POLYTECH'LILLE
DÉPARTEMENT G.I.S.

Modélisation Statistique

Julien JACQUES
http ://labomath.univ-lille1.fr/∼jacques/

Table des matières

Introduction

Pré-requis : la maîtrise des cours de Probabilités et de Statistique Inférentielle (disponible en ligne sur mon site) de troisième année GIS est indispensable à la bonne compréhension de ce cours.

Les modèles

Dans ce cours nous chercherons à modéliser une variable Y (variable à expliquer, réponse) en fonction d'une ou plusieurs variables explicatives $X_1, \ldots, X_p$ (covariables). Lorsque Y sera quantitative (montant d'épargne investit, durée de rémission d'une maladie...), nous parlerons de régression ou encore d'analyse de variance (ou covariance) selon la nature des variables explicatives, qui peuvent être rassemblées sous l'appellation **modèle linéaire**.
Lorsque Y est une variable aléatoire qualitative (défaut de remboursement, achat d'un produit...), nous parlerons généralement de **classification**, **supervisée** lorsque l'on dispose d'observation de Y, et non supervisée dans le cas contraire. Nous verrons dans ce cours deux méthodes de classification supervisée : la régression logistique, qui est une extension du modèle linéaire à la famille des modèles linéaires généralisés, ainsi que l'analyse discriminante probabiliste. Ces notions sont reprises dans la Table 1.

Variable à expliquer	Variables explicatives	Nom de l'analyse
1 quantitative	1 quantitative	régression simple (Section 1)
1 quantitative	plusieurs quantitatives	régression multiple (Section 2)
1 quantitative	plusieurs qualitatives	analyse de variance (Section 3)
1 quantitative	plusieurs qualitatives et quantitatives	analyse de covariance (Section 3.5)
1 qualitative	plusieurs quantitatives et qualitatives	régression logistique (Section 4)
1 qualitative	plusieurs quantitatives (voir quali.)	analyse discriminante probabiliste (Section 5)

TAB. 1 – Les différentes techniques de modélisation étudiées dans ce cours

Remarque. *Concernant la classification supervisée, il existe bien d'autres méthodes que les deux méthodes abordées dans ce cours :*
- *l'analyse factorielle discriminante qui est une méthode géométrique cherchant à construire de nouvelle variables discriminant au mieux les classes (cours Statistique Exploratoire GIS4)*
- *la méthode des k plus proches voisins,*
- *les arbres de décisions (cours Modélisation Avancée GIS4),*
- *ou encore des méthodes qui estiment directement la frontière de classification (SVM, réseaux de neurones).*

Objectifs

Les objectifs d'une modélisation statistique peuvent être de différentes natures, que l'on peut tenter de répartir en deux classes, les objectifs prédictifs et les objectifs explicatifs :
- **prédictifs** : prévoir à partir des renseignements dont on dispose sur un client (âge, catégorie CSP, salaire, situation familiale, statut dans son habitation actuelle...) s'il va ou non souscrire un crédit à la consommation qui lui est proposé. Ces prévisions peuvent également permettre de cibler les bons clients à qui proposer ce crédit.
- **descriptifs**
  - **sélection des variables** pertinentes : parmi l'âge d'un patient, son poids, son taux de cholestérol, le nombre de cigarettes fumées par jour (...), quelles sont les variables qui influent significativement sur la survenue d'un cancer des poumons ?

– **forme du modèle** : comment le montant de l'épargne d'un client évolue-t-il en fonction de son salaire ?

Les étapes

Les différentes étapes d'une modélisation statistique sont les suivantes

(i) **identifier le problème** pour choisir le modèle statistique à utiliser (en fonction de la nature de Y, de X, des résultats attendus...),

(ii) choisir les variables pertinentes (par des études préalables de corrélation par exemple, mais pas seulement),

(iii) estimer les paramètres du modèle (généralement par maximum de vraisemblance),

(iv) évaluer la qualité de la modélisation obtenue (tests statistiques), l'apport des différentes variables, et éventuellement revenir au point (ii) pour remettre en cause le choix des variables, voir en (i) si c'est le modèle qui doit être remis en cause,

(v) utiliser enfin le modèle pour répondre aux objectifs voulus.

Chapitre 1

Régression linéaire simple

*Logiciel **R** : fonction* `lm`.
*Logiciel **SAS** : proc* `reg`.

1.1 Le modèle théorique

Soit Y et X deux variables aléatoires gaussiennes. L'objectif de la régression linéaire est de modéliser la variable aléatoire Y par une certaine fonction de X, $f(X)$, qui soit la meilleure possible au sens de l'erreur quadratique moyenne $E[(Y - f(X))^2]$. Nous avons vu en cours de probabilité que la fonction minimisant cette erreur n'était rien d'autre que l'espérance de Y conditionnellement à X : $E[Y|X]$.
Dans le cas de variables gaussiennes, le calcul de l'espérance conditionnelle donne le résultat suivant :

$$E[Y|X = x] = \beta_0 + \beta_1 x$$

où

$$\beta_0 = E[Y] - \beta_1 E[X] \qquad \text{et} \qquad \beta_1 = \frac{Cov(X,Y)}{V(X)}$$

La meilleure fonction de X permettant de modéliser Y est alors une fonction affine ou linéaire de X, d'où le nom de régression *linéaire*.
Ceci constitue le **postulat de base** de la régression linéaire. Nous chercherons dans ce chapitre à modéliser Y par une fonction linéaire de X, qui est la meilleure modélisation possible lorsque les variables sont gaussiennes.
Il conviendra donc en pratique de s'assurer de la normalité des variables (avec un test de Shapiro-Wilk) avant d'effectuer une régression linéaire. Si une variable n'est pas gaussienne, nous chercherons à la transformer de sorte qu'elle soit *la plus gaussienne* possible.

Remarque 1.1.1. *Si X et Y sont indépendantes, leur covariance est nulle et donc β_1 également. La meilleure modélisation de Y que l'on peut avoir en fonction de X n'est alors que $E[Y]$.*

1.2 Le modèle statistique

Soit un échantillon $(X_i, Y_i)_{i=1,n}$ d'observations indépendantes et identiquement distribuées.
On suppose dans ce cours que les X_i sont déterministes, fixés par l'expérimentation, mais cela ne change rien au modèle et aux estimations si les X_i sont aléatoires.
Le modèle de la régression linéaire suppose :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.1}$$

où β_0 (appelé *intercept*) et β_1 sont des paramètres fixés du modèle ($\beta_0, \beta_1 \in \mathbb{R}$), que nous chercherons à estimer par la suite, et où les résidus ϵ_i vérifient :

- $E[\epsilon_i] = 0$,

- $V(\epsilon_i) = \sigma^2$ (σ^2 étant également un paramètre du modèle). On dit dans ce cas que les résidus sont homoscédastiques (i.e. variance constante),
- $Cov(\epsilon_i, \epsilon_j) = 0$ si $i \neq j$ (ce qui implique la non corrélation des résidus).

Ces hypothèses sont généralement appelées **hypothèses faibles**. Les **hypothèses fortes** supposent en plus la *normalité* des résidus (ce qui implique donc leur indépendance puisqu'ils sont non corrélés), qui nous permettra par la suite d'effectuer des tests sur le modèle de régression linéaire.
D'un point de vue matriciel, le modèle de régression linéaire s'écrit :

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{1.2}$$

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{1.3}$$

1.3 Estimation des paramètres

Comme nous le verrons dans le cas de la régression multiple, l'estimation par maximum de vraisemblance sous les hypothèses fortes est équivalente à l'estimation par moindres carrés (hypothèses faibles). Dans le cadre de l'estimation par moindres carrés, nous cherchons à minimiser les écarts entre les valeurs prédites

$$\mathbf{Y}^* = \mathbf{X}\beta \tag{1.4}$$

et les valeurs observées $\mathbf{Y}$. Nous choisissons traditionnellement le carré de la norme euclidienne comme mesure de l'écart :

$$D(\beta) = ||\mathbf{Y} - \mathbf{Y}^*||_2^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - X_i\beta_1)^2 = \sum_{i=1}^{n} \epsilon_i^2. \tag{1.5}$$

La minimisation de $D(\beta)$ suivant β_0 et β_1 conduit aux estimateurs suivant :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \qquad \hat{\beta}_1 = \frac{S_{XY}}{S_X^2}.$$

où classiquement $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$, $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ et

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

On montre que ces estimateurs de β_0 et β_1 sont des estimateurs sans biais, et de variance minimale parmi les estimateurs fonctions linéaires des Y_i (resp. parmi tous les estimateurs dans le cas gaussien).

A chaque valeur X_i de X correspond donc une valeur prédite $\hat{Y}_i$ de Y :

$$\hat{Y}_i = \hat{\beta}_1 X_i + \hat{\beta}_0.$$

L'écart entre cette prédiction $\hat{Y}_i$ et Y_i est appelé **résidu** : $\hat{\epsilon}_i = \hat{Y}_i - Y_i$.
La variance résiduelle σ^2 est estimée par :

$$S_\epsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

Remarque. *L'utilisation du modèle linéaire dépasse le cadre simple d'une relation linéaire entre X et Y. En effet, de nombreux modèles non linéaires se ramènent facilement au modèle linéaire par des transformations simples :*

- *le modèle $Y = \alpha X^\beta$ très utilisé en économétrie (élasticité constante de Y par rapport à X) devient un modèle linéaire en étudiant le logarithme des variables*
- *le modèle à croissance exponentielle $Y = \alpha e^{\beta X}$ devient un modèle linéaire en travaillant avec $ln(Y)$*
- *... et bien d'autre.*

Un simple nuage de points (X_i, Y_i) pourra aider à identifier une relation non linéaire.

1.4 Tests sur le modèle de régression linéaire

Une fois le modèle de régression linéaire estimé, il convient dans un premier temps de vérifier si les hypothèses faites lors de l'estimation par moindres carrés sont respectées (normalité des variables ou des résidus, non corrélation des résidus, homoscédasticité des résidus). Dans un second temps, nous testerons la validité du modèle de régression et évaluerons sa qualité.
Nous nous plaçons cette fois dans le cas des hypothèses fortes.

1.4.1 Vérification des hypothèses du modèle linéaire

1.4.1.1 Normalité et homoscédasticité des résidus

L'hypothèse de normalité des résidus peut être testée par un test classique de normalité comme le test de Shapiro-Wilk.
L'homoscédasticité peut quant à elle être vérifiée visuellement en représentant le nuage des résidus (X_i, t_i), où t_i sont une normalisation des résidus (résidus studentisés, définis au paragraphe 1.6.2). Ce nuage de point devrait se répartir uniformément de part et d'autre de l'axe des abscisses si les résidus ont bien une variance constante.

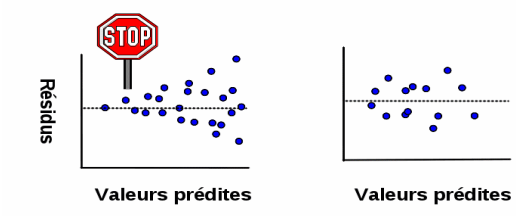


FIG. 1.1 – Homoscédasticité des résidus.

1.4.1.2 Test de non corrélation des résidus

Les propriétés de l'estimation par moindres carrés reposent notamment sur l'hypothèse de non corrélation des résidus. Le test de Durbin-Watson permet de vérifier que les ϵ_i ne sont pas corrélés. La statistique utilisée est

$$d = \frac{\sum_{i=2}^{n}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n}\epsilon_i^2}$$

qui doit être proche de 2 si les résidus sont non corrélées. Cette statistique ne suit pas de loi particulière, mais ses valeurs critiques ont été tabulées.

1.4.2 Tests de la nullité des paramètres du modèle

Sous l'hypothèse de normalité des résidus, les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ des paramètres β_0 et β_1 suivent des lois normales

$$\begin{aligned}\hat{\beta}_1 &\sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{(n-1)S_X^2}\right),\\ \hat{\beta}_0 &\sim \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n} + \frac{\sigma^2\bar{X}^2}{(n-1)S_X^2}\right),\end{aligned}$$

dont on estime la variance en remplaçant σ^2 par son estimation S_ϵ^2.
On peut montrer que

$$\frac{n-2}{\sigma^2}S_\epsilon^2 \sim \chi^2_{n-2}$$

et que

$$\frac{\hat{\beta}_1 - \beta_1}{S_\epsilon\sqrt{\frac{1}{(n-1)S_X^2}}} \sim t_{n-2} \qquad \text{et} \qquad \frac{\hat{\beta}_0 - \beta_0}{S_\epsilon\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}}} \sim t_{n-2}.$$

Ceci permet donc de construire des intervalles de confiance et de tester la nullité de chacun des deux paramètres. A noter que le test portant sur $\hat{\beta}_1$ est équivalent au test sur le coefficient de corrélation linéaire entre X et Y.

1.4.3 Analyse de variance de la régression

Il est d'usage de décomposer la variance totale en la variance expliquée par la régression et la variance résiduelle. La somme des carrés totale (SST) se décompose en la somme des carrés expliqués par la régression (SSReg) et la somme des carrés résiduelles (SSR) :

$$\underbrace{(n-1)S_Y^2}_{SST} = \underbrace{(n-1)\frac{S_{XY}^2}{S_X^2}}_{SSReg} + \underbrace{(n-2)S_\epsilon^2}_{SSR}$$

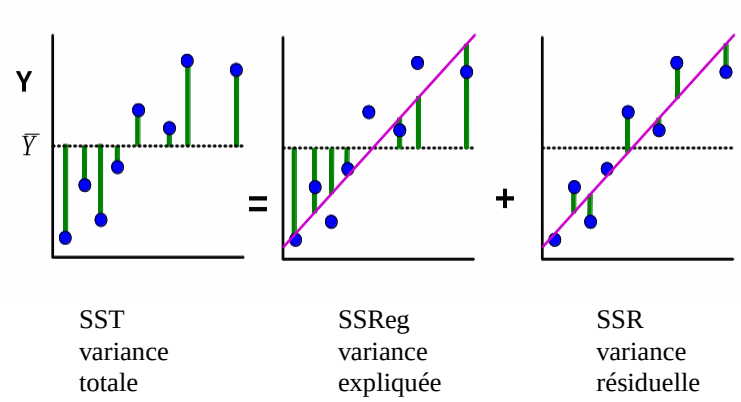


FIG. 1.2 – Analyse de variance de la régression.

Le **coefficient de détermination** R^2 :

$$R^2 = \rho_{XY}^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2} = \frac{SSReg}{SST}$$

exprime le rapport entre la variance expliquée par le modèle de régression et la variance totale (ρ_{XY} étant le coefficient de corrélation linéaire entre X et Y). Il est compris entre 0 et 1 et est un bon indicateur de la qualité de la régression, quoi que très subjectif.
Sous l'hypothèse H_0 de non régression linéaire ($\beta_1 = 0$), la statistique suivante

$$F = (n-2)\frac{R^2}{1-R^2} = (n-2)\frac{SSReg}{SSR}$$

suit une loi de Fisher $F_{1,n-2}$.

1.5 Prédiction

Pour une valeur donnée x^* de X, la prédiction de Y est

$$\hat{y}^* = \hat{\beta}_1 x^* + \hat{\beta}_0.$$

On définit l'**intervalle de confiance d'une prévision** de la façon suivante :

$$IC_{1-\alpha}(E[Y|X=x^*]) \quad = \quad \left[\hat{y}^* + \sigma t_{n-2,\frac{\alpha}{2}}\sqrt{\frac{1}{n}+\frac{(x^*-\bar{X})^2}{(n-1)S_X^2}}\,;\, \hat{y}^* - \sigma t_{n-2,\frac{\alpha}{2}}\sqrt{\frac{1}{n}+\frac{(x^*-\bar{X})^2}{(n-1)S_X^2}}\right]$$

Cette intervalle de confiance définit les limites dans lesquelles se situe probablement une valeur individuelle lue sur la droite de régression : lorsqu'on a construit un modèle qui se présente sous la forme d'une droite de régression,

l'intervalle de confiance en question dit que, pour une valeur donnée x^* de la variable X, la vraie valeur de la variable Y devrait se situer au sein de cet intervalle de confiance.
On définit également l'**intervalle de prédiction d'une prévision** définit les limites dans lesquelles tombera vraisemblablement une nouvelle observation de Y si elle fait partie de la même population statistique que l'échantillon :

$$IC_{1-\alpha}(\hat{y}^*) = \left[\hat{y}^* + \sigma t_{n-2,\frac{\alpha}{2}}\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{X})^2}{(n-1)S_X^2}}\,;\,\hat{y}^* - \sigma t_{n-2,\frac{\alpha}{2}}\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{X})^2}{(n-1)S_X^2}}\right]$$

1.6 Détection d'observations atypiques

Les méthodes d'estimation utilisées sont très sensibles aux observations atypiques (*outliers*). Nous proposons dans cette section quelques outils permettant de détecter de telles observations.
Une fois ces observations détectées, il n'y a pas de remède universel : supprimer une valeur aberrante, sonder si elle est due à une erreur de mesure, ne rien faire... Tout dépend du contexte et doit être négocié avec le commanditaire de l'étude.

1.6.1 Effet levier

Une première façon de détecter un individu atypique est de mesurer l'impact de l'observation Y_i sur la détermination de $\hat{Y}_i$. Pour cela, on montre qu'il est possible d'écrire

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j \qquad \text{où} \qquad h_{ij} = \frac{1}{n} + \frac{(X_i-\bar{X})(X_j-\bar{X})}{\sum_{j=1}^{n}(X_j-\bar{X})^2}.$$

Les h_{ij} forment la matrice H appelée *hat matrix*. Les termes diagonaux h_{ii} mesurent l'impact de Y_i dans l'estimation $\hat{Y}_i$. Cet impact est directement lié à l'éloignement de l'observation X_i à la moyenne des observations $\bar{X}$.

1.6.2 Etude des résidus

Différents types de résidus peuvent être considérés.
- **résidus** : $\epsilon_i = \hat{Y}_i - Y_i$
- **résidus standardisés** (interne) : les résidus bruts ϵ_i n'ayant pas la même variance, on calcule des versions standardisées r_i afin de les rendre comparables :

$$r_i = \frac{\epsilon_i}{S_\epsilon\sqrt{1-h_{ii}}}$$

- **résidus studentisés** (externe) : une autre standardisation (externe) des résidus permet d'obtenir des résidus t_i suivant une loi de Student :

$$t_i = \frac{\epsilon_i}{S_{\epsilon(i)}\sqrt{1-h_{ii}}}$$

où $S_{\epsilon(i)}$ est une estimation de la variance résiduelle ne prenant pas en compte la $i$ième observation (contrairement à S_ϵ ci-dessus) :

$$S_{\epsilon(i)} = \frac{n-2}{n-3}S_\epsilon - \frac{1}{n-3}\frac{\epsilon_i^2}{1-h_{ii}}.$$

En pratique, une observation sera considérée comme atypique (vis-à-vis de son éloignement à $\bar{X}$) si son résidu Studendisé dépasse les bornes ± 2.

1.6.3 Distance de Cook

Les deux indicateurs précédents s'intéressent à l'éloignement d'une observation à la moyenne et à l'importance des résidus. La distance de Cook est un indicateur synthétisant ces deux informations, construit en comparant les prédictions obtenues avec et sans la $i$ième observation :

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_{j(i)} - \hat{Y}_j)^2}{2S_\epsilon^2} = \frac{h_{ii}}{2(1-h_{ii})} r_i^2$$

où $\hat{Y}_{j(i)}$ est l'estimation de Y_j obtenue sans utiliser la $i$ième observation (X_i, Y_i).
Une stratégie de détection classique consiste dans un premier temps à repérer les points atypiques en comparant les distances de Cook à la valeur 1, puis à expliquer cette influence en considérant, pour ces observations, leur résidu ainsi que leur effet levier.

1.7 TP 1 : Régression linéaire simple

Simulation

Cet exercice est à réaliser sous **R**.
On considère dans cet exercice le modèle de régression simple suivant

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

avec $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. On choisit $\beta_0 = 3$ et $\beta_1 = 2$. Les x_i sont supposés être répartis uniformément sur l'intervalle $[0, 1]$.

(i) Simuler les couples $(x_i, y_i)_{i=1,\ldots,n}$ pour une taille d'échantillon $n = 10$ et une variance résiduelle $\sigma^2 = 1$. Stocker vos résultats dans deux vecteurs $\mathbf{x}$ et $\mathbf{y}$.

(ii) Dans l'écriture matricielle du modèle de régression $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ avec $\beta = (\beta_0, \beta_1)$, comment est définie la matrice $\mathbf{X}$? Construisez-la à partir de votre vecteur $\mathbf{x}$.

(iii) Nous avons vu en cours que le meilleur estimateur de β était $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Calculer cet estimateur. Que pensez-vous de vos résultats ? Recommencez la simulation et l'estimation plusieurs fois.
Indication : la fonction `solve(A)` sous R permet de calculer l'inverse de la matrice A.

(iv) Représentez graphiquement le nuage de point (fonction `plot`) ainsi que la droite de régression (avec la fonction `lines` puis avec la fonction `abline`).

(v) Estimer la variance résiduelle σ^2.

(vi) Calculer un intervalle de confiance sur β_0 et β_1, de niveau 95%.

(vii) Créer une fonction `mylm(x,y,plot,alpha)`, qui pour un vecteur `x` et `y` effectue la régression de `y` sur `x`. La fonction devra retourner les estimations des coefficients β_0 et β_1, des intervalles de confiance sur ces derniers de niveau `alpha`, l'estimation de la variance résiduelle, ainsi qu'une représentation graphique du nuage de point et de la régression lorsque l'option `plot` est à `TRUE`.

(viii) Recommencer avec une taille d'échantillon de 100, 1000.

(ix) Retrouvez vos résultats avec la fonction `lm` de R :
```
res=lm(y~x)
summary(res)
```
Explorer toutes les informations que contient le résultat d'une procédure `lm` à l'aide de la fonction `str` :
```
str(res)
```

1.7.1 Revenus immobiliers

Cet exercice est à réaliser sous SAS.
Le fichier `immeublesUSA.dat` contient pour 47 immeubles d'appartements locatifs d'une grande ville américaine, le revenu net en fonction du nombre d'appartements (Jobson, 1991). L'objectif est de modéliser le revenu net des immeubles (première colonne) en fonction du nombre d'appartements (seconde colonne), par une régression linéaire.

Analyse préliminaire

(i) Représenter graphiquement les variables (histogramme, boxplot), et donner une estimation de la densité par la méthode du noyau.

(ii) Les variables vous semblent-elles gaussiennes ?

(iii) Refaire la même chose en transformant les variables (log et racine). Quelles variables choisir pour notre régression linéaire ?

Première modélisation On considère le modèle $revenu = \beta_0 + \beta_1 nb\_appart$.

(i) Estimer les paramètres du modèle.

(ii) Représenter le nuage de points ainsi que la droite de régression.

(iii) Effectuer des tests de significativité des paramètres.

(iv) Calculer les résidus studentisés ainsi que la distance de Cook. Quel est votre diagnostic ?

Seconde modélisation On considère le modèle $log(revenu) = \beta_0 + \beta_1 log(nb\_appart)$.

(i) Estimer les paramètres du modèle.

(ii) Représenter le nuage de points ainsi que la droite de régression.

(iii) Effectuer des tests de significativité des paramètres.

(iv) Calculer les résidus studentisés ainsi que la distance de Cook. Quel est votre diagnostic ?

(v) Comparer la qualité d'ajustement des deux modèles, et conclure.

Chapitre 2

Régression linéaire multiple

*Logiciel **R** : fonction* `lm`*.*
*Logiciel **SAS** : proc* `reg`*.*

Nous cherchons désormais à expliquer une variable aléatoire quantitative Y en fonction de p variables explicatives $X_1, \ldots, X_p$, toutes quantitatives. Nous supposons toujours que les variables explicatives sont déterministes, mais encore une fois cela ne change rien au modèles et aux estimations.

2.1 Le modèle

Soit un échantillon $(X_{i1}, \ldots, X_{ip}, Y_i)_{i=1,n}$ d'observations indépendantes et identiquement distribuées.
Le modèle de la régression linéaire suppose :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i \tag{2.1}$$

où $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ sont les paramètres réels du modèle à estimer, et où les résidus ϵ_i vérifient comme pour la régression simple les *hypothèses faibles* :

- $E[\epsilon_i] = 0$,
- $V(\epsilon_i) = \sigma^2$,
- $Cov(\epsilon_i, \epsilon_j) = 0$ si $i \neq j$.

Nous rappelons que les *hypothèses fortes* supposent de plus la *normalité* des résidus (ce qui implique donc leur indépendance puisqu'ils sont non corrélés).
L'écriture matricielle du modèle (2.1) est la suivante :

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{2.2}$$

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{2.3}$$

La matrice $\mathbf{X}$, déterministe, est souvent appelée matrice de *design*.

2.2 Estimation des paramètres du modèle

2.2.1 Estimation par moindres carrés

On se place sous les hypothèses faibles. Nous cherchons à minimiser les écarts entre les valeurs prédites

$$\mathbf{Y}^* = \mathbf{X}\beta \tag{2.4}$$

et les valeurs observées $\mathbf{Y}$. Nous choisissons traditionnellement le carré de la norme euclidienne comme mesure de l'écart :

$$D(\beta) \quad = \quad ||\mathbf{Y} - \mathbf{Y}^*||_2{}^2 = \sum_{i=1}^{n} \epsilon_i^2. \tag{2.5}$$

L'estimateur par moindres carrés du paramètre β est donc :

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, D(\beta). \tag{2.6}$$

En développant $D(\beta)$ et en prenant le gradient, on obtient

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{2.7}$$

***Exercice*.** Faire la démonstration de l'équation (2.7). Montrer également que l'on a bien un minimum de $D(\beta)$.

Remarque 2.2.1. *Notation : la hat matrix définie dans le chapitre précédent comme la matrice H telle que $\hat{\mathbf{Y}} = H\mathbf{Y}$ est donc $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.*

Remarque 2.2.2. *Nous avons supposé que $\mathbf{X}'\mathbf{X}$ était inversible, ce qui est le cas dès que X est de rang $p+1$. Se reporter au paragraphe 2.6 pour le cas contraire.*

Propriété 2.2.1. *$\hat{\beta}$ est un estimateur sans biais de β.*

***Exercice*.** Faire la preuve.

Propriété 2.2.2. *$\hat{\beta}$ est l'estimateur de variance minimale parmi les estimateurs de β sans biais et linéaires en Y. Sa variance est $V(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$*

L'estimateur non biaisé de σ^2 sera quant à lui :

$$\hat{\sigma}^2 = \frac{||\mathbf{Y} - \hat{\mathbf{Y}}||_2^2}{n-p-1}. \tag{2.8}$$

On notera qu'il est fonction de $\hat{\beta}$.

2.2.2 Estimation par maximum de vraisemblance

On se place sous les hypothèses fortes, c'est-à-dire que les erreurs ϵ_i sont supposées gaussiennes. Nous avons donc

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I_n) \tag{2.9}$$

d'où la vraisemblance du modèle de régression linéaire :

$$L(\beta, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\{-\frac{1}{2\sigma^2}||\mathbf{Y} - \mathbf{X}\beta||_2^2\} \tag{2.10}$$

On montre facilement, après passage à la log-vraisemblance, que la maximisation de (2.10) en fonction de β conduit à l'estimateur (2.7). Quant à σ^2, la maximisation conduit à un estimateur biaisé auquel nous préférerons sa version non biaisée (2.8).

***Exercice*.** Faire la preuve.

Propriété 2.2.3. *Les estimateurs du maximum de vraisemblance de β et σ^2 sont efficaces (de variance minimale). De plus, ils sont indépendants et leur lois sont :*

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \tag{2.11}$$

et

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1} \tag{2.12}$$

2.3 Tests sur le modèle linéaire

Comme pour le modèle linéaire simple, les hypothèses de régression linéaire doivent être vérifiées (normalité des variables ou des résidus, non corrélation des résidus, homoscédasticité des résidus). La démarche est identique à celle de la régression simple (paragraphe 1.4.1).
Nous nous plaçons dans le cadre des hypothèses fortes.

2.3.1 Tests sur les paramètres

Pour chaque paramètre β_j, on peut montrer que son estimateur suit une loi de Student :

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-p-1} \tag{2.13}$$

où $\hat{\sigma}^2_{\hat{\beta}_j}$ est l'estimation de la variance de l'estimateur, égale au $(j+1)$ième terme de la diagonale de la matrice $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$. A partir de cette statistique, il est possible de tester un à un la nullité des différents paramètres du modèle de régression linéaire multiple (penser à maîtriser les risques encourus par une correction de Bonferroni par exemple), ou de construire des intervalles de confiance sur ces paramètres, très utiles lors de la phase d'interprétation du modèle.

Remarque. *Les estimateurs des différents paramètres n'étant pas indépendants, il est possible de tester la nullité de chaque paramètre séparément mais il ne faut rien en conclure conjointement.*

2.3.2 Analyse de variance de la régression

Comme dans le cas de la régression simple (paragraphe 1.4.3), il est possible de tester globalement le modèle ($H_0 : \beta_1 = \ldots = \beta_p = 0$) par une analyse de variance du modèle de régression. Cela consiste à décomposer la dispersion totale (SST) en une part de dispersion expliquée par le modèle de régression ($SSReg$) et une part de dispersion résiduelle (SSR)

$$\underbrace{||\mathbf{Y} - \bar{\mathbf{Y}}||_2^2}_{SST} = \underbrace{||\hat{\mathbf{Y}} - \bar{\mathbf{Y}}||_2^2}_{SSReg} + \underbrace{||\mathbf{Y} - \hat{\mathbf{Y}}||_2^2}_{SSR}. \tag{2.14}$$

L'analyse de variance de la régression est généralement présentée dans un tableau d'analyse de variance

Source	Somme des carrés	degrés de liberté	carré moyen	F
Régression	$SSReg$	p	$MSReg = SSReg/p$	$F = \frac{MSReg}{MSR}$
Erreur	SSR	$n-p-1$	$MSR = SSR/(n-p-1)$	
Total	SST	$n-1$		

La statistique $F = \frac{MSReg}{MSR}$, qui sous H_0 suit une loi de Fisher à p et $n-p-1$ degrés de liberté, permet de tester cette hypothèse.

Remarque. *La statistique F est liée au coefficient de détermination par $F = \frac{R^2}{1-R^2}\frac{n-p-1}{p}$.*

2.4 Prédiction

Pour une valeur $x^* = (1, x_1^*, \ldots, x_p^*)'$ de X, la prévision de Y sera donnée par

$$\hat{y}^* = x^{*\prime}\hat{\beta}. \tag{2.15}$$

Un intervalle de confiance de niveau $1-\alpha$ pour la valeur y^* sera construit à partir de cette prévision ponctuelle :

$$x^{*\prime}\hat{\beta} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{1 + x^{*\prime}(\mathbf{X}'\mathbf{X})^{-1}x^*}. \tag{2.16}$$

2.5 Sélection de variables et choix de modèle

Parmi l'ensemble des p variables disponibles, toutes n'ont pas nécessairement un intérêt dans la modélisation de Y, et il peut alors être néfaste de les utiliser. De plus, il est possible d'hésiter entre l'utilisation d'une variable X_j ou une certaine transformation de cette variable ($\ln X_j$,X_j^2...). Nous sommes alors en présence de différents modèles possibles parmi lesquels il faut faire un choix.
Intuitivement, le fait de ne pas utiliser assez de variables ou bien de trop en utiliser, conduit à une mauvaise estimation de l'espérance conditionnelle $h(X) = E[Y|X]$, notée $\hat{h}(X)$. Il est possible de définir comme mesure de la qualité de l'estimation $\hat{h}(X)$, la moyenne des erreurs quadratiques moyennes ($MEQM$) :

$$MEQM = \frac{1}{n}\sum_{i=1}^{n} E[(\hat{h}(X_i) - h(X_i))^2] \tag{2.17}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{V(\hat{h}(X_i))}_{\text{variance}} + \underbrace{(E[\hat{h}(X_i)] - h(X_i))^2}_{\text{biais}}\right) \tag{2.18}$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n} V(\hat{h}(X_i))}_{\text{moyenne des variances}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(E[\hat{h}(X_i)] - h(X_i))^2}_{\text{moyenne des biais}} \tag{2.19}$$

Un modèle
- trop peu complexe (pas assez de variables) aura un biais fort (et une variance faible),
- trop complexe (trop de variables) aura une variance forte (et un biais faible),

tout l'intérêt étant d'avoir un modèle ayant un $MEQM$ le plus faible possible, c'est-à-dire réalisant le meilleur compromis biais/variance possible.
Malheureusement ce critère théorique n'est pas calculable en pratique ($h(X)$ inconnue) et des critères *approximatifs* doivent être utilisés.

2.5.1 Critères de comparaison de modèle

Remarque. *La sélection de variables par tests d'hypothèses (paragraphe 2.3.1) n'est pas pertinente pour deux raisons : le grand nombre de tests à effectuer rend peu puissante la stratégie globale, et cette stratégie n'est applicable que pour comparer des modèles emboités (l'ensemble des variables d'un modèle doit être inclus dans celui de l'autre).*

Remarque. *Lorsque l'échantillon dont on dispose est de très grande taille, une façon simple d'évaluer la qualité d'un modèle, et donc de choisir parmi plusieurs modèles candidats, est de séparer l'échantillon global en une partie apprentissage ($2/3$ de l'échantillon global) et une partie test (le $1/3$ restant) servant à l'évaluation (par calcul de la somme des carrés des erreurs par exemple). Malheureusement, les échantillons sont souvent de tailles réduites, et ce procédé n'est pas toujours applicable.*

Nous présentons ci-après plusieurs critères évaluant la qualité d'un modèle utilisant d variables parmi les p disponibles ($d \leq p$)

2.5.1.1 Limitation du coefficient de détermination R^2

Le coefficient de détermination est une fonction croissante de la complexité du modèle. Il conduira donc toujours à choisir le modèle qui épouse le mieux les données, autrement dit le modèle le plus complexe. Son utilisation n'est donc pas recommandée sauf dans le cas de modèle à nombres de variables identiques.

2.5.1.2 Coefficient de détermination ajusté $\bar{R}^2$

A partir du coefficient de détermination $R^2 = 1 - \frac{SSR}{SST}$ on définit le coefficient de détermination ajusté :

$$\bar{R}^2 = \frac{(n-1)R^2 - d}{n-d-1} \tag{2.20}$$

qui consiste à pénaliser R^2 par l'augmentation du nombre d de variables utilisées.
Attention : il peut prendre parfois des valeurs négatives.

2.5.1.3 Critère de validation croisée : PRESS (ou CVSS)

La somme des carrés résiduelles $\sum_{i=1}^{n} \epsilon_i^2$ souffre du même problème que le coefficient de détermination. En notant $\epsilon_{(i)}^2$ le $i$ième résidu obtenu en estimant les paramètres du modèle de régression sans utiliser la $i$ième observation, le critère PRESS :

$$\text{PRESS} = \sum_{i=1}^{n} \epsilon_{(i)}^2, \tag{2.21}$$

permet de sélectionner les modèles ayant un bon pouvoir prédictif (on veut le PRESS le plus petit).
Bien qu'étant un des critères à privilégier, ce critère peut parfois être lourd à calculer pour des modèles complexes, et on lui préférera souvent dans ce cas les critères ci-dessous dont le calcul est immédiat.

2.5.1.4 C_p de Mallows

Dans le cas d'un modèle à $d+1$ variables (intercept β_0 y compris), un estimateur de $\frac{MEQM}{\sigma^2}$ est donné par

$$C_p = \frac{SSR_{d+1}}{\sigma_c^2} + 2(d+1) - n \tag{2.22}$$

où
- SSR_{d+1} est la somme des carrés résiduelles pour le modèle restreint à $d+1$ prédicteurs,
- σ_c^2 est l'estimateur de σ^2 obtenu par le modèle le plus complexe.

Selon ces critères, les sous-ensembles de $d+1$ variables fournissant des C_p proches de $d+1$ sont de *bons* sous-ensembles. Parmi ceux-ci, plus C_p est grand, moins bon est le sous-ensemble.

2.5.1.5 Critère AIC

L'utilisation de la vraisemblance souffre également du même problème que le coefficient de détermination. Le critère AIC pénalise la log-vraisemblance du modèle par son nombre de variables :

$$\text{AIC} = -2l + 2(d+1) \tag{2.23}$$

où l est le maximum de la log-vraisemblance. Ce critère est proche du Cp de Mallows.
On retient le modèle ayant le plus petit AIC.

2.5.1.6 Critère bayésien BIC

D'origine théorique différente, le critère BIC pénalise de façon un peu plus forte la log-vraisemblance :

$$\text{BIC} = -2l + (d+1)\ln(n). \tag{2.24}$$

On retient également le modèle ayant le plus petit BIC.

2.5.2 Algorithme de sélection de variables

On recherche le meilleur sous-ensemble de variables au sens d'un des critères précedents.

2.5.2.1 Recherche exhaustive

La façon la plus simple de faire est de tester tous les sous-ensembles de variables possibles. Mais cela devient vite impossible lorsque p est grand.

2.5.2.2 Recherche descendante pas à pas

On part de toutes les variables et on élimine celle qui provoque la plus faible diminution du R^2. On fait cela jusqu'à éliminer toutes les variables, et le nombre de variables est ensuite choisi par un des critères précédents.

2.5.2.3 Recherche ascendante pas à pas

On procède de façon inverse : on part du meilleur modèle à une variable et on introduit ensuite les variables une à une.

2.5.2.4 Recherche stepwise

C'est une recherche ascendante, qui de plus, effectue à chaque pas un test de significativité de toutes les variables utilisées à l'étape courante pour éventuellement en éliminer. L'algorithme s'arrête lorsqu'on ne peut plus ni ajouter ni supprimer de variables.

2.5.2.5 Algorithme de Furnival et Wilson

Cet algorithme est peut être le plus efficace pour sélectionner le meilleur modèle pour un nombre de variables d fixé. Tout l'intérêt de cet algorithme est de rechercher le meilleur modèle (selon les critères précédents) sans avoir à explorer tous les modèles possibles. Il est limité à $p \simeq 15$ sous SAS.

2.6 Multicolinéarité des variables

L'estimation des paramètres nécessite l'inversion de la matrice $\mathbf{X}'\mathbf{X}$. Lorsque des variables sont colinéaires, cette matrice n'est pas de rang plein et n'est donc pas inversible. Ceci n'est rarement le cas en pratique. Par contre, il arrive fréquemment que des variables soit très corrélées et donc *quasi* colinéaires, ce qui rend le déterminent de $\mathbf{X}'\mathbf{X}$ proche de 0 : on dit que le système est *mal conditionné*. L'inversion de la matrice conduit alors à des estimations ayant une variance très importante, voir même parfois à des problèmes numériques. Il est donc important de diagnostiquer de tels problèmes.
Nous nous contenterons ici de donner des outils de diagnostics. Les solutions (régression *ridge, régression sur composante principale*, seront abordées dans le cours de Modélisation avancées (GIS4)).

Matrice de corrélation L'examen de la matrice de corrélation R permet de détecter des fortes corrélations entre deux variables :

$$R = \frac{1}{n-1} S^{-1} \tilde{\mathbf{X}}' \tilde{\mathbf{X}} S^{-1}$$

où $\tilde{\mathbf{X}}$ est la matrice $\mathbf{X}$ sans la première colonne de 1 et à laquelle on a retranché à chaque ligne le vecteur moyen $\bar{X}$, et S la matrice diagonale contenant les écarts-types empiriques des variables X_j.

Facteur d'inflation de la variance VIF On définit le facteur d'inflation de la variance (VIF) par

$$V_j = \frac{1}{1 - R_j^2}$$

où R_j^2 est le coefficient de détermination de la régression de la variable X_j sur les autres variables. Sa racine carré R_j est le coefficient de corrélation multiple entre X_j et les autres variables. Plus X_j est *linéairement* proche des autres variables, plus R_j est proche de 1 et le VIF grand, et donc plus la variance de l'estimateur de β_j est élevée.
L'avantage du VIF par rapport à la matrice de corrélation est qu'il prend en compte des corrélations multiples.

Conditionnement Soit $\lambda_1, \ldots, \lambda_p$ les valeurs propres de R, classées dans l'ordre décroissant. Son déterminent est égal au produit des valeurs propres, et est donc proche de 0 lorsque certaines valeurs propres sont très petites. On définit l'indice de conditionnement comme le rapport :

$$\kappa = \frac{\lambda_1}{\lambda_p}.$$

Lorsque $\kappa < 100$ il n'y a pas de problème, par contre lorsque $\kappa > 1000$ les problèmes de mauvais conditionnement sont importants.

On regardera donc dans un premier temps l'indice de conditionnement, puis on se penchera sur les forts VIF en cas de problème pour détecter la source de la colinéarité.

2.7 TP 2 : Régression linéaire multiple

2.7.1 Simulation

Cet exercice est à réaliser sous **R**.
On considère dans cet exercice le modèle de régression suivant

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \qquad i = 1, \ldots, n$$

avec $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. On choisit $\beta_0 = 3$, $\beta_1 = 2$, $\beta_2 = -2$ et $\beta_3 = 1$. Les x_{ij} sont supposées être répartis uniformément sur l'intervalle $[0, 1]$ et indépendants entre eux.

(i) Simuler les couples $(\mathbf{x}_i, y_i)_{i=1,\ldots,n}$ pour une taille d'échantillon $n = 1000$ et une variance résiduelle $\sigma^2 = 1$. Stocker vos résultats dans une matrice $n \times 3$ $\mathbf{x}$ et un vecteur $\mathbf{y}$.

(ii) Estimer le paramètre $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ par $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Donner un intervalle de confiance sur ces estimations.

(iii) Nous allons maintenant introduire une corrélation entre les variables explicatives. Pour cela, nous allons corréler la seconde variable à la première en remplaçant les x_{i2} par $x_{i2} = x_{i1} + u_i$ où $u_i \sim \mathcal{N}(0, \eta^2)$ est un bruit de variance η. Pour plusieurs valeurs de η (10,1,0.1,0.01 et 0) estimer les paramètres $(\beta_0, \beta_1, \beta_2, \beta_3)$ du modèle et calculer leur variance. Qu'en concluez-vous ?

(iv) Pour chaque valeur de η précédente, calculer les facteurs d'inflation de la variance (VIF). Interpréter les résultats.

2.7.2 Données réelles

Cet exercice est à réaliser sous SAS.
Le fichier `ukcomp1.dat` (Jobson, 1991) contient les résultats comptables de 40 entreprises du Royaume-Uni. Dans ce fichier, la première colonne est la variable RETCAP (Return on capital employed), qui est la variable que nous chercherons à prédire en fonction des 12 autres variables :

- WCFTDT : Ratio of working capital flow to total debt
- LOGSALE : Log to base 10 of total sales
- LOGASST : Log to base 10 of total assets
- CURRAT : Current ratio
- QUIKRAT : Quick ratio
- NFATAST : Ratio of net fixed assets to total assets
- FATTOT : Gross sixed assets to total assets
- PAYOUT : Payout ratio
- WCFTCL : Ratio of working capital flow to total current liabilities
- GEARRAT : Gearing ratio (debt-equity ratio)
- CAPINT : Capital intensity (ratio of total sales to total assets)
- INVTAST : Ratio of total inventories to total assets

L'objectif de ce TP sera de trouver le meilleur modèle de régression en effectuant une sélection parmi les 12 variables explicatives disponibles.

Modèle complet

(i) Vérifier graphiquement que les variables ont une distribution approximativement gaussienne. Si besoin, n'hésitez pas à en transformer certaine.

(ii) Estimer un modèle de régression complet utilisant toutes les variables. Semble-t-il y avoir des points atypiques (résidus studentisés, distance de Cook) ? des problèmes de colinéarité entre variables (VIF) ?

(iii) Calculer le R^2 et sa version ajustée.

Recherche d'un modèle parcimonieux On appelle parcimonieux un modèle dont le nombre de paramètres (et donc ici le nombre de variables explicatives utilisées) est réduit, tout en ayant un bon pouvoir prédictif.
Recherche *backward* :

(i) Itérer à la main le processus suivant :
- choisir la variable dont le test de Student ($H_0 : \beta_j = 0$) est le moins significatif (p-value la plus grande),
- la supprimer et ré-estimer le modèle.

Arrêter la procédure lorsque tous les coefficients sont significatifs (seuil 5%). Attention, on gardera toujours l'intercept (β_0), qui ne doit pas être considéré comme les autres variables.

(ii) Comparer avec la procédure automatique de SAS utilisant l'option `backward`.

(iii) Calculer les critères de choix de modèles (C_p, AIC et BIC, R^2 et R^2 ajusté) pour le meilleur modèle obtenu.

Recherche *forward* :

(i) Itérer à la main le processus suivant : commencer par introduire dans le modèle la variable la plus corrélée avec RETCAP.
- estimer le modèle,
- choisir la variable la plus corrélée avec les résidus du modèle précédent.

Arrêter la procédure lorsque la variable ajoutée n'est plus significative (seuil 5% voir un peu plus).

(ii) Comparer avec la procédure automatique de SAS utilisant l'option `forward`

(iii) Calculer les critères de choix de modèles (C_p, AIC et BIC, R^2 et R^2 ajusté) pour le meilleur modèle obtenu.

Recherche automatique par Furnival et Wilson :

(i) Estimer le meilleur modèle à l'aide de l'algorithme de Furnival et Wilson.

(ii) Calculer les critères de choix de modèles (C_p, AIC et BIC, R^2 et R^2 ajusté) et comparer avec les modèles précédents (complets et ceux obtenus par sélection forward et backward).

Prediction Récupérer le fichier `ukcomp2.dat`.

(i) Estimer la variable RETCAP sur ce fichier à l'aide du modèle complet, du modèle maximisant le R^2 ajusté, celui maximisant le C_p et celui maximisant BIC.

(ii) Pour chaque modèle, calculer la somme des carrés des erreurs de prédiction. Comparer alors les modèles.
Pour ce faire, nous vous proposons l'astuce suivante (si vous avez d'autres idées n'hésitez pas) :
- Concaténer les deux fichiers `ukcomp1.dat` et `ukcomp2.dat`, en appelant différemment la variable RETCAP dans ces deux fichiers (RETCAP1 et RETCAP2 par exemple). Le fichier concaténé contriendra ainsi 80 lignes, dont les 40 premières (correspondant à `ukcomp1.dat`) auront la variable RETCAP1 renseignée tandis que RETCAP2 ne le sera pas, et vice-versa pour les 40 suivantes.
- Estimer le modèle de régression de RETCAP1 en fonction des variables explicatives retenues, et demander à SAS d'effectuer des prédictions (option `p` à indiquer à la suite de la ligne `model`). Ainsi, seules les 40 premières lignes auront servi à estimer le modèle, car seules celles-ci ont une valeur pour RETCAP1, mais les prédictions seront faites pour les 80 lignes (pour lesquelles les variables explicatives sont renseignées).
- Il suffit ensuite de créer une variable résidus, comme la différence entre la prédiction obtenue et la variable RETCAP2. Seules les 40 dernières lignes auront un résidu car seules ces lignes disposent de RETCAP2.
- Il suffit finalement de calculer la moyenne des carrés des résidus (à l'aide d'une `PROC MEANS` par exemple).

Remarquons qu'il est possible de comparer les modèles sur cet échantillon puisqu'il n'a pas servi à estimer le modèle (on parle d'*échantillon test*, alors que l'échantillon `ukcomp1.dat` ayant servi à l'estimation est appelé *échantillon d'apprentissage*). Au contraire, évaluer des modèles sur l'échantillon ayant servi à estimer le modèle conduirait à choisir un modèle trop complexe : on parle de *sur-apprentissage*.

Chapitre 3

Analyse de variance et de covariance

Pour l'ANOVA :
*Logiciel **R** : fonction* `aov`.
*Logiciel **SAS** : proc* `anova` *dans le cas de plans équilibrés (définition ci-après) ou proc* `glm` *dans le cas général.*

Pour l'ANCOVA :
*Logiciel **SAS** : proc* `glm`.

L'analyse de variance (ANOVA) a pour objectif d'expliquer une variable aléatoire quantitative Y à partir de variables explicatives *qualitatives*, appelées *facteur*s et notées dans ce chapitre A, B ... L'objectif est alors de comparer les moyennes empiriques de Y pour les différentes modalités (ou *niveaux*) prises par les facteurs.

Lorsque nous ajoutons des variables explicatives quantitatives, l'analyse s'appelle analyse de covariance (ANCOVA). L'idée générale sera de comparer pour chaque croisement de niveaux des variables qualitatives, le modèle de régression de Y sur les variables quantitatives.

3.1 Analyse de variance à un facteur

3.2 Graphiques préliminaires

Une représentation graphique à l'aide de boîte à moustaches (*boxplot*) des distributions de Y correspondant à chaque niveau d'un facteur permet bien souvent de donner un premier avis sur la situation.

3.2.1 Le modèle

Soit Y une variable quantitative dont on observe les valeurs pour différents niveaux d'un facteur qualitatif A. On suppose disposer de J échantillons indépendants de Y de tailles n_1 à n_J correspondant à chacun des J niveaux du facteur A :

- $Y_{11}, Y_{21}, \ldots, Y_{n_1 1}$ correspondant au niveau A_1 du facteur A,
- $Y_{12}, Y_{22}, \ldots, Y_{n_2 2}$ correspondant au niveau A_2 du facteur A,
- ...
- $Y_{1J}, Y_{2J}, \ldots, Y_{n_J J}$ correspondant au niveau A_J du facteur A.

On note $n = \sum_{j=1}^{J} n_j$ la taille d'échantillon totale.
On suppose que pour chaque niveau de A, les échantillons sont i.i.d. d'espérance μ_j et de variance homogène $\sigma_j^2 = \sigma^2$. On suppose ainsi que le facteur A n'influe que sur l'espérance des échantillons et non sur leur variance. Le modèle peut alors s'écrire :

$$Y_{ij} = \mu_j + \epsilon_{ij} \tag{3.1}$$

où les ϵ_{ij} sont i.i.d., d'espérance nulle et de variance constante σ^2. On supposera de plus que les ϵ_{ij} sont gaussiens pour réaliser des tests sur le modèle d'analyse de variance. Les paramètres du modèle d'analyse de variance sont donc les espérances μ_j ainsi que la variance σ^2.

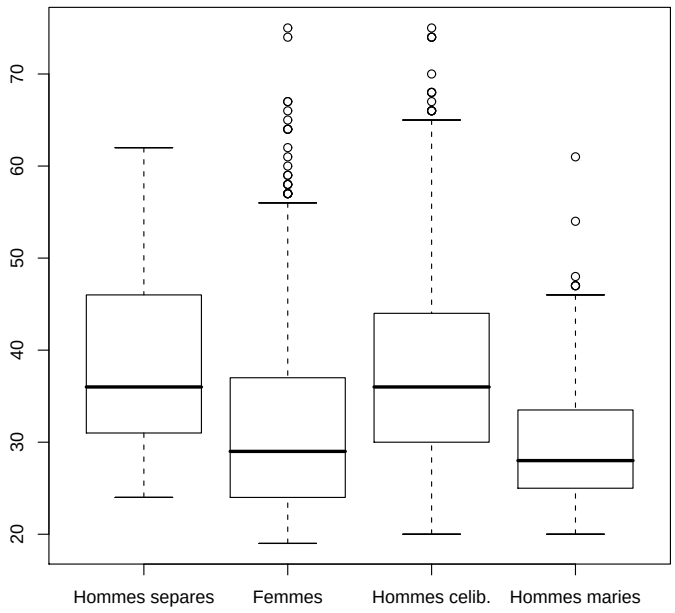


FIG. 3.1 – Boîte à moustaches illustrant la distribution des âges des clients d'une banque allemande suivant les différents statuts maritaux.

On note respectivement

$$\bar{Y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} \quad \text{et} \quad \bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} Y_{ij},$$

la moyenne empirique de l'échantillon correspondant au $j$ième niveau du facteur A et la moyenne empirique globale. De même, on définit la variance empirique au sein du $j$ième niveau de A par :

$$S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{\cdot j})^2.$$

3.2.2 Estimation des effets

Il est possible d'écrire le modèle d'analyse de variance comme un cas particulier de la régression multiple, en considérant une variable *indicatrice* pour chaque niveau du facteur. Le modèle s'écrit alors :

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{1}_1 + \ldots + \beta_J\mathbf{1}_J + \epsilon$$

où $\mathbf{Y} = (Y_{11}, \ldots, Y_{n_1 1}, Y_{12}, \ldots, Y_{n_2 1}, \ldots, Y_{1J}, \ldots, Y_{n_J J})'$ est le vecteur colonne des observations, $\mathbf{1}$ est une colonne de 1, $\mathbf{1}_j$ les variables indicatrices de niveau, et enfin ϵ le vecteur colonne des ϵ_{ij}. Ce modèle s'écrit encore

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

où $\mathbf{X} = (\mathbf{1}, \mathbf{1}_1, \ldots, \mathbf{1}_J)$ et $\beta = (\beta_0, \beta_1, \ldots, \beta_J)'$. Or, la matrice $\mathbf{X}$ n'est pas de plein rang (la première colonne est égale à la somme de toutes les autres). La matrice $\mathbf{X}'\mathbf{X}$ n'est donc pas inversible et le modèle admet une infinité de solution : on dit que les paramètres β_j ne sont donc pas *identifiables*.

Une solution est alors de considérer un sous-ensemble de variables indicatrices de sorte à rendre $\mathbf{X}'\mathbf{X}$ inversible. La façon la plus simple de faire est de ne pas considérer de terme constant :

$$\mathbf{Y} = \beta_1 \mathbf{1}_1 + \ldots + \beta_J \mathbf{1}_J + \epsilon.$$

On a alors $\beta_j = \mu_j$ ($1 \leq j \leq J$), et c'est le modèle considéré en (3.1).
Le paramètre μ_j est estimé sans biais par la moyenne empirique du $j$ième niveau :

$$\hat{\mu}_j \quad = \quad \bar{Y}_{.j},$$

tandis que σ^2 est estimée sans biais (sous l'hypothèse d'homogénéité des variances) par une moyenne pondérée des variances empiriques de chaque niveau :

$$\hat{\sigma}^2 \quad = \quad S^2 = \frac{1}{n-J}\sum_{j=1}^{J}(n_j - 1)S_j^2.$$

Le problème de ce modèle est que les tests découlant consisteront à étudier la nullité des paramètres tandis que nous sommes intéressés par tester leur égalité.

Une autre solution (*base cell model*, adoptée par SAS) et de considérer le modèle

$$\mathbf{Y} = \underbrace{\mu_J}_{\beta_0}\mathbf{1} + \underbrace{(\mu_1 - \mu_J)}_{\beta_1}\mathbf{1}_1 + \ldots + \underbrace{(\mu_{J-1} - \mu_J)}_{\beta_{J-1}}\mathbf{1}_{J-1} + \epsilon.$$

Ainsi, les paramètres β_j estimés seront des différences d'espérance, en adéquation avec ce que l'on cherche à tester par la suite.

3.2.3 Tests

Le principal objectif de l'analyse de variance est de tester si le facteur A a une influence sur la variable Y. Sous les hypothèses précédentes, le problème revient donc à tester

$$H_0 : \mu_1 = \ldots = \mu_J = \mu \qquad \text{contre } H_1 : \exists 1 \leq i, l \leq J \text{ t.q. } \mu_i \neq \mu_l.$$

On montre facilement la **formule d'analyse de variance** :

$$\underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{..})^2}_{SST} = \underbrace{\sum_{j=1}^{J} n_j(\bar{Y}_{.j} - \bar{Y}_{..})^2}_{SSA} + \underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{.j})^2}_{SSR}$$

qui représente la décomposition de la dispersion totale SST en la **dispersion** SSA **due au facteur** A (dispersion **inter**-groupe) et la **dispersion résiduelle** SSR (ou dispersion **intra**-groupe).

Exercice. Écrire la preuve.

En remarquant que $V_R^2 = \frac{SSR}{n} = \frac{1}{n}\sum_{j=1}^{J} n_j V_j^2$ où $V_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{.j})^2$, on montre que $\frac{n}{\sigma^2}V_R^2 = \frac{SSR}{\sigma^2} = \sum_{j=1}^{J}\frac{n_j V_j^2}{\sigma^2}$ suit une loi du χ^2 à $n - J$ degrés de liberté, car chaque $\frac{n_j V_j^2}{\sigma^2}$ suit une loi du χ^2 à $n_j - 1$ degrés de liberté.
De même, sous H_0 cette fois, $\frac{SST}{\sigma^2}$ suit une loi du χ^2 à $n-1$ degrés de liberté (car sous H_0 $\frac{SST}{n}$ est la variance d'un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$) et $\frac{SSA}{\sigma^2}$ suit une loi du χ^2 à $J - 1$ degrés de liberté (car $\frac{SSA}{n}$ peut être vue comme la variance *pondérée* du J-échantillon $(\bar{X}_1, \ldots, \bar{X}_J)$).
L'équation de l'analyse de variance revient alors à $\chi^2_{n-1} = \chi^2_{J-1} + \chi^2_{n-J}$, ce qui permet en outre de conclure via le théorème de Cochran (non abordé dans ce cours) que SSA et SSR sont indépendantes.
La statistique du test est donc

$$F = \frac{\frac{SSA}{J-1}}{\frac{SSR}{n-J}}$$

qui suit sous H_0 une loi de Fisher-Snedecor $F_{J-1,n-J}$, et on rejette l'hypothèse H_0 si la statistique F est supérieure au quantile de la loi $F_{J-1,n-J}$ d'ordre $1 - \alpha$.

Les résultats de l'analyse de variance sont généralement donnés dans un tableau analogue à celui-ci :

Source	Somme des carrés	degrés de liberté	carré moyen	F
Modèle (inter)	SSA	$J-1$	$MSA = SSA/(J-1)$	$F = \frac{MSA}{MSR}$
Erreur (intra)	SSR	$n-J$	$MSR = SSR/(n-J)$	
Total	SST	$n-1$		

Comparaison des moyennes deux à deux
Rejeter H_0 permet de dire que toutes les moyennes ne sont pas égales. Il peut cependant être intéressant de tester l'égalité des moyennes deux à deux.
Pour cela, on effectue un test de comparaison multiple des moyennes (pour $1 \leq j, j' \leq J$) :

$$H_0 : \mu_j = \mu_{j'}.$$

Étant donné le grand nombre de tests que l'on va être amené à faire, la problématique des tests multiples doit être prise en compte (cf. cours Statistique Inférentielle GIS3). Une solution simple peut être d'appliquer une correction de Bonferroni en réalisant chaque test avec un risque de première espèce égal au risque de première espèce global divisé par le nombre de tests effectués.
Une méthode plus conservative due à Scheffé, utilise le fait que

$$p\left(|\bar{X}_j - \bar{X}_{j'} - (\mu_j - \mu_{j'})| \leq S_R\sqrt{(J-1)f_{K-1,n-J,1-\alpha}}\sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}\right) = 1-\alpha$$

où $f_{J-1,n-J,1-\alpha}$ est le quantile de la loi de Fisher de paramètres $J-1$ et $n-J$ d'ordre $1-\alpha$.
On rejette donc l'hypothèse d'égalité des moyennes μ_j et $\mu_{j'}$ si

$$|\bar{X}_j - \bar{X}_{j'}| > S_R\sqrt{(J-1)f_{J-1,n-J,1-\alpha}}\sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}.$$

Remarque. *Attention, l'égalité des moyennes n'est pas transitive.*

3.2.4 Contrôle des hypothèses

Outre la normalité (que l'on peut vérifier classiquement), nous avons supposé l'homogénéité des variances, qu'il peut être intéressant de vérifier. Pour cela, sous l'hypothèse de normalité, Bartlett propose un test permettant de tester

$$H_0 : \sigma_1^2 = \ldots = \sigma_J^2 = \sigma^2 \qquad \text{contre } H_1 : \exists 1 \leq i, l \leq J \text{ t.q. } \sigma_i^2 \neq \sigma_l^2.$$

Posons

$$M = \sum_{j=1}^{J}(n_j - 1)\ln(S^2/S_j^2) \qquad \text{et} \qquad c = \frac{1}{3(J-1)}\left(\frac{\sum_{j=1}^{J}\frac{1}{n_j-1} - 1}{\sum_{j=1}^{J} n_j - 1}\right).$$

Sous H_0, la statistique

$$\frac{M}{c+1} \sim \chi^2_{J-1}$$

permet de réaliser le test.
Dans le cas où l'hypothèse de normalité est violée, une alternative proposée par Levene réalise une analyse de variance sur les variables $Z_{ij} = |Y_{ij} - \bar{Y}_{\cdot j}|$, la statistique de Fisher découlant de l'ANOVA fournissant un bon test de l'homogénéité des variances.

3.3 Analyse de variance à deux facteurs

On suppose désormais que Y est observé en présence de deux facteurs A et B, à respectivement J et K niveaux. En présence de plus d'un facteur, certains problèmes nouveaux apparaissent, parmi lesquels l'interaction entre facteurs. Nous supposons dans cette partie plusieurs hypothèses simplifiant les calculs :

- les niveaux d'un facteur ne sont pas conditionnés par l'autre facteur,
- pour chaque combinaison de facteur, on observe un même nombre (strictement supérieur à 1) de répétitions ($n_{jk} = c > 1$).

Les autres points seront abordés dans la section 3.4.

3.3.1 Le modèle

On note :

- Y_{ijk} la i-ème observation de Y pour les j-ème et k-ème valeurs respectives des facteurs A et B,
- $n_{jk} = c$ le nombre d'observations X_{ijk},
- $n_{j.} = \sum_{k=1}^{K} n_{jk} = Kc$, $n_{.k} = \sum_{j=1}^{J} n_{jk} = Jc$ et $n = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} = JKc$.

Le modèle d'ANOVA s'écrit alors

$$Y_{ijk} = \mu_{..} + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk}, \tag{3.2}$$

où $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, $\mu_{..}$ est l'effet général, α_j est l'effet du niveau j du premier facteur, β_k celui du niveau k de B, et γ_{jk} l'effet de l'interaction entre les niveaux j et k des deux facteurs.

Effet d'interaction L'effet d'interaction existe lorsque le niveau d'un facteur modifie l'influence de l'autre facteur sur Y. Considérons l'exemple suivant : on relève dans différentes villes françaises le taux de fumeur (Y) en fonction de la classe d'âge (facteur A) et du sexe (facteur B). En l'absence d'effet d'interaction, l'effet de la classe d'âge sur le taux de fumeurs serait identique pour les hommes et les femmes. Dans la réalité, il semble (cela reste à prouver par une ANOVA !) que les femmes fument en proportion beaucoup plus à un certain moment de leur vie (de l'adolescence au début de l'âge adulte), tandis que la répartition de fumeurs chez les hommes est plus constante entre les différentes classes d'âge. Ceci semble mettre en évidence un effet d'interaction entre les facteurs âge et sexe : le fait d'être de tel ou tel sexe modifie l'impact qu'à l'âge sur le taux de fumeurs.

3.3.2 Estimation des effets

On considère les moyennes empiriques suivantes :

$$\bar{Y}_{.jk} = \frac{1}{c}\sum_{i=1}^{c} Y_{ijk}, \qquad \bar{Y}_{..k} = \frac{1}{J}\sum_{j=1}^{J} \bar{Y}_{.jk}, \qquad \bar{Y}_{.j.} = \frac{1}{K}\sum_{k=1}^{K} \bar{Y}_{.jk} \quad \text{et} \quad \bar{Y}_{...} = \frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{c} Y_{ijk}.$$

Sous les hypothèses de contraintes (assurant l'unicité des solutions) $\sum_k \alpha_k = \sum_j \beta_j = \sum_k \gamma_{jk} = \sum_j \gamma_{jk} = 0$, les paramètres $\mu_{..}$, α_j, β_k et γ_{jk} de la décomposition (3.2) peuvent être estimés par les relations suivantes :

$$\hat{\mu}_{..} = \bar{Y}_{...}, \qquad \hat{\alpha}_j = \bar{Y}_{.j.} - \bar{Y}_{...}, \qquad \hat{\beta}_k = \bar{Y}_{..k} - \bar{Y}_{...} \quad \text{et} \quad \hat{\gamma}_{jk} = \bar{Y}_{.jk} - \bar{Y}_{.j.} - \bar{Y}_{..k} + \bar{Y}_{...}$$

3.3.3 Tests

Soient les sommes des carrés suivantes :

$$SST = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{c}(Y_{ijk} - \bar{Y}_{...})^2, \qquad SSA = cK\sum_{j=1}^{J}(\bar{Y}_{.j.} - \bar{Y}_{...})^2, \qquad SSB = cJ\sum_{k=1}^{K}(\bar{Y}_{..k} - \bar{Y}_{...})^2,$$

$$SSAB = c\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{Y}_{.jk} - \bar{Y}_{.j.} - \bar{Y}_{..k} + \bar{Y}_{...})^2, \quad \text{et} \quad SSR = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{c}(Y_{ijk} - \bar{Y}_{.jk})^2,$$

où SST est la somme des carrés totale, SSA est la somme des carrés relatifs au facteur A, SSB est la somme des carrés relatifs au facteur B, $SSAB$ est la somme des carrés relatifs à l'interaction entre les facteurs A et B et SSR est la somme des carrés résiduels.

En remarquant que que l'on peut écrire $SST = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{c} Y_{ijk}^2 - n\bar{Y}_{...}^2$, on obtient l'équation d'analyse de la variance à deux facteurs :

$$SST = SSA + SSB + SSAB + SSR.$$

Exercice. Écrire la preuve.

Comme en analyse de variance à un facteur, sous l'hypothèse $H_0 : \alpha_j = 0$, les quantités SSA et SSR suivent à σ^2 près des lois du χ^2 indépendantes à $J-1$ et $n-JK$ degrés de liberté. La statistique suivante est donc de loi de Fisher de paramètres $J-1$ et $K-1$:

$$F_A = \frac{SSA/(J-1)}{SSR/(n-JK)}.$$

De même, sous les hypothèses respectives $H_0 : \beta_k = 0$ et $H_0 : \gamma_{jk} = 0$, les statistiques

$$F_B = \frac{SSB/(K-1)}{SSR/(n-JK)} \quad \text{et} \quad F_{AB} = \frac{SSAB/(K-1)(J-1)}{SSR/(n-JK)}$$

suivent des lois de Fisher de paramètres $K-1$ et $n-JK$ pour F_B, $(K-1)(J-1)$ et $n-JK$ pour F_{AB}.
Ainsi, on peut donc tester l'existence des effets principaux des deux facteurs et de leur interaction en comparant ces statistiques aux quantiles de la loi de Fisher : si les valeurs observées de ces statistiques sont supérieures au quantile de la loi de Fisher d'ordre $1-\alpha$ on conclura à un effet significatif.
On présente usuellement l'analyse de variance sous la forme du tableau suivant :

Facteur	Somme des carrés	degrés de liberté	carré moyen	F
A	SSA	$J-1$	$SSA/(J-1)$	$F_A = \frac{SSA/(J-1)}{SSR/(n-JK)}$
B	SSB	$K-1$	$SSB/(K-1)$	$F_B = \frac{SSB/(K-1)}{SSR/(n-JK)}$
Interaction AB	$SSAB$	$(J-1)(K-1)$	$SSAB/(K-1)(J-1)$	$F_{AB} = \frac{SSAB/(K-1)(J-1)}{SSR/(n-JK)}$
Résidu	SSR	$n-JK$	$SSR/(n-JK)$	
Total	SST	$n-1$		

3.4 Problèmes spécifiques

3.4.1 ANOVA pour mesures répétées

Dans de nombreuses applications médicales, les mesures de Y sont réalisées plusieurs fois sur un même patient. Les répétitions ne sont plus indépendantes et la méthodologie classique n'est plus valide. L'idée consiste alors à introduire un facteur supplémentaire : un facteur individu. Ainsi, cela permet, en incorporant un effet sujet aléatoire, d'incorporer la corrélation intra-unité et de mieux estimer la résiduelle.

3.4.2 Plan sans répétition

Dans le cas où une seule observation est disponible pour chaque croisement de niveau, l'effet d'interaction est alors confondu avec l'effet résiduel et ne peut donc pas être évalué.

3.4.3 Plans déséquilibrés ou incomplets

Le cas de plans déséquilibrés (n_{jk} non constant) ou incomplets ($\exists j,k : n_{jk} = 0$) conduit à des modèles beaucoup plus compliqués, le cas $n_{jk} = c$ simplifiant grandement les calculs lors des décompositions des variances.

La solution consiste alors à écrire le modèle d'ANOVA comme un modèle de régression, de façon similaire à ce qui a été fait dans le cas de l'ANOVA à un facteur.
Ceci ne sera pas abordé dans ce cours, mais nous précisons néanmoins que la procédure `glm` de SAS permet de traiter ce cas (se référer aux résultats de `type III`).

3.5 Analyse de covariance

Nous cherchons à expliquer une variable quantitative Y en fonction de plusieurs variables explicatives, certaines qualitatives et d'autre quantitatives. L'idée générale sera de comparer pour chaque croisement de niveaux des variables qualitatives, le modèle de régression de Y sur les variables quantitatives.

Nous nous plaçons dans le cas d'un **unique facteur qualitatif** A, à J niveaux, et d'une **unique variable quantitative** X. La procédure `glm` de SAS permet de considérer des situations beaucoup plus complexes.

Pour chaque niveau j de A on observe les couples $(X_{ij}, Y_{ij})_{1\leq i\leq n_j}$. Soit $n = \sum_{j=1}^{J} n_j$ le nombre total d'observations.

3.5.1 Graphiques préliminaires

Comme pour l'ANOVA, une représentation graphique du nuage de points $(X_{ij}, Y_{ij})_{1\leq i\leq n_j, 1\leq j\leq n_J}$ en différenciant les couleurs pour chaque niveau du facteur permet de donner un premier avis permettant de guider l'analyse.

3.5.2 Le modèle

On considère un modèle de régression par niveau du facteur A :

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij} \qquad j = 1, \ldots, J \quad i = 1, \ldots, n_j \tag{3.3}$$

où ϵ_{ij} sont i.i.d. centrés de variance σ^2 et supposés de loi normale pour réaliser les tests.
La résolution simultanée des J modèles peut être obtenue en écrivant le système de façon matricielle :

$$\mathbf{Y} = \tilde{\mathbf{X}}\beta + \epsilon \tag{3.4}$$

avec les notations suivantes :
- $\mathbf{Y}$ et ϵ sont les vecteurs colonnes des Y_{ij} et ϵ_{ij},
- $\beta = (\beta_{01}, \beta_{11}, \ldots, \beta_{0J}, \beta_{1J})'$,
- $\tilde{\mathbf{X}}$ est la matrice $n \times 2J$ constituée des J blocs $[\mathbf{1}_j | \mathbf{X}.\mathbf{1}_j]$ où $\mathbf{1}_j$ est l'indicatrice de niveau, $\mathbf{X}$ est le vecteur colonnes des X_{ij} , et $\mathbf{X}.\mathbf{1}_j$ correspond au produit terme à terme des deux vecteurs.

Afin d'obtenir directement les bonnes hypothèses pour les tests que nous chercherons à effectuer, des logiciels comme SAS utilisent une reparamétrisation du modèle (3.4) faisant intervenir des effets différentiels par rapport au dernier niveau. Le modèle considéré s'écrit alors

$$\mathbf{Y} = \beta_{0J}\mathbf{1} + \underbrace{\beta_{1J}\mathbf{X}}_{\text{effet de } X} + \underbrace{(\beta_{01}-\beta_{0J})\mathbf{1}_1 + \ldots + (\beta_{0J-1}-\beta_{0J})\mathbf{1}_{J-1}}_{\text{effet de } A} + \underbrace{(\beta_{11}-\beta_{1J})\mathbf{X}.\mathbf{1}_1 + \ldots + (\beta_{1J-1}-\beta_{1J})\mathbf{X}.\mathbf{1}_{J-1}}_{\text{effet d'interaction}} \tag{3.5}$$

Nous pourrons alors tester directement :
- l'effet de $\mathbf{X}$ sur $\mathbf{Y}$,
- l'égalité des intercepts des J modèles de régression en testant l'effet de A,
- l'égalité des pentes des J modèles de régression en testant l'effet de l'interacton entre A et $\mathbf{X}$.

3.5.3 Tests

Des tests de Fisher peuvent être mis en place en comparant le modèle complet (3.5) à des modèles réduits n'intégrant que l'effet de X, que l'effet de A ou que l'effet d'interaction. Ces tests permettent de tester les trois hypothèses suivantes :

- $H_0^{(1)}$: $\beta_{11} = \ldots = \beta_{1J}$: il n'y a pas d'interaction, les pentes de la régression de Y sur X sont toutes identiques à celle β_{1J} du dernier niveau du facteur A,
- $H_0^{(2)}$: $\beta_{1J} = 0$,
- $H_0^{(3)}$: $\beta_{01} = \ldots = \beta_{0J}$: les ordonnées à l'origine de la régression de Y sur X sont toutes identiques à celle β_{0J} du dernier niveau du facteur A.

La démarche d'analyse de ces tests est la suivante :

- on commence par tester l'interaction avec $H_0^{(1)}$.
- si l'interaction n'est pas significative, on teste $H_0^{(2)}$, qui, s'il n'est pas non plus significatif, conduit à conclure à l'absence d'effet de X,
- toujours si $H_0^{(1)}$n'est pas significative, on teste $H_0^{(3)}$ pour juger de l'effet du facteur A.

3.6 TP 3 : Analyse de variance et de covariance

3.6.1 Analyse de variance à deux facteurs

A faire sous R.

Le fichier `milk.dat` contient les résultats d'une étude visant à évaluer l'impact sur la consommation de lait de quatre campagnes de publicité. Quatre villes, une par campagne, ont été choisies dans cinq régions différentes. Les données mesurent les consommations de lait (en $) après deux mois de campagne.

Le fichier comporte 6 colonnes (région, consommation pour la première campagne publicitaire, la deuxième, la troisième, la quatrième et taille de la famille).

Analyser cette étude en commençant par effectuer des représentations graphiques adéquates, puis en réalisant une ANOVA afin d'évaluer l'effet des différents facteurs présents dans cette étude.

3.6.2 Analyse de covariance

A faire sous SAS à l'aide de la `proc GLM`.

Nous considérons le même jeu de données que précédemment, mais en prenant en compte désormais la taille de la famille. L'objectif de l'étude est alors de tester l'impact des différentes campagnes publicitaires.

(i) A partir du fichier de données, construire un fichier à plat :

```
data milk1; set milk;
array c{4} consommation1-consommation4;
do pub=1 to 4;
consom=c{pub};
output;
end;
drop consommation1-consommation4;
run;
```

(ii) Réaliser une analyse de covariance étudiant l'impact de la taille de la famille et de la campagne publicitaire sur la consommation :

```
proc glm data=milk1 plot;
class pub;
model consom=pub taille pub*taille/ solution;
run;
```

Interpréter les différents effets.

(iii) Nous avons vu dans l'ANOVA à deux facteurs, que le facteur région avait un effet. Refaites l'analyse précédentes par région (on n'oublira pas de trier la table de données au préalable).

3.6.3 Analyse de variance à mesures répétées

A faire sous SAS.

Le fichier `health.dat` contient des données d'une étude sur l'impact du régime alimentaire sur les capacités physiques. Pour cela, on a mesuré le rythme cardiaque de 18 sportifs après des exercices d'échauffement, après un jogging léger et après une course à pied intense (respectivement PULSE1, PULSE2 et PULSE 3). Pour chaque personne, on a noté son régime alimentaire (DIET : 1 pour carnivore et 2 pour végétarien), ainsi que le type d'exercice qu'elle pratique habituellement (EXERTYPE : 1 pour aerobic (step), 2 pour tennis ou squash et 3 pour fitness).

(i) Créer un fichier à plat, qui contiendra entre autre une variable `ind` identifiant de l'individu et une variable `time` indiquant le numéro de la mesure effectuée (`time`=1,2 et 3 pour PULSE1, PULSE2 et PULSE 3).

(ii) Donner des représentations graphiques signifiantes (boxplot). Certains facteurs vous semblent-ils influencer le rythme cardiaque ?

(iii) Analyser l'impact des différents facteurs intervenant dans l'étude, à l'aide d'une proc mixed.

```
proc mixed data=health_plat;
class time EXERTYPE DIET ind;
model PULSE=EXERTYPE DIET EXERTYPE*DIET;
repeated time /subject=ind;
run;
```

Le modèle est-il significatif ? Si oui, quels effets sont significatifs ?

3.7 Un exemple d'application de l'ANOVA et l'ANCOVA

Le fichier `milk.dat` contient les résultats d'une étude visant à évaluer l'impact sur la consommation de lait de quatre campagnes de publicité. Quatre villes, une par campagne, ont été choisies dans cinq régions différentes. Les données mesurent les consommations de lait (en $) après deux mois de campagne au sein de plusieurs familles de tailles différentes.

Afin d'organiser le fichier sous une forme habituelle individus / variables, nous commençons pas créer un fichier à plat :

```
data milk1; set milk;
array c{4} consommation1-consommation4;
do pub=1 to 4;
consom=c{pub};
output;
end;
drop consommation1-consommation4;
run;
```

Nous réalisons ensuite une ANOVA à deux facteurs, campagne publicitaire et région, à l'aide de la commande suivante :

```
proc glm data=milk1 plot;
class region pub;
model consom=pub region pub*region;
run;
```

Les résultats obtenus sont les suivants :

Source	DF	Type III SS	Mean Square	F value	Pr>F
pub	3	4585.680487	1528.560162	3.61	0.0160
region	4	4867.511417	1216.877854	2.87	0.0268
region*pub	12	8937.917430	744.826453	1.76	0.0658

Ils indiquent un effet région et un effet campagne publicitaire (au risque 5%), alors que l'effet d'interaction est plus contrasté.

Intégrons désormais la variable taille de la famille à l'étude, et concentrons nous sur l'effet des campagnes publicitaires. La taille de la famille étant une variable quantitative, nous réalisons une ANCOVA :

```
proc glm data=milk1 plot;
class region pub;
model consom=pub taille pub*taille/ solution;
run;
```

L'option `solution` permet d'afficher les coefficients des modèles estimés (cf ci-après). Les résultats sont les suivants (on se réfère bien toujours aux résultats de type III) :

Source	DF	Type III SS	Mean Square	F value	Pr>F
pub	3	227.18067	75.72689	0.57	0.6377
taille	1	40926.01565	40926.01565	306.57	<.0001
taille*pub	3	309.84511	103.28170	0.77	0.5111

La seconde ligne indique qu'il y a un effet significatif de la taille. L'examen des valeurs des coefficients (tableau ci-dessous), montre qu'en effet la consommation augmente globalement de façon assez forte ($\beta \simeq 12$) avec la taille de la famille.

La première ligne indique qu'il n'y a pas de différence significative entre les intercepts des 4 modèles de régression de la consommation en fonction de la taille, ce qui ce traduit par l'absence d'effet campagne de publicité.

De même, la dernière ligne indique l'absence de différence significative entre les pentes des 4 modèles de régression de la consommation en fonction de la taille, ce qui ce traduit par l'absence d'interaction entre le type de campagne de publicité et la taille.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	8.27253333	4.81033834	1.72	0.0882
pub 1	-6.65546667	6.80284572	-0.98	0.3300
pub 2	-7.44426667	6.80284572	-1.09	0.2762
pub 3	-7.51253333	6.80284572	-1.10	0.2718
pub 4	0.00000000	.	.	.
taille	12.21651429	1.23518086	9.89	<.0001
taille*pub 1	-2.03891429	1.74680952	-1.17	0.2456
taille*pub 2	-1.12554286	1.74680952	-0.64	0.5207
taille*pub 3	-2.44765714	1.74680952	-1.40	0.1639
taille*pub 4	0.00000000			

La figure 3.2 représente les 4 modèles de régression correspondants aux 4 campagnes de publicités

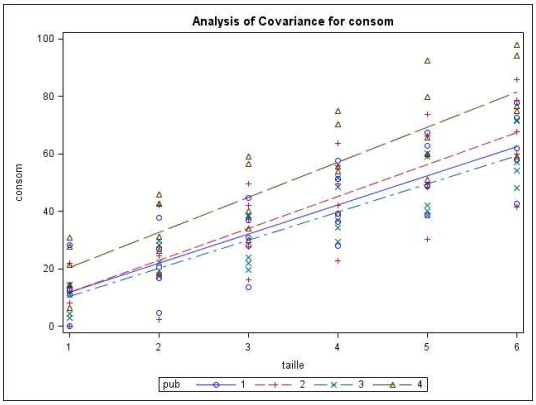


FIG. 3.2 – Régression de la consommation en fonction de la taille pour les différentes campagnes publicitaires

Néanmoins, étant donné l'effet région détecté dans l'analyse de variance, nous avons envie d'aller plus en avant dans l'analyse en réalisant la même ANCOVA mais région par région cette fois :

```
proc glm data=milk1 plot;
by region;
class pub;
model consom=pub taille pub*taille;
run;
```

On obtient alors les résultats suivants :

Région	Source	DF	Type III SS	Mean Square	F value	Pr>F
1	pub	3	72.029738	24.009913	4.62	0.0164
	taille	1	7178.321423	7178.321423	1380.25	<.0001
	taille*pub	3	217.370477	72.456826	13.93	<.0001
2	pub	3	231.734221	77.244740	30.36	<.0001
	taille	1	8655.252009	8655.252009	3402.34	<.0001
	taille*pub	3	50.150687	16.716896	6.57	0.0042
3	pub	3	79.546880	26.515627	6.01	0.0061
	taille	1	6993.301603	6993.301603	1585.35	<.0001
	taille*pub	3	173.193053	57.731018	13.09	0.0001
4	pub	3	415.666636	138.555545	15.23	<.0001
	taille	1	9743.378300	9743.378300	1071.32	<.0001
	taille*pub	3	361.395564	120.465188	13.25	0.0001
5	pub	3	15.354936	5.118312	0.79	0.5168
	taille	1	8513.285160	8513.285160	1314.71	<.0001
	taille*pub	3	52.751193	17.583731	2.72	0.0793

On constate alors, en réalisant les analyses région par région, que les différences d'intercept et de pentes sont toujours significatives (sauf pour la région 5 concernant l'intercept). Le type de campagne publicitaire influe donc

sur le lien entre la consommation et la taille. La figure 3.3 illustre les différences entre les différentes droites de régression.

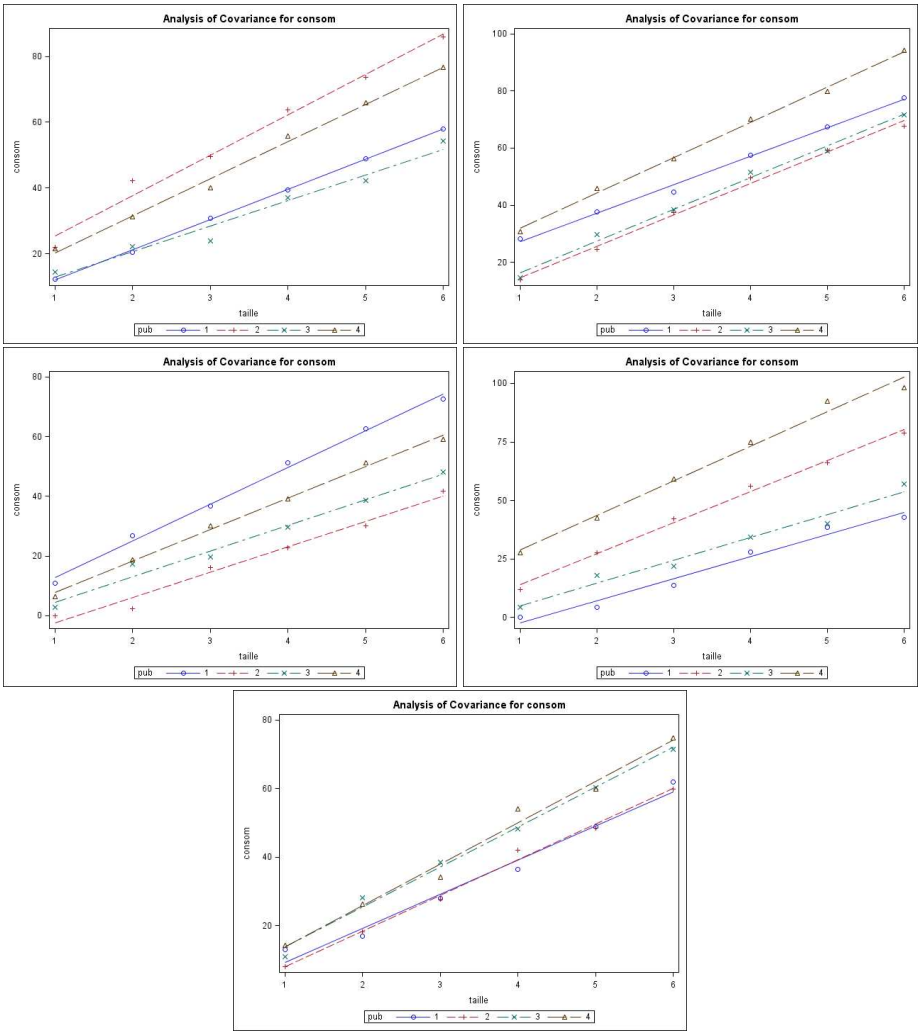


FIG. 3.3 – Régression de la consommation en fonction de la taille pour les différentes campagnes publicitaires, région par région.

L'analyse globale faite précédemment, prenant en compte toutes les régions ensemble, avait eu pour effet de cacher les différences d'influence des campagnes publicitaires, qui ne sont décelables qu'en concentrant l'étude région par région.

Chapitre 4

Régression logistique

*Logiciel **SAS** : proc* `logistic`.
*Logiciel **R** : fonction* `glm`.

La fin de ce cours est désormais consacré à modéliser une variable Y qualitative, à K modalités, à partir de p variables explicatives $\mathbf{X} = (X_1, \ldots, X_p)$ qualitatives ou quantitatives. On parle généralement dans ce cadre de classification (chaque modalité de Y représentant une classe d'individus). Nous verrons deux méthodologies, la régression logistique ainsi que l'analyse discriminante probabiliste (Chapitre 5).
Comme dans le reste de ce chapitre, nous supposons disposer d'un échantillon d'observations conjointes de Y et de $\mathbf{X}$: on parle alors d'apprentissage supervisé, et plus particulièrement ici de classification supervisée.

Nous supposons dans ce chapitre, pour simplicité de présentation, que les variables explicatives sont quantitatives. Dans le cas de variables qualitatives, il suffira de considérer les variables indicatrices correspondantes. Attention : par soucis d'identifiabilité, nous ne considérerons que $J-1$ indicatrices pour une variable à J modalités.

4.1 Le modèle logistique dichotomique (K=2)

On se place dans le cas où Y prend deux modalités (0 ou 1, présence ou absence d'une maladie, panne ou non d'un composant électronique, bon ou mauvais client...). Nous représenterons ces deux modalités par 0 et 1 dans la suite. La modalité 1 est généralement utilisée pour le caractère que l'on cherche à étudier (achat d'un produit, présence d'une maladie, panne...). Les modèles de régression vus précédemment ne s'appliquent plus puisque le régresseur linéaire habituel $\mathbf{X}\beta$ ne prend pas des valeurs simplement binaire.

4.1.1 Le modèle

L'idée est alors de ne plus modéliser Y, mais les probabilités d'avoir $Y = 0$ et $Y = 1$ conditionnellement à la connaissance des variables explicatives $\mathbf{X} = \mathbf{x}$:

$$\pi(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x}) \qquad \text{et} \qquad 1 - \pi(\mathbf{x}) = P(Y = 0|\mathbf{X} = \mathbf{x}).$$

Même si π n'est plus binaire, elle est toujours bornée dans l'intervalle $[0, 1]$, ce qui ne convient toujours pas à un régresseur linéaire $\mathbf{X}\beta$ qui prendra a priori des valeurs sur tout $\mathbb{R}$. La régression logistique consiste donc à modéliser une certaine transformation de π, appelée transformation *logit*, par une fonction linéaire des variables explicatives :

$$logit(\pi(\mathbf{x})) = \ln \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

Ce modèle s'écrit également

$$\pi(\mathbf{x}) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)}. \tag{4.1}$$

Dans la suite, nous noterons parfois $\pi(\mathbf{x};\beta)$ pour signifier que la probabilité $\pi(\mathbf{x})$ est paramétrée par β, et de même $P(Y=1|\mathbf{X}=\mathbf{x};\beta)$.

Remarque. *Justification du modèle : dans le cas d'une unique variable explicative X, on modélise la probabilité $\pi(x)=P(Y=1|X=x)$ par une fonction de la forme $\frac{\exp\beta x}{1+\exp\beta x}$ dont l'allure correspond bien à la représentation du nuage de point (x_i,y_i) dans le cas d'observation y_i binaire (cf Figure 4.1).*

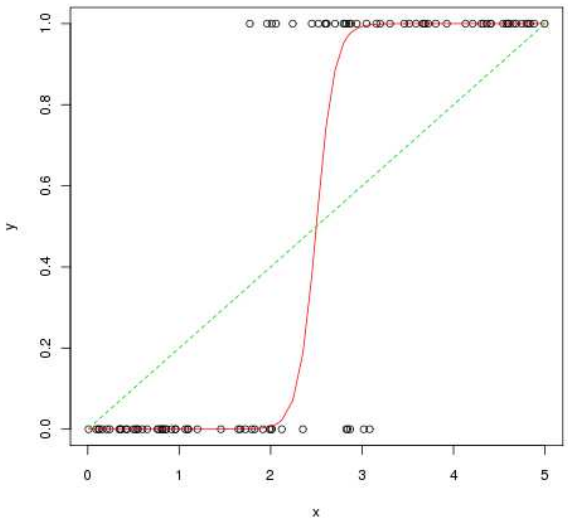


FIG. 4.1 – Modélisation de Y par une fonction de la forme $\frac{\exp\beta x}{1+\exp\beta x}$ (rouge) et par une fonction linéaire de x (pointillé vert).

4.1.2 Odds et odds-ratio

Le succès de la régression logistique, très utilisée en entreprise (finance, assurance, médecine, marketing...), est en partie dû aux capacités d'interprétabilité du modèle.
On définit par ***odds*** le rapport

$$\text{odds}(\mathbf{x})=\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}$$

qui représente combien de fois on a plus de chance d'avoir $Y=1$ au lieu d'avoir $Y=0$ lorsque $\mathbf{X}=\mathbf{x}$.
On définit de même les ***odds-ratio*** par le rapport

$$\text{odds-ratio}(\mathbf{x}_i,\mathbf{x}_j)=\frac{\text{odds}(\mathbf{x}_i)}{\text{odds}(\mathbf{x}_j)}$$

qui représente combien de fois on a plus de chance d'avoir $Y=1$ au lieu d'avoir $Y=0$ lorsque $\mathbf{X}=\mathbf{x}_i$ au lieu de $\mathbf{X}=\mathbf{x}_j$.

Remarque. *Bien que l'on ait défini les odds et odds-ratio pour une variable explicative $\mathbf{X}$ multidimensionnelle, on ne fait généralement varier qu'une seule dimension entre les deux valeurs $\mathbf{x}_i$ et $\mathbf{x}_j$, et on définit donc autant d'odds et odds-ratio qu'il y a de dimensions.*

Exemple On considère comme variable à prédire Y la présence ou l'absence d'un cancer des poumons, et comme variable explicative (qualitative) le fait d'être fumeur ou non fumeur. Les données sont fictives bien que pas si éloignées que cela de la réalité :

- La probabilité d'avoir un cancer du poumon chez un fumeur est $P(Y=1|X=\text{fumeur})=0.01$, d'où $P(Y=0|X=\text{fumeur})=0.99$. On a alors $\text{odds}(X=\text{fumeur})=1/99$. On dit que l'on a une *chance* sur 99 d'avoir un cancer des poumons lorsque l'on est fumeur.

– Chez les non fumeurs, la prévalence du cancer du poumons n'est que de $P(Y = 1|X = \text{non fumeur}) = 10^{-4}$. On a donc odds-ratio(fumeur, non fumeur) $= \frac{1/99}{1/9999} = 101$, d'où 101 fois plus de chance d'avoir un cancer des poumons pour un fumeur que pour un non fumeur.

4.2 Estimation des paramètres et prédiction

4.2.1 Estimation des β_j

Les paramètres à estimer sont $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$. Si on dispose d'un échantillon $(y_i, \mathbf{x}_i)_{i=1,n}$, où $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, telle que les y_i soient indépendants conditionnellement aux $\mathbf{x}_i$, on peut estimer β par maximum de vraisemblance. Les probabilités de Y étant exprimées conditionnellement aux variables explicatives $\mathbf{X}$, nous maximisons la vraisemblance conditionnelle :

$$L(\beta) = \prod_{i=1}^{n} P(Y = y_i|\mathbf{X} = \mathbf{x}_i).$$

Or, en utilisant la notation habituelle $\tilde{\mathbf{x}}_i = (1\mathbf{x}_i)'$, on a :

$$P(Y = y_i|\mathbf{X} = \mathbf{x}_i) = \begin{cases} \frac{\exp\beta'\tilde{\mathbf{x}}_i}{1+\exp\beta'\tilde{\mathbf{x}}_i} & \text{si } y_i = 1 \\ 1 - \frac{\exp\beta'\tilde{\mathbf{x}}_i}{1+\exp\beta'\tilde{\mathbf{x}}_i} & \text{si } y_i = 0 \end{cases} = \left(\frac{\exp\beta'\tilde{\mathbf{x}}_i}{1+\exp\beta'\tilde{\mathbf{x}}_i}\right)^{y_i} \left(1 - \frac{\exp\beta'\tilde{\mathbf{x}}_i}{1+\exp\beta'\tilde{\mathbf{x}}_i}\right)^{1-y_i}$$

d'où la log-vraisemblance

$$l(\beta) = \sum_{i=1}^{n} \ln P(Y = y_i|\mathbf{X} = \mathbf{x}_i) = \sum_{i=1}^{n} y_i\beta'\tilde{\mathbf{x}}_i - \ln(1 + \exp\beta'\tilde{\mathbf{x}}_i).$$

Exercice. Refaire le calcul.

La maximisation de cette vraisemblance se fait en dérivant par rapport au vecteur β. On obtient

$$\frac{\partial l(\beta)}{\partial\beta} = \sum_{i=1}^{n} y_i\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_i\frac{\exp\beta'\tilde{\mathbf{x}}_i}{1+\exp\beta'\tilde{\mathbf{x}}_i} = \sum_{i=1}^{n} \tilde{\mathbf{x}}_i(y_i - \pi(\mathbf{x}_i))$$

qui n'est pas une équation linéaire en β. Sa résolution peut être réalisée numériquement par un algorithme de type Newton-Raphson.
D'après les propriétés du maximum de vraisemblance, la matrice de variance de l'estimateur β est donnée par l'inverse de la matrice d'information de Fisher. Ainsi :

$$\hat{V}(\hat{\beta}) = \left[\frac{-\partial^2 l(\hat{\beta})}{\partial\beta^2}\right]^{-1} = (\tilde{\mathbf{X}}'\hat{V}\tilde{\mathbf{X}})^{-1} \tag{4.2}$$

où $\tilde{\mathbf{X}}$ est la matrice $n \times (p + 1)$ dont les lignes sont composées des $\tilde{\mathbf{x}}_i$ et $\hat{V}$ est la matrice diagonale $n \times n$ des $\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$.

4.2.2 Estimation des odds-ratio

Dans le cas d'une seule variable explicative X, on a

$$\begin{aligned} \ln \text{odds-ratio}(x_i, x_j) &= \ln \frac{\text{odds}(x_i)}{\text{odds}(x_j)} \\ &= logit(\pi(x_i)) - logit(\pi(x_j)) \\ &= \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x_j) \\ &= \beta_1(x_i - x_j), \end{aligned}$$

d'où $\widehat{\text{odds-ratio}}(x_i, x_j) = \exp(\hat{\beta}_1(x_i - x_j))$

4.2.3 Redressement dans le cas d'une modalité rare

Nous avons supposé que l'échantillon utilisé pour l'estimation respectait les proportions réelles des deux modalités (échantillonnage simple classique). Or il est très fréquent en pratique, lorsqu'une des deux modalités est rare (présence d'une maladie, client à risque...), d'utiliser un échantillonnage stratifié : on sur-représente artificiellement dans l'échantillon la modalité rare.
Cette modification du schéma d'échantillonnage n'a un impact que sur l'estimation de β_0, qu'il suffit alors de *redresser* en ajoutant le terme $\ln \frac{p_0}{p_1}$ où p_0 et p_1 sont les taux de sondage des modalités $Y = 0$ et $Y = 1$ (p_0 est donc le rapport de la probabilité d'avoir $Y = 0$ après ré-échantillonnage sur cette même probabilité dans la population initiale).

4.2.4 Prévisions

4.2.4.1 Classement d'une nouvelle observation

Pour une nouvelle observation $\mathbf{x}^*$, on cherche à prédire y^*. Il existe plusieurs façons d'effectuer la prédiction.
La règle du *maximum a posteriori* (MAP) consiste à affecter l'observation à la classe la plus probable : on prédit donc la valeur de y par la modalité k maximisant la probabilité $P(Y = k|\mathbf{X} = \mathbf{x}_i; \hat{\beta})$:

$$\hat{y}^*_{MAP} = \underset{k\in\{0,1\}}{\operatorname{argmax}} P(Y = k|\mathbf{X} = \mathbf{x}^*; \hat{\beta}).$$

Puisqu'on est en présence de deux classes, une observation sera classée dans la classe $Y = 1$ si sa probabilité d'être dans cette classe est supérieur à $1/2$. Or, ce choix est totalement arbitraire et peut être remis en cause, notamment lorsque les risques encourus en cas de mauvais classement ne sont pas symétriques (coûte-t-il aussi cher d'accepter un mauvais client que de ne pas en accepter un bon ?). On définira plus généralement la prédiction, ou règle de classement, au seuil s de la façon suivante :

$$\hat{y}^*_s = \begin{cases} 1 & \text{si } P(Y = 1|\mathbf{X} = \mathbf{x}^*; \hat{\beta}) \geq s \\ 0 & \text{sinon} \end{cases}$$

4.2.4.2 Notions de score

Dans de nombreux domaines, comme le *credit-scoring* ou la médecine, ce n'est pas tant la prédiction $\hat{y}^*$ qui nous intéresse que la probabilité $\pi(\mathbf{x}^*)$ que Y prenne la modalité 1. Cette probabilité est appelée *score*. Elle pourra représenter la probabilité qu'un client achète un produit, la probabilité pour un patient de contracter une maladie, etc.

4.2.4.3 Tableau de classement ou matrice de confusion

Le résultat d'un procédé de classification est souvent représenté sous la forme d'un tableau de classement (ou matrice de confusion) obtenu en appliquant la méthode de classification sur des observations pour lesquelles la variable Y (i.e. la classe d'appartenance) est connue et en comparant aux classes prédites :

		prédit		total
		$Y = 0$	$Y = 1$	
réel	$Y = 0$	VN	FP	N
	$Y = 1$	FN	VP	P
total		$\hat{\text{N}}$	$\hat{\text{P}}$	n

TAB. 4.1 – Matrice de confusion contenant les effectifs de vrais négatifs (VN), vrais positifs (VP), faux négatifs (FN) et faux positifs (FP)

Dans ce tableau figurent les effectifs des observations en fonction de leur classe réelle et de la prédiction de celle-ci. On parle parfois d'observations classées comme *positives* lorsqu'elles ont la modalité 1 de Y (car bien souvent on associe à la modalité $Y = 1$ le caractère que l'on cherche à détecter : maladie, achat...), et *négatives* dans le cas contraire. Avec ces appellations, le contenu des cases du tableau peut être décrit de la façon suivante :

- vrai négatif (VN) : nombre d'observations pour lesquelles la modalité 0 de Y a correctement été prédite,
- vrai positif (VP) : nombre d'observations pour lesquelles la modalité 1 de Y a correctement été prédite,
- faux négatif (FN) : nombre d'observations détectées à tort comme négatives,
- faux positifs (FP) : nombre d'observations détectées à tort comme positives,
- N,P,$\hat{\text{N}}$ et $\hat{\text{P}}$ respectivement les nombres de négatif et positif réels et prédits.

En général, les fréquences sous forme de pourcentage figurent également dans ce type de tableau.

Sensibilité et spécificité On appelle *sensibilité* du modèle le pourcentage de vrais positifs, et *spécificité* le pourcentage de vrais négatifs.

4.3 Tests, intervalles de confiance et choix de modèle

Nous présentons ici les tests permettant d'évaluer l'apport des différentes variables explicatives, ainsi que des intervalles de confiance, notamment sur les odds-ratio, utilisés dans l'interprétation du modèle logistique.

4.3.1 Tests sur β_j

On cherche à tester si une composante β_j du paramètre β est nulle :

$$H_0 : \beta_j = 0 \qquad \text{contre} \qquad H_1 : \beta_j \neq 0$$

Plusieurs tests sont disponibles :

- le *test du rapport des vraisemblances maximales* : sous H_0

$$-2 \ln \frac{\max_\beta L_{H_0}(\beta)}{\max_\beta L_{H_1}(\beta)} \to \chi^2_1$$

  où L_{H_0} et L_{H_1} sont respectivement les vraisemblances du modèle sans et avec la variable X_j,
- le *test de Wald* : sous H_0

$$\frac{\hat{\beta}_j^2}{\hat{\sigma}_j^2} \to \chi^2_1$$

  où $\hat{\sigma}_j^2$ est la variance de l'estimateur de β_j, donnée par (4.2),
- et enfin le *test du score*,

$$U(\hat{\beta}_{H_0})' \hat{V}(\hat{\beta}_{H_0}) U(\hat{\beta}_{H_0}) \to \chi^2_1$$

  où $\hat{V}(\hat{\beta}_{H_0})$ est l'inverse de la matrice d'information de Fisher, et $U(\hat{\beta}_{H_0})$ est le vecteur des dérivées partielles de la log-vraisemblance estimée sous H_0.

Pour tout ces tests, on rejettera l'hypothèse de nullité du coefficient β_j si la statistique du test est supérieure au quantile $\chi^2_{1,1-\alpha}$.

Remarque.

- *Si on conclut à la nullité d'un coefficient, tous les autres coefficients doivent être ré-estimés.*
- *Bien souvent, le test du rapport des vraisemblances est le plus puissant, mais nécessite l'estimation de β sous H_0, ce qui n'est pas le cas pour le test de Wald.*

4.3.2 Intervalles de confiance

Sachant que $\hat{\beta}_j$ est asymptotiquement distribué suivant une loi normale, centrée en β_j, et de variance donnée par (4.2), il est facile d'en déduire des intervalles de confiance asymptotiques sur les β_j.
En pratique, ces intervalles de confiance ne sont que peu souvent utilisés car les β_j ne sont que rarement interprétés, au contraire des odds-ratio. Les intervalles de confiance sur les odds-ratio sont construits à partir de résultats sur la normalité asymptotique du logarithme d'un odds-ratio.
Un intervalle de confiance sur un odds-ratio qui contient la valeur 1 ne permettra pas de conclure à un effet quelconque de la variable en question.

4.3.3 Choix de modèle

Comme pour tout modèle statistique, le principe général de ne pas évaluer un modèle sur les données qui ont servi à estimer le modèle doit être respecté.

4.3.3.1 Algorithme de sélection de variables

Comme en régression multiple, il existe des algorithmes de sélection (forward, backward, stepwise...) dont le principe est à chaque étape de comparer un modèle avec un sous-modèle et d'évaluer l'apport de termes supplémentaires. Le critère utilisé est généralement la statistique issue des tests de Wald ou du rapport des vraisemblances maximales.

4.3.3.2 Critères de choix de modèles

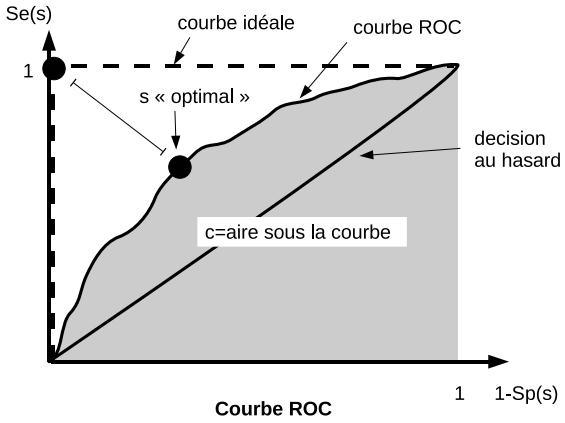
Les algorithmes de sélection de variables précédents favorisent la qualité d'ajustement du modèle. Afin de s'intéresser au pouvoir prédictif, d'autres critères classiques comme les critères AIC, BIC, ou de validation croisée peuvent être utilisés.
Ce dernier critère, dans le cas d'une validation croisée *leave-one-out*, s'écrit :

$$CV = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{y}_{(i)}=y_i}$$

où $\hat{y}_{(i)}$ est l'estimation de y_i obtenue sans utiliser la $i$ième observation $(y_i, \mathbf{x}_i)$.

4.4 Un outil d'interprétation : la courbe ROC

Nous avons défini précédemment les notions de sensibilité (pourcentage de vrais positifs) et spécificité (pourcentage de vrai négatif). La courbe ROC (*Receiver Operaor Characteristic curve*) donne l'évolution du taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (1-spécificité) lorsqu'on fait bouger le seuil s utilisé pour la prédiction.



Cette courbe permet de voir l'évolution des sensibilité et spécificité en fonction du seuil s choisi. Le praticien pourra alors choisir le seuil :

- à la main en fonction d'une sensibilité ou spécificité souhaitée,
- de façon à minimiser l'erreur totale de classement (sans différencier les FP et FN), c'est-à-dire le seuil s minimisant :

$$p_0(1 - Se(s)) + p_1(1 - Sp(s))$$

où $Se(s)$ et $Sp(s)$ sont les sensibilité et spécificité (en fonction du seuil s), et p_0 et p_1 sont les proportions de négatifs et de positifs dans la population totale,

– en cherchant à être le plus près possible du point idéal de coordonnées $(0, 1)$ ($Se = Sp = 1$), c'est-à-dire en minimisant :

$$(1 - Se(s))^2 + (1 - Sp(s))^2.$$

La courbe ROC permet également d'évaluer la qualité du modèle. Pour cela, on calcule l'aire sous cette courbe, notée AUC (*Area Under Curve*) :

$$AUC = \int_0^1 Se(s)d(1 - Sp(s)).$$

Le meilleur modèle sera celui qui se rapprochera le plus de l'AUC maximale égale à 1. Cette aire correspond à la probabilité de détecter un positif d'un négatif.

4.5 Le modèle logistique polytomique (K>2) et ordinal

Le modèle logistique présenté précédemment se généralise au cas d'une variable Y à K modalités (K>2). Lorsque ces dernières sont ordonnées on parle de régression logistique ordinale.
Notons $\pi_k(\mathbf{x}) = P(Y = k|\mathbf{X} = \mathbf{x})$. Dans cette situation, on se fixe une modalité de référence ($Y = K$ par exemple), et on réalise $K - 1$ régressions logistiques de $\pi_k(\mathbf{x})$ versus $\pi_K(\mathbf{x})$:

$$\ln \frac{\pi_k(\mathbf{x})}{\pi_K(\mathbf{x})} = \beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_j \qquad \forall 1 \leq k \leq K - 1.$$

Cette procédure ne dépend pas du choix du groupe de référence (dans les logiciels le groupe de référence est généralement soit le premier soit le Kième).
Lorsque la variable est ordinale, on modélise généralement des *logits cumulatifs* :

$$\ln \frac{\pi_{k+1}(\mathbf{x}) + \ldots + \pi_K(\mathbf{x})}{\pi_1(\mathbf{x}) + \ldots + \pi_k(\mathbf{x})} = \beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_j \qquad \forall 1 \leq k \leq K - 1.$$

Ce dernier modèle comportant un grand nombre de paramètres, les β_{jk} sont souvent supposés constants par classe $\beta_{jk} = \beta_j \quad \forall 1 \leq k \leq K - 1$.

4.6 TP 4 : Régression logistique

4.6.1 Simulation

A faire sous R.

Soit Y une variable binaire que l'on va chercher à prédire, X_1 et X_2 deux variables aléatoires uniformes sur $[-4, 5]$.

(i) Simuler un lien de type logit entre Y et (X_1, X_2) :

```
n = 100; a = -2; b = 2; c = 3
x1 = runif(n, -4, 5); x2 = runif(n, -4, 5)
y = exp(a*x1+b*x2+c + rnorm(n))
y = y/(1+y)
y = rbinom(n,1,y)
```

(ii) Représenter graphiquement le nuage de point formé par les variables explicatives, en représentant les points d'une couleur différente selon la modalité de Y. Représenter également Y en fonction de X_1, et en fonction de X_2.

(iii) Estimer le modèle de régression logistique à l'aide de la fonction `glm` :
`glm.res <- glm(y∼ x1+x2, family=binomial)`
Affichez et commentez les résultats à l'aide de la commande `summary(glm.res)` et `plot(glm.res)`. Analyser l'apport de chaque variable explicative.

(iv) Effectuer les prédictions de Y pour votre échantillon de simulation à l'aide de la commande :
`predict(glm.res,data.frame(x1=x1,x2=x2),type='response')`
et représenter les résultats à l'aide d'une matrice de confusion :
`table(ychap,y)`
Les prédictions seront réalisées à l'aide de la règle du MAP (règle *du seuil* avec $s = 0.5$)

(v) Simuler un nouvel échantillon de données de taille 100. Evaluer la sensibilité et la spécificité pour `s=seq(0,1,0.01)`. Tracer la courbe ROC (sous la forme d'une fonction en escalier).

(vi) Faites la même chose en utilisant une seule variable explicative dans le modèle logistique. Superposez les deux courbes ROC et choisissez le meilleur modèle.

4.6.2 Cancer du sein

A faire sous R.

Ce jeu de données classique est disponible dans le fichier `BreastCancer.dat`.
L'objectif est de prédire si la tumeur est maligne ou bénigne à partir de plusieurs variable explicatives.

(i) Découper aléatoirement le fichier en une partie apprentissage et une partie test, à l'aide de la fonction `sample`.

(ii) Estimer le modèle complet. Analyser l'apport de chaque variable explicative. Calculer le critère AIC à l'aide de la commande `summary`.
`glm.res <- glm(Class ∼ ., family=binomial, data=data_app)`.

(iii) Estimer un premier modèle simplifié en intégrant que les variables significative lors de la précédente régression ($\alpha = 5\%$). Calculer AIC.

(iv) Estimer un modèle simplifié à l'aide de l'algorithme forward suivant :

```
pr1.glm = glm(Class∼1,family=binomial,data=data_app)
pr1.step <- step(pr1.glm, direction="forward", scope=list(lower=∼1,
upper=∼Cl.thickness+Cell.size+Cell.shape+Marg.adhesion+Epith.c.size+Bare.nuclei+
Bl.cromatin+Normal.nucleoli+Mitoses), trace = TRUE)
```

Examiner l'ordre d'introduction des variables.

(v) Estimer un modèle simplifié à l'aide de l'algorithme forward/backward suivant :

```
pr2.glm = glm(Class∼1,family=binomial,data=data_app)
pr2.step <- step(pr2.glm, direction="both", scope=list(lower=∼1,
upper=∼Cl.thickness+Cell.size+Cell.shape+Marg.adhesion+Epith.c.size+Bare.nuclei+
Bl.cromatin+Normal.nucleoli+Mitoses), trace = TRUE)
```

Examiner l'ordre d'introduction des variables.

(vi) Quel est le meilleur des modèles, au sens de AIC ?

(vii) Et selon l'échantillon test ?

(viii) Tracer la courbe ROC pour chaque modèle. Quel est le meilleur ?

4.6.3 Cancer de la prostate

A faire sous SAS.

Les données sont dans le fichier `prostate.dat`.

Il y a encore quelques années, le traitement du cancer de la prostate dépendait de son extension au niveau des ganglions du système lymphatique. Afin d'éviter une intervention chirurgicale, des médecins ont cherché à prédire cette extension à partir de plusieurs variables explicatives : l'âge du patient, le niveau de *serum acid phosphatase*, le résultat d'une radiographie (0 : négatif, 1 : positif), la taille de la tumeur (0 : petite, 1 : grande), le résultat d'une biopsie (0 : moins sérieux, 1 : sérieux). En plus de ces variables, le jeu de données contient une dernière variable exprimant la contamination (1) ou non (0) du système lymphatique.

L'objectif de cet exercice est de trouver le meilleur modèle possible permettant de prédire la contamination du système lymphatique.

(i) Étudier graphiquement les liaisons entre les variables explicatives et la variable à expliquer.

(ii) Estimer le modèle complet, avec variables qualitatives et quantitatives, sans interaction.

```
proc logistic data=prostate;
class radio taille gravite lymph;
model lymph (REF=FIRST) = age acid radio taille;
run;
```

(iii) Interpréter les odds-ratio (`rapports de cotes` en français) obtenus. Si SAS ne les donne pas automatiquement, calculez-les à partir de l'estimation des coefficients du modèle de régression logistique.

(iv) Rechercher un modèle plus simple par la méthode stepwise. Pour cela, il suffit d'indiquer l'option `selection=stepwise` à la fin de l'instruction `model`.

(v) Comparer le modèle complet au modèle simplifié en fonction du pourcentage de biens classés.

(vi) Rechercher d'éventuels points particulièrement influents. La suppression du point le plus influent a-t-elle un effet favorable sur le pourcentage de bonnes classifications ?

Chapitre 5

Analyse discriminante probabiliste

*Logiciel **SAS** : proc* `discrim`*.*
*Logiciel **R** : fonction* `lda` *et* `qda` *du package MASS.*

Seconde méthode de classification supervisée abordée dans ce cours, l'analyse discriminante probabiliste a pour objectif d'affecter une observation $\mathbf{x}$ de $\mathbf{X} \in \mathbb{R}^p$ (le cas de variable qualitative peut également être traiter, cf Section 5.3) à une des K classes connues, que l'on notera $G_1, \ldots, G_K$, et qui correspondent aux modalités $1, \ldots, K$ de la variable Y.
L'objectif est donc identique à celui de la régression logistique, mais l'approche est différente. En régression logistique on modélise directement la probabilité $P(Y = k|\mathbf{X} = \mathbf{x})$, autrement dit la probabilité que l'observation $\mathbf{x}$ soit dans la classe G_k, tandis que l'analyse discriminante probabiliste consiste à modéliser la distribution de $\mathbf{X}$ conditionnellement à la classe.

5.1 Formalisme de la discrimination probabiliste

5.1.1 Définitions

Proportion d'une classe On note $p_k = P(Y = k)$ la probabilité qu'un individu a de provenir de la classe G_k. Cette probabilité est aussi appelée *proportion* de la classe G_k, et vérifie $\sum_{k=1}^{K} p_k = 1$.

Densité conditionnelle à une classe $\mathbf{X}$ a pour densité de probabilité $f_k(\mathbf{x})$ s'il provient de la classe G_k :

$$X_{|Y=k} \sim f_k(\mathbf{x}).$$

Densité marginale de $\mathbf{X}$ C'est une densité mélange

$$\mathbf{X} \sim \sum_{k=1}^{K} p_k f_k(\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}).$$

Probabilité conditionnelle La probabilité qu'une observation $\mathbf{x} \in \mathbb{R}^p$ provienne de la classe G_k est donnée par le théorème de Bayes :

$$t_k(\mathbf{x}) = P(Y = k|\mathbf{X} = \mathbf{x}) = \frac{p_k f_k(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}.$$

Remarque. *Nous supposons dans cette section que toutes les caractéristiques des lois sont connues : proportions, densités... Nous verrons dans la section suivante les méthodes d'estimations de ces quantités.*

5.1.2 Règle d'affectation et probabilité d'erreur

Une *règle d'affectation* (ou de classement, de décision...) r associe à une observation $\mathbf{x}$ une des K classes :

$$r : \mathbf{x} \in \mathbb{R}^p \rightarrow r(\mathbf{x}) \in \{1, \ldots, K\}.$$

La définition de r revient à partitionner $\mathbb{R}^p$ en K régions Ω_k telle que

$$\mathbf{x} \in \Omega_k \Leftrightarrow r(\mathbf{x}) = k.$$

La probabilité de classer un individu de G_k dans G_l ($l \neq k$) avec la règle r est :

$$e_{kl}(r) = P(r(\mathbf{X}) = l | Y = k) = \int_{\Omega_l} f_k(\mathbf{x}) d\mathbf{x}.$$

La probabilité qu'un individu de G_k soit mal classé avec la règle r est :

$$e_k(r) = P(r(\mathbf{X}) \neq k | Y = k) = \sum_{l \neq k} e_{kl}(r) = \int_{\bar{\Omega}_k} f_k(\mathbf{x}) d\mathbf{x}.$$

Et finalement la probabilité de mauvais classement global (ou erreur global de classement) :

$$e(r) = \sum_{k=1}^{K} p_k e_k(r).$$

5.1.3 Règle de classement optimale de Bayes

L'objectif est de définir la meilleure règle de classement possible.
On définit le *coût de mauvais classement* de classer un individu de G_k dans G_l :

$$C : (k, l) \in \{1, \ldots, K\} \times \{1, \ldots, K\} \rightarrow C(k, l) \in \mathbb{R}^+,$$

où par convention $C(k, k) = 0$.
Les fonctions de coût ne sont généralement pas symétriques. Comme nous l'avons déjà dit, classer un individu sain comme malade n'a pas le même coût que l'erreur inverse. Ces coûts seront à définir :

- avec le praticien en fonction de son expérience,
- en testant plusieurs systèmes de coûts possibles et en comparant les résultats obtenus,
- en les fixant tous à 1 lorsque l'on a aucune idée.

On définit le *risque conditionnel* associé à $\mathbf{x}$ par le coût moyen de classement :

$$R(r(\mathbf{X})|\mathbf{X} = \mathbf{x}) = E[C(r(\mathbf{X}), Y)|\mathbf{X} = \mathbf{x}] = \sum_{k=1}^{K} C(r(\mathbf{x}), k) t_k(\mathbf{x}),$$

et le *risque moyen* comme le coût moyen de classement inconditionnel

$$R(r) = E_{\mathbf{X}}[R(r(\mathbf{X})|\mathbf{X} = \mathbf{x})] = \sum_{k=1}^{K} p_k \sum_{l=1}^{K} C(l, k) \int_{\Omega_l} f_k(\mathbf{x}) d\mathbf{x}.$$

Exercice. Faire le calcul.

On cherche donc la règle de classement optimale r^* qui minimise le risque moyen, ce qui revient à minimiser le risque conditionnel pour chaque individu car :

$$R(r^*) = \min_r E_{\mathbf{X}}[R(r(\mathbf{X})|\mathbf{X} = \mathbf{x})] \geq E_{\mathbf{X}}[\min_r R(r(\mathbf{X})|\mathbf{X} = \mathbf{x})].$$

La règle optimale affecte donc $\mathbf{x}$ à G_k si

$$R(r(\mathbf{X}) = k|\mathbf{X} = \mathbf{x}) < R(r(\mathbf{X}) = l|\mathbf{X} = \mathbf{x}) \qquad \forall l \neq k.$$

Comme

$$R(r(\mathbf{X}) = k|\mathbf{X} = \mathbf{x}) = E[C(k, Z)|\mathbf{X} = \mathbf{x}] = \sum_{l=1}^{K} C(k, l)t_l(\mathbf{x}) = \sum_{l \neq k}^{K} C(k, l)t_l(\mathbf{x}),$$

la règle optimale de Bayes est :

$$r^*(\mathbf{x}) = k \qquad \text{si } \sum_{l \neq k}^{K} C(k, l)t_l(\mathbf{x}) < \sum_{l \neq k'}^{K} C(k', l)t_l(\mathbf{x}) \quad \forall k' \neq k.$$

Cas de l'égalité des coûts Si tous les coûts sont égaux à c, le risque conditionnel est alors

$$R(r(\mathbf{X}) = k|\mathbf{X} = \mathbf{x}) = c\sum_{l \neq k}^{K} t_l(\mathbf{x}) = c(1 - t_k(\mathbf{x})),$$

et donc $r^*(\mathbf{x}) = k$ si $c(1 - t_k(\mathbf{x})) < c(1 - t_{k'}(\mathbf{x})) \quad \forall k' \neq k$ ou encore

$$r^*(\mathbf{x}) = k \qquad \text{si } t_k(\mathbf{x}) < t_{k'}(\mathbf{x}) \quad \forall k' \neq k.$$

L'observation $\mathbf{x}$ est donc affectée à la classe conditionnellement la plus probable (règle du *maximum a posteriori*). Les coûts étant égaux, en posant $c = 1$, le risque moyen de classement

$$R(r) = \sum_{k=1}^{K} p_k \sum_{l \neq k} \int_{\Omega_l} f_k(\mathbf{x})d\mathbf{x} = \sum_{k=1}^{K} p_k \int_{\bar{\Omega}_l} f_k(\mathbf{x})d\mathbf{x} = \sum_{k=1}^{K} p_k e_k(r) = e(r)$$

est égal à l'erreur globale de classement.

Cas de deux classes On a

$$\begin{array}{lll} & r^*(\mathbf{x}) = 1 & \text{si} \quad C(1,2)t_2(\mathbf{x}) < C(2,1)t_1(\mathbf{x}), \\ \text{et} & r^*(\mathbf{x}) = 2 & \text{si} \quad C(2,1)t_1(\mathbf{x}) < C(1,2)t_2(\mathbf{x}), \end{array}$$

soit en posant $g(\mathbf{x}) = \frac{C(2,1)t_1(\mathbf{x})}{C(1,2)t_2(\mathbf{x})}$

$$\begin{array}{lll} & r^*(\mathbf{x}) = 1 & \text{si} \quad g(\mathbf{x}) > 1, \\ \text{et} & r^*(\mathbf{x}) = 2 & \text{si} \quad g(\mathbf{x}) < 1. \end{array}$$

L'équation de la *surface discriminante* (ou *frontière de classement*) est $g(\mathbf{x}) = 1$.

5.2 Discrimination paramétrique gaussienne

Lorsque les variables sont continues, une des lois les plus répandues est la loi gaussienne. Nous allons donc dans ce chapitre supposer que les variables explicatives $\mathbf{X}$ suivent des lois normales p-variées, dont les paramètres sont conditionnés par la classe k. Ainsi, la densité $f_k(\mathbf{x})$ du groupe G_k est :

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k)\}$$

où μ_k et Σ_k sont respectivement les espérance et variance de la classe k.

5.2.1 Règle de classement théorique

On se place ici dans le cas de 2 classes, la généralisation ne posant aucun problème.
L'équation de la surface discriminante est $g(\mathbf{x}) = 1$, ou encore $\ln g(\mathbf{x}) = 0$. On a :

$$\begin{aligned}
\ln g(\mathbf{x}) &= \ln \frac{C(2,1)p_1 f_1(\mathbf{x})}{C(1,2)p_2 f_2(\mathbf{x})} \\
&= \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} + \underbrace{\ln \frac{C(2,1)p_1}{C(1,2)p_2}}_{s} \\
&= \frac{1}{2}\left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} + (\mathbf{x}-\mu_2)'\Sigma_2^{-1}(\mathbf{x}-\mu_2) - (\mathbf{x}-\mu_1)'\Sigma_1^{-1}(\mathbf{x}-\mu_1) \right) + s.
\end{aligned}$$

Cette équation étant quadratique en $\mathbf{x}$, on dit que la frontière de classement est *quadratique*. On parle alors d'*analyse discriminante quadratique* (QDA).
Lorsque les matrices de variances sont identiques $\Sigma_1 = \Sigma_2 = \Sigma$ (cas *homoscédastique* par opposition au cas *hétéroscédastique* $\Sigma_1 \neq \Sigma_2$), l'équation de la surface discriminante est

$$(\mu_1 - \mu_2)'\Sigma^{-1}(\mathbf{x} - \frac{\mu_1 + \mu_2}{2}) + s = 0,$$

qui est une équation linéaire en $\mathbf{x}$. On dit que la frontière de classement est *linéaire* ou plus correctement que la séparation entre les classes est un hyperplan. On parle d'*analyse discriminante linéaire* (LDA).

5.2.2 Taux d'erreur théorique

On se place ici dans le cas de 2 classes, avec hypothèse d'homoscédasticité.
On affecte une observation $\mathbf{x}$ à la classe 1 (règle r) si $G(\mathbf{x}) = \ln g(\mathbf{x}) > 0$, ce qui est équivalent à

$$\begin{aligned}
& (\mu_1 - \mu_2)'\Sigma^{-1}(\mathbf{x} - \frac{\mu_1 + \mu_2}{2}) + s > 0 \\
\Leftrightarrow \quad & (\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) + s > 0.
\end{aligned}$$

La probabilité qu'un individu de G_2 soit mal classé avec cette règle est :

$$e_2(r) = P(G(\mathbf{X}) > 0 | Y = 2) = P(G(\mathbf{X}) > 0 | \mathbf{X} \sim \mathcal{N}(\mu_2, \Sigma)).$$

Il nous faut donc connaître la loi de $G(X)$ pour calculer cette probabilité. Or $G(X)$ est une combinaison linéaire de loi normale à une dimension (produit $(\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x}$) donc suit également une loi normale à une dimension, dont il nous suffit de calculer les moments.

$$\begin{aligned}
E[G(\mathbf{X})] &= (\mu_1 - \mu_2)'\Sigma^{-1}\mu_2 - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) + s \\
&= -\frac{1}{2}\underbrace{(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)}_{D^2} + s
\end{aligned}$$

où D^2 est la *distance de Mahalanobis* entre les deux classes. La variance est quant à elle

$$\begin{aligned}
V(G(\mathbf{X})) &= V((\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{X}) \\
&= (\mu_1 - \mu_2)'\Sigma^{-1}V(\mathbf{X})\Sigma^{-1}(\mu_1 - \mu_2) \\
&= (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \\
&= D^2
\end{aligned}$$

On a donc $G(\mathbf{X}) \sim \mathcal{N}(-D^2/2 + s, D^2)$ d'où

$$e_2(r) = 1 - \Phi\left(\frac{D}{2} - \frac{s}{D}\right)$$

où Φ est la fonction de répartition de la $\mathcal{N}(0,1)$.
On obtient de même

$$e_1(r) = \Phi\left(-\frac{D}{2} - \frac{s}{D}\right),$$

et on en déduit la probabilité globale d'erreur :

$$e(r) = p_1\Phi\left(-\frac{D}{2} - \frac{s}{D}\right) + p_2\left(1 - \Phi\left(\frac{D}{2} - \frac{s}{D}\right)\right).$$

Remarque. *Lorsque les coûts et les proportions sont égales, on obtient $e(r) = \Phi\left(-\frac{D}{2}\right)$, et donc plus les classes sont séparées, plus leur distance de Mahalanobis est grande, et plus l'erreur globale de classement est petite.*

5.2.3 Estimation de la règle de classement

On suppose qu'on dispose d'un échantillon $(\mathbf{x}_i, y_i)_{1\leq i\leq n}$ de réalisations indépendantes et identiquement distribuées.
A partir de cet échantillon on veut estimer le paramètre $\theta = (p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$ et en déduire l'estimation de la règle de classement r_θ (qui dépend de θ).
La méthode du maximum de vraisemblance peut être utilisée. De la vraisemblance :

$$L(\theta) = \prod_{k=1}^{K}\prod_{\mathbf{x}_i\in G_k} p_k f_k(\mathbf{x}_i),$$

on déduit la log-vraisemblance

$$l(\theta) = \sum_{k=1}^{K}\sum_{\mathbf{x}_i\in G_k} \ln p_k - \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\mathbf{x}_i-\mu_k)'\Sigma_k^{-1}(\mathbf{x}_i-\mu_k).$$

En dérivant puis égalant à 0 on obtient les estimateurs du maximum de vraisemblance suivant :

$$\begin{aligned}
\hat{p}_k &= \frac{n_k}{n} \qquad \text{où } n_k \text{ est le nombre d'observations de } G_k\\
\hat{\mu}_k &= \frac{1}{n_k}\sum_{\mathbf{x}_i\in G_k}\mathbf{x}_i,\\
\hat{\Sigma}_k &= \begin{cases} \hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{K}\sum_{\mathbf{x}_i\in G_k}(\mathbf{x}_i-\mu_k)'(\mathbf{x}_i-\mu_k) & \text{dans le cas homoscédastique}\\ \hat{\Sigma}_k = \frac{1}{n_k}\sum_{\mathbf{x}_i\in G_k}(\mathbf{x}_i-\mu_k)'(\mathbf{x}_i-\mu_k) & \text{dans le cas hétéroscédastique}\end{cases}
\end{aligned}$$

Les estimateurs de Σ_k étant biaisé, on en déduit les estimateurs sans biais suivants :

$$\begin{aligned}
\hat{\Sigma} &= \frac{1}{n-K}\sum_{k=1}^{K}\sum_{\mathbf{x}_i\in G_k}(\mathbf{x}_i-\mu_k)'(\mathbf{x}_i-\mu_k),\\
\hat{\Sigma}_k &= \frac{1}{n_k-1}\sum_{\mathbf{x}_i\in G_k}(\mathbf{x}_i-\mu_k)'(\mathbf{x}_i-\mu_k).
\end{aligned}$$

5.2.4 Estimation du taux d'erreur

Quelle que soit la méthode de classification utilisée, l'estimation de l'erreur de classement est intéressante puisqu'elle permet d'évaluer la qualité de la discrimination.

Taux d'erreur apparent $\hat{e}^a$ Cela consiste à appliquer la règle de classement sur les observations qui ont servies à estimer le modèle. On montre que cet estimateur est en général biaisé et optimiste ;

$$E_{\underline{Y}}[\hat{e}^a|\underline{X}] \leq E_{\underline{Y}}[e(r_{\hat{\theta}})|\underline{X}]$$

où $\underline{Y} = (Y_1, \ldots, Y_n)$.
Cet estimateur est donc à proscrire.

Méthode de la partition $\hat{e}^p$ Cela consiste à diviser l'échantillon en un échantillon d'apprentissage (environ $2/3$ à 75% de l'échantillon global) et un échantillon test. L'erreur de classement pourra alors être estimée sans biais sur l'échantillon test.

Remarque. *Cette technique demande une taille d'échantillon suffisamment grande.*

Méthode de la validation croisée $\hat{e}^{cv}$ On définit l'estimateur *validation croisée leave-one-out* de l'erreur par

$$\hat{e}^{cv} = \frac{1}{n}\sum_{i=1}^{n} \hat{e}^p_{(i)}$$

où $\hat{e}^p_{(i)}$ est l'évaluation de l'erreur sur une partition test constituée d'uniquement la $i$ième observation $(\mathbf{x}, y)_i$.
On parle de *validation croisée v-fold* lorsque l'échantillon initial est partagé en v sous-échantillons servant chacun tour à tour d'échantillon test tandis que le reste des échantillons est utilisé pour l'apprentissage.
On montre que l'on obtient un estimateur sans biais de l'erreur, ayant une variance plus faible que $\hat{e}^p$ avec une partition test réduite à une seule observation.

Remarque. *Cette technique demande de ré-estimer les paramètres pour chaque échantillon test considéré. Dans le cas de la validation croisée leave-one-out, les paramètres du modèle sont donc estimés n fois.*

Remarque. *Cette technique est à privilégier dans le cas de petits échantillons.*

5.2.5 Sélection de variables

Les taux d'erreurs précédents (sauf l'erreur apparente) peuvent être utilisés afin de choisir les variables intéressante (rappelons nous le principe biais-variance vu précédemment).
Afin d'éviter de comparer toutes les combinaisons de variables, on peut utiliser des algorithmes de sélection similaires à ceux utilisés en régression.

5.2.6 Choix de modèle

Il s'agît de choisir entre le modèle homoscédastique et hétéroscédastique.
On peut comme précédemment utiliser les taux d'erreurs, ou encore des critères classiques comme le critère BIC que l'on veut le plus petit possible :

$$BIC = -2\ln L(\hat{\theta}) + \nu \ln n$$

où μ est le nombre de paramètres du modèle :

$$\begin{aligned} \nu &= K - 1 + Kp + \frac{p(p+1)}{2} \qquad \text{dans le cas homoscédastique,} \\ \nu &= K - 1 + Kp + K\frac{p(p+1)}{2} \qquad \text{dans le cas hétéroscédastique.} \end{aligned}$$

5.3 Analyse discriminante pour variables qualitatives

L'analyse discriminante probabiliste peut facilement être étendue au cas de p variables explicatives qualitatives $\mathbf{X}(X_1, \ldots, X_p)$, à respectivement $m_1, \ldots, m_p$ modalités.
Pour cela, il suffit de considérer comme loi de $\mathbf{X}$ un modèle multinomiale complet, en définissant une probabilité pour chaque croisement de modalité possible. Ainsi, en notant $1, \ldots, m_j$ les modalités de la $j$ème variable, on a :

$$P(\mathbf{X} = \mathbf{x}) = P(X_1 = j_1, \ldots, X_p = j_p) = \alpha_{j_1,\ldots,j_p}$$

avec la contrainte $\sum_{j_1=1}^{m_1} \ldots \sum_{j_p=1}^{m_p} \alpha_{j_1,\ldots,j_p} = 1$.

5.4 Mise en oeuvre informatique

5.4.1 SAS : `PROC DISCRIM`

La procédure `DISCRIM` de SAS permet d'effectuer une analyse discriminante linéaire ou quadratique. La syntaxe est la suivante :

```
PROC DISCRIM DATA=...CROSSVALIDATE OUTSTAT=Dis_Func POOL = TEST;
CLASS indiquer ici la variable définissant les classes ;
PRIORS PROPORTIONAL;
VAR indiquer ici les variables explicatives ;
RUN;
```

L'option `CROSSVALIDATE` donne une estimation du taux d'erreur par validation croisée.
L'option `OUTSTAT=Dis_Func` permet de sauvegarder la fonction discriminante dans le but de classer de futures observations.
L'option `POOL = TEST` permet de tester entre l'égalité des matrices de variances, et donc de choisir entre une analyse discriminante quadratique ou linéaire. Pour imposer l'utilisation d'une analyse discriminante linéaire, indiquer `POOL = YES` (option par défaut), et pour l'analyse discriminante quadratique il faut indiquer `POOLED = NO`.
L'instance `PRIORS PROPORTIONAL` conduit à estimer les proportions des classes. Il est possible de les imposer à être égale grâce à `PRIORS EQUAL` (option par défaut).

Pour ensuite classer de nouvelles observations, il faut procéder de la façon suivante :

```
PROC DISCRIM DATA=Dis_Func TESTDATA=NEWDATA TESTLIST;
CLASS indiquer ici la variable définissant les classes dans Dis_Func ;
RUN;
```

L'option `DATA=Dis_Func` utilise la fonction discriminante précédemment estimée pour classer les individus spécifié dans `TESTDATA=NEWDATA`.
L'option TESTLIST affiche chaque nouvelle observation ainsi que sa classe estimée.

5.4.2 R : fonctions `lda` et `qda` du package MASS

Tout est dans le titre ! L'aide de ces deux fonctions (`lda` pour *linear discriminant analysis* et `qda` pour *quadratic discriminant analysis*) est très bien faite sous R.

5.5 TP 5 : Analyse discriminante probabiliste

5.5.1 Simulation

A faire sous R.

Soit $Y \sim \mathcal{B}(1/2)$ une variable binaire que l'on va chercher à prédire, X_1 et X_2 deux variables aléatoires gaussiennes, dont les paramètres des lois dépendent de la modalité de Y, et que l'on va utiliser pour prédire Y :

- $X_1 \sim \mathcal{N}(1,1)$ et $X_2 \sim \mathcal{N}(3,1)$ si $Y = 0$,
- $X_1 \sim \mathcal{N}(2,1)$ et $X_2 \sim \mathcal{N}(2,1)$ si $Y = 1$.

(i) Simuler un échantillon de taille $n = 100$ de réalisations du triplet (Y, X_1, X_2).

(ii) Représenter graphiquement le nuage de point formé par les variables explicatives, en représentant les points d'une couleur différente selon la modalité de Y. Représenter également les distributions marginales de X_1 et X_2 (histogramme et estimation non paramétrique de la densité).

(iii) Estimer les paramètres du modèle LDA (proportions, moyennes et matrices de variance de chaque classe). Vérifiez ensuite vos résultats à l'aide de la fonction `lda` :

```
library('MASS')
lda1=lda(y~x)
plot(lda1)
```

(iv) Simuler un échantillon *test* de taille 100. Prédire la variable Y par la règle du maximum a posteriori, et évaluer le taux de bon classement.

(v) Evaluer le taux de bon classement par validation croisée leave-one-out : `lda1_CV=lda(y~x,CV=TRUE)`

```
table(y,lda1_CV$class)
```

(vi) Faites de même avec QDA. Comparer les deux modèles (selon le taux d'erreur sur l'échantillon test et la validation croisée). Commentaires ?

(vii) A l'aide de l'échantillon test, comparer aux résultats de la régression logistique. Les prédictions à l'aide des modèles LDA/QDA se font également à l'aide de la fonction `predict` :

```
pred=predict(lda1,data.frame(x),type='response')
```

5.5.2 Iris

A faire sous R.

Le fichier de données *iris* est disponible sous R. Ce célèbre jeu de données donne les mesures en centimètres des longueur et largeur des sépales et des longueur et largeur des pétales pour 150 fleurs réparties en trois espèces d'iris.

(i) En croisant les variables explicatives deux à deux, représenter les nuages de point avec des couleurs différentes selon les espèces.

```
plot(iris[,1 :4],col=iris$Species)
```

Certaines variables semblent-elles plus discriminantes que d'autres ?

(ii) Calculer les matrices de variance de chaque groupe. Sont-elles semblables ?

(iii) Estimer les modèles QDA et LDA utilisant les 4 variables.

(iv) Calculer les taux d'erreurs de classement par validation croisée leave-one-out. Quel est le meilleur modèle ?

(v) Estimer maintenant les deux modèles QDA et LDA sous SAS à l'aide de la `proc discrim`. Existe-t-il des procédures pré-définies permettant de sélectionner les variables ?

Chapitre 6

Annexes

6.1 Dérivées de matrice et de vecteurs

Nous donnons ici quelques formules de dérivée par rapport à un vecteur ou une matrice, sachant que
- la dérivée d'un réel x par rapport à un vecteur $\mathbf{a}$ est un vecteur dont les composantes sont les dérivées de x par rapport aux composantes de $\mathbf{a}$: $\left(\frac{\partial x}{\partial \mathbf{a}}\right)_i = \frac{\partial x}{\partial a_i}$,
- inversement $\left(\frac{\partial \mathbf{a}}{\partial x}\right)_i = \frac{\partial a_i}{\partial x}$,
- et $\left(\frac{\partial \mathbf{a}}{\partial \mathbf{a}}\right)_{ij} = \frac{\partial a_i}{\partial b_i}$.

Soient $\mathbf{a}$ et $\mathbf{x}$ deux vecteurs :

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{a}'\mathbf{x} = \frac{\partial}{\partial \mathbf{x}}\mathbf{x}'\mathbf{a} = \mathbf{a}$$

Soit A et B deux matrices :

$$\begin{aligned}
\frac{\partial}{\partial A}Tr(AB) &= B' \\
\frac{\partial}{\partial A}Tr(A'B) &= B \\
\frac{\partial}{\partial A}Tr(A) &= I \\
\frac{\partial}{\partial A}Tr(ABA') &= A(B+B') \\
\frac{\partial}{\partial A}\ln|A| &= (A^{-1})'
\end{aligned}$$

où Tr est la trace de la matrice.

6.2 Lois de probabilités

6.2.1 Loi multinomiale

On répète n fois une expérience à K résultats possibles, de probabilités $p_1, \ldots, p_K$ ($\sum_{k=1}^{K} p_k$).
On appelle $\mathbf{Y}$ le vecteur de dimension K tel que sa $k$ième composante Y_k soit égale au nombre de résultats d'expériences ayant conduit au résultat k.
Alors $\mathbf{Y}$ suit une loi multinomiale d'ordre n de paramètres $p_1, \ldots, p_K$

$$Y \sim \mathcal{M}(n, p_1, \ldots, p_K).$$

La probabilité d'avoir $Y = (y_1, \ldots, y_K)$ est

$$P(Y = y_1, \ldots, y_K) = \frac{n!}{y_1! \ldots y_K!} p_1^{y_1} \ldots p_K^{y_K}.$$

Son espérance est

$$E[Y] = (np_1, \ldots, np_K),$$

et sa matrice de variance $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$ définie par :

$$\begin{aligned} \sigma_{ii} &= np_i(1-p_i), \\ \sigma_{ij} &= -np_ip_j \quad \text{si } i \neq j. \end{aligned}$$

6.2.2 Loi gaussienne multivariée

Une variable aléatoire $\mathbf{X} \in \mathbb{R}^p$ de loi normale p-variée, d'espérance $\mu \in \mathbb{R}^p$ et de matrice de variance Σ, a pour densité de probabilité

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\}$$

où $|\Sigma|$ est le déterminant de la matrice Σ.

Bibliographie

[1] P. Besse. *Pratique de la modélisation statistique*, Publications du Laboratoire de Statistique et Probabilités, 2003.
Disponible sur `http ://www.math.univ-toulouse.fr/~besse/pub/modlin.pdf`

[2] P. Besse. *Apprentissage Statistique & Data mining*, Publications du Laboratoire de Statistique et Probabilités, 2009.
Disponible sur `http ://www.math.univ-toulouse.fr/~besse/pub/Appren_stat.pdf`

[3] G.J. McLachlan. *Discriminant analysis and Statistical Pattern Recognition*. Wiley, New-York, 1992.

[4] J-P. Nakache et J. Confais. *Statistique explicative appliquée*. Editions Technip, 2003.

[5] G. Saporta. *Probabilités, analyse de données et statistique*. 2ème édition, Editions Technip, 2006.