Généralités sur le modèle linéaire

Résumé

L'objectif du chapitre 2 est uniquement de mettre en place les principaux éléments du modèle linéaire (essentiellement gaussien), à savoir l'estimation ponctuelle, l'estimation par intervalle de confiance et les tests.

Retour au plan du cours

1 Introduction

Pour des compléments bibliographiques, nous renvoyons essentiellement à six ouvrages : trois en français et trois autres en langue anglaise. Azaïs & Bardet (2005) est un ouvrage consacré spécifiquement au modèle linéaire et constitue un excellent complément de ce cours ; Monfort (1997) propose une approche très mathématique, de la statistique en général et du modèle linéaire en particulier ; Saporta (2006) est d'un abord plus simple, le modèle linéaire ne constituant qu'une petite partie de cet ouvrage très complet et très intéressant ; Jorgensen (1993) couvre bien les chapitres 2 et 3 de ce cours ; Milliken & Johnson (1984) en couvre la presque totalité ; enfin, Rencher & Schaalje (2008) est notre ouvrage de référence sur le modèle linéaire. Cela étant, signalons que le nombre d'ouvrages consacrés, au moins partiellement, au modèle linéaire est considérable.

Précisons l'écriture du modèle linéaire pour tout individu i ($i = 1, \ldots, n$) d'un échantillon de taille n :

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^j + U_i \, .$$

Y_i est la variable aléatoire réelle réponse et U_i est la variable aléatoire réelle erreur, supposée $\mathcal{N}(0, \sigma^2)$, les U_i étant indépendantes (et donc i.i.d.). Les β_j sont des coefficients, des paramètres inconnus, à estimer. Les X_i^j sont les valeurs des variables explicatives qui ne sont en général pas considérées comme aléatoires : on suppose qu'il s'agit de valeurs choisies, contrôlées.

Matriciellement, on peut réécrire

$$Y = \mathbf{X}\beta + U \, ,$$

où Y et U sont des vecteurs aléatoires de $\mathbb{R}^n$, $\mathbf{X}$ est une matrice $n \times p$ et β est le vecteur de $\mathbb{R}^p$ des paramètres.

Si l'estimation ponctuelle est possible sans aucune hypothèse de distribution sur les erreurs du modèle, grâce à la méthode des moindres carrés, il n'en va pas de même pour l'estimation par intervalle de confiance et pour les tests : dans ce cas, l'hypothèse de normalité des erreurs (l'hypothèse gaussienne) est indispensable. De manière souvent implicite, l'hypothèse gaussienne sera faite dans tout ce cours car elle est quasiment partout indispensable.

L'estimation ponctuelle du vecteur des paramètres β, que ce soit par moindres carrés ou par maximum de vraisemblance dans le cas gaussien, conduit au résultat suivant :

$$\hat{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y.$$

On appelle valeurs prédites les $\hat{Y}_i$, coordonnées du vecteur aléatoire

$$\hat{Y} = \mathbf{X}\hat{B} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y = \mathbf{H}Y,$$

où $\mathbf{H}$ est la matrice de projection orthogonale sur le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de $\mathbf{X}$.

On appelle résidus les $\hat{U}_i$, coordonnées du vecteur aléatoire

$$\hat{U} = Y - \hat{Y} = \mathbf{H}^{\perp}Y,$$

où $\mathbf{H}^{\perp} = \mathbf{I}_n - \mathbf{H}$ est la matrice de projection orthogonale sur le sous-espace vectoriel de $\mathbb{R}^n$ supplémentaire orthogonal au précédent.

L'estimateur de la variance du modèle (σ^2), après correction de biais, est donnée par :

$$\hat{\Sigma}^2 = \frac{\sum_{i=1}^{n} \hat{U}_i^2}{n-p} = \frac{\|\hat{U}\|^2}{n-p} \, .$$

L'estimation par intervalle de confiance d'une fonction linéaire des paramètres, $c'\beta = \sum_{j=1}^{p} c_j\beta_j$, conduit à l'intervalle

$$c'\hat{\beta} \ \pm \ t \, [\hat{\sigma}^2 c'(\mathbf{X}'\mathbf{X})^{-1}c]^{1/2} \, ,$$

où $t = t_{n-p}(1-\frac{\alpha}{2})$ est le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi de Student à $n-p$ degrès de liberté. Le coefficient de sécurité de cet intervalle est $1-\alpha$, autrement dit son risque est α.

Le test d'une hypothèse nulle $\{H_0 : \mathbf{C}'\beta = 0\}$, linéaire en β, contre l'alternative opposée, se fait au moyen de la statistique de Fisher (ou Fisher-Snedecor) qui s'écrit :

$$F = \frac{NUM}{q\hat{\Sigma}^2} ,$$

où q est le nombre de contraintes définies par H_0 (autrement dit, le rang de $\mathbf{C}$, matrice de dimension $p \times q$, avec $1 \leq q < p$) et où le numérateur NUM peut s'écrire sous l'une des formes suivantes

$$\begin{aligned} NUM &= \|\hat{U}_0\|^2 - \|\hat{U}\|^2 = \|\hat{U}_0 - \hat{U}\|^2 = \|\hat{Y}_0 - \hat{Y}\|^2 = \|\hat{B}_0 - \hat{B}\|^2_{X'X} \\ &= \hat{B}'\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'\hat{B} , \end{aligned}$$

$\hat{B}_0$, $\hat{Y}_0$ et $\hat{U}_0$ désignant respectivement le vecteur des estimateurs, celui des valeurs prédites et celui des résidus dans le modèle sous H_0.

2 Définitions et notations

2.1 Le modèle linéaire

DÉFINITION 1 *On appelle modèle linéaire un modèle statistique qui peut s'écrire sous la forme*

$$Y = \sum_{j=1}^{p} \beta_j X^j + U.$$

Dans la définition ci-dessus, les éléments intervenant ont les caractéristiques suivantes :

- Y est une variable aléatoire réelle (v.a.r.) que l'on observe et que l'on souhaite expliquer, ou prédire (ou les deux à la fois) ; on l'appelle variable à expliquer, ou **variable réponse** (parfois aussi variable dépendante, ou variable endogène).
- Chaque variable X^j est une variable réelle (éventuellement ne prenant que les valeurs 0 et 1), non aléatoire dans le modèle de base, également observée ; l'ensemble des X^j est censé expliquer Y, en être la cause (au moins partiellement) ; les variables X^j sont appelées variables explicatives, ou **prédicteurs** (parfois variables indépendantes, ou variables exogènes).
  Pour chaque variable X^j, l'expérimentateur est supposé choisir diverses valeurs caractéris- tiques (au moins deux) pour lesquelles il réalise une ou plusieurs expériences en notant les valeurs correspondantes de Y : il contrôle donc les variables X^j, pour cette raison appelées aussi **variables contrôlées** ; en réalité, dans la pratique, ce n'est pas toujours exactement le cas.
- Les β_j $(j = 1, \ldots, p)$ sont des coefficients, des **paramètres**, non observés ; on devra donc les estimer au moyen de techniques statistiques appropriées.
- U est le terme d'erreur du modèle ; c'est une v.a.r. non observée pour laquelle on fait systématiquement les hypothèses suivantes :

$$\mathbb{E}(U) = 0 \; ; \; \mathrm{Var}(U) = \sigma^2 > 0$$

  (σ^2 est un paramètre inconnu, également à estimer). Lorsqu'on répète les observations de Y et des X^j, on suppose que la variance de U est constante (σ^2) ; c'est ce que l'on appelle l'hypothèse d'**homoscédasticité**.
- Les hypothèses faites sur U entraînent les conséquences suivantes sur Y :

$$\mathbb{E}(Y) = \sum_{j=1}^{p} \beta_j X^j \; ; \; \mathrm{Var}(Y) = \sigma^2.$$

- L'espérance mathématique de Y s'écrit donc comme une combinaison linéaire des X^j : la liaison entre les X^j et Y est de nature linéaire (linéaire en moyenne). C'est la raison pour laquelle ce modèle est appelé le *modèle linéaire*.

2.2 Le modèle linéaire gaussien

C'est un modèle linéaire dans lequel on fait l'hypothèse supplémentaire que la v.a.r. U est gaussienne, c'est-à-dire normale. On pose donc :

$$U \sim \mathcal{N}(0, \sigma^2),$$

cette hypothèse entraînant la normalité de Y.

Si l'on veut, dans un modèle linéaire, pouvoir construire des intervalles de confiance ou faire des tests concernant les paramètres (les β_j et σ^2), cette hypothèse gaussienne est indispensable. Sauf indication contraire, elle sera faite dans toute la suite de ce cours.

2.3 Notations

Pour pouvoir faire, au minimum, l'estimation ponctuelle des paramètres β_j et σ^2, il est indispensable de répliquer, de manières indépendantes, les observations simultanées des variables X^j et Y.

Nous supposerons donc par la suite que n observations indépendantes sont réalisées et nous écrirons le modèle, pour la i-ième observation ($i = 1, \ldots, n$), sous la forme :

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^j + U_i \quad \text{(égalité entre v.a.r.)}.$$

Les valeurs observées des variables seront notées par des minuscules, de sorte qu'on écrira :

$$y_i = \sum_{j=1}^{p} \beta_j x_i^j + u_i \quad \text{(égalité entre nombres réels)}.$$

Par ailleurs, on notera $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ le vecteur aléatoire de $\mathbb{R}^n$ correspondant à l'ensemble de l'échantillon des v.a.r. réponses (la notation Y est identique à celle introduite en 2.1.1 pour une seule v.a.r. réponse, mais cela ne devrait pas entraîner de confusion puisqu'on travaillera dorénavant avec un échantillon), $\mathbf{X} = (x_i^j)$ la matrice réelle, $n \times p$, des valeurs contrôlées des prédicteurs, $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ le vecteur des paramètres dans $\mathbb{R}^p$ et $U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$ le vecteur aléatoire de $\mathbb{R}^n$ contenant les erreurs du modèle (même remarque que ci-dessus).

Matriciellement, le modèle linéaire s'écrit donc

$$Y = \mathbf{X}\beta + U,$$

avec, dans le cas gaussien,

$$U \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n) \text{ et } Y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n),$$

$\mathbf{I}_n$ désignant la matrice identité d'ordre n.

Par la suite, on supposera $n > p$ (le nombre d'observations est au moins égal au nombre de paramètres à estimer), $p \geq 1$ (il y a au moins une variable explicative dans le modèle) et $\mathbf{X}$ de rang p (les variables X^j sont linéairement indépendantes).

Remarque. — On notera que les v.a.r. U_i sont i.i.d. (indépendantes et identiquement distribuées) par hypothèse, alors que les v.a.r. Y_i sont indépendantes, de même variance, normales dans le cas gaussien, mais n'ont pas toutes la même moyenne (elles ne sont donc pas i.i.d.).

Remarque. — Dans le modèle linéaire, et plus particulièrement dans l'analyse de variance, la matrice $\mathbf{X}$ est souvent appelée **matrice d'incidence**.

2.4 Trois exemples basiques

2.4.1 Le modèle constant, ou modèle "blanc"

Il s'écrit :

$$Y_i = \beta + U_i \quad (Y = \beta \mathbb{1}_n + U).$$

Autrement dit, $p = 1$ et $\mathbf{X} = \mathbb{1}_n$: l'unique prédicteur est la variable constante et égale à 1. Ce modèle n'a pas d'intérêt pratique, mais il est utilisé comme modèle de référence, celui par rapport auquel on comparera d'autres modèles.

2.4.2 Le modèle de régression linéaire simple

C'est le modèle suivant :

$$Y_i = \beta_1 + \beta_2 X_i^2 + U_i.$$

Ici, $p = 2$ et $\mathbf{X} = (\mathbb{1}_n \; X^2)$: on a rajouté un "vrai" prédicteur quantitatif (X^2) à la constante.

2.4.3 Le modèle d'analyse de variance à un facteur à deux niveaux

Ce modèle s'écrit :

$$Y_i = \beta_j + U_i,$$

lorsque la i-ième observation de Y est réalisée au niveau j ($j = 1, 2$) du facteur (la variable explicative est ici qualitative à deux modalités ; dans le contexte du modèle linéaire, on parle plutôt de *facteur* à deux *niveaux*). En fait, chaque niveau du facteur est remplacé par une variable indicatrice, de sorte que $p = 2$.

Matriciellement, ce modèle peut s'écrire

$$Y = \mathbf{X}\beta + U,$$

avec

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \text{ et } \mathbf{X} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}.$$

Dans la matrice $\mathbf{X}$ ci-dessus, les n_1 premières lignes sont $(1\ 0)$ s'il y a n_1 observations réalisées au niveau 1 du facteur, les n_2 suivantes étant $(0\ 1)$ s'il y a n_2 observations réalisées au niveau 2 du facteur ($n_1 + n_2 = n$).

3 Estimation des paramètres

3.1 Estimation de β dans le cas général

En l'absence d'hypothèse sur la distribution de U, on estime β par la méthode des moindres carrés. Elle consiste à poser :

$$\hat{\beta} = \text{ Arg min } \|y - \mathbf{X}\beta\|^2 \ , \ \beta \in \mathbb{R}^p. \tag{1}$$

(Cette écriture suppose que $\mathbb{R}^n$ est muni de la norme euclidienne classique, autrement dit que l'on utilise le critère dit des *moindres carrés ordinaires*.)

On montre alors que ce problème admet la solution unique

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \text{ (estimation)},$$

valeur observée du vecteur aléatoire

$$\hat{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \text{ (estimateur)}.$$

3.1.1 Propriétés de $\hat{B}$

- $\mathbb{E}(\hat{B}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(Y) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta$: $\hat{B}$ est un estimateur sans biais de β.
- $\text{Var}(\hat{B}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \dfrac{\sigma^2}{n}\mathbf{S}_n^{-1}$, avec $\mathbf{S}_n = \dfrac{1}{n}\mathbf{X}'\mathbf{X}$ (matrice des variances-covariances empiriques lorsque les variables X^j sont centrées). On obtient un estimateur convergent, sous réserve que :
$$\lim_{n\to\infty} \det(\mathbf{S}_n) = d > 0.$$

3.2 Moindres carrés ordinaires et moindres carrés généralisés

Dans le point 2.1.3, on a posé $\text{Var}(U) = \sigma^2\mathbf{I}_n$. Supposons maintenant, de façon plus générale, que $\text{Var}(U) = \sigma^2\mathbf{V}$, où $\mathbf{V}$ est une matrice *connue*, carrée d'ordre n, symétrique et strictement définie-positive. On peut alors se ramener au cas précédent en faisant intervenir la matrice $\mathbf{V}^{-1}$ dans le critère des moindres carrés. Pour cela, on cherche le vecteur $\hat{\beta}$ de $\mathbb{R}^p$ solution de :

$$\hat{\beta} = \text{ Arg min } \|y - \mathbf{X}\beta\|^2_{\mathbf{V}^{-1}}. \tag{2}$$

La solution est donnée par :

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}Y).$$

Le critère (1) est appelé critère des moindres carrés ordinaires (MCO), alors que le critère (2) est appelé critère des moindres carrés généralisés (MCG) (voir, par exemple, Monfort, 1997, chapitre 26). Le critère des moindres carrés généralisés sera utilisé au chapitre 6.

3.3 Estimation de β dans le cas gaussien

3.3.1 Densité d'une loi multinormale

Soit Z un vecteur aléatoire à valeurs dans $\mathbb{R}^n$, de densité gaussienne, admettant μ comme vecteur des moyennes ($\mu \in \mathbb{R}^n$) et $\mathbf{\Sigma}$ comme matrice des variances-covariances ($\mathbf{\Sigma}$ est carrée d'ordre n, symétrique, strictement définie-positive). On rappelle la densité de Z :

$$f(z) = \frac{1}{(2\pi)^{n/2}} \frac{1}{(\det \mathbf{\Sigma})^{1/2}} \exp[-\frac{1}{2}(z-\mu)'\mathbf{\Sigma}^{-1}(z-\mu)].$$

3.3.2 Vraisemblance d'un échantillon gaussien de taille n

Dans le cadre du modèle linéaire gaussien, le vecteur aléatoire Y admet pour espérance le vecteur $\mathbf{X}\beta$ et pour matrice des variances-covariances $\mathbf{\Sigma} = \sigma^2\mathbf{I}_n$. Sa vraisemblance s'écrit donc :

$$L(y,\beta,\sigma^2) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp[-\frac{1}{2\sigma^2}(y-\mathbf{X}\beta)'(y-\mathbf{X}\beta)].$$

3.3.3 Log-vraisemblance

Le logarithme (népérien) de la fonction ci-dessus s'écrit :

$$\begin{aligned} l(y,\beta,\sigma^2) &= \log[L(y,\beta,\sigma^2)] \\ &= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}(y-\mathbf{X}\beta)'(y-\mathbf{X}\beta) \\ &= \text{constante} - n\log(\sigma) - \frac{1}{2\sigma^2}\|y-\mathbf{X}\beta\|^2. \end{aligned}$$

3.3.4 Conséquences

Maximiser $l(y,\beta,\sigma^2)$ selon β, pour trouver l'estimateur maximum de vraisemblance, revient donc à minimiser $\|y-\mathbf{X}\beta\|^2$ selon β, et redonne l'estimateur $\hat{B}$ introduit en 2.2.1. Ainsi, estimateurs moindres carrés ordinaires et maximum de vraisemblance sont identiques dans le modèle linéaire gaussien.

3.3.5 Propriétés

L'estimateur $\hat{B}$ de β demeure d'une part sans biais, d'autre part convergent, sous la même condition que précédemment. De plus, on peut, dans le cadre gaussien, préciser sa distribution : comme transformée linéaire d'un vecteur gaussien, elle est gaussienne, donc $\mathcal{N}_p(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. Enfin, on peut vérifier que $\hat{B}$ est un estimateur efficace de β (sa variance est égale à la borne inférieure de l'inégalité de Cramér-Rao).

Remarque. — Si les prédicteurs X^j sont deux à deux orthogonaux, alors $\mathbf{X}'\mathbf{X} = \text{diag}\,(\alpha_1 \cdots \alpha_p)$, avec $\alpha_j = \sum_{i=1}^n (x_i^j)^2 > 0$ (sinon, la j-ième colonne de $\mathbf{X}$ serait nulle et $\mathbf{X}$ ne serait pas de rang p). Il vient donc $(\mathbf{X}'\mathbf{X})^{-1} = \text{diag}\,(\frac{1}{\alpha_1} \cdots \frac{1}{\alpha_p})$ et l'on en déduit $\hat{B}_j \sim \mathcal{N}(\beta_j, \frac{\sigma^2}{\alpha_j})$, les $\hat{B}_j$ étant donc mutuellement indépendants. Cette situation se rencontre, dans certains cas particuliers, en analyse de variance (voir chapitre 3).

3.4 Estimation d'une fonction linéaire de β

On considère maintenant un vecteur non nul c de $\mathbb{R}^p$ et la forme linéaire $c'\beta$. On vérifie simplement, dans le modèle gaussien, que l'estimateur maximum de vraisemblance de $c'\beta$ est $c'\hat{B}$ et que $c'\hat{B} \sim \mathcal{N}(c'\beta, \sigma^2 c'(\mathbf{X}'\mathbf{X})^{-1}c)$. Il s'agit d'un estimateur sans biais, convergent (toujours sous la même condition) et efficace.

On utilise ce résultat pour estimer l'un des paramètres β_j, une différence entre deux paramètres $\beta_j - \beta_k$, etc.

3.5 Valeurs prédites et résidus

3.5.1 Valeurs prédites

On appelle vecteur des valeurs prédites le vecteur $\hat{y}$ de $\mathbb{R}^n$ défini par :

$$\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y.$$

Il s'agit du vecteur des prédictions (ou approximations) $\hat{y}_i$ des y_i réalisées avec le modèle linéaire considéré ; on parle aussi de *valeurs ajustées*.

En fait, en posant $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, on remarque que $\mathbf{H}$ est la matrice de la projection orthogonale (au sens de la métrique usuelle) sur le sous-espace

vectoriel F_X de $\mathbb{R}^n$ engendré par les colonnes de $\mathbf{X}$. Par suite, $\hat{y} = \mathbf{H}y$ est la projection orthogonale de y sur F_X.

Dans le modèle gaussien, on obtient

$$\hat{Y} = \mathbf{H}Y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2\mathbf{H});$$

en effet, $\hat{Y}$ est gaussien comme transformé linéaire de Y gaussien, $\mathbf{H}\mathbf{X}\beta = \mathbf{X}\beta$ (le vecteur $\mathbf{X}\beta$ étant, par définition, dans le sous-espace F_X) et $\sigma^2\mathbf{H}\mathbf{H}' = \sigma^2\mathbf{H}^2$ ($\mathbf{H}$ est symétrique) $= \sigma^2\mathbf{H}$ ($\mathbf{H}$ est idempotente).

3.5.2 Erreur-type (standard error) d'une valeur prédite

De façon usuelle, on note h_i le i-ième terme de la diagonale de $\mathbf{H}$ ($i = 1, \ldots, n$). On obtient ainsi $\hat{Y}_i \sim \mathcal{N}((\mathbf{X}\beta)_i, \sigma^2 h_i)$. L'écart-type (*standard deviation*) de $\hat{Y}_i$ est donc $\sigma\sqrt{h_i}$ et on l'estime par $\hat{\sigma}\sqrt{h_i}$ (voir le point suivant pour l'expression de $\hat{\sigma}^2$, donc de $\hat{\sigma}$). La quantité $\hat{\sigma}\sqrt{h_i}$ est appelée erreur-type de $\hat{Y}_i$ et sera utilisée par la suite.

3.5.3 Résidus

On appelle résidu le vecteur $\hat{u}$ de $\mathbb{R}^n$ défini par $\hat{u} = y - \hat{y}$. C'est l'écart entre l'observation du vecteur aléatoire Y et sa prédiction (son approximation) par le modèle considéré. Autrement dit, c'est une approximation du vecteur des erreurs U.

On obtient ainsi

$$\hat{U} = Y - \hat{Y} = (\mathbf{I}_n - \mathbf{H})Y = \mathbf{H}^{\perp}Y,$$

où $\mathbf{H}^{\perp}$ est le projecteur orthogonal sur le sous-espace vectoriel $F_X^{\perp}$ de $\mathbb{R}^n$ supplémentaire orthogonal à F_X.

Dans le modèle gaussien, on obtient :

$$\hat{U} = \mathbf{H}^{\perp}Y \sim \mathcal{N}_n(0, \sigma^2\mathbf{H}^{\perp}).$$

3.5.4 Indépendance de $\hat{U}$ avec $\hat{Y}$ et avec $\hat{B}$

On a :

$$\mathrm{Cov}(\hat{U}, \hat{Y}) = \mathrm{Cov}(\mathbf{H}^{\perp}Y, \mathbf{H}Y) = \sigma^2\mathbf{H}^{\perp}\mathbf{H} = 0.$$

Par suite, $\hat{Y}$ et $\hat{U}$ sont non correlés, donc indépendants dans le cas gaussien. Il en est de même pour $\hat{U}$ et $\hat{B}$.

3.5.5 Résidus studentisés

Dans le cas gaussien, pour tout i ($i = 1, \ldots, n$), on a $\hat{U}_i \sim \mathcal{N}(0, \sigma^2(1-h_i))$. L'écart-type de $\hat{U}_i$ est donc $\sigma\sqrt{1-h_i}$ et son estimation, appelée erreur-type de $\hat{U}_i$, est $\hat{\sigma}\sqrt{1-h_i}$.

On appelle alors i-ième résidu studentisé la quantité $\hat{s}_i = \dfrac{\hat{u}_i}{\hat{\sigma}\sqrt{1-h_i}}$. Il s'agit de l'approximaton de l'observation d'une loi $\mathcal{N}(0,1)$, utilisée dans la validaton du modèle.

Remarque. — On notera que si la construction de $\hat{s}_i$ rappelle celle d'une observation de loi de Student, ce n'est pas ici le cas puisqu'il n'y a pas indépendance entre $\hat{U}_i$ et $\hat{\Sigma}^2 = \dfrac{\sum_{i=1}^n \hat{U}_i^2}{n-p}$ (voir l'expression de $\hat{\Sigma}^2$ ci-dessous). Pour cette raison, on trouve dans la littérature statistique d'autres expressions pour les résidus studentisés ; nous ne les introduisons pas ici car elles nous semblent peu utiles.

3.6 Estimation de σ^2 dans le cas général

Sans hypothèse gaussienne, on ne peut envisager d'utiliser le maximum de vraisemblance. Par ailleurs, les moindres carrés ne permettent pas d'estimer σ^2, dans le mesure où ce paramètre n'est pas lié à l'espérance de Y. On doit donc avoir recours à une estimation empirique (souvent appelée *plug-in*) : le paramètre σ^2 représentant la variance de la variable erreur U, on l'estime par la variance empirique des résidus $\hat{U}_i$, soit $\Sigma^{*2} = \dfrac{1}{n}\sum_{i=1}^n \hat{U}_i^2$ (la moyenne empirique des $\hat{U}_i$ est nulle).

On peut alors vérifier que cet estimateur est biaisé et le corriger en posant $\hat{\Sigma}^2 = \dfrac{1}{n-p}\sum_{i=1}^n \hat{U}_i^2$, estimateur sans biais de σ^2. On ne peut toutefois rien dire ni sur sa variance ni sur sa convergence.

3.7 Estimation de σ^2 dans le cas gaussien

Dans ce cas, on applique la méthode du maximum de vraisemblance qui consiste à maximiser, selon σ^2, l'expression de $l(y, \beta, \sigma^2)$ donnée en 2.2.3.

On peut vérifier que cela conduit à la même expression Σ^{*2} que celle fournie par la méthode empirique. On utilise donc encore l'estimateur corrigé $\hat{\Sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{U}_i^2$, de façon à disposer d'un estimateur sans biais.

De plus, l'hypothèse gaussienne permet maintenant de montrer :

$$\frac{(n-p)\hat{\Sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\hat{U}_i^2}{\sigma^2} = \frac{\|\hat{U}\|^2}{\sigma^2} \sim \chi^2_{n-p}.$$

On déduit de ce résultat :
- $\mathbb{E}(\hat{\Sigma}^2) = \sigma^2$ (résultat déjà connu) ;
- $\mathrm{Var}(\hat{\Sigma}^2) = \frac{2\sigma^4}{n-p}$: $\hat{\Sigma}^2$ est donc un estimateur convergent ;
- par ailleurs, on peut vérifier que $\hat{\Sigma}^2$ n'est pas efficace, mais est asymptotiquement efficace ; de plus, il s'agit d'un estimateur optimal pour σ^2, c'est-à-dire de variance minimum parmi les estimateurs sans biais (propriété générale de la famille exponentielle) ;
- enfin, dans le cas gaussien, on peut vérifier que les estimateurs $\hat{B}$ et $\hat{\Sigma}^2$ sont indépendants.

3.8 Intervalle de confiance pour une fonction linéaire de β

On ne peut envisager un tel intervalle que dans le cadre du modèle gaussien. Soit donc c un vecteur non nul de $\mathbb{R}^p$ et $c'\beta$ la forme linéaire associée. On a vu en 2.2.4 :

$$c'\hat{B} \sim \mathcal{N}(c'\beta, \sigma^2 c'(\mathbf{X}'\mathbf{X})^{-1}c).$$

La variance ci-dessus faisant intervenir le paramètre inconnu σ^2, on utilise $\hat{\Sigma}^2$ et l'indépendance de $c'\hat{B}$ et de $\hat{\Sigma}^2$ pour obtenir une loi de Student, dont on déduit l'intervalle de confiance suivant, de coefficient de sécurité $1-\alpha$:

$$c'\hat{\beta} \pm \hat{\sigma}[c'(\mathbf{X}'\mathbf{X})^{-1}c]^{1/2}\, t_{n-p}(1-\frac{\alpha}{2}).$$

Dans l'expression ci-dessus, on notera que :
- $c'\hat{\beta}$ est l'estimation ponctuelle de $c'\beta$;
- $\hat{\sigma}[c'(\mathbf{X}'\mathbf{X})^{-1}c]^{1/2}$ est l'erreur-type de $c'\hat{\beta}$;
- $t_{n-p}(1-\frac{\alpha}{2})$ est le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi de Student à $n-p$ degrés de liberté (d.d.l.).

Remarque. — On peut tester l'hypothèse nulle $\{H_0 : c'\beta = 0\}$ à partir de l'intervalle de confiance défini ci-dessus. Il suffit de regarder si l'intervalle contient, ou non, la valeur 0. En fait, cette démarche est équivalente au test de Student de cette hypothèse nulle (voir la remarque 8).

3.9 Intervalles de confiance conjoints : méthode de Bonferroni

En considérant $c' = (0,\ldots,0,1,0,\ldots,0)$, où le 1 est situé en j-ième position ($j = 1,\ldots,p$), on obtient, par la méthode ci-dessus, un intervalle de confiance de risque α (c'est-à-dire de coefficient de sécurité $1-\alpha$) pour le paramètre β_j.

Pour construire simultanément des intervalles de confiance pour les p paramètres β_j, de risque inconnu mais majoré par α ($\alpha \in]0,1[$), on peut utiliser la méthode de Bonferroni. Elle consiste à construire un intervalle, pour chacun des paramètres β_j, selon la formule indiquée ci-dessus, en utilisant pour risque non pas α mais $\frac{\alpha}{p}$. Toutefois, il faut noter que, dès que p vaut 5 ou plus, cette méthode est trop conservative : elle a tendance à ne pas rejeter l'hypothèse nulle d'égalité des paramètres β_j, autrement dit à regrouper la plupart des niveaux du facteur.

Nous donnons quelques développements de cette méthode dans l'Annexe A.

4 Test d'une hypothèse linéaire en β

Dans le modèle linéaire, on est souvent amené à tester une hypothèse nulle, linéaire en β, du type $\{H_0 : \mathbf{C}'\beta = 0\}$, où $\mathbf{C}$ est une matrice $p \times q$ de rang q, $(1 \leq q < p)$, ce qui revient à tester la réalité de q contraintes linéaires sur le paramètre β (par exemple, $\beta_1 = 0$, $\beta_2 = \beta_3$, etc.). Le but est, en fait, de simplifier le modèle. On notera que cela revient à tester $\{H_0 : \beta \in E_0\}$, où E_0 est un sous-espace vectoriel de $\mathbb{R}^p$ de dimension $p-q$, ou encore $\mathbb{E}(Y) = \mathbf{X}\beta \in F_0$, où F_0 est un sous-espace vectoriel de $\mathbb{R}^n$ de dimension $p-q$.

On a vu :

$$\frac{(n-p)\hat{\Sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n \hat{U}_i^2}{\sigma^2} = \frac{\|\hat{U}\|^2}{\sigma^2} \sim \chi^2_{n-p}.$$

De la même manière, si H_0 est vraie, on peut vérifier que

$$\frac{\|\hat{U}_0\|^2 - \|\hat{U}\|^2}{\sigma^2} \sim \chi^2_q,$$

avec $\|\hat{U}_0\|^2 = \sum_{i=1}^n \hat{U}_{i0}^2$, $\hat{U}_{i0} = Y_i - \hat{Y}_{i0}$ et $\hat{Y}_{i0} = \mathbf{X}\hat{B}_0$, $\hat{B}_0$ étant l'estimateur maximum de vraisemblance de β sous la contrainte $\mathbf{C}'\beta = 0$. De plus, sous H_0, les deux statistiques de khi-deux définies ci-dessus sont indépendantes.

On en déduit le test de H_0 : rejet de H_0 ssi (si, et seulement si)

$$F = \frac{\|\hat{U}_0\|^2 - \|\hat{U}\|^2}{\|\hat{U}\|^2} \times \frac{n-p}{q} > f_{q\ ;\ n-p}\,(1-\alpha),$$

où $f_{q\ ;\ n-p}\,(1-\alpha)$ est le quantile d'ordre $1-\alpha$ d'une loi de Fisher à q et $n-p$ d.d.l. Ce test est de niveau α.

Autres expressions de F

On peut écrire la statistique F sous la forme $\dfrac{NUM}{q\,\hat{\Sigma}^2}$, puisque $\hat{\Sigma}^2 = \dfrac{\|\hat{U}\|^2}{n-p}$; le numérateur peut alors prendre les expressions suivantes :

$$\begin{aligned} NUM &= \|\hat{U}_0\|^2 - \|\hat{U}\|^2 = \|\hat{U}_0 - \hat{U}\|^2 = \|\hat{Y}_0 - \hat{Y}\|^2 = \|\hat{B}_0 - \hat{B}\|^2_{X'X} \\ &= \hat{B}'\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'\hat{B}. \end{aligned}$$

La quantité $\|\hat{U}\|^2$ correspond à ce qui est souvent appelé, dans les logiciels, *error sum of squares* (dans le modèle complet).

Remarque. — Ce test est en général appelé test de Fisher, parfois test de Fisher-Snedecor, voire test de Snedecor.

Remarque. — Dans la pratique, les logiciels calculent la valeur observée f de la statistique F (sur les données considérées), puis la probabilité $P[F_{q\ ;\ n-p} > f]$ ($F_{q\ ;\ n-p}$ désigne une loi de Fisher à q et $n-p$ d.d.l.), en général appelée *p-value*. On rejette alors H_0 ssi la *p-value* est inférieure à α.

Remarque. — Si $q = 1$, le test de Fisher ci-dessus peut se ramener à un test de Student, lui-même équivalent à l'intervalle de confiance construit en 2.2.8.

Critère de choix de modèle : le C_p de Mallows

Lorsqu'on hésite à prendre en compte un effet faiblement significatif (dont la *p-value* est proche de α), on peut utiliser le critère C_p (voir l'Annexe D) pour décider : on calcule ce critère pour chacun des deux modèles (avec et sans cet effet) et on retient celui des deux qui minimise le C_p.

5 Contrôles d'un modèle linéaire

À l'issue de différents traitements statistiques (études exploratoires élémentaires, puis multidimensionnelles, modélisations avec tests et estimations des paramètres...), lorsqu'un modèle linéaire semble convenir à un jeu de données, un certain nombre de contrôles sont nécessaires avant de le retenir effectivement. Ces contrôles ont pour but d'apprécier la *qualité* et la *validité* du modèle envisagé. Ils peuvent, bien sûr, conduire à en changer.

5.1 Contrôles de la qualité d'un modèle

- *Significativité.* Le test de significativité du modèle est le test de l'hypothèse nulle correspondant au modèle constant (ou modèle blanc) au sein du modèle retenu (autrement dit, à la nullité de tous les paramètres β_j, à l'exception de celui correspondant au vecteur constant). Ce test doit être très significatif (c'est la condition minimale).
- *Valeur du R^2.* Le coefficient $R^2 = \dfrac{\|\hat{Y}\|^2}{\|Y\|^2}$, compris entre 0 et 1, mesure la qualité globale du modèle et doit être suffisamment proche de 1.
- *Graphique des valeurs prédites contre les valeurs observées.* En axes orthonormés, on repré-sente le nuage des points ayant pour abscisses les valeurs observées (y_i) et pour ordonnées les valeurs prédites par le modèle ($\hat{y}_i$). Plus le nuage obtenu est proche de la première bissectrice, plus le modèle est globalement bon. On peut également faire figurer la première bissectrice sur ce graphique pour préciser les choses. Ce graphique fournit, d'une autre manière, une information analogue à celle fournie par le coefficient R^2. Mais, il permet aussi de contrôler que la forme générale du nuage (donc l'ensemble des observations de Y) n'a rien de particulier. On en trouvera des exemples au chapitre 3 (Figures 3.1 et 3.3).

5.2 Contrôles de la validité d'un modèle

Ces contrôles se font à partir de ce qu'il est convenu d'appeler le **graphique des résidus**. C'est le graphique donnant le nuage des points ayant pour abscisses les valeurs prédites ($\hat{y}_i$) et pour ordonnées les résidus studentisés ($\hat{s}_i$), et dont on trouvera aussi des exemples au chapitre 3 (Figures 3.2 et 3.4).

Trois éléments sont contrôlés à travers ce graphique.

- *Le caractère linéaire des données.* Les données ayant été ajustées par un modèle linéaire, si leur structure est réellement linéaire, on ne doit retrouver aucune structure dans les résidus. Si on retrouve une forme en "U", on pourra essayer de remplacer Y par $\log(Y)$ ou par $\sqrt{Y}$ (à condition que Y soit à valeurs positives) ; pour une forme en "U renversé", on pourra essayer de remplacer Y par $\exp(Y)$ ou par Y^2 ; etc.
- *L'homoscédasticité.* La variance de la variable erreur U étant supposée constante d'une observation à l'autre, la variabilité des résidus studentisés doit être de même amplitude quelles que soient les valeurs $\hat{y}_i$, ce que l'on peut contrôler sur le graphique des résidus. Là encore, en cas de croissance des résidus en fonction des valeurs $\hat{y}_i$, on peut envisager la transformation de Y en $\log(Y)$ ou en $\sqrt{Y}$ (toujours sous la même condition).
- *La normalité.* Enfin, si les données sont réellement gaussiennes, les résidus studentisés sont approximativement distribués selon une loi normale réduite, et pas plus de 5% d'entre eux ne doivent sortir de l'intervalle $[-2, +2]$, ce qui est très facile à contrôler sur le graphique.

Il est donc conseillé de n'utiliser un modèle linéaire que s'il a passé avec succès l'ensemble des contrôles de qualité et de validité indiqués ci-dessus.

6 Panorama sur le modèle linaire

6.1 Le modèle linéaire gaussien de base

Il s'agit du modèle développé dans les paragraphes précédents.

Précisons que si tous les prédicteurs X^j sont quantitatifs, on obtient ce que l'on appelle la *régression linéaire*. Celle-ci ne sera pas développée dans ce cours et nous renvoyons pour cela aux enseignements de première année de Master ou à la bibliographie mentionnée en début de chapitre.

Lorsque tous les prédicteurs sont qualitatifs, on parle alors de facteurs et le modèle linéaire recouvre ce que l'on appelle l'*analyse de variance*, ou ANOVA (acronyme anglais de *ANalysis Of VAriance*), ou encore les *plans factoriels*. Les cas les plus simples seront traités au chapitre 3, tandis que des cas plus particuliers seront abordés au chapitre 4.

Enfin, lorsqu'il y a mélange de prédicteurs quantitatifs et qualitatifs, on parle d'*analyse de covariance*, pour laquelle nous renvoyons encore aux enseignements de première année de Master ou à la bibliographie.

6.2 Le modèle linéaire gaussien général

C'est l'objet principal de ce cours. Il s'agit de diverses généralisations du modèle linéaire gaussien de base.

- Lorsque la variable réponse Y est multidimensionnelle, on obtient le modèle linéaire multivarié. Dans le chapitre 5, on s'intéressera au cas de prédicteurs X^j qualitatifs, ce qui nous donnera l'analyse de variance multivariée, ou MANOVA.
- Avec une variable réponse Y unidimensionnelle, on peut introduire, parmi les prédicteurs X^j, des variables aléatoires (et plus seulement des prédicteurs contrôlés). On définit ainsi les modèles à effets aléatoires et les modèles mixtes que nous traiterons au chapitre 6.
- On peut enfin considérer, pour chaque individu i pris en compte, des observations de Y_i répétées dans le temps. Ces observations sont naturellement correlées, ce qui nécessite l'introduction de modèles spécifiques : les modèles pour données répétées, étudiés au chapitre 7.

6.3 Le modèle linéaire généralisé

Il s'agit d'une extension du modèle linaire qui ne sera pas abordée dans ce cours. Pour mémoire, indiquons qu'il s'agit toujours d'expliquer une variable Y au moyen de prédicteurs X^j, en utilisant un échantillon de taille n, mais qu'il y a généralisation à deux niveaux :

- chaque v.a.r. Y_i de l'échantillon est distribuée selon une même loi de la *famille exponentielle* (normale, binomiale, Poisson, gamma...) ;
- la relation linéaire entre $\mathbb{E}(Y_i)$ et les prédicteurs X^j se fait au moyen d'une fonction particulière g, monotone et dérivable, appelée *fonction*

lien, de la façon suivante :

$$g[\mathbb{E}(Y_i)] = \sum_{i=1}^{p} \beta_j X^j.$$

6.3.1 Exemples

- Si l'on prend la loi normale comme loi de la famille exponentielle et la fonction identité comme fonction lien, on retrouve le modèle linéaire gaussien de base : le modèle linéaire généralisé en constitue donc bien une généralisation.
- Si l'on suppose maintenant $Y_i \sim \mathcal{B}(n_i, p_i)$, qu'on modélise $\frac{Y_i}{n_i}$ et qu'on choisit la fonction *logit* comme fonction lien ($g(x) = \log(\frac{x}{1-x}),\ x \in]0,1[$), on obtient la régression logistique :

$$\mathbb{E}(\frac{Y_i}{n_i}) = p_i \ ;\ g(p_i) = \log(\frac{p_i}{1-p_i}) = \sum_{j=1}^{p} \beta_j x_i^j.$$