Chapitre 1

Le modèle linéaire

1.1 Introduction

Le modèle statistique de base que l'on utilise pour analyser une expérience où l'on étudie sur n unités expérimentales les variations d'une *variable réponse* y en fonction de facteurs qualitatifs ou quantitatifs (appelés aussi *variables explicatives*), peut s'écrire :

$$Y_i = m_i + E_i$$

où

- i est le numéro de l'unité expérimentale,

  m_i est l'espérance de Y_i et inclut l'effet de variables explicatives,

- E_i est une variable aléatoire résiduelle, appelée erreur, incluant la variabilité du matériel expérimental, celle due aux variables explicatives non incluses dans le modèle, et celle due aux erreurs de mesure.

Selon la nature des variables incluses dans la partie explicative m_i du modèle, on distingue trois grandes catégories de modèle linéaire :

- Lorsque les variables explicatives sont quantitatives, le modèle est appelé modèle de régression : simple s'il n'y a qu'une seule variable explicative, multiple sinon. Des exemples sont présentés dans le paragraphe 1.2.1, p. 10 et dans deux exemples détaillés dans les parties 2.1, p. 44 et 2.2, p. 50.
- Lorsque les variables explicatives sont qualitatives, elles sont appelées **facteurs** et le modèle ainsi construit est un modèle d'analyse de la variance. Ce modèle est construit sur un exemple dans le paragraphe 1.2.2, p. 11 ci-dessous, puis étudié en détail dans les exemples 2.3, p. 59, 2.4, p. 67, 2.5, p. 79.
- Lorsque les variables explicatives sont à la fois de nature quantitatives et qualitatives, le modèle ainsi construit est un modèle d'analyse de la covariance. Il est brièvement présenté dans le paragraphe 1.2.3, p. 12, puis étudié en détail au travers d'un exemple 2.6, p. 91.

1.2 Modélisation

1.2.1 Modèle de régression

Les brochets sont des prédateurs supérieurs qui cumulent l'ensemble des pesticides présents aux différents niveaux trophiques. Une étude cherche à comprendre si de plus, ils cumulent ces pesticides au cours de leur vie. Dans cet objectif, on souhaite quantifier le lien entre la concentration en DDT et l'âge, variable $x^{(1)}$. Un modèle de régression simple qui étudie le lien entre ces variables s'écrit :

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + E_i, \quad E_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2),$$

où *i.i.d.* signifie que les variables sont indépendantes et identiquement distribuées. Ce modèle étant linéaire en ses paramètres, il peut se mettre sous une forme matricielle. Le vecteur $Y = (Y_1, \ldots, Y_n)'$ est le vecteur des variables à expliquer, le vecteur $E = (E_1, \ldots, E_n)'$ est le vecteur des erreurs résiduelles. Le vecteur de paramètres θ est défini par $\theta = (\beta_0, \beta_1)'$. Enfin la variable explicative et le terme constant sont stockés dans la matrice d'incidence, parfois appelée matrice de design, X, qui s'écrit donc

$$X = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_2^{(1)} \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_n^{(1)} \end{bmatrix}.$$

Le **modèle de régression simple** s'écrit alors sous la forme

$$Y = X\theta + E, \quad E \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

Si l'on souhaite inclure davantage de variables explicatives, on se trouve dans le cadre d'un modèle de régression linéaire multiple qui s'écrit

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)} + E_i, \quad E_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2).$$

En écrivant $\theta = (\beta_0, \beta_1, \ldots, \beta_p)'$ et

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(p)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(p)} \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(p)} \end{bmatrix},$$

on écrit **le modèle de régression multiple** sous sa forme matricielle

$$Y = X\theta + E, \quad E \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

1.2.2 Modèle d'analyse de la variance (anova)

Pour comparer les rendements de cinq variétés de blé, 4 parcelles sont ensemencées pour chacune des cinq variétés étudiées, puis le rendement final y est mesuré. La variable explicative variété est qualitative, elle est souvent appelé facteur explicatif. Ce facteur possède cinq niveaux. Le modèle d'analyse de la variance à un facteur s'écrit alors, sous sa forme régulière :

$$Y_{ik} = \mu_i + E_{ik}, \quad i = 1, \dots I, k = 1, \dots n_i, \quad E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \tag{1.1}$$

où i désigne le niveau i du facteur et k le numero de l'observation au sein de ce niveau i. I désigne le nombre total de niveaux de ce facteur, n_i le nombre d'observations pour le niveau i et $n = \sum_{i=1}^{I} n_i$ le nombre total d'observations. Dans le cas présent, on a $I = 5, n_i = 4,\ i = 1, \dots, 5$ et $n = 20$. Lorsqu'on veut dissocier un effet commun à toutes les variétés et un effet différentiel de chaque espèce par rapport à un comportement de référence, le modèle peut s'écrire sous sa forme singulière (forme singulière qui permettra une généralisation plus simple au cas à plus de deux facteurs) :

$$Y_{ik} = \mu + \alpha_i + E_{ik}, \quad E_{ik} \sim \mathcal{N}(0, \sigma^2). \tag{1.2}$$

Sous cette forme, le modèle possède un paramètre supplémentaire et n'est plus identifiable[1]. Ce problème est abordé dans le paragraphe 1.3.1, p. 18 de ce chapitre.

En posant

$$Y = (Y_{11} \dots, Y_{1n_1}, Y_{21} \dots, Y_{2n_2}, Y_{I1} \dots, Y_{In_I})',$$

$$E = (E_{11} \dots, E_{1n_1}, E_{21} \dots, E_{2n_2}, E_{I1} \dots, E_{In_I})',$$

puis $\theta = (\mu, \alpha_1, \dots, \alpha_I)'$,

et en notant $\mathbf{1}_{n_i}$ le vecteur de taille n_i ne contenant que des 1.

1. Un modèle est identifiable si pour deux jeux de paramètres différents θ_1 et θ_2 la loi des observations sous θ_1 est différente de la loi des observations sous θ_2 .

$$X = \begin{bmatrix} 1 & & 0 & 0 & 0 & 0 & 0 \\ \vdots & \mathbf{1}_{n_i} & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & & 0 & \vdots & \vdots & \vdots & \vdots \\ \vdots & 0 & & \vdots & \vdots & \vdots & \\ \vdots & \vdots & \mathbf{1}_{n_2} & \vdots & \vdots & \vdots & \\ \vdots & \vdots & & 0 & \vdots & \vdots & \\ \vdots & \vdots & 0 & & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \mathbf{1}_{n_3} & \vdots & \vdots & \\ \vdots & \vdots & \vdots & & 0 & \vdots & \\ \vdots & \vdots & \vdots & 0 & & \vdots & \\ \vdots & \vdots & \vdots & \vdots & \mathbf{1}_{n_4} & \vdots & \\ \vdots & \vdots & \vdots & \vdots & & 0 & \\ \vdots & \vdots & \vdots & \vdots & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \mathbf{1}_{n_5} & \\ 1 & 0 & 0 & 0 & 0 & & \end{bmatrix},$$

le **modèle d'analyse de la variance** se met sous la forme matricielle suivante :

$$Y = X\theta + E, \quad E \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

1.2.3 Modèle d'analyse de la covariance (ancova)

Si dans l'expérience précédente, on veut prendre en compte une quantité d'azote x différente dans chacune des parcelles de l'expérience, il s'agit alors de proposer un modèle qui permet d'utiliser à la fois une variable quantitative et un facteur pour expliquer la variabilité du rendement et savoir si la réponse à l'azote est la même ou non pour toutes les variétés.
La forme régulière du modèle d'analyse de la covariance est donnée par

$$Y_{ik} = \mu_i + \beta_i x_{ik} + E_{ik}, \quad E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2).$$

En écrivant $Y = (Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{In_I})'$,
$E = (E_{11}, \ldots, E_{1n_1}, E_{21}, \ldots, E_{I1}, \ldots, E_{In_I})'$, $\theta = (\mu_1, \ldots, \mu_I, \beta_1, \ldots, \beta_I)'$ et

$$X = \begin{bmatrix} 1 & 0 & \ldots & 0 & x_{11} & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & \ldots & 0 & x_{1n_1} & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & x_{21} & \ldots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & x_{In_I} \end{bmatrix},$$

le **modèle d'analyse de la covariance** se met alors sous la forme matricielle suivante :

$$Y = X\theta + E, \quad E \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

Le modèle d'Ancova peut aussi s'écrire sous forme singulière :

$$Y_{ik} = \mu + \alpha_i + \beta x_{ik} + \gamma_i x_{ik} + E_{ik}, \quad E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2).$$

1.2.4 Présentation unifiée du modèle linéaire

Ainsi quel que soit le modèle linéaire considéré et la nature des variables explicatives qui y sont incluses, l'écriture matricielle du modèle linéaire est :

$$Y = X\ \theta + E, \tag{1.3}$$

où

- Y, de dimension $(n, 1)$, contient les variables aléatoires représentant la variable à expliquer pour les n expériences. C'est un vecteur aléatoire.
- E, de dimension $(n, 1)$, contient les variables aléatoires résiduelles du modèle, rangées dans le même ordre que Y. Les E_i sont indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$ ou autrement dit le vecteur E, de dimension n suit une loi normale n-dimensionnelle centrée de matrice de variance $\sigma^2 I_n$.
- θ, de dimension $(p + 1, 1)$, contient $p + 1$ paramètres fixes et inconnus.
- X, de dimension $(n, p + 1)$, est une matrice (fixe et connue) contenant les valeurs des variables explicatives. La ligne i contient les variables explicatives concernant l'individu i, la colonne j contient la variable explicative j pour les individus 1 à n. Dans le cas de facteurs qualitatifs, ces valeurs sont des 1 ou des 0. X s'appelle "matrice du plan d'expérience" ou "matrice de design".
- x_i est le vecteur ligne correspondant à la ième ligne de X.

Attention : l'aspect linéaire du modèle linéaire n'est pas aussi réducteur qu'on peut le penser, c'est la linéarité en chacun des paramètres qui est essentielle. Ainsi, le modèle $Y = \theta_0 + \theta_1 x + \theta_2 x^2 + E$ est encore un modèle linéaire. "Modèle linéaire" signifie que $\mathbb{E}[Y]$ est une combinaison linéaire des paramètres du modèle et les coefficients de ces combinaisons sont quelconques.

1.3 Estimation des paramètres

Une fois le modèle posé, la question qui se pose ensuite est l'estimation des paramètres inconnus du modèle. Les paramètres sont de deux sortes, ceux qui relèvent de l'espérance et sont contenus dans le vecteur θ, et le paramètre σ^2 qui mesure la variabilité qui subsiste lorsque l'on a enlevé à la variabilité totale des observations tout ce qui est expliqué par le modèle. Il met donc en jeu $p + 1$ paramètres pour l'espérance (p pour l'effet des p variables ou des p niveaux et 1 pour la constante) et 1 paramètre pour la variance (σ^2).
La méthode d'estimation des paramètres classiquement utilisée est la méthode du maximum de vraisemblance. Les variables aléatoires Y_i ayant été supposées indépendantes

et de loi gaussienne, la vraisemblance de l'échantillon $y = (y_1, \ldots, y_n)$ s'écrit :

$$\mathcal{L}(y;\theta,\sigma^2)\,^2 = \prod_{i=1}^{n} f(y_i;\theta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y_i - x_i\theta)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{\sum_{i=1}^n (y_i - x_i\theta)^2}{2\sigma^2}},$$

et la log-vraisemblance

$$\ell(y;\theta,\sigma^2) \quad = \quad -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\theta)^2\,.$$

Les valeurs de θ et σ^2 qui maximisent cette log-vraisemblance sont solutions du système d'équations aux dérivées partielles suivant :

$$\begin{cases} \dfrac{\partial \ell(y;\theta_j,\sigma^2)}{\partial \theta_j} = 0 \quad \text{pour } j = 0,\ldots,p \\ \dfrac{\partial \ell(y;\theta,\sigma^2)}{\partial \sigma^2} = 0 \end{cases} \tag{1.4}$$

Remarquons que la maximisation de la log-vraisemblance en θ peut se faire indépendamment de σ^2. En effet, maximiser la log-vraisemblance en θ revient à minimiser $\sum_{i=1}^n (y_i - x_i\theta)^2$ qui peut s'écrire sous la forme suivante :

$$\|y - X\theta\|^2,$$

où $\|.\|^2$ correspond à la norme euclidienne de $\mathbb{R}^n$: si u est un vecteur de $\mathbb{R}^n$, $u = (u_1, \ldots, u_n)$, la norme de u vaut $\|u\|^2 = \sum_{i=1}^n u_i^2$. D'un point de vue géométrique, cette norme s'interprète comme la distance séparant l'origine d'un repère O, et le point U le point de coordonnées $(u_1, \ldots, u_n)$.

Ainsi nous nous intéresserons dans un premier temps à l'estimation de θ (les paramètres de l'espérance) puis à celle de la variance σ^2.

Dans la suite, pour simplifier l'écriture, l'estimateur et l'estimation des paramètres θ seront notés de la même façon $\widehat{\theta}$.

1.3.1 Estimation des paramètres de l'espérance θ

Cette méthode d'estimation est connue sous le nom de méthode des moindres carrés ordinaires (MCO) et l'estimateur résultant porte alors le nom d'estimateur des moindres carrés. Remarquons que dès que la distribution est supposée gaussienne, la méthode du maximum de vraisemblance est équivalente à la méthode des moindres carrés pour l'estimation du paramètre de la moyenne. L'estimateur $\widehat{\theta}$ est tel qu'il rend

$$\|Y - X\theta\|^2 \quad \text{minimale.}$$

2. La vraisemblance dépend également de $x = (x_1, \ldots, x_n)$ mais dans toute la suite on travaille conditionnellement aux variables explicatives et cette dépendance sera donc omise dans toutes les notations.

Théorème 1.3.1. *Soit $< X >$ le sous-espace linéaire de $\mathbb{R}^n$ engendré par les vecteurs colonnes de la matrice X. L'estimateur des moindres carrés du paramètre θ, noté $\widehat{\theta}$, est tel que :*

$$\begin{aligned} \widehat{Y} &= P\,Y \quad \textit{est le projeté orthogonal de } Y \textit{ sur } < X > \\ &= X\widehat{\theta}, \end{aligned}$$

où P est le projecteur orthogonal sur $< X >$. L'estimateur $\widehat{\theta}$ vérifie le système

$$X'X\widehat{\theta} = X'Y, \tag{1.5}$$

où X' est la transposée de la matrice X. Ce système s'appelle traditionnellement ***système des équations normales.***

Ce système correspond exactement au système des dérivées partielles (1.4).

Preuve théorème 1.3.1. $< X >$ est le sous-espace linéaire de $\mathbb{R}^n$ engendré par les vecteurs colonnes de la matrice X, c'est-à-dire que tout élément de $< X >$ s'écrit comme une combinaison linéaire de ces vecteurs. D'après la proposition A.4.3 de l'Annexe A, p. 297, la projection orthogonale sur $< X >$ minimise l'écart entre n'importe quel élément de $< X >$ et Y c'est-à-dire $\|Y - X\widehat{\theta}\|^2 = min_{U \in < X >} \|Y - U\|^2$. Cette projection est illustrée dans la figure 1.1. Pour obtenir une forme explicite de $\widehat{\theta}$, il suffit d'écrire les relations d'orthogonalité

$$\forall k = 1, .., p+1 \qquad \left\langle X^k, Y - X\widehat{\theta} \right\rangle = X^{k\prime}(Y - X\widehat{\theta}) = 0,$$

où X^k désigne la colonne k de X. Cette égalité s'écrit plus synthétiquement

$$X'(Y - X\widehat{\theta}) = 0,$$

□

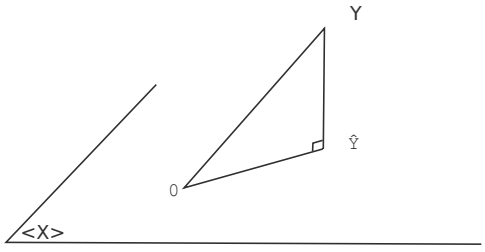

FIGURE 1.1 – Représentation de la projection de Y sur $< X >$.

On ne peut résoudre le système des équations normales (1.5, p. 15) que si la matrice carrée $X'X$ de dimension $(p+1) \times (p+1)$ admet une inverse $(X'X)^{-1}$. Or cette dernière n'existe que si la matrice $X'X$ est de plein rang, c'est-à-dire que son rang vaut exactement $p+1$. Noter que le rang de la matrice $X'X$ est le même que celui de X et

rappelons qu'il est égal au nombre de colonnes de la matrice X qui sont linéairement indépendantes (qui ne sont pas combinaisons linéaires des autres colonnes). Notons

$$r = \text{rang de } X$$

Cela revient à dire que le système (1.5, p. 15) est un système linéaire de r équations indépendantes à $p+1$ inconnues. Si r est strictement inférieur à $p+1$, le système ne peut pas avoir une solution unique.

Dans les deux sous-sections suivantes, nous distinguons les cas où X est de plein rang ou non et donnons la forme de la matrice du projecteur P associée.

Cas où $r = p+1$

C'est le cas des modèles de régression linéaire simple, de régression polynomiale et de régression linéaire multiple à condition que les variables explicatives ne soient pas liées linéairement entre elles et qu'elles ne soient pas non plus liées au vecteur constant. Dans ce cas, la matrice $X'X$ est inversible et la solution du système des équations normales est unique. Son expression est :

$$\widehat{\theta} = (X'X)^{-1}X'Y. \tag{1.6}$$

On a donc que

$$\widehat{Y} = X\widehat{\theta} = X(X'X)^{-1}X'Y.$$

Grâce à cette dernière égalité, on peut retrouver la matrice du projecteur orthogonal P :

$$P = X(X'X)^{-1}X'.$$

On peut vérifier que P possède bien toutes les propriétés d'un projecteur orthogonal, en particulier que $P^2 = P$ et qu'il est symétrique ($P' = P$).

Remarquons que dans un modèle de régression multiple, les fortes corrélations entre variables peuvent rendre la matrice $X'X$ difficile à inverser, pour des raisons d'instabilité numérique (liée au mauvais conditionnement de la matrice $X'X$). Dans ce cas une approche possible consiste à choisir certaines variables comme représentantes du groupe des variables avec lesquelles elles sont fortements corrélées et à travailler uniquement avec ces représentantes.

Cas où $r < p+1$: résolution par inverse généralisé

Ce cas correspond aux situations où les vecteurs colonnes de la matrice X sont liées par $p+1-r$ relations linéaires indépendantes. C'est le cas des modèles d'analyse de la variance et d'analyse de la covariance écrits sous leurs formes singulières. Dans ce cas, la matrice $X'X$ n'est pas inversible et donc il existe une infinité de solutions $\widehat{\theta}$ vérifiant le système. Nous sommes dans un cas d'indétermination du système. La stratégie consiste alors à choisir, parmi cette infinité de solutions, une solution particulière en ajoutant des contraintes sur les paramètres de θ. Le nombre de contraintes indépendantes à ajouter est égale au nombre d'équations indépendantes manquantes dans le système. Cependant, les contraintes choisies donneront un sens particulier aux estimations des