Régression linéaire et logistique

Frédéric Lavancier

Université de Nantes
M1 Ingénierie Statistique

2021/2022

Références

- "Régression avec R", P-A. Cornillon, E. Matzner-Løber
- "An introduction to statistical learning with applications in R", G. James, D. Witten, T. Hastie, R. Tibshirani.
- ESL : "The elements of statistical learning", T. Hastie, R. Tibshirani, J. Friedman.

Table des matières

Chapitre 1

Analyse bivariée

1.1 Lien quanti-quanti

1.2 Lien quanti-quali

1.3 Lien quali-quali

$\longrightarrow$ Voir les slides associés dans Madoc

Chapitre 2

Régression linéaire

On s'intéresse au lien entre une variable **quantitative** Y et p variables **quantitatives** $X_1, \ldots, X_p$. A-t-on $Y \approx f(X_1, \ldots, X_p)$ pour une certaine fonction f ?

En régression linéaire, on suppose que f est linéaire, ce qui conduit au modèle

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

pour certains paramètres $\beta_1, \ldots, \beta_p$ inconnus et où ϵ représente l'erreur de modélisation. Dans cette écriture, l'une des variables (disons X_1) est souvent la constante $X_1 = 1$, ce qui conduit en réalité à un modèle affine entre Y et $X_2, \ldots, X_p$.

En pratique, on observe les variables Y et $X_1, \ldots, X_p$ auprès de n individus et on souhaite :

- estimer les paramètres $\beta_1, \ldots, \beta_p$,
- valider la relation précédente.

2.1 Modélisation

Pour chaque individu i ($i = 1, \ldots, n$), on note :

y_i : valeur de la variable Y pour l'individu i
$x_{1,i}$: valeur de la variable X_1 pour l'individu i
$\vdots$
$x_{p,i}$: valeur de la variable X_p pour l'individu i
ϵ_i : l'erreur de modélisation associé à l'individu i.

Pour chaque individu i, le modèle de régression linéaire s'écrit donc :

$$y_i = \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \epsilon_i. \tag{2.1}$$

On regroupe les observations dans les vecteurs

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{et} \quad X_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1n} \end{pmatrix}, \quad \cdots \quad , \; X_p = \begin{pmatrix} x_{p1} \\ \vdots \\ x_{pn} \end{pmatrix}.$$

Attention, il y a ici un changement dans les notations : Y désigne à présent le vecteur des valeurs $y_1, \ldots, y_n$ alors qu'il désignait jusqu'alors la "variable" Y dont est issue ces observations, de même pour $X_1, \ldots, X_p$.

On introduit la matrice X de taille (n, p) regroupant toutes les variables explicatives, appelée également "matrice de design" :

$$X = \big(X_1 \ldots X_p\big) = \begin{pmatrix} x_{11} & \ldots & x_{p1} \\ \vdots & \vdots & \vdots \\ x_{1n} & \ldots & x_{pn} \end{pmatrix}.$$

On note enfin $\beta \in \mathbb{R}^p$ le vecteur des paramètres et $\epsilon \in \mathbb{R}^n$ celui des erreurs de modélisation de chaque individu (même changement de notation que précédemment) :

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Avec ces notations, le modèle de régression (2.1) sur les n individus s'écrit donc :

$$Y = X\beta + \epsilon.$$

Nous ferons les **hypothèses** suivantes :

- Les erreurs de modélisations ϵ_i sont aléatoires, d'espérance nulle et de même variance σ^2 (homoscédasticité). Elles sont de plus non corrélées 2 à 2. Autrement dit

  $$\mathbb{E}(\epsilon) = 0 \quad \text{et} \quad Var(\epsilon) = \sigma^2 I_n,$$

  où I_n désigne la matrice identité de taille n.
- La matrice de design X est non aléatoire et de plein rang ($p \leq n$ et $rg(X) = p$). Cela signifie qu'aucune colonne X_j n'est combinaison linéaire des autres.

Remarque 2.1.1.

- *La première hypothèse implique que Y est aléatoire avec $\mathbb{E}(Y) = X\beta$ et $Var(Y) = \sigma^2 I_n$.*
- *Si X n'était pas de plein rang, alors le modèle ne serait pas identifiable, dans le sens où une infinité de paramètres β donneraient le même modèle. → Exemple en cours.*

Exemple : Régression simple (→ schéma en cours) : $y_i = \beta_1 + \beta_2 x_i + \epsilon_i$ pour $i = 1, \dots, n$. Dans ce cas $p = 2$, $x_{i,1} = 1$, $x_{i,1} = x_i$. Autrement dit

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

2.2 Estimation des paramètres

2.2.1 Estimation de β par les moindres carrés ordinaires (MCO)

Les MCO consistent à trouver la valeur du vecteur β qui minimise

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_{1,i} - \cdots - \beta_p x_{p,i})^2.$$

Soit $[X]$ l'espace vectoriel engendré par les vecteurs $X_1, \dots, X_p$, i.e.

$$[X] = \{X\alpha,\ \alpha \in \mathbb{R}^p\} = \{v \in \mathbb{R}^n,\ \exists \alpha \in \mathbb{R}^p, v = X\alpha\}.$$

L'élément $X\hat{\beta} \in [X]$ qui minimise $\|Y - X\beta\|^2$ est la projection orthogonale de Y sur $[X]$. On note $\hat{Y} = X\hat{\beta}$ et le vecteur des résidus $\hat{\epsilon} = Y - \hat{Y}$.

→ Voir Figure 2.1 et schéma en cours.

Theorem 2.2.1. *Si $rg(X) = p$, $\hat{\beta} = (X'X)^{-1}X'Y$.*

Démonstration. Cf cours □

Remarque 2.2.2. *La matrice $P_{[X]} = X(X'X)^{-1}X'$ est la matrice de projection sur $[X]$. Sachant cela, on retrouve le résultat car par définition $\hat{Y} = P_{[X]}Y = X(X'X)^{-1}X'Y$, ce qui signifie que $\hat{Y} = X\hat{\beta}$ avec $\hat{\beta} = (X'X)^{-1}X'Y$.*

Soit $[X]^\perp$ l'espace vectoriel orthogonal à $[X]$ dans $\mathbb{R}^n$, c'est à dire $[X]^\perp = \{v \in \mathbb{R}^n,\ X'v = 0\}$. La matrice de projection sur $[X]^\perp$ est $P_{[X]^\perp} = I_n - P_{[X]} = I_n - X(X'X)^{-1}X'$.

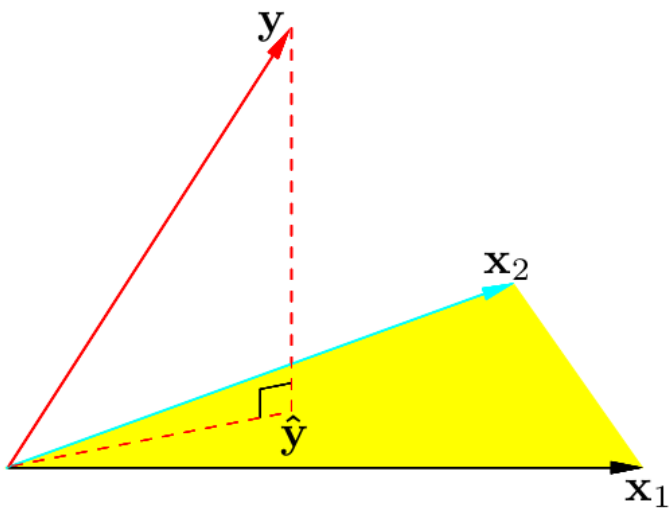

FIGURE 2.1 – Figure extraite de l'ouvrage ESL. Le plan en jaune représente $[X]$ lorsque $p = 2$. Le vecteur Y est projeté sur $[X]$ pour donner $\hat{Y}$.

Proposition 2.2.3. *Si* $rg(X) = p$, $\mathbb{E}(\epsilon) = 0$ *et* $Var(\epsilon) = \sigma^2 I_n$, *alors*

$$\mathbb{E}(\hat{\beta}) = \beta \quad (\hat{\beta} \text{ est un estimateur sans biais})$$
$$Var(\hat{\beta}) = (X'X)^{-1}\sigma^2.$$

Démonstration. Cf cours □

Theorem 2.2.4 (Théorème de Gauss-Markov)**.** *Si* $rg(X) = p$, $\mathbb{E}(\epsilon) = 0$ *et* $Var(\epsilon) = \sigma^2 I_n$, *alors* $\hat{\beta}$ *est le meilleur estimateur linéaire sans biais de* β, *au sens du coût quadratique.*

Cela signifie qu'étant donné un autre estimateur linéaire $\tilde{\beta} = MY$ pour une certaine matrice M non aléatoire, avec $\tilde{\beta}$ sans biais ($\mathbb{E}(\tilde{\beta}) = \beta$), alors on a nécessairement que $Var(\hat{\beta}) \leq Var(\tilde{\beta})$ au sens où la différence $Var(\tilde{\beta}) - Var(\hat{\beta})$ est semi-définie positive (les estimateurs étant des vecteurs, leur variance est une matrice).

Démonstration. Cf cours □

2.2.2 Estimation de σ^2

On introduit les **résidus** : $\hat{\epsilon}_i = y_i - \hat{y}_i$ pour $i = 1, \dots, n$ et on note le vecteur des résidus

$$\hat{\epsilon} = \begin{pmatrix} \hat{\epsilon}_1 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix}.$$

Proposition 2.2.5.

- $\hat{\epsilon} = Y - \hat{Y} = Y - P_{[X]}Y = P_{[X]^\perp}Y = P_{[X]^\perp}\epsilon$.
- *Si* $rg(X) = p$, $\mathbb{E}(\epsilon) = 0$ *et* $Var(\epsilon) = \sigma^2 I_n$, *alors* $\mathbb{E}(\hat{\epsilon}) = 0$ *et* $Var(\hat{\epsilon}) = \sigma^2 P_{[X]^\perp} = \sigma^2(I_n - X(X'X)^{-1}X')$
- *Si le modèle contient une constante, typiquement* $X_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, *alors* $\bar{\hat{\epsilon}} = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i = 0$, *autrement dit* $\bar{\hat{Y}} = \bar{Y}$.

Démonstration. Cf cours □

Proposition 2.2.6.

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{1}{n-p}\|\hat{\epsilon}\|^2$$

est un estimateur sans biais de σ^2.

Démonstration. Cf cours □

2.2.3 Cas Gaussien

Dans cette partie on suppose que $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Cela implique que $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$.

Proposition 2.2.7. *En notant* $\hat{\beta}_{MV}$ *et* $\hat{\sigma}^2_{MV}$ *les estimateurs du maximum de vraisemblance de* β *et* σ^2 *(et* $\hat{\beta}$ *et* $\hat{\sigma}^2$ *les estimateurs précédents) on a*

- $\hat{\beta}_{MV} = \hat{\beta}$ *et* $\hat{\sigma}^2_{MV} = \frac{1}{n}\|\hat{\epsilon}\|^2 = \frac{n-p}{n}\hat{\sigma}^2$.
- $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$
- $\frac{n-p}{\sigma^2}\hat{\sigma}^2 = \frac{n}{\sigma^2}\hat{\sigma}^2_{MV} \sim \chi^2(n-p)$
- $\hat{\beta}$ *et* $\hat{\sigma}^2$ *sont indépendants*

Démonstration. L'expression de $\hat{\beta}_{MV}$ et $\hat{\sigma}^2_{MV}$ est admise (cf le module de statistique inférentielle). Pour le reste, cf cours □

Remarque 2.2.8. *Le fait de connaître la loi de $\hat{\beta}$ et de $\hat{\sigma}^2$ permet de construire des intervalles de confiance, de faire des tests, etc. Si le modèle n'est pas Gaussien, la loi de $\hat{\beta}$ et de $\hat{\sigma}^2$ n'est pas connue à n fixé, mais elle le devient asymptotiquement (pour n grand), sous certaines conditions assez faibles de régularité, et coïncide avec le cas Gaussien (admis).*

Theorem 2.2.9. *Dans le modèle Gaussien, $\hat{\beta}$ est un estimateur efficace de β, c'est à dire qu'il s'agit du meilleur estimateur sans biais possible de β.*

Démonstration. Admis. La notion d'efficacité d'un estimateur sera vue dans le module de statistique inférentielle. □

2.2.4 Tests et intervalles de confiances pour β_j

On rappelle que $\hat{\sigma}^2 = \frac{1}{n-p}\|\hat{\epsilon}\|^2$.

Proposition 2.2.10. *Dans le modèle Gaussien (i.e. $rg(X) = p$ et $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$), pour tout $j = 1, \dots, p$,*

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}} \sim St(n-p)$$

où $(X'X)^{-1}_{jj}$ désigne le j-ème élément de la diagonale de la matrice $(X'X)^{-1}$.

Remarque 2.2.11. *Dans la formule ci-dessus $\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}$ n'est autre qu'une estimation de l'écart-type de $\hat{\beta}_j$, car $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$ implique que $Var(\hat{\beta}_j) = \sigma^2(X'X)^{-1}_{jj}$. On note parfois cette variance $\sigma^2_{\hat{\beta}_j}$ de telle sorte que dans le modèle Gaussien*

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim St(n-p).$$

Démonstration. Cf cours □

Grâce à la proposition 2.2.10, on peut construire des tests et des intervalles de confiance sur chaque paramètre β_j :

1. Test de significativité : $H_0 : \ \beta_j = 0$ contre $H_1 : \ \beta_j \neq 0$. La région critique au niveau α est

$$RC_\alpha = \{\frac{|\hat{\beta}_j|}{\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}} > t_{n-p}(1-\alpha/2)\}.$$

2. L'intervalle de confiance au niveau $1-\alpha$ pour β_j est

$$IC_{1-\alpha}(\beta_j) = \left[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}; \hat{\beta}_j + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}\right].$$

On vérifie en effet facilement que d'après la proposition 2.2.10

$$\mathbb{P}(\beta_j \in IC_{1-\alpha}(\beta_j)) = 1-\alpha.$$

Remarque 2.2.12. *Ces tests et intervalles de confiance sont ceux calculés par les logiciels (R en particulier). En toute rigueur, ils ne sont valables que sous l'hypothèse Gaussienne. Néanmoins, ils restent valables dès que n est grand (quelques dizaines), conformément à la remarque 2.2.8.*

2.2.5 Prévision

On suppose qu'on a estimé β et σ^2 par $\hat{\beta}$ et $\hat{\sigma}^2$ à partir des observations des variables Y et $X_1, \ldots, X_p$ auprès de n individus.

On souhaite prédire Y pour un nouvel individu $n+1$, c'est à dire prédire y_{n+1}, connaissant les valeurs prises par cet individu pour les variables $X_1, \ldots, X_p$, c'est à dire connaissant les valeurs $x_{1,n+1}, \ldots, x_{p,n+1}$.

On suppose que ce nouvel individu suit exactement le même modèle de régression linéaire que les autres individus, associé à une erreur de modélisation ϵ_{n+1} qui lui est propre, centré, de même variance σ^2 et non corrrélées avec les erreurs des autres individus. En notant x_{n+1} le vecteur de taille p

$$x_{n+1} = \begin{pmatrix} x_{1,n+1} \\ \vdots \\ x_{p,n+1} \end{pmatrix}.$$

cela signifie que

$$y_{n+1} = x'_{n+1}\beta + \epsilon_{n+1} = \beta_1 x_{1,n+1} + \cdots + \beta_p x_{p,n+1} + \epsilon_{n+1},$$

où $\mathbb{E}(\epsilon_{n+1}) = 0$, $V(\epsilon_{n+1}) = \sigma^2$ et $Cov(\epsilon_{n+1}, \epsilon_i) = 0$ pour tout $i = 1, \dots, n$.

Etant donné ce modèle, la prévision naturelle de y_{n+1} est

$$\hat{y}_{n+1} = x'_{n+1}\hat{\beta}.$$

Deux erreurs se cumulent dans cette prévision : celle due à "l'oubli" de ϵ_{n+1} et celle due à l'estimation de β par $\hat{\beta}$.

L'erreur de prévision vaut $y_{n+1} - \hat{y}_{n+1}$. On a (cf cours) :

- $\mathbb{E}(y_{n+1} - \hat{y}_{n+1}) = 0$,
- $Var(y_{n+1} - \hat{y}_{n+1}) = \sigma^2(x'_{n+1}(X'X)^{-1}x_{n+1} + 1)$.

L'erreur de prévision est donc nulle en moyenne, tandis que sa variance intègre les deux types d'erreurs évoquées ci-dessus : la première est liée à $\hat{\beta}$ et devient négligeable lorsque n est grand, dès que $(X'X)^{-1} \to 0$ (ce qui est généralement le cas) ; la seconde est liée à ϵ_{n+1} et vaut toujours σ^2, cette erreur est incompressible.

Intervalles de prédiction :

Si l'on suppose que le modèle est Gaussien (c'est à dire que les erreurs suivent une loi Gaussienne comme dans la section 2.2.3), alors $y_{n+1} - \hat{y}_{n+1} \sim \mathcal{N}(0, \sigma^2(x'_{n+1}(X'X)^{-1}x_{n+1} + 1))$. On en déduit que

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma}\sqrt{x'_{n+1}(X'X)^{-1}x_{n+1} + 1}} \sim St(n-p)$$

et on peut donc fournir un intervalle de prédiction pour y_{n+1} :

$$IP_{1-\alpha}(y_{n+1}) = \hat{y}_{n+1} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{x'_{n+1}(X'X)^{-1}x_{n+1} + 1},$$

dans le sens où $\mathbb{P}(y_{n+1} \in IP_{1-\alpha}(y_{n+1})) = 1 - \alpha$.

Attention : si n est grand, $\hat{\beta}$ suit approximativement une loi Gaussienne, même si le modèle n'est pas Gaussien, mais par contre ϵ_{n+1} suit sa propre loi, qui n'est pas Gaussienne si le modèle n'est pas Gaussien. Ainsi, $y_{n+1} - \hat{y}_{n+1}$ ne suit pas une loi Gaussienne si le modèle n'est pas Gaussien, même si n est grand. Les intervalles de prédiction ne sont donc valables que pour les modèles Gaussiens.

2.3 Validation

2.3.1 Qualité explicative globale

On définit le R^2, appelé également coefficient de détermination ou coefficient de corrélation multiple, à l'aide du théorème de Pythagore.

On distingue deux cas, selon que le modèle contient une constante (le vecteur $\mathbb{1}$ appartient à $[X]$, par exemple $X_1 = \mathbb{1}$) ou non.

Si $\mathbb{1} \in [X]$, on a (voir schéma en cours) :

$$\|Y - \bar{Y}\mathbb{1}\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{Y}\mathbb{1}\|^2$$

qu'on écrit généralement : $SCT = SCR + SCE$, où SCT est la "somme des carré totaux", SCR est la "somme des carrés des résidus" et SCE est la "somme des carrés expliqués".

Dans le cas général (même si $\mathbb{1} \notin [X]$), on a :

$$\|Y\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y}\|^2.$$

Définition 2.3.1. *Le R^2 est défini ainsi :*

- *si $\mathbb{1} \in [X]$,*
$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT},$$
- *si $\mathbb{1} \notin [X]$,*
$$R^2 = \frac{\|\hat{Y}\|^2}{\|Y\|^2} = 1 - \frac{SCR}{\|Y\|^2}.$$

Remarque 2.3.2.

- *Cela n'a aucun sens de comparer le R^2 d'un modèle avec constante et le R^2 d'un modèle sans constante (les définitions diffèrent).*
- *On a toujours $0 \leq R^2 \leq 1$, le modèle étant d'autant "meilleur" que R^2 est proche de 1.*
- *En régression linéaire simple ($y = \beta_1 + \beta_2 x + \epsilon$), R^2 correspond simplement à la corrélation empirique (au carré) entre y et x : $R^2 = \hat{\rho}^2$ (voir cours).*

Le R^2 a un défaut important : il augmente nécessairement lorsqu'on ajoute une variable explicative, même si cette dernière n'est pas significative. En effet, ajouter une variable explicative grossit l'espace $[X]$, ce qui diminue automatiquement la SCR issue de la projection. Utiliser le R^2 pour choisir

entre deux modèles possibles conduira donc toujours à prendre le modèle le plus gros. Pour palier ce problème, on introduit le R^2 ajusté.

Définition 2.3.3. *Le R^2 ajusté, noté R^2_a est défini ainsi :*

- *si $\mathbb{1} \in [X]$,*
$$R^2_a = 1 - \frac{n-1}{n-p}\frac{SCR}{SCT},$$
- *si $\mathbb{1} \notin [X]$,*
$$R^2_a = 1 - \frac{n}{n-p}\frac{SCR}{\|Y\|^2}.$$

On remarque que lorsqu'on ajoute une variable explicative, le SCR diminue nécessairement, mais dans le même temps $n-p$ passe à $n-(p+1)$ donc R^2_a n'augmente pas nécessairement. Le R^2_a n'augmente que si la SCR diminue de façon significative.

Remarque 2.3.4. *L'idée dans la définition du R^2_a est la suivante. Puisque $R^2 = 1 - SCR/SCT$, on peut le voir comme un estimateur de $1-\sigma^2/Var(Y)$. En utilisant à la place les estimateurs corrigés de la variance $\hat{\sigma}^2 = SCR/(n-p)$ et $\widehat{Var}(Y) = SCT/(n-1)$, on obtient le R^2_a.*

2.3.2 Tests de contraintes linéaires sur les coefficients

On désire tester q contraintes linéaires sur le coefficient β (vecteur de taille p). Cela s'écrit

$$H_0 : \ R\beta = 0 \quad \text{contre} \quad H_1 : \ R\beta \neq 0,$$

où R est une matrice de taille (q,p) encodant les contraintes.

Exemples

a. *Test de Student* : si $R = (0,\dots,0,1,0,\dots,0)$ est de taille $(1,p)$ dont les coefficients valent tous 0 sauf le j-ème qui vaut 1, on retrouve le test de Student $H_0 : \ \beta_j = 0$ contre $H_1 : \ \beta_j \neq 0$ discuté dans la section 2.2.4.

b. *Test de Fisher global* : on suppose qu'il existe une constante dans le modèle (disons $X_1 = \mathbb{1}$) et on teste si au moins une variable (autre que la constante) est significative, c'est à dire $H_0 : \ \beta_2 = \cdots = \beta_p = 0$ contre H_1 : il existe au moins un $j \in \{2,\dots,p\}$ tel que $\beta_j \neq 0$.

L'hypothèse nulle s'écrit $H_0 : \ R\beta = 0$ avec R de taille $(p-1, p)$ donné par

$$R = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}.$$

La statistique de test dans ce cas est souvent écrite

$$F = \frac{n-p}{p-1}\frac{SCE}{SCR} = \frac{n-p}{p-1}\,\frac{R^2}{1-R^2} \tag{2.2}$$

et il s'agit d'un cas particulier traité dans le théorème ci-dessous.

c. *Test de modèles emboités* : on veut tester le modèle global

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

contre le sous-modèle

$$Y = \beta_1 X_1 + \cdots + \beta_{p-q} X_{p-q} + \epsilon$$

dans lequel on n'a pas pris en compte les q dernières variables. Cela revient à tester dans le modèle global $H_0 : \ \beta_{p-q+1} = \cdots = \beta_p = 0$ contre H_1 : le contraire, ce qui revient à $H_0 : \ R\beta = 0$ avec R de taille (q, p) donné par $R = \begin{pmatrix} 0_{q,p-q} & | & I_q \end{pmatrix}$ où $0_{q,p-q}$ désigne la matrice nulle de taille $(q, p-q)$.

Theorem 2.3.5. *Si $rg(X) = p$ et $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, alors sous $H_0 : \ R\beta = 0$*

$$F = \frac{n-p}{q}\ \frac{SCR_c - SCR}{SCR} \sim F(q, n-p)$$

où $F(q, n-p)$ désigne la loi de Fisher à $(q, n-p)$ degré de liberté. Dans la formule précédente, SCR correspond à la SCR dans le modèle global, tandis que SCR_c correspond à la SCR dans le modèle contraint, c'est à dire le sous-modèle vérifiant $R\beta = 0$.

On en déduit la région critique du test au niveau α :

$$RC_\alpha = \{F > f_{q,n-p}(1-\alpha)\}$$

où $f_{q,n-p}(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ d'une $F(q, n-p)$.

Démonstration. Cf cours □

Remarque 2.3.6. *Si le modèle global et le modèle contraint contiennent une constante, ou si aucun des deux ne la contient, alors F s'écrit également*

$$F = \frac{n-p}{q} \frac{R^2 - R_c^2}{1 - R^2}$$

où R^2 est le R^2 dans le modèle global, tandis que R_c^2 est le R^2 dans le modèle contraint (cf TD).

Retour sur le exemples

a. *Test de Student* : ici $q = 1$ et $F = (n-p)(SCR_c - SCR)/SCR$. On peut montrer que dans ce cas $F = T^2$ où $T = \hat{\beta}_j/(\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}})$ correspond à la statistique du test de Student présenté dans la section 2.2.4. Le test correspond donc exactement au test de Student.
b. *Test de Fisher global* : on peut montrer que la statistique F du théorème correspond exactement à celle donnée en (2.2) (voir cours).
c. *Test de modèles emboités* : on calcule la statistique F où SCR correspond à la SCR du modèle global et SCR_c correspond à la SCR du sous-modèle sans les q variables.

Sous R : soit on estime les deux modèles (avec et sans contraintes) et on compare leur SCR via la formule définissant F, soit on utilise la fonction `linearHypothesis` de la librairie `car`.

2.3.3 Vérification des hypothèses du modèle

Pour rappel, les hypothèses sont : $Y = X\beta + \epsilon$ avec $rg(X) = p$ et $\mathbb{E}(\epsilon) = 0$, $V(\epsilon) = \sigma^2 I_n$.

Il s'agit donc de vérifier si le lien linéaire est adéquat, s'il n'y a pas de colinéarité entre les variables explicatives ($rg(X) = p$), et si l'erreur de modélisation ϵ vérifie bien $\mathbb{E}(\epsilon) = 0$ et $V(\epsilon) = \sigma^2 I_n$.

a. Lien linéaire

Avant la modélisation, on peut représenter les nuages de points (X_j, Y) entre chaque variable explicative X_j et la variable à expliquer Y : un lien linéaire doit apparaitre.

Après la modélisation, on peut analyser le vecteur des résidus $\hat{\epsilon}$ qui, si le lien linéaire est mis en défaut, n'aura pas le comportement attendu (cf la partie c. plus bas).

Autre outil (non présenté en détail dans ce cours) : l'analyse des résidus partiels.

Si le lien linéaire ne semble pas approprié : on peut éventuellement essayer de transformer les variables X_j et/ou Y (par exemple via une transformation logarithmique) pour faire apparaitre un lien linéaire. Sinon, il faut se tourner vers des modèles non linéaires.

b. Non-colinéarité des variables explicatives ($rg(X) = p$)

Quel est le problème en présence de variables explicatives colinéaires ? Comme nous l'avons évoqué dans la remarque (2.1.1), si deux variables sont linéairement liées, le paramètre β n'est pas identifiable. Mathématiquement, la matrice $X'X$ n'est pas inversible et la formule de $\hat{\beta}$ n'a donc pas de sens.

Mais de façon moins extrême, si deux variables explicatives sont presque colinéaires (c'est à dire que leur corrélation empirique $|\hat{\rho}|$ est élevée, sans pour autant valoir 1), cela pose également un problème. En effet dans ce cas la matrice $X'X$ est inversible, mais son inverse est très instable dans le sens où si on enlève un individu au jeu de données (on enlève une ligne à X), alors le résultat de $(X'X)^{-1}$ peut devenir radicalement différent. Cela signifie que $\hat{\beta}$ peut donc varier énormément à cause d'un seul individu, ce qui n'est pas souhaitable d'un point de vue statistique.

On peut détecter ce phénomène en calculer les VIF (Variance Inflation Factor) pour chaque variable X_j :

1. on régresse X_j par rapport aux autres variables X_k $(k \neq j)$;
2. on calcule le R^2 dans cette régression, que l'on note R_j^2 (il s'agit donc d'une mesure de la corrélation entre X_j et les autres variables) ;
3. le VIF pour la variable X_j vaut

$$VIF_j = \frac{1}{1 - R_j^2}.$$

Le VIF est toujours supérieur à 1. Plus X_j est corrélé aux autres variables et plus R_j^2 sera proche de 1, et plus VIF_j sera élevé. On considère généralement que VIF_j devient trop élevé lorsque sa valeur dépasse 5 (ce qui correspond à $R_j^2 > 0.8$).

En pratique, si VIF_j est élevé pour une variable X_j, il l'est également pour au moins une autre (la variable fortement corrélée à X_j). De ce point de vue, X_j apporte une redondance d'informations peu pertinente à la modélisation, mais perturbante pour l'estimation. Face à ce genre de situation, on peut

- Enlever du modèle une des variables dont le VIF est élevé, en recommançant jusqu'à ce que tous les VIF soient faibles.
- Ou faire appel à des méthodes d'estimation robuste comme la régression ridge (cf le cours de statistique en grande dimension en M2), qui évite d'avoir à sélectionner les variables "à la main".

Sous R : pour calculer les VIF de chaque variable d'un modèle de régression : fonction `vif` du package `car`. Par exemple, pour le modèle `reg` issu d'un ajustement avec la fonction `lm` : `vif(reg)`.

c. Analyse des résidus

Pour rappel, le vecteur des résidus est $\hat{\epsilon} = Y - \hat{Y} = P_{[X]^\perp}\epsilon$.

On sait d'après la proposition 2.2.5 que si $Y = X\beta + \epsilon$ avec $rg(X) = p$, $\mathbb{E}(\epsilon) = 0$ et $Var(\epsilon) = \sigma^2 I_n$, alors

- $\mathbb{E}(\hat{\epsilon}) = 0$ et $Var(\hat{\epsilon}) = \sigma^2 P_{[X]^\perp}$.
- $Cov(\hat{\epsilon}, \hat{Y}) = 0$ (cf justification en cours)
- et si le modèle contient une constante ($\mathbb{1} \in [X]$), alors $\bar{\hat{\epsilon}} = 0$.

Puisque $Cov(\hat{\epsilon}, \hat{Y}) = 0$, un nuage de points entre $\hat{Y}$ et $\hat{\epsilon}$ ne devrait pas faire apparaître de structures particulières. A défaut, cela peut témoigner d'un lien non linéaire initiale entre Y est les variables explicatives (cf illustration en cours).

i) Vérification de l'homoscédasticité (cad $Var(\epsilon_i) = \sigma^2$ pour tout i)

On souhaite baser cette vérification sur les résidus $\hat{\epsilon}_i$, mais ces derniers ne sont pas homoscédastiques. En effet $Var(\hat{\epsilon}) = \sigma^2 P_{[X]^\perp} = \sigma^2(I_n - P_{[X]})$ donc en notant h_{ij} les éléments de la matrice $P_{[X]}$ (h comme "hat matrix", le nom donné à $P_{[X]}$ en anglais) :

$$Var(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$$

dépend de i. Mais $Var(\hat{\epsilon}_i/(\sigma\sqrt{1-h_{ii}})) = 1$ ne dépend pas de i. Cela motive l'utilisation des résidus standardisés :

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

Le fait de remplacer σ par $\hat{\sigma}$ rend la variance de t_i différente de 1, mais néanmoins $Var(t_i)$ ne dépend pas de i (admis) et reste proche de 1.

Pour vérifier graphiquement l'homoscédasticité, on peut ainsi tracer les résidus standardisés t_i en fonction de i : le nuage de points devrait être dans une même bande. On peut de même représenter le nuage de points des $(\hat{y}_i, t_i)$, qui devrait aussi rester dans une même bande. Si le nuage de points s'élargit avec $\hat{y}_i$ (par exemple), cela témoigne d'une hétéroscédasticité, la variance des résidus augmentant avec les valeurs des $\hat{y}_i$ (cf illustrations en cours).

De façon plus formelle, on peut appliquer le test de Breusch-Pagan. Ce dernier suppose que dans le modèle de régression linéaire, l'erreur ϵ_i a une variance $\sigma_i^2 = \sigma^2 + z_i'\gamma$ où z_i est un vecteur de k variables à choisir qui pourraient expliquer l'hétéroscédasticité (si elle est présente) et γ est un paramètre inconnu, de dimension k, à estimer. Le choix par défaut sous R est de prendre $z_i = X_{.i}$ c'est à dire les mêmes variables explicatives que dans le modèle de régression.

Si $\gamma = 0$, on retrouve le modèle linéaire classique dont le bruit est homoscédastique ($\sigma_i^2 = \sigma^2$ pour tout i). Mais si $\gamma \neq 0$, le bruit est hétéroscédastique. Le test de Breusch-Pagan consiste donc à tester

$$H_0 : \gamma = 0 \quad \text{contre} \quad H_1 : \gamma \neq 0.$$

La procédure de test n'est pas détaillée ici.

Sous R : fonction `bptest` de la librairie `lmtest`. L'option `studentize=FALSE` est approprié pour les modèles Gaussiens, tandis que l'option `studentize=TRUE` (par défaut) est adapté à un cas plus général.

Si on observe un problème d'hétéroscédasticité :

- on peut essayer de transformer Y pour "stabiliser" la variance
- on peut aussi essayer de modéliser cette hétéroscédasticité. Par exemple, si on pense que la variance diffère selon que l'individu est ou non dans le groupe A, cela donnerait : $\sigma_i^2 = \sigma_1^2$ si $i \in A$, et $\sigma_i^2 = \sigma_2^2$ si $i \notin A$. On utilise alors les MCG (moindres carrés généralisés) qui est une méthode généralisant les MCO et permettant d'estimer conjointement β, σ_1^2 et σ_2^2 (voir TD).

ii) Non-corrélation des erreurs (cad $Var(\epsilon)$ matrice diagonale)

La corrélation entre deux ϵ_i survient généralement lorsque les données sont temporelles (le "i" représente le temps).

Exemple : Y_i : résultats des ventes d'un produit le jour i. On peut s'attendre à ce que Y_i soit corrélé à Y_{i-1} (il peut y avoir des périodes de fortes ventes), et de même ϵ_i à ϵ_{i-1}.

Les résidus $\hat{\epsilon}_i$ ne sont pas décorrélés, même si les erreurs ϵ_i le sont, car $Var(\hat{\epsilon}) = \sigma^2 P_{[X]^\perp} = I_n - P_{[X]}$ n'est pas une matrice diagonale. Cependant, $P_{[X]} = X(X'X)^{-1}X'$ tend vers 0 dès que $(X'X)^{-1}$ tend vers 0, qui est la condition usuelle (et faible) pour que $\hat{\beta}$ converge en moyenne quadratique. Ainsi, si les ϵ_i sont décorrélés, les résidus $\hat{\epsilon}_i$ le sont aussi asymptotiquement.

On présente ci-dessous deux tests qui permettent de vérifier la non-corrélation des erreurs.

1. Test de Durbin-Watson

Dans le modèle linéaire, on suppose que $\epsilon_i = \rho\epsilon_{i-1} + \eta_i$ où $|\rho| < 1$ et les η_i sont iid suivant une $\mathcal{N}(0, \sigma^2)$. On dit dans ce cas que les ϵ_i sont "auto-corrélés" à l'ordre 1. La condition $|\rho| < 1$ assure l'existe d'un tel modèle. Si $\rho = 0$ dans cette relation, les $\epsilon_i = \eta_i$ sont non corrélés, sinon ils le sont. On teste donc

$$H_0 : \rho = 0 \quad \text{contre} \quad H_1 : \rho \neq 0.$$

La statistique de test de Durbin-Watson est

$$d = \frac{\sum_{i=2}^n (\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^n \hat{\epsilon}_i^2}.$$

Cette statistique est toujours comprise entre 0 et 4, et elle estime $2(1-\rho)$ (cf cours). Ainsi l'hypothèse H_0 sera rejetée lorsque la valeur de d est éloignée de 2. Le problème de cette statistique est que sous H_0, sa loi dépend de tous les paramètres du modèle, y compris la valeur de la matrice de design X. Il n'est donc pas possible de proposer des quantiles en toute généralité. En pratique, des approximations de ces quantiles, indépendantes de X, sont utilisées. La règle de décision du test est de la forme :

- si $|d-2| > q^+(1-\alpha/2)$ alors on rejette H_0 au niveau α,
- si $|d-2| < q^-(1-\alpha/2)$ on ne rejette pas H_0,

où $q^+(1-\alpha/2) > q^-(1-\alpha/2)$ sont deux approximations (une par le haut, l'autre pas le bas) du vrai quantile. En particulier, il existe une zone où on ne sait pas conclure, lorsque $q^-(1-\alpha/2) < |d-2| < q^+(1-\alpha/2)$.

Sous R : fonction `dwtest` de la librairie `lmtest`

2. Test de Breusch-Godfrey

Ce test ne souffre pas du défaut d'approximation du test de Durbin-Watson. De plus, il intègre une plus grande variété de corrélations possibles.

On suppose en effet que dans le modèle linéaire $\epsilon_i = \rho_1\epsilon_{i-1}+\cdots+\rho_r\epsilon_{i-r}+\eta_i$ où les η_i sont iid suivant une $\mathcal{N}(0,\sigma^2)$. Les ϵ_i sont donc "auto-corrélés" à l'ordre r (des conditions sur les coefficients ρ_k sont nécessaires pour assurer l'existence d'un tel modèle, elles seront détaillées dans le module de Séries Temporelles de M2). La valeur de r est choisi par l'utilisateur. Si $r = 1$, cela revient à l'hypothèse de Durbin-Watson. On souhaite tester

$$H_0 : \rho_1 = \cdots = \rho_r = 0 \quad \text{contre} \quad H_1 : \text{le contraire.}$$

Sous R : fonction `bgtest` de la librairie `lmtest`. Par défaut, l'option `type=''chisq''` utilise une statistique qui suit une loi $\chi^2(r)$ lorsque n est grand, sans hypothèse de loi sur les ϵ_i. On peut aussi choisir l'option `type=''F''` qui est dédié au modèle Gaussien et met en place un test de Fisher de contraintes sur les coefficients. La stat de test dans ce cas suit une $F(r,(n-r)-(p+r))$.

Que faire si une corrélation est détectée ?

Cela n'est pas forcément une mauvaise nouvelle : cela signifie que le modèle peut être enrichi en incluant de l'information contenue dans le passé des variables. Exemple : $y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \alpha_1 y_{i-1} + \alpha_2 x_{1,i-1} + \epsilon_i$. Avec cette démarche, les nouvelles erreurs ϵ_i du modèle peuvent devenir non-corrélées.

A défaut, la présence de corrélations dans les ϵ_i rend l'estimateur par MCO de β moins performant, mais il reste consistant si $(X'X)^{-1}$ tend vers 0. Pour améliorer l'estimation, on peut faire appel aux MCG (moindres carrés généralisés) qui peuvent tenir compte de la corrélation dans la procédure d'estimation. Cela nécessite néanmoins de bien spécifier le type de corrélations (par exemple une auto-corrélation d'ordre r), qui sera estimée en même temps que β. Mais attention, utiliser les MCG en se trompant sur la forme des corrélations peut conduire à des performances pires que l'utilisation simple des MCO.

iii) Normalité des erreurs

On rappelle que cette hypothèse n'est pas indispensable, dès lors que n est suffisamment grand. Tous les tests énoncés pour le modèle Gaussien restent valables asymptotiquement pour les modèles non Gaussiens. Le seul résultat qui exploite vraiment l'hypothèse Gaussienne est la formule de l'intervalle de prédiction de la section 2.2.5.

Pour vérifier la normalité des erreurs, on s'intéresse à la normalités des résidus $\hat{\epsilon}_i$. En effet, puisque $\hat{\epsilon} = P_{[X]^\perp}\epsilon$, les résidus sont Gaussiens si ϵ l'est.

Pour cela, on peut tracer la droite de Henry (qqplot ou qqnorm dans le cas d'une loi normale) des résidus, qui consiste à comparer les quantiles empirique des $\hat{\epsilon}_i$ aux quantiles théoriques de la loi normale. Si la représentation est (à peu près) une droite, l'hypothèse de normalité est acceptée.

Sous R : `qqnorm(residus)` si `residus` désigne le vecteur des résidus.

On peut également mettre en oeuvre un test de normalité. Le plus utilisé est le test de Shapiro-Wilk dont l'hypothèse nulle est l'hypothèse de normalité.

Sous R : `shapiro.test(residus)`.

2.3.4 Analyse des individus atypiques et/ou influents

Un individu est atypique dans la mesure où
i) il est très mal expliqué par le modèle,
ii) et/ou il influence énormément l'estimation des coefficients.

i) On identifie ces individus à l'aide de la valeur de leur résidu standardisé :

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

Si la valeur de t_i est trop extrême par rapport aux autres résidus standardisés, on considère que l'individu est aberrant. Dans ce cas, il est important de comprendre pourquoi et d'évaluer si cet individu a une forte influence sur l'estimation, cf le point suivant.

Remarque 2.3.7. *Le seuil de détection s'appuie généralement sur les quantiles de la loi de Student à $n-p$ degré de libertés. Néanmoins, en toute rigueur, il est faux d'affirmer que t_i suit cette loi, même pour un modèle Gaussien, car $\hat{\sigma}$ au dénominateur dépend de $\hat{\epsilon}_i$ au numérateur, or les deux quantités doivent être indépendantes pour que t_i suive une loi de Student. La loi de t_i est néanmoins très proche d'une $St(n-p)$.*

Il est à noter qu'on peut considérer les résidus "studentisés" t_i^* qui correspondent à la formule de t_i dans laquelle $\hat{\sigma}$ est remplacé par l'estimateur $\hat{\sigma}_{(-i)}$ de σ calculé à partir du modèle de régression ne faisant pas intervenir l'individu i. Cette démarche rend $\hat{\sigma}_{(-i)}$ et $\hat{\epsilon}_i$ indépendant, et cette fois (dans un modèle Gaussien) t_i^* suit bien une $St(n-p)$. On peut montrer (admis) que $t_i^* = t_i\sqrt{(n-p-1)/(n-p-t_i^2)}$.

Le logiciel R utilise t_i et non t_i^* dans les représentations graphiques proposées par la fonction `plot.lm`.

ii) Les points influents ne sont pas forcément des points aberrants (cf illustration en cours). On les détecte grâce à leur éventuel "effet levier", défini ci-dessous.

Définition 2.3.8. *Le poids de l'individu i sur sa propre estimation $\hat{y}_i$ est h_{ii}, où on rappelle que h_{ii} correspond au i-ème élément dans la diagonale de $P_{[X]} = X(X'X)^{-1}X'$.*

Cette définition provient du fait que

$$\hat{y}_i = [X\hat{\beta}]_i = [X(X'X)^{-1}X'Y]_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i}^{n} h_{ij}y_j$$

et donc y_i contribue au calcul de $\hat{y}_i$ avec le poids h_{ii}.

On sait que $tr(P_{[X]}) = p$, c'est à dire $\sum_{i=1}^{n} h_{ii} = p$. On peut donc s'attendre à ce qu'en moyenne $h_{ii} \approx p/n$. Si h_{ii} est beaucoup plus grande que cette valeur, alors l'individu est "levier".

Définition 2.3.9. *Un individu i est dit levier si $h_{ii} >> p/n$, typiquement $h_{ii} > 2p/n$ ou $h_{ii} > 3p/n$.*

Un individu levier influence beaucoup l'estimation de β donc il faut les détecter, les analyser, et éventuellement les enlever de l'étude.

La distance de Cook quantifie l'influence de i sur $\hat{\beta}$:

$$C_i = \frac{\|\hat{Y} - \hat{Y}_{(-i)}\|^2}{p\hat{\sigma}^2}$$

où $\hat{Y}_{(-i)} = X\hat{\beta}_{(-i)}$ avec $\hat{\beta}_{(-i)}$ l'estimation de β sans utiliser l'individu i. On peut montrer que

$$C_i = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}}t_i^2.$$

Cette dernière formule montre que C_i cumule l'effet "aberrant" de l'individu, au travers de la présence de t_i, et son effet levier, via h_{ii}.

Sous R : la fonction `cooks.distance` permet le calcul de C_i pour tous les individus. La dernière représentation graphique proposée avec `plot(reg)` (si `reg` est le nom du modèle de régression estimé avec `lm`) représente le nuage de points (h_{ii}, t_i). La valeur limite de la distance de Cooks au delà de laquelle on peut considérer que le point est très influent apparait selon une ligne hyperbolique : le point a en effet d'autant plus de chances d'être influent qu'il cumule une valeur élevée de h_{ii} et de t_i (en valeur absolue).

2.4 Critères de sélection de modèles

Objectif : choisir entre deux modèles concurrents, voire choisir parmi tous les sous-modèles possibles avec les variables explicatives à disposition.

2.4.1 Les critères

De nombreux critères existent. Les plus classiques sont :

i) le R_a^2 : On choisit le modèle ayant le R_a^2 le plus élevé.

ii) Si les deux modèles sont emboités, on peut effectuer un test de contraintes de Fisher, comme dans la section 2.3.2 (exemple c.).

iii) le C_p de Mallows : On suppose disposer de $p_{\max}$ variables explicatives, formant les colonnes de la matrice de design $X_{\max}$. On suppose par ailleurs que le vrai modèle (inconnu) expliquant Y s'écrit $Y = X^*\beta^* + \epsilon$ où X^* est la sous-matrice de $X_{\max}$ formée de $p^* \leq p_{\max}$ de ses colonnes. Autrement dit, parmi les $p_{\max}$ variables explicatives disponibles, seules p^* sont pertinentes pour le modèle. En pratique on ne connait pas la valeur p^* et encore moins de quelles variables il s'agit. Pour trouver ces variables, et pouvoir estimer le modèle au mieux, on calcule un score pour chaque sous-modèle candidat.

Soit $Y = X\beta + \epsilon$ un modèle candidat contenant p variables explicatives (qui est potentiellement faux). Le C_p de Mallows vise à estimer l'erreur $\mathbb{E}(\|Y - X\hat{\beta}\|^2)$ où $\hat{\beta}$ est l'estimateur des MCO dans ce modèle. L'expression de C_p est

$$C_p = \frac{SCR}{\hat{\sigma}^2} - n + 2p$$

où SCR est la SCR dans le modèle testé et $\hat{\sigma}^2$ est l'estimation de σ^2 dans le plus gros modèle (celui contenant les $p_{\max}$ variables explicatives).

Selon ce critère, parmi tous les modèles testés, on retient celui qui a le C_p de Mallows le plus faible.

iv) Dans le même esprit que le critère précédent, on peut évaluer la qualité du "modèle candidat" à p variables en calculant le critère AIC (Akaike Information Criterion), défini par

$$AIC = n \log \frac{SCR}{n} + 2(p+1),$$

où SCR est la SCR dans le modèle testé.

Ce critère est très proche du C_p de Mallows : la différence est qu'au lieu d'utiliser la distance quadratique $\mathbb{E}(\|Y - X\hat{\beta}\|^2)$ pour mesurer la qualité du modèle, il utilise la distance de Kullback. On retient au final le modèle ayant le plus petit AIC.

v) Le critère BIC (Bayesian Information Criterion) est motivé différemment mais conduit à un score relativement proche du précédent, à ceci près que la pénalité associée à la taille p du modèle est plus important ($log(n)$ au lieu de 2) :

$$BIC = n \log \frac{SCR}{n} + (p+1) \log n.$$

On retient au final le modèle ayant le plus petit BIC.

2.4.2 Lien entre les critères

Lors de la sélection de variables dans un modèle de régression linéaire, les critères précédents s'ordonnent de la manière suivante en fonction de leur propension à sélectionner le modèle le plus parcimonieux (celui ayant le moins de variables) :

$$BIC < F\,test < C_p \approx AIC < R_a^2$$

Le critère BIC est donc celui qui aura tendance à retenir les plus petits modèles. Voir TD pour une justification.

2.4.3 Aspects théoriques

Supposons que le vrai modèle (inconnu) appartienne aux sous-modèles testés. Sous des hypothèses standards, on a les résultats asymptotiques suivants.

Pour le critère BIC :

La probabilité qu'il sélectionne un modèle plus petit que le vrai modèle tend vers 0 lorsque $n \to \infty$.

La probabilité qu'il sélectionne un modèle plus gros que le vrai modèle tend vers 0 lorsque $n \to \infty$.

La probabilité qu'il sélectionne le bon modèle tend vers 1 lorsque $n \to \infty$.

Pour les autres critères (C_p, AIC, R_a^2) :

La probabilité qu'ils sélectionnent un modèle plus petit que le vrai modèle tend vers 0 lorsque $n \to \infty$.

La probabilité qu'ils sélectionnent un modèle plus gros que le vrai modèle **ne tend pas** vers 0 lorsque $n \to \infty$.

La probabilité qu'ils sélectionnent le bon modèle **ne tend pas** vers 1 lorsque $n \to \infty$.

2.4.4 Algorithme de sélection automatique

Si on dispose de $p_{\max}$ variables, il y a $2^{p_{\max}}$ modèles possibles (ex : pour $p_{\max} = 10$, 1024 modèles possibles).

- Si $p_{\max}$ n'est pas trop grand, on peut effectuer une procédure de sélection automatique exhaustive.

Sous R : fonction `regsubsets` de la librairie `leaps`

Cette fonction renvoie le meilleur modèle à 1 variable, à 2 variables, ..., et à $p_{\max}$ variables. Il n'y a pas d'ambiguité sur la notion de meilleur ici car à nombre de variables fixé, le meilleur modèle est celui qui minimise la SCR (tous les critères précédents sont d'accord là-dessus). La comparaison finale entre tous ces "meilleurs" modèles se fait finalement avec le critère de notre choix (parmi ceux exposés ci-dessus), qui accorde plus ou moins d'importance au nombre de variables.

- Si $p_{\max}$ est trop grand pour effectuer une rechercher exhaustive, on peut utiliser une procédure "pas à pas" (procédure stepwise), selon le critère de notre choix (BIC par exemple). Il en existe plusieurs :

- Procédure stepwise backward : on part du plus gros modèle contenant $p_{\max}$ variables et on élimine la variable la moins significative (au sens où son retrait optimise le critère choisi, par exemple conduit au BIC le plus faible). On élimine ainsi successivement les variables les unes après les autres, jusqu'à ce que plus aucun retrait n'améliore le modèle (chaque retrait détériore le critère choisi).
- Procédure stepwise forward : on part du plus petit modèle (celui ne contenant que la constante) et on ajoute la meilleure variable (au sens où son ajout optimise le critère choisi). On ajoute ainsi successivement les variables jusqu'à ce que plus aucun ajout n'améliore le modèle.
- Procédure stepwise backward hybride : idem que la procédure backward, sauf qu'à chaque étape, on tente d'éliminer une variable du modèle mais aussi d'ajouter une variable éliminée précédemment (on choisit l'opération la plus bénéfique au sens du critère choisi).

- Procédure stepwise forward hybride : idem que la procédure forward, sauf qu'à chaque étape, on tente d'ajouter une variable au modèle mais aussi d'éliminer une variable ajoutée précédemment (on choisit l'opération la plus bénéfique au sens du critère choisi).

Sous R : fonction `step` avec l'option `direction` égale à ``backward'' ou ``forward'' ou ``both''.

Chapitre 3

Analyse de la variance (ANOVA) et de la covariance (ANCOVA)

Dans le chapitre précédent (Régression linéaire), on a considéré que :

- la variable à expliquer Y est une variable quantitative
- les variables explicatives X_j $(j = 1, \ldots, p)$ sont des variables quantitatives.

Dans ce chapitre, on suppose toujours que Y est une variable quantitative mais les variables explicatives peuvent être qualitatives et/ou quantitatives.

- Si toutes les variables explicatives X_j $(j = 1, \ldots, p)$ sont qualitatives, on parle d'ANOVA (analyse de la variance)
- Si les variables explicatives mêlent à la fois des variables quantitatives et des variables quantitatives, on parle d'ANCOVA (analyse de la covariance).

La première partie explique l'ANOVA à un facteur : il s'agit de la situation où il n'y a qu'une seule variable explicative, cette dernière étant qualitative. La seconde partie montre comment cette approche s'étend à plusieurs facteurs en considérant l'ANOVA à deux facteurs (la généralisation à plus de deux facteurs s'en déduit facilement). Enfin le cas général de l'ANCOVA est un mélange de la régression linéaire et de l'ANOVA : sa présentation fera l'objet d'une activité en ligne.

3.1 Analyse de la variance à 1 facteur

3.1.1 Ecriture du modèle

Notations :

Y : variable à expliquer observée auprès de n individus

A : variable explicative qualitative (on dit aussi "facteur") composée de I modalités notées $A_1, \dots, A_I$, observée sur les mêmes individus.

n_i : effectif dans la modalité A_i, $i = 1, \dots, I$.

$y_{i,j}$: valeur de Y pour l'individu j appartenant à la modalité A_i, pour $i = 1, \dots, I$ et $j = 1, \dots, n_i$.

y_k : valeur de Y pour l'individu k, $k = 1, \dots, n$ (cette notation ne tient pas compte de l'appartenance de l'individu à sa modalité pour A).

μ_i : espérance de Y dans la classe A_i, i.e. $\mu_i = \mathbb{E}(Y|A_i)$.

Question : le facteur A a-t-il une influence sur Y ? Plus précisément a-t-on $\mu_1 = \cdots = \mu_I$?

Le modèle de base s'écrit, pour tout $i = 1, \dots, I$ et $j = 1, \dots, n_i$:

$$y_{i,j} = \mu_i + \epsilon_{i,j}$$

où $\epsilon_{i,j}$ sont des variables Gaussiennes i.i.d suivant une $\mathcal{N}(0, \sigma^2)$. On suppose donc que le comportement de Y dans chaque modalité est Gaussien, qu'il varie autour d'une moyenne μ_i propre à la modalité, et que les variations autour de cette moyenne sont similaires quelle que soit la modalité (la variance σ^2 est commune à tous). En utilisant la notation y_k au lieu de $y_{i,j}$, ce modèle s'écrit également, pour tout $k = 1, \dots, n$:

$$y_k = \sum_{i=1}^{I} \mu_i \mathbb{1}_{A_i}(k) + \epsilon_k$$

où les ϵ_k sont iid suivant une $\mathcal{N}(0, \sigma^2)$ et où $\mathbb{1}_{A_i}$ est la variable indicatrice dont chaque entrée vaut 1 ou 0 selon que l'individu appartient à A_i. De façon matricielle, le modèle s'écrit donc

$$Y = X\mu + \epsilon \tag{3.1}$$

où $\mu = (\mu_1, \dots, \mu_I)'$ et X est la matrice de taille (n, I) $X = [\mathbb{1}_{A_1} \dots \mathbb{1}_{A_I}]$ ne contenant que des 0 et des 1. Il s'agit d'un modèle de régression linéaire standard dans lequel toutes les variables sont quantitatives.

Le modèle général contenant une constante s'écrit, pour tout $k = 1, \dots, n$

$$y_k = m + \sum_{i=1}^{I} \alpha_i \mathbb{1}_{A_i}(k) + \epsilon_k$$

où les ϵ_k sont iid suivant une $\mathcal{N}(0, \sigma^2)$. Ce modèle n'est pas de plein rang car la variable $\mathbb{1}$ associée à la constante est une combinaison linéaire des autres variables : $\mathbb{1} = \sum_{i=1}^{I} \mathbb{1}_{A_i}$. Il faut donc ajouter une contrainte pour le rendre identifiable.

Exemples de contraintes

- $m = 0$: on retrouve alors le modèle initial sans constante. Les paramètres α_i s'identifient alors avec les μ_i ($\alpha_i = \mu_i$) car ils correspondent bien à $E(Y|A_i)$. Sous R, on peut imposer cette contrainte avec la commande `lm(Y~A-1)`.
- $\alpha_1 = 0$: dans ce cas l'interprétation des coefficients est différente : $m = \mu_1$ et $\alpha_i = \mu_i - \mu_1$ pour tout $i = 2, \dots, I$. Sous R, il s'agit de la contrainte par défaut choisie par la commande `lm(Y~A)`.

Proposition 3.1.1. *Dans le modèle précédent, quelle que soit la contrainte linéaire choisie, l'estimation par MCO conduit à*

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{i,j} = \bar{y_i}, \quad i = 1, \dots, I$$

et

$$\hat{\sigma}^2 = \frac{1}{n - I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{i,j} - \bar{y_i})^2.$$

Démonstration. Cf cours □

3.1.2 Significativité du facteur

On rappelle que d'après la formule d'analyse de la variance (cf début du cours) :

$$S_T^2 = S_{inter}^2 + S_{intra}^2$$

où $S_T^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{i,j} - \bar{y})^2$, $S_{inter}^2 = \sum_{i=1}^{I} n_i (\bar{y_i} - \bar{y})^2$ et $S_{intra}^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{i,j} - \bar{y_i})^2$.

On souhaite tester $H_0 : \mu_1 = \cdots = \mu_I$.

Proposition 3.1.2. *Si $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, alors sous $H_0 : \mu_1 = \cdots = \mu_I$,*

$$F = \frac{S^2_{inter}/(I-1)}{S^2_{intra}/(n-I)} \sim F(I-1, n-I)$$

d'où la région critique au niveau α :

$$RC_\alpha = \{F > f_{I-1,n-I}(1-\alpha)\}$$

où $f_{I-1,n-I}(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ d'une $F(I-1, n-I)$.

Démonstration. Il s'agit d'effectuer un test de contraintes linéaires sur les paramètres dans le modèle (3.1), comme on l'a vu dans le chapitre précédent. Il y a $I-1$ contraintes à tester ($\mu_1 = \mu_2$,..., $\mu_1 = \mu_I$) sur le paramètre μ de taille I. La statistique de test s'écrit donc

$$F = \frac{n-I}{I-1} \frac{SCR_c - SCR}{SCR}$$

et on montre qu'elle correspond exactement à la quantité de la proposition. □

Sous R, le test précédent s'effectue à l'aide de la commande `anova(lm(Y∼A))` ou `aov(Y∼A)` suivi du `summary` du résultat. La sortie se présente sous la forme du tableau suivant

	dll	SC	mean SC	F	p-value
Facteur A	$I-1$	S^2_{inter}	$S^2_{inter}/(I-1)$	$\frac{S^2_{inter}/(I-1)}{S^2_{intra}/(n-I)}$	...
Résidus	$n-I$	S^2_{intra}	$S^2_{intra}/(n-I)$		

Remarque 3.1.3. *Le test précédent est vraiment ce qui nous intéresse concernant le lien entre Y et le facteur A. Il est assez rare de s'intéresser de près à la sortie du modèle* (3.1) *ou à ses variantes selon les contraintes choisies (en particulier le choix par défaut `lm(Y∼A)`). En effet la signification des coefficients dans ce modèle dépend de la contrainte choisie et les tests de significativité de Student ($H_0 : \alpha_i = 0$) n'ont pas forcément d'intérêt. A l'inverse, le test ANOVA précédent ne dépend pas de la contrainte choisie et répond à la question initiale, c'est à dire tester $H_0 : \mu_1 = \cdots = \mu_I$.*

Remarque 3.1.4. *Le test ANOVA est valide sous l'hypothèse* $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. *Comme d'habitude, le caractère Gaussien n'est pas restrictif pourvu que* n *soit grand. La vraie hypothèse contraignante ici est l'homoscédasticité. En particulier, cela implique que la variance de* Y *doit être la même dans chaque modalité* A_i*, ce qui n'est pas toujours une hypothèse raisonnable. Cette égalité des variances peut par exemple se tester avec le test de Bartlett ou de Levene (*`bartlett.test` *ou* `leveneTest` *de la librairie* `car` *sous R). A défaut d'égalité des variances, on peut notamment envisager une transformation de* Y *pour stabiliser la variance (par exemple étudier* $ln(Y)$ *au lieu de* Y*).*

3.1.3 Analyse post-hoc

Si d'après le test précédent le facteur A est significatif, on cherche souvent à savoir quelle(s) modalité(s) diffère(nt) des autres. Pour cela, on désire effectuer tous les tests

$$H_0^{i,j} : \mu_i = \mu_j \quad \text{versus} \quad H_1^{i,j} : \mu_i \neq \mu_j$$

pour tout $i \neq j$ dans $\{1, \dots, I\}$, ce qui correspond à $I(I-1)/2$ tests.

Pour i et j fixés, on peut tester $H_0^{i,j}$ au niveau α par un test de Student d'égalité des moyennes, dont la région critique au niveau α est

$$RC_\alpha = \left\{ \frac{|\bar{y}_i - \bar{y}_j|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} > t_{n-I}(1-\alpha/2) \right\} \tag{3.2}$$

où $t_{n-I}(1-\alpha/2)$ désigne le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-I$ degrés de liberté. Ce test garantit que la probabilité de l'erreur de première espèce vaut α, c'est à dire que $\mathbb{P}_{\mu_i=\mu_j}(H_1^{i,j}) = \alpha$.

Question : Si on fait tous les tests précédents, quelle est la probabilité d'annoncer $\mu_i \neq \mu_j$ pour un certain couple (i,j) alors que $\mu_1 = \cdots = \mu_I$? Autrement dit, quelle est la probabilité de détecter au moins une différence entre modalités, alors qu'il n'y en a aucune.

Cette probabilité est

$$\mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\text{conclure } H_1^{i,j} \text{ pour au moins un couple } (i,j)\right) = \mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\bigcup_{(i,j)} H_1^{i,j}\right).$$

Elle est en générale beaucoup plus grande que α.

Exemple : Si tous les tests sont indépendants entre eux, alors

$$\begin{aligned}\mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\bigcup_{(i,j)} H_1^{i,j}\right) &= 1-\mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\bigcap_{(i,j)} H_0^{i,j}\right)\\ &= 1-\prod_{(i,j)}\mathbb{P}_{\mu_1=\cdots=\mu_I}\left(H_0^{i,j}\right)\\ &= 1-\prod_{(i,j)}(1-\alpha) = 1-(1-\alpha)^{I(I-1)/2}\end{aligned}$$

Cette probabilité vaut pratiquement 1 dès que I est grand, autrement dit on est quasiment certain d'annoncer à tort $\mu_i \neq \mu_j$ pour un certain couple (i,j) même si en réalité $\mu_1 = \cdots = \mu_I$.

Il s'agit d'un problème bien connu des tests multiples : à force de chercher, on trouve toujours des faux positifs ! Pour corriger ce problème, il faut apporter une correction aux multiples tests précédents.

Solution 1 : *Correction de Bonferroni.* Au lieu d'effectuer chaque test au niveau α, on les effectue au niveau $\alpha/(I(I-1)/2)$ (i.e. on divise α par le nombre de tests effectués). Cela garantit que

$$\mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\bigcup_{(i,j)} H_1^{i,j}\right) \leq \alpha.$$

En effet la probabilité d'une union est toujours inférieure à la somme des probabilités, donc $\mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\bigcup_{(i,j)} H_1^{i,j}\right) \leq \sum_{(i,j)} \mathbb{P}_{\mu_1=\cdots=\mu_I}\left(H_1^{i,j}\right)$. Chacune de ces probabilités vaut $\alpha/(I(I-1)/2)$ si le choix de Bonferroni a été opéré, et il y a $I(I-1)/2$ termes dans la somme, d'où le résultat.

L'avantage de la correction de Bonferroni est qu'elle est facilement applicable et qu'elle est toujours valable, sans aucune hypothèse. Le défaut est que le niveau de chaque test peut devenir tellement petit (si I est grand), qu'aucune détection n'a lieu. Autrement dit, la probabilité de conclure à un faux positif est bien contrôlé par α, mais au risque qu'aucun vrai positif ne soit détecté.

Solution 2 : *Correction de Benjamin Hochberg.* Il s'agit d'une procédure plus puissante que celle de Bonferroni pour gérer les problèmes de tests multiples, mais qui repose sur quelques hypothèses. Elle est très populaire. Cf l'an prochain pour une présentation.

Solution 3 : *Test de Tukey.* Contrairement aux deux précédentes solutions, ce test n'est pas une solution générale à la problématique des tests multiples. Il s'agit d'un test qui répond à la problématique de l'analyse post-hoc de l'ANOVA.

Au lieu d'utiliser la statistique de Student dans les régions critiques (3.2), on s'appuie sur la statistique

$$Q = \sqrt{2} \max_{(i,j)} \frac{|\bar{y}_i - \bar{y}_j|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}.$$

Sous $H_0 : \mu_1 = \cdots = \mu_I$ et en supposant $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, cette statistique suit la loi particulière $Q_{I,n-I}$ appelée loi de Tuckey (ou Studentized range distribution) à $(I, n-I)$ degrés de liberté. Pour être précis, la loi est exactement la loi de Tukey si tous les n_i sont égaux et est approximativement la loi de Tukey sinon.

Pour tester chaque $H_0^{i,j}$, on utilise alors les régions critiques

$$RC_\alpha = \left\{ |\bar{y}_i - \bar{y}_j| > \frac{\hat{\sigma}}{\sqrt{2}}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\, Q_{I,n-I}(1-\alpha) \right\},$$

où $Q_{I,n-I}(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ d'une loi $Q_{I,n-I}$. L'utilisation de ces régions critiques assure un niveau *simultané* de première espèce ("family-wise error rate" en anglais) α, au sens où la probabilité de faux positif est inférieur à α. En effet

$$\begin{aligned}
&\mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\bigcup_{(i,j)} H_1^{i,j}\right) \\
&\qquad = \mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\exists (i,j),\ \sqrt{2}\frac{|\bar{y}_i - \bar{y}_j|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} > Q_{I,n-I}(1-\alpha)\right) \\
&\qquad = \mathbb{P}_{\mu_1=\cdots=\mu_I}\left(\max_{(i,j)} \sqrt{2}\frac{|\bar{y}_i - \bar{y}_j|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} > Q_{I,n-I}(1-\alpha)\right) \\
&\qquad = \mathbb{P}_{\mu_1=\cdots=\mu_I}\left(Q > Q_{I,n-I}(1-\alpha)\right) \\
&\qquad = \alpha.
\end{aligned}$$

En utilisant la même statistique, on peut construire de façon similaire des intervalles de confiance de niveau *simultané* $1-\alpha$ pour les différences $\mu_i - \mu_j$:

$$IC_{1-\alpha}(\mu_i - \mu_j) = \left[\bar{y}_i - \bar{y}_j \pm \frac{\hat{\sigma}}{\sqrt{2}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\, Q_{I,n-I}(1-\alpha) \right].$$

Le niveau $1-\alpha$ est simultané dans le sens où (notez la présence du $\forall$)

$$\mathbb{P}(\forall (i,j),\ \mu_i - \mu_j \in IC_{1-\alpha}(\mu_i - \mu_j)) = 1 - \alpha.$$

Donc avec grande probabilité *toutes* les différences appartiennent aux intervalles de confiance.

En pratique, on utilise donc le test de Tukey pour identifier les couples (i,j) de modalités ayant des moyennes significativement différentes, i.e. dont l'IC précédent ne contient pas 0.

Sous R : fonction `TukeyHSD` puis `plot` du résultat.

3.2 Analyse de la variance à 2 facteurs

3.2.1 Modèle

Notations :

Y : variable à expliquer observée auprès de n individus

A : variable explicative qualitative (on dit aussi "facteur") composée de I modalités notées $A_1, \ldots, A_I$, observée sur les mêmes individus.

B : variable explicative qualitative (on dit aussi "facteur") composée de J modalités notées $B_1, \ldots, B_J$, observée sur les mêmes individus.

n_{ij} : effectif dans la modalité $A_i \cap B_j$, $i = 1, \ldots, I$, $j = 1, \ldots, J$.

$n_{i.}$: effectif dans A_i. On a $n_{i.} = \sum_{j=1}^{J} n_{ij}$

$n_{.j}$: effectif dans B_j. On a $n_{.j} = \sum_{i=1}^{I} n_{ij}$

y_{ijk} : valeur de Y pour l'individu k appartenant à la modalité A_i et à la modalité B_j, pour $i = 1, \ldots, I$, $j = 1, \ldots, J$ et $k = 1, \ldots, n_{ij}$.

y_k : valeur de Y pour l'individu k, $k = 1, \ldots, n$ (cette notation ne tient pas compte de l'appartenance de l'individu aux modalités de A et B).

Le modèle général liant le comportement de la variable Y en fonction des 2 facteurs A et B s'écrit :

$$y_{ijk} = m + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \tag{3.3}$$

pour tout $i = 1, \dots, I$, $j = 1, \dots, J$ et $k = 1, \dots, n_{ij}$. Dans cette écriture, les ϵ_{ijk} sont iid suivant une $\mathcal{N}(0, \sigma^2)$ et les $1 + I + J + IJ$ paramètres sont m, les α_i, les β_j et les γ_{ij}. Le paramètre m peut-être vu comme l'effet moyen de Y (sans tenir compte de A et B), le paramètre α_i comme l'effet dû à A, le paramètre β_j comme l'effet dû à B, et le paramètre γ_{ij} comme l'effet dû à l'interaction entre A et B.

On peut récrire ce modèle sous la forme plus standard d'un modèle de régression linéaire en introduisant des variables indicatrices : pour tout $k = 1, \dots, n$,

$$y_k = m + \sum_{i=1}^{I} \alpha_i \mathbb{1}_{A_i}(k) + \sum_{j=1}^{J} \beta_j \mathbb{1}_{B_j}(k) + \sum_{i=1}^{I} \sum_{j=1}^{J} \gamma_{ij} \mathbb{1}_{A_i \cap B_j}(k) + \epsilon_k$$

où les ϵ_k sont iid suivant une $\mathcal{N}(0, \sigma^2)$.

Il y a plusieurs problèmes de colinéarité dans ce modèle. En fait, le problème initial de l'ANOVA à 2 facteurs fait intervenir $I \times J$ inconnues (les espérances de Y dans les $A_i \cap B_j$), or le modèle précédent contient $1+I+J+IJ$ paramètres. Il faut donc $1 + I + J$ contraintes pour rendre le modèle identifiable. Une autre manière d'identifier ce problème de colinéarité est de déterminer le noyau de la matrice X contenant les $1 + I + J + IJ$ variables présentes dans le modèle précédent. On se rend compte que ce noyau est de dimension $1 + I + J$ donc le rang de X vaut IJ. Il faut donc bien $1 + I + J$ contraintes pour rendre X de plein rang. Une infinité de choix sont possibles.

Exemple : la commande `lm(Y~A+B)` sous R fixe par défaut les contraintes $\alpha_1 = 0$, $\beta_1 = 0$, $\gamma_{1j} = 0$ pour tout $j = 1, \dots, J$, et $\gamma_{i1} = 0$ pour tout $i = 1, \dots, I$. Cela fait bien $1 + I + J$ (la contrainte $\gamma_{11} = 0$ apparaît deux fois).

Proposition 3.2.1. *Dans le modèle précédent, quelle que soit les contraintes linéaires choisies, l'estimation par MCO conduit à la prévision, pour tout $i = 1, \dots, I$, $j = 1, \dots, J$ et $k = 1, \dots, n_{ij}$,*

$$\hat{y}_{ijk} = \bar{y}_{ij}$$

où $\bar{y}_{ij}$ désigne la moyenne empirique dans la modalité croisée $A_i \cap B_j$ ($\bar{y}_{ij} = n_{ij}^{-1} \sum_{k=1}^{n_{ij}} y_{ijk}$), et à l'estimation de la variance résiduelle

$$\hat{\sigma}^2 = \frac{1}{n - IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2.$$

3.2.2 Tests

On souhaite tester si les effets marginaux dus à A et à B, et si l'effet d'interaction entre A et B sont significatifs. Pour cela on part du modèle complet (3.3) et on commence par tester la présence de l'effet d'interaction : A-t-on $\gamma_{ij} = 0$ pour tout i, j ? Dans ce dernier cas, on dit que le modèle est additif car (3.3) devient $y_{ijk} = m + \alpha_i + \beta_j + \epsilon_{ijk}$.

Analyse graphique : La présence d'interaction peut se détecter graphiquement grâce aux "interaction plots" : on représente la moyenne de Y par modalités croisées en plaçant un facteur en abscisse et l'autre en ordonnée. Par exemple la figure 3.1 représente l'évolution des moyennes de Y selon les 5 modalités de A (en abscisse) et les 5 modalités de B (en ordonnée). Chaque courbe représente donc les moyennes de Y associées à une modalité de B : elles évoluent selon qu'on parcourt les modalités de A, celle de B étant fixée. On peut de la même manière inverser le rôle joué par A et B pour mettre B en abscisse, cf la figure 3.2. Sous R : `interaction.plot(A,B,Y)` met le facteur A en abcisse.

S'il n'y a pas d'interaction, les courbes doivent être plus ou moins parallèles entre elles, comme justifié ci-dessous. C'est la situation que l'on observe dans les figures 3.1 et 3.2. Inversement, la situation de la figure 3.3 montre une interaction entre A et B.

Pourquoi les courbes sont-elles parallèles en absence d'interaction ? La courbe j est composée des valeurs $\bar{y}_{ij}$ pour i variant de 1 à I. Si les courbes sont parallèles cela signifie qu'il existe, pour tout $j = 1, \dots, J$, un coefficient λ_j tel que pour tout $i = 1, \dots, I$,

$$\bar{y}_{ij} = \bar{y}_{i1} + \lambda_j,$$

autrement dit la courbe j est à un facteur λ_j de la courbe 1. En moyennant sur tous les i, cela implique

$$\bar{y}_{.j} = \bar{y}_{.1} + \lambda_j$$

en notant $\bar{y}_{.j}$ la moyenne empirique de Y dans la modalité B_j. On en déduit $\lambda_j = \bar{y}_{.j} - \bar{y}_{.1}$ et donc

$$\bar{y}_{ij} = \bar{y}_{i1} + \bar{y}_{.j} - \bar{y}_{.1}.$$

Le premier terme ne dépend que de i : il correspond à l'effet α_i dans (3.3), le second ne dépend que de j : il correspond à l'effet β_j, et le dernier est

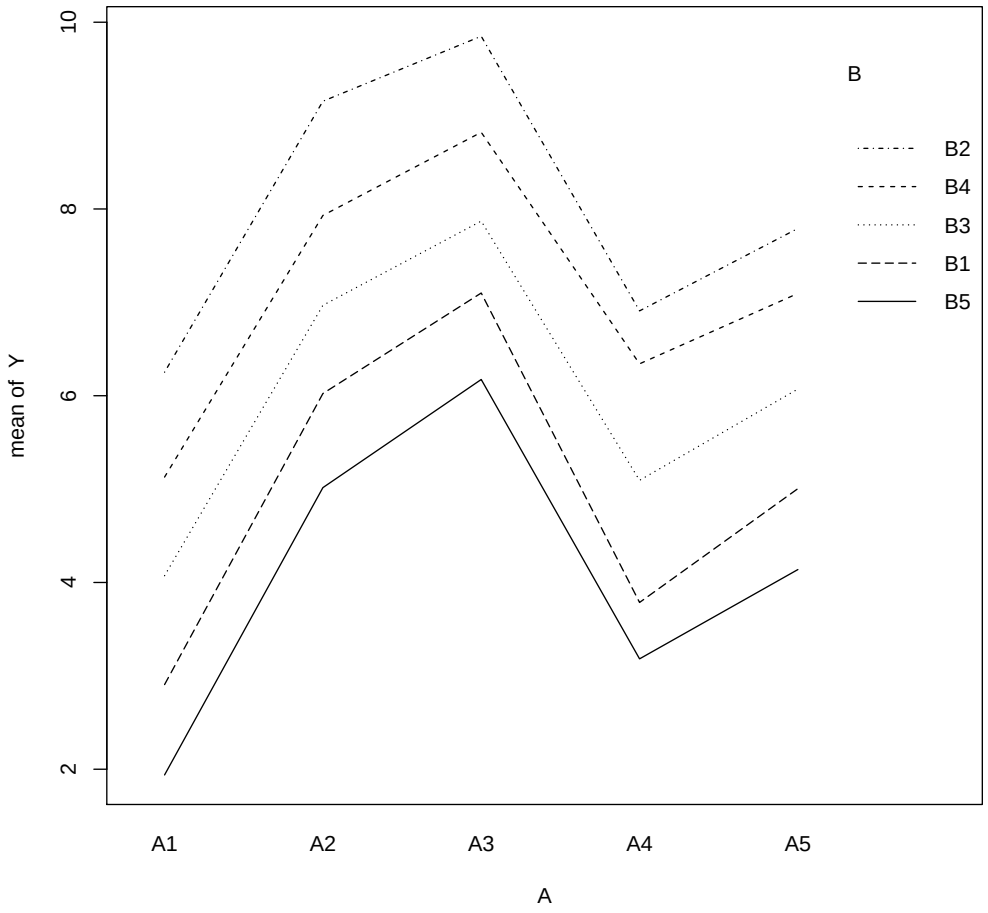

FIGURE 3.1 – Interaction plot : évolution des moyennes de Y selon les modalités croisées de A (en abscisse) et de B (en ordonnée). Les courbes sont à peu près parallèles, ce qui témoigne de l'absence d'interaction entre A et B.

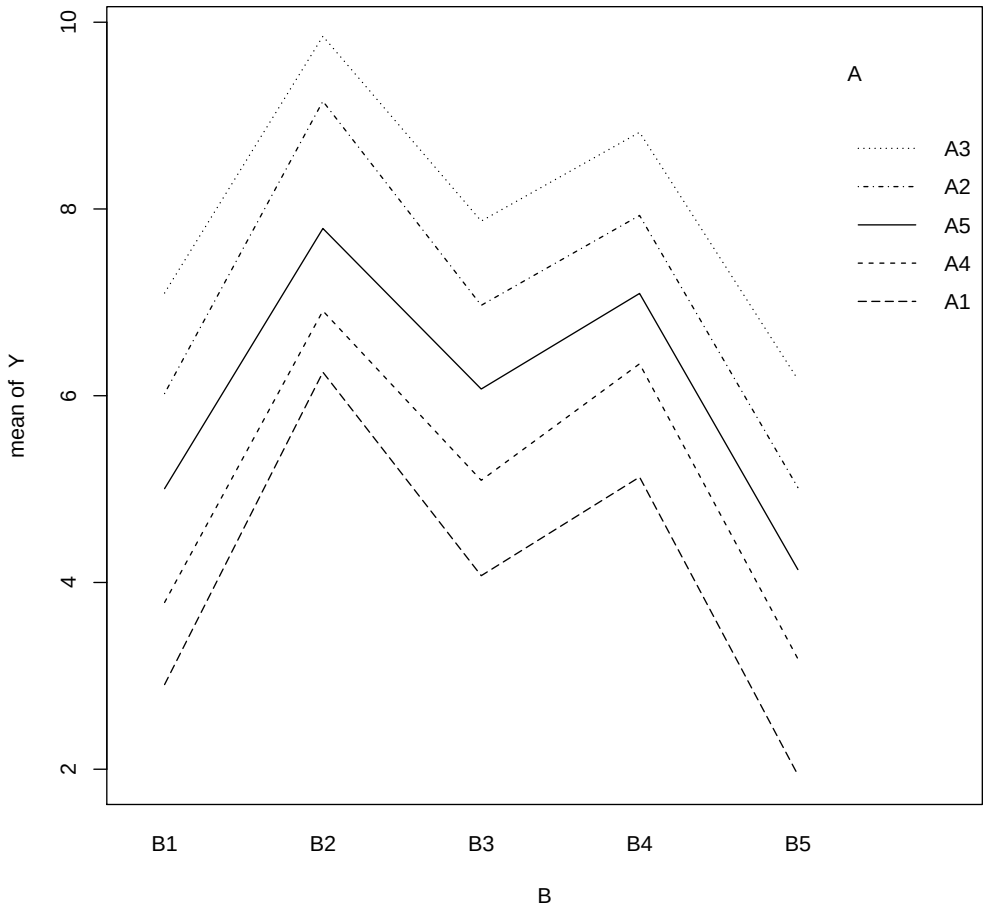

FIGURE 3.2 – Interaction plot : évolution des moyennes de Y selon les modalités croisées de B (en abscisse) et de A (en ordonnée). Il s'agit des mêmes données que dans la figure 3.1.

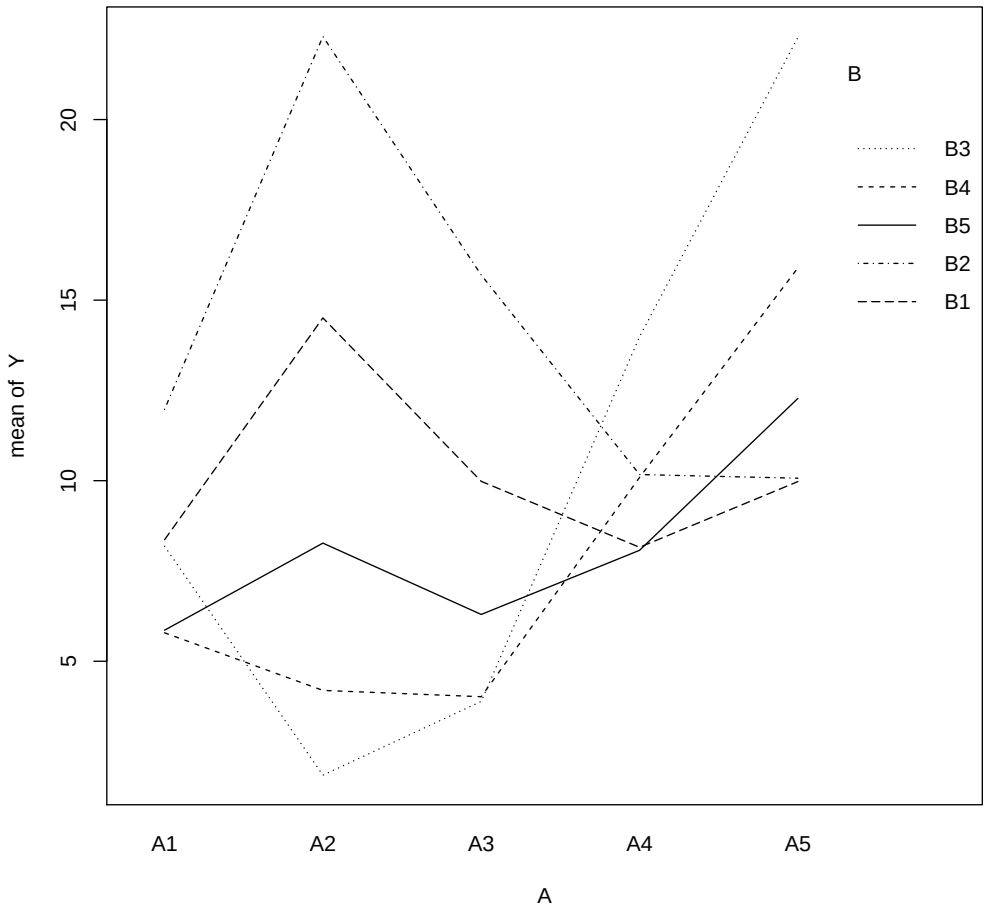

FIGURE 3.3 – Interaction plot, cas d'une présence d'interaction entre A et B : contrairement aux deux figures précédentes, les courbes ne sont pas parallèles.

constant : il correspond à m. Ainsi dans la relation (3.3), $\gamma_{ij} = 0$. Il n'y a donc pas d'interaction.

Tests ANOVA : On suppose dans la suite que "le plan est équilibré", c'est à dire que les effectifs sont les mêmes dans chaque modalité croisée. Cette hypothèse implique donc $n_{ij} = n/IJ$. Lorsque ce n'est pas le cas, il existe des adaptations à ce qui est présenté ci-dessous, mais les détails sont omis.

Sous l'hypothèse précédente, on a la formule d'analyse de la variance :

$$S_T^2 = S_A^2 + S_B^2 + S_{AB}^2 + S_R^2$$

où

- $S_T^2 = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y})^2$,
- $S_A^2 = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (\bar{y}_{i.} - \bar{y})^2$ est l'équivalent de S_{inter}^2 dans le cas de l'ANOVA à 1 facteur dont le facteur est A,
- $S_B^2 = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (\bar{y}_{.j} - \bar{y})^2$ est l'équivalent de S_{inter}^2 dans le cas de l'ANOVA à 1 facteur dont le facteur est A,
- $S_{AB}^2 = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2$
- $S_R^2 = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2$ est l'équivalent de S_{intra}^2 dans l'ANOVA à 1 facteur.

Partant de cette formule d'analyse de la variance, on peut construire différents tests de Fisher de significativité des effets de A, de B, et de l'interaction AB, de la même manière que cela est fait pour l'ANOVA à 1 facteur. On commence par tester la présence de l'interaction :

$$H_0^{(AB)} \ : \ \gamma_{ij} = 0 \text{ pour tout } i, j.$$

Pour cela on utilise la statistique

$$F^{(AB)} = \frac{S_{AB}^2/(I-1)(J-1)}{S_R^2/(n-IJ)}$$

qui suit sous H_0 une loi $F((I-1)(J-1), n-IJ)$. On en déduit la région critique au niveau α

$$RC_\alpha^{(AB)} = \{F^{(AB)} > f_{(I-1)(J-1),n-IJ}(1-\alpha)\}.$$

Si $H_0^{(AB)}$ est accepté, cela signifie que le modèle est additif et on peut tester si l'effet dû à A et à B est significatif, exactement comme dans l'ANOVA à 1 facteur. Pour tester

$$H_0^{(A)} \ : \ \alpha_i = 0 \text{ pour tout } i,$$

on utilise la statistique

$$F^{(A)} = \frac{S_A^2/(I-1)}{S_R^2/(n-IJ)}$$

qui suit sous $H_0^{(A)}$ une loi $F(I-1, n-IJ)$. Et pour tester

$$H_0^{(B)} \ : \ \beta_j = 0 \text{ pour tout } j,$$

on utilise la statistique

$$F^{(B)} = \frac{S_B^2/(J-1)}{S_R^2/(n-IJ)}$$

qui suit sous $H_0^{(B)}$ une loi $F(J-1, n-IJ)$.

Tous ces tests sont résumés dans un tableau sous R suite à la commande `anova(lm(Y~ A+B))` ou `aov(Y~ A+B)` suivi de `summary`.

	dll	SC	mean SC	F	p-value
A	$I-1$	S_A^2	$S_A^2/(I-1)$	$\frac{S_A^2/(I-1)}{S_R^2/(n-IJ)}$	...
B	$J-1$	S_B^2	$S_B^2/(J-1)$	$\frac{S_B^2/(J-1)}{S_R^2/(n-IJ)}$	...
AB	$(I-1)(J-1)$	S_{AB}^2	$S_{AB}^2/(I-1)(J-1)$	$\frac{S_{AB}^2/(I-1)(J-1)}{S_R^2/(n-IJ)}$	...
Résidus	$n-IJ$	S_R^2	$S_R^2/(n-IJ)$		

3.3 Analyse de la variance à k facteurs

L'approche de l'ANOVA à deux facteurs s'étend à la présence de plus de deux facteurs : en toute généralité, on peut supposer que l'espérance de Y dépend de chaque facteur, et des interactions 2 à 2 des facteurs, et des interactions triples, etc. Par exemple dans le cas de 3 facteurs A, B et C, on pourrait avoir en toute généralité les effets marginaux de A, B et C, les effets dus aux interactions doubles AB, AC et BC, et l'effet dû à l'interaction triple ABC. Pour k facteurs, cela représente $2^k - 1$ effets possibles.

On peut tester chacun de ces effets par un test d'analyse de la variance comme on l'a présenté ci-dessus dans le cas de 2 facteurs. Néanmoins, si k est grand, cela fait trop de tests à réaliser, et surtout l'effectif dans chaque

modalité croisée à k facteurs risque d'être très faible, de l'ordre de 0 ou 1 individu, rendant ces tests inefficaces.

On est donc amené en pratique à faire des choix sur la présence possible des interactions : on se limite par exemple aux interactions doubles sans inclure les interactions supérieures, on peut n'inclure de plus que certaines de ces interactions doubles et non toutes, voire on se limite qu'aux effets marginaux sans inclure d'interactions.

3.4 Analyse de la covariance (ANCOVA)

Il s'agit de la situation dans laquelle les variables explicatives incluent à la fois des facteurs et des variables quantitatives. Il s'agit donc d'un mélange de la régression linéaire standard telle que vue dans le chapitre précédent, et de l'ANOVA. Le modèle pourra ainsi inclure : les effets de chaque variable quantitative (via chaque coefficient de régression β_j associé), les effets des facteurs et des interactions entre les facteurs (comme a l'a vu dans les parties précédentes), mais aussi les effets des interactions entre les facteurs et les variables quantitatives. Par exemple on peut imaginer que le coefficient β_j de la variable quantitative X_j prend en réalité deux valeurs différentes selon qu'on est dans la première modalité du facteur A ou dans la seconde : il s'agit d'une interaction entre X et A.

Une présentation de l'ANCOVA ainsi que sa mise en oeuvre sous R font l'objet d'une activité à réaliser sous Madoc.

Chapitre 4

Régression logistique

Dans les chapitres précédents, la variable à expliquer Y était quantitative. On suppose dans ce chapitre que Y est une variable qualitative (facteur). On se concentrera principalement sur le cas le plus courant, qui est celui où Y a deux modalités, appelons-les 0 et 1.

Exemples :

Y : acheter un produit ou non ($Y_i = 1$ si l'individu i achète le produit, 0 sinon)

Y : être ou non sujet à risque pour une maladie particulière

Y : être ou non un spam (les individus i sont des messages électroniques)

Il existe de nombreuses méthodes dont l'objectif est d'expliquer au mieux l'occurence 0 ou 1 de Y en fonction de variables explicatives $X_1, \ldots, X_p$ (il s'agit de l'essentiel des méthodes de machine learning). La régression logistique est un des modèles les plus simples et les plus utilisés pour cela. La dernière partie expliquera comment étendre la démarche au cas où Y admet plus que 2 modalités.

4.1 Modélisation

4.1.1 Régression logistique simple

Supposons pour commencer qu'il n'y a qu'une seule variable explicative X. La variable Y prenant les valeurs 0 ou 1 (ses deux modalités possibles), le nuage de points entre Y et X ressemble au graphique de gauche dans la figure 4.1, auquel on a ajouté la droite des moindres carrés. Comme le

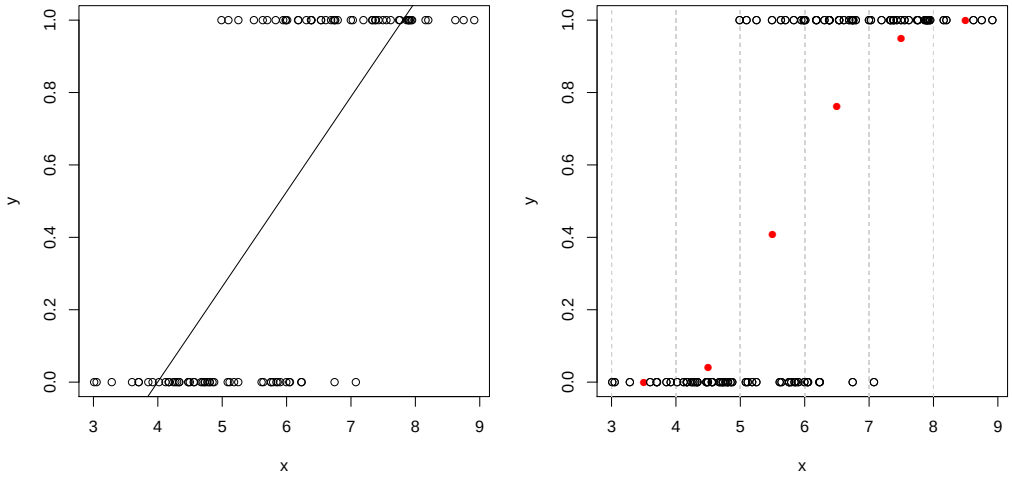

FIGURE 4.1 – Exemple de nuage de point de (x, y) lorsque $y \in \{0, 1\}$. Le graphique de gauche contient la droite des moindres carrés. Les points rouges du graphique de droite montrent la fréquence de $y = 1$ dans chaque intervalle de x délimité par les pointillés.

montre ce graphique, ajuster un modèle de régression linéaire pour expliquer Y à l'aide de X n'est clairement pas adapté.

Une idée naturelle pour exploiter ce nuage de points est de calculer les fréquences d'apparition de $Y = 1$ selon l'appartenance de X à des classes de valeurs. Par exemple, dans le graphique de droite dans la figure 4.1, chaque point rouge indique cette fréquence lorsque X appartient aux intervalles $[3; 4]$ ou $[4; 5]$ ou ... ou $[8; 9]$ (symbolisés par les pointillés verticaux). Ces points rouges sont une approximation de la probabilité que $Y = 1$ sachant que X appartient à la classe de valeurs correspondante.

Suivant cette démarche, on s'intéresse de façon plus précise à la probabilité

$$\pi(x) = \mathbb{P}(Y = 1|X = x).$$

Estimer cette fonction revient à interpoler d'une façon ou d'une autre les points rouges dans la figure 4.1 pour obtenir une valeur pour tout x. On observe que cette fonction a une forme en "S". Plusieurs fonctions en forme de "S" sont envisageables pour essayer d'interpoler les points rouges. La plus naturelle, comme cela est justifié dans la remarque ci-dessous, est la fonction

logistique (appelée également fonction sigmoïde) :

$$\pi(x) = \frac{e^{ax+b}}{1+e^{ax+b}}$$

dont il restera à estimer au mieux les paramètres a et b. Cette fonction est bien à valeurs dans $[0,1]$ et sa représentation est en forme de "S", conformément à ce que l'on souhaite pour $\pi(x)$.

Définition 4.1.1. *On appelle logit la fonction définie sur* $[0,1]$ *et à valeurs dans* $\mathbb{R}$

$$logit(x) = \ln\left(\frac{x}{1-x}\right).$$

Sa réciproque est définie sur $\mathbb{R}$ *et à valeurs dans* $[0,1]$

$$logit^{-1}(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}.$$

Avec ces notations, le modèle logistique simple consiste à supposer qu'il existe a et b tels que

$$logit(\pi(x)) = ax + b.$$

Cette forme pour $\pi(x) = \mathbb{P}(Y=1|X=x)$ n'est en générale qu'une approximation, mais elle est rigoureusement exacte dans le cas important suivant.

Remarque 4.1.2. *Si* X *est un mélange de deux lois Gaussiennes ayant même variance, i.e. :*

- *la loi de* X *sachant que* $Y=0$ *est une* $\mathcal{N}(m_0,\sigma^2)$,
- *la loi de* X *sachant que* $Y=1$ *est une* $\mathcal{N}(m_1,\sigma^2)$,

alors on peut montrer (cf TD) que

$$\mathbb{P}(Y=1|X=x) = \frac{e^{ax+b}}{1+e^{ax+b}}$$

pour un certain couple (a,b) *qui dépend de* m_0, m_1 *et* σ^2. *Autrement dit dans ce cas, la forme logistique que l'on a supposée ci-dessus est exacte. Cette situation est en fait celle de la figure 4.1 dans laquelle on a choisi* $m_0 = 5$, $m_1 = 7$ *et* $\sigma^2 = 1$.

4.1.2 Modèle de régression logistique général

En présence de p variables explicatives $X_1, \dots, X_p$, on définit

$$\pi(x_1, \dots, x_p) = \mathbb{P}(Y = 1 | X_1 = x_1, \dots, X_p = x_p)$$

ce qui, en notant x le vecteur $(x_1, \dots, x_p)$, revient à l'écriture précédente

$$\pi(x) = \mathbb{P}(Y = 1 | X = x).$$

Le modèle de régression logistique général s'écrit, pour $x \in \mathbb{R}^p$,

$$\pi(x) = \frac{e^{\beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_1 x_1 + \dots + \beta_p x_p}} = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

ou encore $logit(\pi(x)) = x'\beta$, dans lequel $\beta = (\beta_1, \dots, \beta_p)'$ est le vecteur des paramètres qu'il s'agira d'estimer.

Comme en régression linéaire, une des variables $X_1, \dots, X_p$ peut être la variable constante $\mathbb{1}$, ce qui est la situation standard. C'est par exemple le cas de la régression logistique simple présentée dans la partie précédente pour laquelle $p = 2$, $X_1 = \mathbb{1}$ et $X_2 = X$.

On observe les variables auprès de n individus. Avec les notations usuelles de la régression linéaire, on regroupe ces observations dans les vecteurs et matrices :

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{et} \quad X = (X_1 | \dots | X_p) \quad \text{où} \quad X_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1n} \end{pmatrix}, \dots, X_p = \begin{pmatrix} x_{p1} \\ \vdots \\ x_{pn} \end{pmatrix}.$$

Les y_i valent 0 ou 1 et les x_{ji}, pour $i = 1, \dots, n$, sont les réalisations de la variable X_j. La matrice X est de taille (n, p) et ne doit pas être confondue avec la variable X considérée dans la présentation précédente : la matrice X contient des réalisations de la variable X, tout comme le vecteur Y contient des réalisations de la variable Y (c'est le même abus de notation qu'en régression linéaire).

Ainsi, pour un individu i de $1, \dots, n$, les valeurs des variables explicatives pour i sont contenues dans la i-ème ligne de X et donc on a d'après le modèle logistique

$$\mathbb{P}(Y_i = 1 | X_{.i} = x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}.$$

Remarque 4.1.3. *Ce modèle reste valable en présence de variables explicatives de type facteur. Dans ce cas certaines variables X_j sont les indicatrices d'appartenir à telle ou telle modalité du facteur (exactement comme pour l'ANOVA ou l'ANCOVA).*

Interprétation des coefficients, les "odds ratio" :

Définition 4.1.4. *On appelle côte ("odds" en anglais) d'un évènement la quantité $odds = p/(1-p)$ où p est la probabilité de l'évènement.*

C'est une notion très utilisée dans les pays anglo-saxons et qu'on retrouve par exemple dans les paris sportifs. Par exemple une côte de 2 contre 1 signifie que $odds = 2/1$ d'où $p/(1-p) = 2$ et $p = 2/3$.

Dans un modèle logistique, e^{β_j} correspond au *rapport de côtes* (odds ratio) entre l'évènement $Y = 1|X_j = x$ et $Y = 1|X_j = x+1$ (vérifier ce résultat en guise d'exercice). Autrement dit e^{β_j} nous indique de quel pourcentage la côte de $Y = 1$ augmente lorsqu'on augmente la variable X_j de 1 unité.

Si la variable X_j correspond à la modalité d'un facteur, l'interprétation est très simple : e^{β_j} correspond au rapport de côtes entre l'absence ($X_j = 0$) et la présence ($X_j = 1$) de cette modalité.

Pour cette raison, les sorties d'un modèle logistique de certains logiciels se concentrent davantage sur les odds ratio e^{β_j} que sur les coefficients bruts β_j. En particulier, tester l'effet d'une variable ($\beta_j = 0\,?$) revient à tester si son odds ratio vaut 1 ou non.

4.2 Inférence du modèle

4.2.1 Estimation des coefficients

Dans le modèle logistique, on estime $\beta \in \mathbb{R}^p$ par maximum de vraisemblance. Pour tout i dans $1, \dots, n$, on a, pour tout $x \in \mathbb{R}^p$:

$$\begin{cases} \mathbb{P}(Y_i = 1 | X_{.i} = x) = \pi(x), \\ \mathbb{P}(Y_i = 0 | X_{.i} = x) = 1 - \pi(x). \end{cases}$$

Donc la variable $Y_i|X_{.i} = x$ suit une loi de Bernoulli de paramètre $\pi(x)$, i.e. $(Y_i|X_{.i} = x) \sim \mathcal{B}(\pi(x))$. Ainsi, en notant y_i l'observation de Y_i ($y_i \in \{0, 1\}$)

et $x_{.i}$ l'observation du vecteur $X_{.i}$ ($x_{.i} \in \mathbb{R}^p$), la vraisemblance basée sur l'observation de n individus indépendants est

$$V = \prod_{i=1}^{n} \pi(x_{.i})^{y_i}(1-\pi(x_{.i}))^{1-y_i}.$$

Donc la log-vraisemblance est

$$L = \sum_{i=1}^{n} y_i \ln(\pi(x_{.i})) + (1-y_i)\ln(1-\pi(x_{.i}))$$

qui, compte tenu du fait que $\pi(x) = e^{x'\beta}/(1+e^{x'\beta})$ et $1-\pi(x) = 1/(1+e^{x'\beta})$, s'écrit

$$L = \sum_{i=1}^{n} y_i x'_{.i}\beta - \ln(1+e^{x'_{.i}\beta}). \tag{4.1}$$

Son gradient par rapport à β vaut

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} x_{.i}\left(y_i - \frac{e^{x'_{.i}\beta}}{1+e^{x'_{.i}\beta}}\right).$$

Si la matrice X est de plein rang, l'équation $\frac{\partial L}{\partial \beta} = 0$ admet une solution unique $\hat{\beta}$ (voir la proposition suivante) mais celle-ci ne s'écrit pas de façon explicite. On utilise donc un algorithme d'optimisation (de type Newton-Raphson) en pratique pour obtenir le maximum de vraisemblance $\hat{\beta}$.

Calculons à présent l'information de Fisher du modèle pour n observations. En notant $\mathbb{E}_{|X}$ l'espérance conditionnelle sachant X, elle vaut

$$I(\beta) = -\mathbb{E}_{|X}\left(\frac{\partial^2 L}{\partial \beta^2}\right)$$

Or

$$\begin{aligned}
-\frac{\partial^2 L}{\partial \beta^2} &= \sum_{i=1}^{n} x_{.i}\frac{x'_{.i}e^{x'_{.i}\beta}(1+e^{x'_{.i}\beta}) - x'_{.i}e^{2x'_{.i}\beta}}{(1+e^{x'_{.i}\beta})^2} \\
&= \sum_{i=1}^{n} x_{.i}x'_{.i}\frac{e^{x'_{.i}\beta}}{(1+e^{x'_{.i}\beta})^2} \\
&= \sum_{i=1}^{n} x_{.i}x'_{.i}\pi(x_{.i})(1-\pi(x_{.i})).
\end{aligned}$$

Cette dernière expression peut s'écrire $X'WX$ où W est la matrice diagonale dont le i-ème élément sur la diagonale vaut $\pi(x_{.i})(1-\pi(x_{.i}))$, i.e.

$$W = \begin{bmatrix} \pi(x_{.1})(1-\pi(x_{.1})) & & 0 \\ & \ddots & \\ 0 & & \pi(x_{.n})(1-\pi(x_{.n})) \end{bmatrix}.$$

On obtient donc $I(\beta) = X'WX$ et la proposition suivante.

Proposition 4.2.1. *Dans le modèle de régression logistique, si X est une matrice de plein rang, alors le maximum de vraisemblance $\hat{\beta}$ existe et est unique, mais il ne peut s'exprimer de façon explicite. Sous certains hypothèses de régularité (non précisées), $\hat{\beta}$ suit approximativement la loi, lorsque $n \to \infty$,*

$$\hat{\beta} \sim \mathcal{N}(\beta, (X'WX)^{-1}).$$

Démonstration. La log-vraisemblance est une fonction strictement concave car sa matrice Hessienne $\partial^2 L/\partial\beta^2 = -X'WX$ (cf calcul précédent) est définie négative. En effet pour tout $u \in \mathbb{R}^p$ avec $u \neq 0$, en notant $v = Xu$, on a $v \neq 0$ car X est de plein rang, et on a $-u'X'WXu = -v'Wv = -\sum_{i=1}^n v_i^2 \pi(x_{.i})(1-\pi(x_{.i})) < 0$ car par ailleurs $\pi(x) \in]0,1[$ pour tout x. La log-vraisemblance admet donc un maximum unique $\hat{\beta}$. Ce maximum est solution de l'équation $\frac{\partial L}{\partial \beta}(\beta) = 0$ mais on ne connait pas sa forme explicite.

Pour le résultat de convergence en loi, on rappelle le résultat suivant (cf le cours de statistique inférentielle) : pour des observations iid suivant une loi (régulière) paramétrée par θ, l'estimateur du maximum de vraisemblance $\hat{\theta}$ vérifie $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{\mathcal{L}} N(0, I(\theta)^{-1})$ où $I(\theta)$ est l'information de Fisher associée à une observation. Autrement dit, lorsque $n \to \infty$, $\hat{\theta} \sim \mathcal{N}(\theta, I_n(\theta)^{-1})$ où $I_n(\theta) = nI(\theta)$ est l'information de Fisher du modèle pour n observations.

Le résultat de la proposition est tout à fait similaire, excepté que dans notre contexte les observations y_i ne sont pas iid : elles sont indépendantes mais elles ne suivent pas la même loi car $\mathcal{B}(\pi(x_{.i}))$ dépend de i. On admet que la convergence reste cependant valide dans ce contexte. □

Sous R : on lance l'estimation d'un modèle de régression logistique avec la commande `glm(Y~X,family='binomial')`.

4.2.2 Tests de significativité et intervalles de confiance

Comme en régression linéaire, nous souhaitons tester la significativité de chaque coefficient : $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$. On utilise pour cela la proposition précédente, qui nous indique que sous H_0 et pour $n \to \infty$, l'estimateur du maximum de vraisemblance $\hat{\beta}$ vérifie

$$\frac{\hat{\beta}_j}{\sqrt{(X'WX)^{-1}_{jj}}} \sim N(0,1).$$

Puisque W est inconnue (elle dépend de β au travers de π), on considère l'estimateur

$$\hat{W} = \begin{bmatrix} \hat{\pi}(x_{.1})(1-\hat{\pi}(x_{.1})) & & 0 \\ & \ddots & \\ 0 & & \hat{\pi}(x_{.n})(1-\hat{\pi}(x_{.n})) \end{bmatrix}$$

où $\hat{\pi}(x) = e^{x'\hat{\beta}}/(1+e^{x'\hat{\beta}})$. Par application du lemme de Slutsky, la convergence ci-dessus reste vraie en remplaçant W par $\hat{W}$, et donc on a la région critique au niveau asymptotique α :

$$RC_\alpha = \left\{ \frac{|\hat{\beta}_j|}{\sqrt{(X'\hat{W}X)^{-1}_{jj}}} > q(1-\alpha/2) \right\}$$

où $q(1-\alpha/2)$ désigne le quantile d'ordre $1-\alpha/2$ d'une $\mathcal{N}(0,1)$. Ce test, basé sur la convergence de l'estimateur vers une loi Gaussienne, s'appelle **test de Wald**.

De la même manière, on peut construire un intervalle de confiance de niveau asymptotique $1-\alpha$ pour β_j :

$$IC_{1-\alpha}(\beta_j) = \left[\hat{\beta}_j \pm q(1-\alpha/2)\sqrt{(X'\hat{W}X)^{-1}_{jj}} \right].$$

Sous R Le test de significativité de Wald est celui présenté par défaut pour chaque coefficient estimé dans le `summary` de la sortie `glm`.

4.2.3 Prévision

On souhaite prédire la probabilité que Y vale 1 pour un nouvel individu dont on connaît la valeur des variables $X_1, \dots, X_p$. Notons $x \in \mathbb{R}^p$ le vecteur des observations de ces variables. On cherche donc à prédire $\pi(x) = \mathbb{P}(Y = 1|X = x)$ pour cet individu. En supposant qu'il suive le même modèle logistique que les n individus observés, cette prévision est simplement

$$\hat{\pi}(x) = \frac{e^{x'\hat{\beta}}}{1 + e^{x'\hat{\beta}}}$$

où $\hat{\beta}$ est l'estimateur du maximum de vraisemblance calculé à partir des n individus observés.

On peut construire un intervalle de confiance de la manière suivante : puisque d'après la proposition 4.2.1 et le lemme de Slutsky $\hat{\beta} \sim \mathcal{N}(\beta, (X'\hat{W}X)^{-1})$ lorsque $n \to \infty$, on en déduit que $x'\hat{\beta} \sim \mathcal{N}(x'\beta, x'(X'\hat{W}X)^{-1}x)$ lorsque $n \to \infty$. Ainsi lorsque $n \to \infty$,

$$\mathbb{P}\left(x'\beta \in \left[x'\hat{\beta} \pm q(1-\alpha/2)\sqrt{x'(X'\hat{W}X)^{-1}x}\right]\right) \to 1 - \alpha.$$

Puisque $\pi(x) = logit^{-1}(x'\beta)$, par application de la fonction croissante $logit^{-1}$, on en déduit que

$$\mathbb{P}\Big(logit^{-1}\left(x'\hat{\beta} - q(1-\alpha/2)\sqrt{x'(X'\hat{W}X)^{-1}x}\right) \leq \pi(x) \leq logit^{-1}\left(x'\hat{\beta} + q(1-\alpha/2)\sqrt{x'(X'\hat{W}X)^{-1}x}\right)\Big) \to 1 - \alpha.$$

Ainsi un intervalle de confiance au niveau asymptotique $1 - \alpha$ pour $\pi(x)$ est

$$IC_{1-\alpha}(\pi(x)) = \Big[logit^{-1}\left(x'\hat{\beta} - q(1-\alpha/2)\sqrt{x'(X'\hat{W}X)^{-1}x}\right); logit^{-1}\left(x'\hat{\beta} + q(1-\alpha/2)\sqrt{x'(X'\hat{W}X)^{-1}x}\right)\Big]. \quad (4.2)$$

Il est à noter que cet intervalle est bien inclus dans $[0, 1]$, l'ensemble de définition de $\pi(x)$. En conséquence il n'est pas symétrique autour de $\hat{\pi}(x)$, la prévision de $\pi(x)$, contrairement aux intervalles de confiance qu'on a l'habitude de rencontrer. Il est cependant possible de proposer un intervalle de

confiance symétrique du type $[\hat{\pi}(x) \pm q(1-\alpha/2)\hat{\sigma}_{\hat{\pi}(x)}]$ où $\hat{\sigma}^2_{\hat{\pi}(x)}$ est un estimateur de la variance de $\hat{\pi}(x)$, en exploitant la normalité asymptotique de $\hat{\pi}(x)$ (que l'on peut obtenir grâce à la delta-méthode). Néanmoins les bornes de cet intervalle de confiance ne restent pas nécessairement dans $[0,1]$, c'est pourquoi le choix (4.2) est préférable.

Remarque 4.2.2. *On ne prédit pas Y mais $\mathbb{P}(Y=1|X=x)$ c'est à dire $\mathbb{E}_{|X}(Y)$. A titre de comparaison cela revient en régression linéaire à prédire $\mathbb{E}_{|X}(Y) = X\beta$ et non $Y = X\beta + \epsilon$.*

Sous R : la fonction `predict` prédit par défaut la combinaison linéaire $x'\beta$. Pour obtenir une prévision de $\pi(x)$, il faut ajouter l'option `type='response'`. Il n'y a pas d'option permettant d'obtenir un intervalle de confiance pour ces prévisions (contrairement à `predict.lm`). Par contre on peut obtenir l'estimation de l'écart type $\hat{\sigma}^2_{\hat{\pi}(x)}$ de $\hat{\pi}(x)$ en ajoutant l'option `se.fit=TRUE` (dans le cas où `type='response'`) , ce qui permet d'en déduire facilement un intervalle de confiance symétrique autour de $\hat{\pi}(x)$ (cf ci-dessus). Mieux, on peut obtenir l'écart-type d'estimation de $x'\beta$, c'est à dire $\sqrt{x'(X'\hat{W}X)^{-1}x}$ avec la même option dans le cas où `type='link'` (choix par défaut), ce qui permet d'obtenir l'intervalle (4.2) en utilisant par ailleurs la fonction `plogis` qui correspond à $logit^{-1}$.

4.3 Validation

Tout comme en régression linéaire, on cherche à définir la notion de résidus et de SCR, afin d'évaluer la qualité de la modélisation et d'effectuer des tests de comparaison de modèles. La notion suivante joue le rôle de la SCR.

Définition 4.3.1. *La déviance du modèle est*

$$D = -2L(\hat{\beta})$$

où L est la log-vraisemblance du modèle.

On remarque que $D \geq 0$ car la vraisemblance est toujours dans $[0,1]$ donc $L \leq 0$. Si on appliquait cette définition au cas d'un modèle de régression linéaire Gaussien de variance connue (disons 1), on obtiendrait $D = SCR +$ *constante* (le facteur 2 dans D est nécessaire à cette cohérence). Pour un

modèle de régression logistique, la déviance peut donc s'interpréter comme la SCR et sera utilisée comme telle dans la suite.

D'après l'expression (4.1) de L, on a

$$D = -2\left(\sum_{i=1}^{n} y_i x'_{.i}\beta - \ln(1+e^{x'_{.i}\beta})\right) = \sum_{i=1}^{n}\left(2\ln(1+e^{x'_{.i}\beta}) - 2y_i x'_{.i}\beta\right).$$

Cette écriture motive les définitions suivantes.

Définition 4.3.2. *La déviance de l'individu i vaut*

$$d_i = 2\ln(1+e^{x'_{.i}\beta}) - 2y_i x'_{.i}\beta$$

et le résidu de type déviance de l'individu i vaut

$$\pm\sqrt{d_i}$$

où le signe vaut $+$ si $y_i = 1$ et $-$ si $y_i = 0$.

Cette dernière définition est bien cohérente avec l'interprétation de D comme une somme de carrés des résidus puisque $D = \sum(\pm\sqrt{d_i})^2$.

Il existe par ailleurs d'autres types de résidus pour la régression logistique

Définition 4.3.3. *Le résidu brut est $y_i - \hat{\pi}(x_{.i})$. Le résidu de Pearson est $(y_i - \hat{\pi}(x_{.i}))/\sqrt{\hat{\pi}(x_{.i})(1-\hat{\pi}(x_{.i}))}$.*

Tout comme en régression linéaire, ces différents résidus permettent de détecter des individus atypiques.

Test de déviance de contraintes linéaires

Comme en régression linéaire, on souhaite tester q contraintes linéaires sur le paramètre β : $H_0 : R\beta = 0$ versus $H_1 : R\beta \neq 0$ où R est une matrice de contraintes de taille (q, p).

Sous H_0, on peut montrer que $D_c - D \sim \chi^2(q)$ lorsque $n \to \infty$, où D désigne la déviance du modèle sans contrainte et D_c désigne la déviance du modèle contraint. On en déduit donc la région critique au niveau asymptotique α :

$$RC_\alpha = \{D_c - D > \chi^2_q(1-\alpha)\}$$

où $\chi^2_q(1-\alpha)$ est le quantile d'ordre $1-\alpha$ d'une $\chi^2(q)$.

On peut décliner ce test comme en régression linéaire :

- Pour tester la significativité de la variable X_j : $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. Dans ce cas $q = 1$ et D_c correspond simplement à la déviance du modèle sans X_j.
- Pour tester deux modèles emboités : dans ce cas q correspond au nombre de variables en moins dans le sous-modèle par rapport au modèle global.
- Pour tester la significativité globale du modèle : H_0 : tous les β_j sauf la constante sont nuls, versus H_1 : au moins un coefficient autre que la constante n'est pas nul. Dans ce cas le modèle contraint ne contient que la constante et sa déviance s'appelle la "déviance nulle".

Choix de modèle de régression logistique

Les critères de sélection standards s'appliquent :

$$AIC = D + 2p, \qquad BIC = D + \ln(n)p.$$

Les procédures de sélection automatique de type forward, backward et leur version hybride, sont disponibles. Le critère de sélection des variables à chaque étape peut être basé sur le test de Wald de significativité, ou sur les critères précédents.

Sous R : Si `fit` désigne le modèle logistique estimé, `deviance(fit)` fournit la valeur de sa déviance D (celle-ci est également donné en bas de la sortie du `summary`). Pour comparer deux modèles emboités par le test de déviance, on peut faire `anova(fit1,fit2,test='Chisq')`. L'AIC et le BIC se calculent avec les fonctions `AIC` et `BIC`.

4.4 Classification

Le but d'une régression logistique est en général d'être capable de classer en $Y = 0$ ou $Y = 1$ un nouvel individu pour lequel on a observé la valeur $x \in \mathbb{R}^p$ de ses variables explicatives $X_1, \ldots, X_p$. On sait prédire $\pi(x) = \mathbb{P}(Y = 1|X = x)$ par $\hat{\pi}(x) = e^{x'\hat{\beta}}/(1 + e^{x'\hat{\beta}})$. Pour un seuil à choisir $s \in [0, 1]$, on utilise donc la règle de classement :

$$\begin{cases} \text{si} \quad \hat{\pi}(x) > s, \quad \hat{Y} = 1, \\ \text{si} \quad \hat{\pi}(x) < s, \quad \hat{Y} = 0. \end{cases}$$

Le choix par défaut du seuil est $s = 0.5$ mais ce choix peut être optimisé.

Pour évaluer la qualité de la classification, on procède par validation croisée : on sépare l'échantillon en deux parties, l'une est appelée *échantillon d'apprentissage* ("train") et est utilisé pour l'estimation de β, l'autre est appelée *échantillon test* et sert à former la *matrice de confusion*. Cette dernière recense le nombre d'individus bien classés de l'échantillon test.

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	VN	FP
$Y=1$	FN	VP

- VN : nombre de *vrais négatifs* : nombre d'individus ayant été classés négatifs ($\hat{Y}=0$), et étant réellement négatifs ($Y=0$)
- FN : nombre de *faux négatifs* : nombre d'individus ayant été classés négatifs ($\hat{Y}=0$), et étant en fait positifs ($Y=1$)
- FP : nombre de *faux positifs* : nombre d'individus ayant été classés positifs ($\hat{Y}=1$), et étant en fait négatifs ($Y=0$)
- VP : nombre de *vrais positifs* : nombre d'individus ayant été classés positifs ($\hat{Y}=1$), et étant réellement positifs ($Y=1$)

La classification est d'autant meilleure que cette matrice est diagonale. Différents scores permettent de quantifier cette qualité. Les deux plus utilisés sont la spécificité et la sensibilité, terminologies issues de la médecine où l'enjeu est par exemple de tester la réponse à un traitement.

Définition 4.4.1.

- *La sensibilité estime* $\mathbb{P}(\hat{Y}=1|Y=1)$ *par* $VP/(VP+FN)$.
- *La spécificité estime* $\mathbb{P}(\hat{Y}=0|Y=0)$ *par* $VN/(VN+FP)$.
- *La précision estime* $\mathbb{P}(Y=1|\hat{Y}=1)$ *par* $VP/(VP+FP)$.

L'objectif d'une bonne règle de classification est de maximiser à la fois la sensibilité et la spécificité. Mais ces deux quantités n'évoluent pas de la même manière : par exemple si on choisit un seuil $s=0$, on prédit tout le monde à $\hat{Y}=1$ et donc la sensibilité sera maximale (elle vaut 1) alors que la spécificité vaut 0 ; inversement si on choisit un seuil $s=1$, tout le monde est prédit à $\hat{Y}=0$, la sensibilité vaut 0 et la spécificité vaut 1.

Le choix optimal du seuil dépend du contexte. Il peut être plus important de privilégier une bonne sensibilité qu'une bonne spécificité (ou le contraire). Par exemple dans le cas d'une classification visant à détecter une maladie très grave, on préfère privilégier la sensibilité (ne pas passer à côté des personnes atteintes pour pouvoir les soigner rapidement), quitte à avoir quelques faux positifs.

Les critères usuels d'optimisation pour choisir le seuil s sont :

- minimiser l'erreur total FP+FN (souvent par défaut) ;
- maximiser sensibilité+spécificité ou une moyenne pondérée des deux ;
- maximiser la F-mesure, qui correspond à la moyenne harmonique entre la sensibilité et la précision.

Courbe ROC

Un outil très utilisé dans les problèmes de classification à 2 classes ($Y = 0$ ou $Y = 1$) est la courbe ROC ("Receiver Operating Characteristics"). Cette courbe représente la sensibilité en fonction de 1-spécificité lorsque le seuil s varie de 0 à 1. On parle aussi de *True positive rate* (=sensibilité) en fonction de *False positive rate* (=1-spécificité).

La figure 4.2 en montre un exemple. Lorsque $s = 0$, la sensibilité vaut 1 et la spécificité 0, donc le point associé sur la courbe ROC est $(1, 1)$. Lorsque $s = 1$, c'est l'exact contraire et le point associé est $(0, 0)$. Le point idéal serait celui pour lequel à la fois la sensibilité et la spécificité valent 1, c'est à dire le point $(0, 1)$.

La diagonale montre la courbe ROC que l'on aurait si on effectuait une classification complètement aléatoire sans tenir compte des variables explicatives : $\hat{Y} = 1$ avec proba $1 - s$ et $\hat{Y} = 0$ avec proba s. Le modèle tenant compte des variables explicatives, on s'attend à ce que ses performances soient meilleures que le pur hasard et donc que sa courbe ROC soit au-dessus de la diagonale.

Un score de qualité globale du modèle est l'AUC (area under curve) : plus elle est proche de 1 est meilleur est le modèle. La valeur AUC=0.5 correspond à la diagonale et on s'attend donc à ce que l'AUC d'un modèle vale au moins 0.5 (sinon il fait pire qu'une classification au hasard !). On trouve ainsi un autre indicateur appelé indice de Gini qui vaut simplement $2 \times AUC - 1$.

Sous R : Si `pred` désigne le vecteur des $\hat{\pi}(x)$ pour l'échantillon test et `ytest` le vecteur des vraies valeurs de Y pour cet échantillon, on obtient la matrice de confusion avec `table(pred>s,ytest)` pour un seuil s donné. Pour obtenir la courbe ROC, on peut utiliser la libraire `ROCR` : on commence par

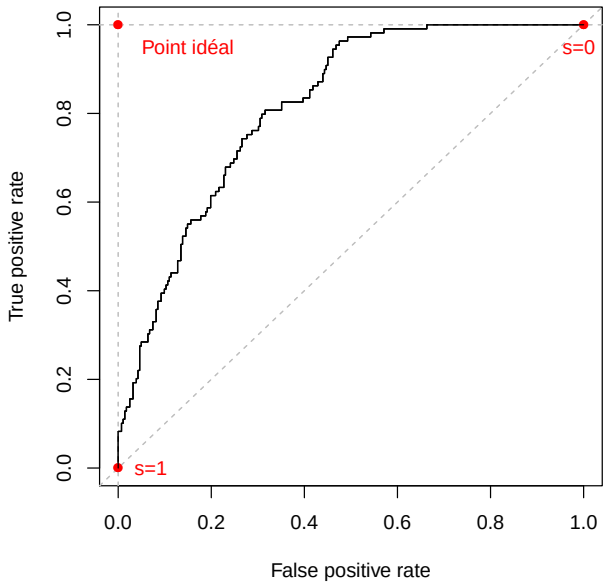

FIGURE 4.2 – Exemple de courbe ROC.

créer l'objet `pr=predicition(pred,ytest)`, puis on obtient la courbe ROC avec `roc=performance(pr, measure = "tpr", x.measure = "fpr")` que l'on peut représenter avec un `plot`. La fonction `performance` permet par ailleurs de calculer de nombreux scores en fonction du seuil s (voir l'aide), comme par exemple le taux d'erreur $FP+FN$ avec `performance(pr, measure = "err")`, ce qui est précieux en vue de choisir le s optimal.

4.5 Extension au cas multinomial

Si Y n'est pas binaire mais admet K modalités, notées $0, \dots, K-1$, on généralise la régression logistique grâce à la *régression logistique multinomiale* qui consiste à modéliser chaque ratio $\mathbb{P}(Y=k|X=x)/\mathbb{P}(Y=0|X=x)$ pour $k=1,\dots,K-1$ par

$$\frac{\mathbb{P}(Y=k|X=x)}{\mathbb{P}(Y=0|X=x)} = e^{x'\beta_k} \tag{4.3}$$

où $\beta_k \in \mathbb{R}^p$. Il y a donc $K-1$ paramètres β_k dans $\mathbb{R}^p$ à estimer.

De l'équation précédente, on en déduit la forme de $\mathbb{P}(Y=k|X=x)$. En

effet

$$\sum_{k=0}^{K-1} \mathbb{P}(Y=k|X=x) = \mathbb{P}(Y=0|X=x) + \sum_{k=1}^{K-1} \frac{\mathbb{P}(Y=k|X=x)}{\mathbb{P}(Y=0|X=x)} \mathbb{P}(Y=0|X=x)$$
$$= \mathbb{P}(Y=0|X=x) \left(1 + \sum_{k=1}^{K-1} e^{x'\beta_k}\right)$$

et puisque $\sum_{k=0}^{K-1} \mathbb{P}(Y=k|X=x) = 1$, on en déduit $\mathbb{P}(Y=0|X=x)$ et donc

$$\mathbb{P}(Y=k|X=x) = \frac{e^{x'\beta_k}}{1+\sum_{k=1}^{K-1} e^{x'\beta_k}}. \tag{4.4}$$

Si $K=2$, on retrouve bien la forme de la régression logistique standard.

Chaque équation (4.3) contient un paramètre $\beta_k \in \mathbb{R}^p$ à estimer, ce qui fait $(K-1)\times p$ coefficients à estimer en tout. L'inférence, la prévision et la validation du modèle suivent les mêmes lignes que pour la régression logistique :

- L'estimation des paramètres se fait par maximum de vraisemblance (Y suit une loi multinomiale) ;
- Les tests de Wald sont disponibles pour évaluer la significativité des variables ;
- On peut de façon similaire calculer et exploiter la déviance du modèle ;
- On prédit $\mathbb{P}(Y=k|X=x)$ en remplaçant β_k par $\hat{\beta}_k$ dans (4.4) ;
- Suite à une prévision, on peut classer un individu dans la modalité de Y de probabilité prédite la plus élevée (ce qui revient à prendre le seuil $s=0.5$ dans le cas binaire).

Sous R : fonction `multinom` de la librairie `nnet`.