\sum y_i^2 = 7244$

On cherche la droite d'ajustement de X en Y. On a

$$V(X) = \left(\frac{1}{8}\sum_{i=1}^{8} x_i^2\right) - \bar{x}^2 = \frac{204}{8} - (4.5)^2 = 5.25,$$

$$V(Y) = \left(\frac{1}{8}\sum_{i=1}^{8} y_i^2\right) - \bar{y}^2 = \frac{7244}{8} - (26.5)^2 = 203.25,$$

$$COV(X,Y) = \left(\frac{1}{8}\sum_{i=1}^{8} x_i y_i\right) - \bar{x}\bar{y} = \frac{1213}{8} - 4.5 \times 26.5 = 32.375,$$

alors

$$a = \frac{COV(X,Y)}{V(X)} = 6.166,$$

et

$$b = \bar{y} - a\bar{x} = -1.25,$$

d'où

$$Y = 6.166x - 1.25.$$

D'autre part

$$\sigma_X = \sqrt{V(X)} = 2.29, \quad \sigma_Y = \sqrt{V(Y)} = 14.25,$$

d'où

$$r = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = 0.99,$$

et par suite il y a une corrélation positive forte entre X et Y.

3.2.2 régression linéaire multiple

La régression linéaire multiple est une généralisation, à p variables explicatives, de la régression linéaire simple.

Nous sommes toujours dans le cadre de la régression mathématique : nous cherchons à prédire, avec le plus de précision possible, les valeurs prises par une variable y, dite endogène, à partir d'une série de variables explicatives $x_1, x_2, \ldots, x_p$.

Dans le cas de la régression linéaire multiple, la variable endogène et les variables exogènes sont toutes quantitatives (continues), et le modèle de prédiction est linéaire.

Equation de régression et objectifs

Nous disposons de n observations ($i = 1, 2, \ldots, n$). L'équation de régression s'écrit

$$y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \ldots + a_p x_{i,p} + \epsilon_i.$$

où ϵ_i est l'erreur du modèle, elle exprime, ou résume, l'information manquante dans l'explication linéaire des valeurs de y à partir des x_j. $a_0, a_1, \ldots, a_p$ sont les coefficients (paramètres) du modèle à estimer.

Notation matricielle

Nous pouvons adopter une écriture condensée qui rend la lecture et la manipulation de l'ensemble plus facile. Les équations suivantes

$$\begin{cases} y_1 &= a_0 + a_1 x_{1,1} + a_2 x_{1,2} + \ldots + a_p x_{1,p} + \epsilon_1 \\ y_2 &= a_0 + a_1 x_{2,1} + a_2 x_{2,2} + \ldots + a_p x_{2,p} + \epsilon_2 \\ &\vdots \\ y_n &= a_0 + a_1 x_{n,1} + a_2 x_{n,2} + \ldots + a_p x_{n,p} + \epsilon_n \end{cases}$$

peuvent être résumées avec la notation matricielle $Y = Xa + \epsilon$, avec

— *Y est un vecteur colonne de dimension $(n, 1)$.*
— *X est la matrice des observations de dimension $(n, p + 1)$.*
— *a est un vecteur des coefficients de dimension $(p + 1, 1)$.*
— *ϵ est le vecteur des erreurs de dimension $(n, 1)$.*

La matrice X est égale à

$$\begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{pmatrix}.$$

*Les paramètres de la régression linéaire multiple peuvent être estimés en utilisant **la méthode des moindres carrés**. L'estimation des coefficients $a_0, a_1, \ldots, a_p$ se fait en résolvant le système d'équations obtenu par cette minimisation.*

La solution en notation matricielle est donnée par la formule :

$$\bar{a} = \left(X^t X\right)^{-1} X^t Y.$$

où $\bar{a}$ est le vecteur des coefficients estimés, X est la matrice des variables explicatives, et Y est le vecteur des observations de la variable dépendante.

Exemple 3.2.3

Supposons que nous voulons modéliser le prix de vente d'une maison Y en fonction de deux variables explicatives :

1. *X_1 : La surface de la maison en m^2.*
2. *X_2 : Le nombre de pièces.*

   Nous avons 5 observations (maisons) avec les informations suivantes :

Maison(i)	*Surface X_1*	*Nombre de pièces X_2*	*Prix de vente Y*
1	50	3	200
2	60	4	250
3	80	5	300
4	100	6	400
5	120	7	500

L'équation de régression multiple est :

$$y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \epsilon_i,$$

où :

— *y_i : Le prix de vente de la maison i.*

— *$x_{i,1}$: la surface de la maison i.*

— *$x_{i,2}$: Le nombre de pièces de la maison i.*

— *a_0, a_1, a_2 : Les coefficients à estimer.*

Nous pouvons écrire cela sous forme matricielle :

$$Y = Xa + \epsilon.$$

Où

$$Y = \begin{pmatrix} 200 \\ 250 \\ 300 \\ 400 \\ 500 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 50 & 3 \\ 1 & 60 & 4 \\ 1 & 80 & 5 \\ 1 & 100 & 6 \\ 1 & 120 & 7 \end{pmatrix}, \quad a = \begin{pmatrix} a_0 \\ a_1 \\ a_1 \end{pmatrix}.$$

La solution en notation matricielle est donnée par la formule :

$$\bar{a} = \left(X^t X\right)^{-1} X^t Y,$$

où

$$X^t = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 80 & 100 & 120 \\ 3 & 4 & 5 & 6 & 7 \end{pmatrix}.$$

On a

$$X^t X = \begin{pmatrix} 1 & 50 & 3 \\ 1 & 60 & 4 \\ 1 & 80 & 5 \\ 1 & 100 & 6 \\ 1 & 120 & 7 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 80 & 100 & 120 \\ 3 & 4 & 5 & 6 & 7 \end{pmatrix} = \begin{pmatrix} 5 & 410 & 25 \\ 410 & 36900 & 2230 \\ 25 & 2230 & 135 \end{pmatrix}.$$

$$\det\left(X^t X\right) = \begin{vmatrix} 5 & 410 & 25 \\ 410 & 36900 & 2230 \\ 25 & 2230 & 135 \end{vmatrix}$$

$$= 5[36900\times135-(2230)^2]-410[410\times135-25\times2230]+25[410\times2230-25\times36900] = 2000.$$

$$\left(X^tX\right)^{-1} = \frac{1}{det\,(X^tX)}\left(com\left(X^tX\right)\right)^t$$

$$= \frac{1}{2000}\begin{pmatrix} 8600 & 400 & -8200 \\ 400 & 50 & -900 \\ -8200 & -900 & 16400 \end{pmatrix}^t = \begin{pmatrix} 4.3 & 0.2 & -4.1 \\ 0.2 & 0.025 & -0.45 \\ -4.1 & -0.45 & 8.2 \end{pmatrix}.$$

$$X^tY = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 80 & 100 & 120 \\ 3 & 4 & 5 & 6 & 7 \end{pmatrix}\begin{pmatrix} 200 \\ 250 \\ 300 \\ 400 \\ 500 \end{pmatrix} = \begin{pmatrix} 1650 \\ 149000 \\ 9000 \end{pmatrix}.$$

Alors

$$\bar{a} = \left(X^tX\right)^{-1}X^tY = \begin{pmatrix} 4.3 & 0.2 & -4.1 \\ 0.2 & 0.025 & -0.45 \\ -4.1 & -0.45 & 8.2 \end{pmatrix}\begin{pmatrix} 1650 \\ 149000 \\ 9000 \end{pmatrix} = \begin{pmatrix} -5 \\ 5 \\ -15 \end{pmatrix}.$$

Pour calculer le prix de vente prédit des maisons, on va utiliser l'équation de régression multiple

$$\hat{y}_i = -5 + 5x_{i,1} - 15x_{i,2},$$

alors, on obtient le tableau suivant

Maison	*Surface* X_1	*Nombre de pièce* X_2	*Prix observé* Y	*Prix estimé* $\hat{Y}$
1	50	3	200	200
2	60	4	250	235
3	80	5	300	320
4	100	6	400	405
5	120	7	500	490

Remarquons que :

— *Pour la Maison* 1, *le prix estimé correspond exactement au prix observé* (200).

— *Pour les Maisons* 2, 3, 4 *et* 5, *les prix estimés sont relativement proches des prix observés, avec une différence maximale de* 20 *unités (Maison* 3*).*

Cela montre que le modèle de régression multiple fournit des prédictions raisonnablement proches des valeurs réelles, indiquant une bonne qualité d'ajustement dans l'ensemble. Toutefois, il y a une légère marge d'erreur pour certaines observations, ce qui est attendu dans tout modèle statistique.