Régression linéaire et généralisée

Frédéric Lavancier

ENSAI 2A

2024/2025

Table des matières

Références

- "Régression avec R", P-A. Cornillon, E. Matzner-Løber
  $\longrightarrow$ *Livre en français, très accessible, en lien avec les 3 premiers chapitres*

- "Le modèle linéaire par l'exemple", J.-M. Azais, J.-M. Bardet.
  $\longrightarrow$ *Livre en français, en lien avec les 3 premiers chapitres : des discussions intéressantes sur l'enjeu des hypothèses, et des résultats théoriques fins.*

- "An introduction to statistical learning with applications in R", G. James, D. Witten, T. Hastie, R. Tibshirani.
  $\longrightarrow$ *Grand classique sur les méthodes de machine learning, y compris les méthodes vues dans ce cours. Exemples avec R.*

- ESL : "The elements of statistical learning", T. Hastie, R. Tibshirani, J. Friedman.
  $\longrightarrow$ *Grand classique également. Version plus théorique (et plus complète) que le précédent.*

- Agresti, A. Foundations of Linear and Generalized Linear Models, Wiley.
  $\longrightarrow$ *Livre classique sur le sujet, en lien avec le chapitre 4*

- Antoniadis, A. Berruyer J. et Carmona R. Régression non linéaire et applications, Economica.
  $\longrightarrow$ *Résultats théoriques complets, en lien avec le chapitre 4*

- Dobson, A.J., Barnett, A.G. An Introduction to Generalized Linear Models, CRC Press.
  $\longrightarrow$ *Des exemples en R, en lien avec le chapitre 4*

- Hosmer, D. et Lemeshow S. Applied Logistic Regression, Wiley.
  $\longrightarrow$ *La régression logistique en applications, en long et en large*

Chapitre 1

Introduction

Ce cours traite des modèles élémentaires de régression. L'objectif est d'expliquer une grandeur Y en fonction de p grandeurs $X^{(1)}, \ldots, X^{(p)}$. Pour cela on dispose des observations de chacune de ces grandeurs auprès de n individus.

Exemples :
• Y : consommation électrique quotidienne en France
$X = X^{(1)}$: température moyenne journalière.
Données : un historique de Y et X sur n jours.
Question : a-t-on $Y \approx f(X)$ pour une certaine fonction f ?
En simplifiant : a-t-on $Y \approx aX + b$ pour certaines valeurs a et b ?
Si oui, que valent a et b ? La relation est-elle "fiable" ?
Il s'agit d'un modèle de régression linéaire.

• Y : qualité d'un client ("*bon*" ou "*pas bon*")
$X^{(1)}$: revenu du client
$X^{(2)}$: catégorie socio professionnelle (6-7 possibilités)
$X^{(3)}$: âge
Données : n clients.
On modélise dans ce cas $p = \mathbb{P}(Y = \text{"}bon\text{"})$.
A-t-on $p \approx f(X^{(1)}, X^{(2)}, X^{(3)})$ pour une fonction f à valeurs dans $[0, 1]$?
Cela rentre dans le cadre des modèles de régression généralisée.

La relation approximative que l'on cherche à établir entre Y et $X^{(1)}, \ldots, X^{(p)}$ est un **modèle**.

Pourquoi chercher à établir un tel modèle ? Deux raisons principales :

- Objectif descriptif : quantifier l'effet marginal de chaque variable. Par exemple, si $X^{(1)}$ augmente de 10%, comment évolue Y ?
- Objectif prédictif : étant données des nouvelles valeurs pour $X^{(1)}, \dots, X^{(p)}$, prédire la grandeur Y (approximative) associée.

Dans le cas particulier où $p = 1$, il s'agit de décrire la relation entre 2 variables : Y et $X = X^{(1)}$. Il s'agit donc d'une analyse bivariée. Nous rappelons dans la partie suivante les éléments principaux d'une telle analyse (vus dans le cours de statistique descriptive de 1A). Elle constitue l'étape préliminaire avant toute modélisation entre Y et X. De façon générale ($p > 1$), il est indispensable d'effecter une analyse descriptive préalable à toute modélisation, lorsque cela est possible, pour mettre en évidence la nature des liens éventuels (linéaires, non-linéaires, forts, faibles,...) : cela peut s'appuyer sur des analyses bivariées entre les variables Y et $X^{(j)}$, pour $j = 1, \dots, p$, ou sur des outils multidimensionnels tels que l'ACP.

Comme cela apparaît dans la partie suivante consacrée à l'analyse bivariée, les outils sont différents selon la nature des variables considérées, que ces dernières soient quantitatives ou qualitatives. Cette distinction sera également valable dans la manière de modéliser les relations entre les variables, comme nous le verrons dans la partie 1.2 plus générale ci-dessous, ainsi que dans la structure même du cours.

Nous distinguerons en effet les approches suivantes :

- Régression linéaire : modélisation de Y quantitative dans $\mathbb{R}$ (exemple typique : Y suit une loi Gaussienne), en fonction de $X^{(1)}, \dots, X^{(p)}$ quantitatives également
- Analyse de la variance et de la covariance (ANOVA et ANCOVA) : généralisation du cas précédent à la situation où une ou plusieurs des variables $X^{(1)}, \dots, X^{(p)}$ sont qualitatives.
- Modèles linéaires généralisés : modélisation d'une variable Y qui est qualitative, ou qui prend des valeurs particulières (par exemple uniquement entières), en fonction de $X^{(1)}, \dots, X^{(p)}$ quantitatives et/ou qualitatives.

1.1 Analyse bivariée (rappels)

On considère deux variables Y et X. On rappelle ci-dessous les outils d'analyse bivariée classiques, selon que Y est quantitatif ou qualitatif, et de

même pour X. Ces outils visent à identifier si un lien semble présent entre ces 2 variables, avant une éventuelle modélisation de ce dernier. Pour chaque situation, trois éléments seront rappelés :

1. Comment visualiser graphiquement le lien entre les deux variables ?
2. Comment le quantifier ?
3. Comment tester sa significativité ?

La description ci-dessous est sommaire. Plus de détails sont présents dans les slides accompagnant ce cours. Voir également le cours de "Statistique Descriptive" de 1A. Les justifications théoriques, notamment concernant la validité des tests présentés, feront partie de la suite du cours, en tant que cas particuliers de situations plus générales.

1.1.1 Lien quanti-quanti

On note $(x_1, y_1) \ldots, (x_n, y_n)$ les valeurs observées du couple de variables quantitatives (X, Y).

1. Comment visualiser graphiquement le lien entre les deux variables ?
   On visualise le lien entre X et Y grâce au nuage des points (x_i, y_i). Si la forme du nuage est étirée, cela nous indique la présence d'un lien linéaire entre les variables. Une autre structure peut témoigner d'un lien non linéaire.
2. Comment quantifier le lien ?
   On calcule la corrélation linéaire de Pearson :

$$\hat{\rho} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2 \ \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y}_n)^2}}$$

   en notant *var* et *cov* la variance et la covariance empirique, et où $\bar{x}_n$ (resp. $\bar{y}_n$) désigne la moyenne empirique de X (resp. Y).

   La corrélation $\hat{\rho}$ est toujours comprise entre -1 et 1. Si $\hat{\rho} = 1$, il y a un lien linéaire "parfait" positif, i.e., $\hat{\rho} = 1$ ssi il existe $a > 0$ et b tels que $y_i = ax_i + b$ pour tout $i = 1, \ldots, n$. Si $\hat{\rho} = -1$, il y a un lien linéaire "parfait" négatif (idem avec $a < 0$). Si $\hat{\rho} = 0$, il n'y a aucun lien linéaire (mais il peut exister un lien non-linéaire).

   Sous R : fonction `cor`.

3. Comment tester la significativité du lien ?

   $\hat{\rho}$ est un estimateur de la corrélation théorique ρ entre X et Y défini par

$$\rho = \frac{\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)]}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}.$$

   On peut tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$.

   Si (X, Y) est Gaussien et que les observations sont i.i.d, on peut montrer que $T \sim St(n-2)$ sous H_0, où

$$T = \sqrt{n-2}\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}$$

   et $St(n-2)$ désigne la loi de Student à $n-2$ degrés de liberté. On en déduit la région critique au niveau $\alpha \in [0, 1]$ pour ce test :

$$RC_\alpha = \{|T| > t_{n-2}(1 - \alpha/2)\}.$$

   Si (X, Y) n'est pas Gaussien, cette région critique reste valable asymptotiquement, moyennant des hypothèses très faibles sur la loi de (X, Y).

   Sous R : fonction `cor.test`.

Lorsqu'un lien linéaire semble apparent, on peut chercher à estimer la droite des moindres carrés, cf la figure 1.1.1. Il s'agit de la droite qui passe "le mieux" au milieu des points (x_i, y_i), au sens où la somme des distances en rouge prises au carré est minimale. Il s'agit de la régression linéaire de Y sur X. L'équation de la droite recherchée est donc $y = \hat{a}x + \hat{b}$ où $\hat{a}$ et $\hat{b}$ vérifient :

$$(\hat{a}, \hat{b}) = \underset{(a,b)}{argmin} \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

On trouve (à savoir faire, et résultat à connaître !), si $var(X) \neq 0$:

$$\hat{a} = \frac{cov(X, Y)}{var(X)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

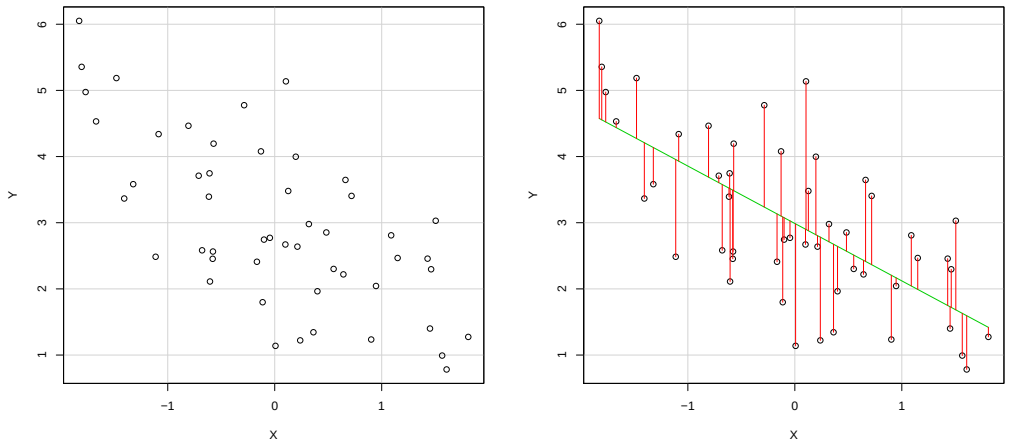

FIGURE 1.1 – Exemple d'un nuage de points (gauche) et droite des moindres carrés associée (droite)

1.1.2 Lien quali-quali

On suppose que les deux variables sont qualitatives. On les appelle parfois également des "facteurs". Notations :

X : premier facteur à I modalités

Y : second facteur à J modalités.

n_{ij} : nombre d'individus ayant la modalité i pour X et j pour Y.

$n_{i.}$: nombre d'individus ayant la modalité i pour X.

$n_{.j}$: nombre d'individus ayant la modalité j pour Y.

$$n_{i.} = \sum_{j=1}^{J} n_{ij}, \quad n_{.j} = \sum_{i=1}^{I} n_{ij}, \quad n = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}.$$

Les effectifs n_{ij} sont résumés dans un tableau de contingence.

1. Comment visualiser graphiquement le lien entre les deux variables ?

   On peut résumer le tableau de contingence par des diagrammes en batons "croisés", soit par empilement, soit côte à côte, cf la figure 1.

   Sous R : fonction `barplot`.

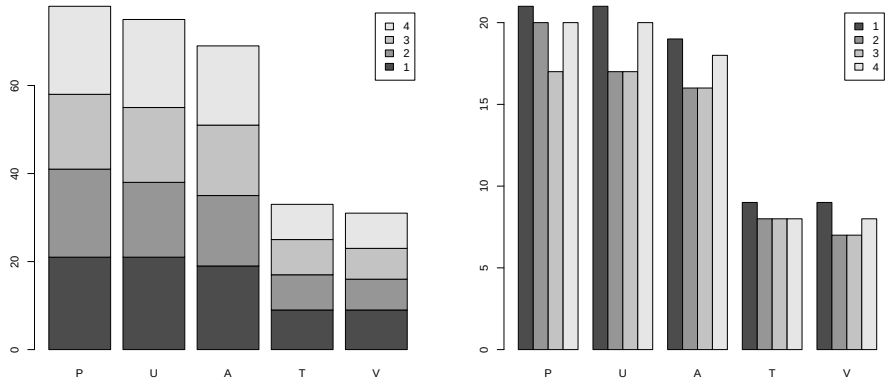

FIGURE 1.2 – Exemple d'un diagramme en batons croisés par empilement (gauche) et côte à côte (droite). Il s'agit du croisement entre une variable prenant les modalités 1, 2, 3, ou 4, et une autre variable prenant les modalités P, U, A, T ou V.

2. Comment quantifier le lien ?
   On calcule la distance du χ^2 (khi-deux) :

$$d^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

   Cette distance mesure la différence entre les effectifs observés n_{ij} et les effectifs théoriques s'il y avait indépendance : dans ce cas la fréquence observée dans i et j, $\frac{n_{ij}}{n}$, vaudrait le produit des fréquences marginales $\frac{n_{i.}}{n}\frac{n_{.j}}{n}$ et donc d^2 serait nulle.
3. Comment tester la significativité du lien ?
   Par un test du χ^2 : H_0 : X et Y indépendants, contre H_1 : le contraire. Sous H_0, $d^2 \sim \chi^2((I-1)(J-1))$ lorsque $n \to \infty$ d'où

$$RC_\alpha = \{d^2 > \chi^2_{(I-1)(J-1)}(1-\alpha)\}$$

   est une région critique au niveau asymptotique α, avec $\chi^2_{(I-1)(J-1)}(1-\alpha)$ le quantile d'ordre $1-\alpha$ d'une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté.
   Sous R : fonction `chisq.test`

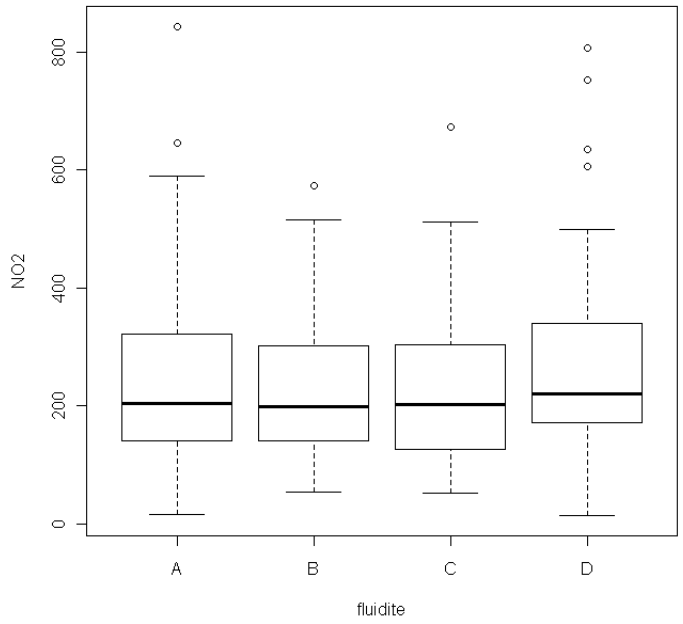

FIGURE 1.3 – Exemple de répartition de la concentration en NO2 (variable quantitative) en fonction des modalités A, B, C ou D (de fluide à congestionné) de la variable "fluidite" du trafic routier (variable qualitative).

1.1.3 Lien quanti-quali

On suppose qu'une variable est quantitative tandis que l'autre est qualitative (facteur). Notations :

X : facteur à I modalités contenant chacune n_i individus ($\sum_{i=1}^{I} n_i = n$).

Y : variable quantitative

y_{ij} : valeur de Y pour l'individu j se trouvant dans la modalité i de X.

On note $\bar{y}_i$ la moyenne de Y dans la modalité i et $\bar{y}$ la moyenne totale, i.e.

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{I} n_i \bar{y}_i.$$

1. Comment visualiser graphiquement le lien entre les deux variables ?
   On effectue des boxplots de Y par modalité de X, cf la figure 1.
   Sous R : `boxplot(Y~X)`
2. Comment quantifier le lien ?

On utilise la formule de décomposition de la variance : La variance totale est la somme de la variance inter-modalités et de la variance intra-modalités, ce qui s'écrit :

$$\frac{1}{n}\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}_{S_T^2} = \frac{1}{n}\underbrace{\sum_{i=1}^{I} n_i(\bar{y}_i-\bar{y})^2}_{S_{inter}^2} + \frac{1}{n}\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}_{S_{intra}^2}$$

autrement dit $S_T^2 = S_{inter}^2 + S_{intra}^2$.
Le lien entre X et Y est parfois mesuré par le rapport de corrélation :

$$\hat{\eta}^2 = \frac{S_{inter}^2}{S_T^2} = \frac{\sum_{i=1}^{I} n_i(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}.$$

On a ainsi $0 \leq \hat{\eta}^2 \leq 1$.

3. Comment tester la significativité du lien ?
Le coefficient $\hat{\eta}^2$ estime son équivalent théorique η^2 défini par

$$\eta^2 = \frac{\mathbb{V}(\mathbb{E}(Y|X))}{\mathbb{V}(Y)}.$$

On peut effectuer un test d'analyse de la variance. En notant $\mu_i = \mathbb{E}(Y|X=i)$ pour $i=1,\dots,I$, on souhaite tester

$$H_0 : \mu_1 = \cdots = \mu_I \quad (\Leftrightarrow \eta^2 = 0)$$

contre $H_1 : Y$ est différent, en espérance, dans au moins deux modalités de X.
Si les y_{ij} sont issus d'une loi Gaussienne de même variance pour tout i,j, alors

$$F = \frac{S_{inter}^2/(I-1)}{S_{intra}^2/(n-I)} = \frac{\hat{\eta}^2/(I-1)}{(1-\hat{\eta}^2)/(n-I)} \sim F(I-1,n-I) \quad \text{sous } H_0.$$

D'où la région critique au niveau α

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

où $f_{I-1,n-I}(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ d'une loi de Fisher $F(I-1,n-I)$.
Sous R : fonction `aov(Y~X)` pour obtenir S_T^2, S_{inter}^2 et S_{intra}^2, et `summary` du résultat pour effectuer le test.
Pour $I=2$, cela correspond au test de Student d'égalité des moyennes (`t.test` de R).

1.2 Aspects généraux sur la modélisation

On rappelle que l'objectif est d'établir un lien entre une grandeur Y et p grandeurs $X^{(1)}, \dots, X^{(p)}$, sur la base d'observations auprès de n individus.

Pour chaque individu i, on observe donc Y_i ainsi que $X_i^{(1)}, \dots, X_i^{(p)}$.

1.2.1 Où est l'alea ?

Dans une grande majorité des cas, les observations auprès de chaque individu ne sont pas contrôlées, dans la sens où on ne connait pas a priori les valeurs de Y_i, ni de $X_i^{(1)}, \dots, X_i^{(p)}$. Elles deviennent connues uniquement une fois qu'on les a observées. Dans ce contexte, toutes les quantités d'intérêt, que ce soit Y ou $X^{(1)}, \dots, X^{(p)}$, peuvent être vues comme étant des variables aléatoires. C'est par exemple le cas lorsque l'on observe des mesures environnementales en certains sites, ou lorsqu'on observe des caractéristiques individuelles de clients.

Cependant, dans certaines situations, on maitrise les valeurs de $X^{(1)}, \dots, X^{(p)}$. C'est par exemple le cas lorsque ces variables correspondent à des doses de médicaments, que l'expérimentateur choisit différemment suivant les patients, afin d'observer l'effet sur une réponse Y (une mesure physiologique). Dans ce contexte, la réponse Y n'est pas maitrisée et peut donc être vue comme étant aléatoire, mais les variables d'entrée $X^{(1)}, \dots, X^{(p)}$ ne sont pas aléatoires. Un autre exemple est l'effet du dosage de certains produits phytosanitaires $X^{(1)}, \dots, X^{(p)}$ sur la productivité Y d'une culture. Ici aussi Y est aléatoire, mais pas $X^{(1)}, \dots, X^{(p)}$ (dans l'hypothèse où l'expérimentateur décide des dosages).

Ainsi, de façon générale, la réponse Y est toujours vue comme étant une variable aléatoire (les Y_i en sont les réalisations), mais, selon le contexte, les variables $X^{(1)}, \dots, X^{(p)}$ sont vues comme des variables aléatoires (les $X_i^{(1)}, \dots, X_i^{(p)}$ en sont alors les réalisations) ou comme des variables déterministes (les $X_i^{(1)}, \dots, X_i^{(p)}$ sont alors des valeurs connues, déterministes).

Néanmoins, dans les deux cas, nous cherchons à établir un lien entre les entrées $X^{(1)}, \dots, X^{(p)}$ et la réponse Y. On se pose donc la question : étant données des valeurs pour $X^{(1)}, \dots, X^{(p)}$, quelle réponse Y est attendue ? On s'intéresse donc à Y **sachant** $X^{(1)}, \dots, X^{(p)}$. Ce point de vue est commun aux deux situations, que les variables d'entrée soient aléatoires ou non.

1.2.2 Quel lien cherche-t-on à établir ?

On suppose dans cette partie que Y et les $X^{(1)}, \dots, X^{(p)}$ sont des variables aléatoires réelles (donc quantitatives). Les autres cas (qui s'y ramènent) sont discutées dans la partie suivante.

On chercher à expliquer au mieux Y par $X^{(1)}, \dots, X^{(p)}$, ce qui se traduit par chercher la meilleure fonction $f(X^{(1)}, \dots, X^{(p)})$ qui approche Y. On peut imaginer plusieurs notions d'optimalité pour préciser ce qu'on entend par "meilleur". La plus naturelle est le coût quadratique : on cherche la fonction f qui minimise $\mathbb{E}[(Y - f(X^{(1)}, \dots, X^{(p)}))^2]$.

La solution théorique est connue : il s'agit de l'espérance conditionnelle de Y sachant $X^{(1)}, \dots, X^{(p)}$. On cherche donc

$$f(X^{(1)}, \dots, X^{(p)}) = \mathbb{E}(Y|X^{(1)}, \dots, X^{(p)}). \tag{1.1}$$

Cela est cohérent avec la dernière remarque de la section précédente. Proposer la meilleure valeur de Y (qui est aléatoire) sachant $X^{(1)}, \dots, X^{(p)}$ consiste à en donner la moyenne : l'espérance de Y sachant $X^{(1)}, \dots, X^{(p)}$.

Evidemment la fonction f dans (1.1) est inconnue en pratique puisqu'on ne connait pas les lois jointes (ni marginales) de $(Y, X^{(1)}, \dots, X^{(p)})$. Son estimation, à partir de l'observation auprès de n individus, est l'objectif de la plupart des modèles de machine learning. Néanmoins le problème est complexe puisqu'il s'agit non pas d'estimer un unique paramètre réel ou vectoriel, comme cela a été abordé dans le cours de statistique inférentielle, mais d'estimer toute une fonction, dépendant de p variables. En pratique, on n'attaque pas ce problème en toute généralité, mais on le simplifie, en fonction de certaines hypothèses que l'on est prêt à prendre. En voici des exemples :

Exemple 1. Si on suppose que la loi de $(Y, X^{(1)}, \dots, X^{(p)})$ est Gaussienne, en notant X le vecteur colonne $X = (X^{(1)}, \dots, X^{(p)})'$ et Σ sa matrice de covariance, alors on sait que, pourvu que Σ soit inversible (cf le cours de probabilités de 1A) :

$$\mathbb{E}(Y|X^{(1)}, \dots, X^{(p)}) = \mathbb{E}(Y) + (X - \mathbb{E}(X))'\beta$$

où $\beta = \Sigma^{-1}(\mathrm{Cov}(Y, X^{(1)}), \dots, \mathrm{Cov}(Y, X^{(p)}))'$. Autrement dit, dans le cas où toutes les variables sont Gaussiennes, la fonction recherchée f est simplement une fonction affine en $X^{(1)}, \dots, X^{(p)}$, i.e. $f(X^{(1)}, \dots, X^{(p)}) = \beta_0 + \beta_1 X^{(1)} +$

$\cdots + \beta_p X^{(p)}$ si l'on note $\beta = (\beta_1, \ldots, \beta_p)$ et $\beta_0 = \mathbb{E}(Y) - \mathbb{E}(X)'\beta$. Les paramètres β_0, $\beta_1, \ldots, \beta_p$ sont inconnus a priori, puisqu'ils dépendent des covariances (inconnues) entre les variables, mais le problème d'estimation devient beaucoup plus simple : estimer la fonction f dans ce contexte revient simplement à estimer ces paramètres.

Exemple 2. Plutôt que de faire une hypothèse sur la loi de $(Y, X^{(1)}, \ldots, X^{(p)})$, on peut directement faire une hypothèse sur la fonction f recherchée. On peut par exemple supposer que cette fonction appartient à un ensemble de fonctions donné : l'ensemble des fonctions dérivables, ou l'ensemble des fonctions engendrées par une certaine base, ou l'ensemble des polynômes, etc. La famille de fonctions la plus simple est la famille des fonctions affines. Dans ce dernier cas, on suppose que $f(X^{(1)}, \ldots, X^{(p)}) = \beta_0 + \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)}$ pour certains paramètres β_0, $\beta_1, \ldots, \beta_p$ inconnus, que l'on cherchera à estimer. Comme on l'a vu dans l'exemple précédent, il s'agit exactement de la forme de f dans le cas Gaussien. Pour d'autres lois, il s'agit d'une approximation, qui peut être plus ou moins bonne. Ce cadre est exactement celui de la **régression linéaire** que l'on traitera dans le prochain chapitre.

Exemple 3. Si Y ne prend que 2 valeurs (disons 0 et 1), alors cela contraint f. En effet dans ce cas $\mathbb{E}(Y|X^{(1)}, \ldots, X^{(p)}) = \mathbb{P}(Y = 1|X^{(1)}, \ldots, X^{(p)})$ et donc f est à valeurs dans $[0, 1]$. Il convient alors de chercher une fonction f qui respecte cette contrainte. En particulier, il n'est pas envisageable de supposer que f est affine. Par contre, il est possible de supposer que $g(f(X^{(1)}, \ldots, X^{(p)}))$ est affine, pourvu qu'on ait préalablement choisi une transformation g qui va de $[0, 1]$ dans $\mathbb{R}$. Cette approche, consistant à approcher de façon affine une *transformation* de f (rendant la forme affine acceptable), est l'objet de la **régression linéaire généralisée**.

1.2.3 Compléments

Dans le cas où on suppose que les $X^{(1)}, \ldots, X^{(p)}$ ne sont pas aléatoires, la démarche précédente est mal adaptée car l'espérance conditionnelle devient triviale. Ceci dit, dans ce cas, la loi de Y, en particulier son espérance, est supposée dépendre des $X^{(1)}, \ldots, X^{(p)}$ (sinon aucun lien n'est à attendre). La démarche précédente se traduit ainsi : on cherche à estimer l'espérance de la loi de Y, qui est une fonction de $X^{(1)}, \ldots, X^{(p)}$. En ce sens, l'objectif est similaire à la recherche de f dans (1.1).

Evoquons pour terminer le cas des variables qualitatives. Dans ce cas, on se ramène à des variables quantitatives en introduisant des indicatrices (on parle de "one-hot encoding" en machine learning). Ainsi, si la variable réponse Y est une variable qualitative à deux modalités ("A" ou "pas A"), on modélise $\tilde{Y} = \mathbb{1}_{Y="A''}$ qui vaut 1 si $Y =$ "A'' et 0 sinon. Cette variable $\tilde{Y}$ est binaire et sa modélisation rentre dans le cadre de l'exemple 3 ci-dessus, autrement dit dans celui des modèles linéaires généralisés.

De façon plus générale, si Y est qualitative et prend K modalités $A_1, \ldots, A_K$, on modélise les variables $\tilde{Y}_k = \mathbb{1}_{Y=A_k}$ qui sont chacune binaire. Il est à noter qu'il y aura donc a priori autant de modèles que de modalités, moins une. Le "moins une" provient du fait que la dernière modalité se déduit des autres (elle vaut 1 lorsque toutes les autres valent 0).

La même transformation s'applique aux variables explicatives $X^{(1)}, \ldots, X^{(p)}$: si l'une d'entre elles est qualitative, elle est transformée en autant de variables binaires qu'il y a de modalités, et on revient ainsi au cadre de la section précédente. Cette transformation amène toutefois quelques spécificités dans l'écriture et les interprétations du modèle, sur lesquelles nous reviendrons dans le chapitre consacré à l'ANOVA et l'ANCOVA.

Chapitre 2

Régression linéaire

On s'intéresse dans ce chapitre au lien entre une variable **quantitative** Y et p variables **quantitatives** $X^{(1)}, \ldots, X^{(p)}$. Pour rappel on cherche à trouver une fonction f telle que $Y \approx f(X^{(1)}, \ldots, X^{(p)})$. La fonction idéale est l'espérance conditionnelle de Y sachant $X^{(1)}, \ldots, X^{(p)}$, cf la discussion dans la partie 1.2.2 du chapitre précédent. Pour chaque individu $i = 1, \ldots, n$, on observe Y_i ainsi que $X_i^{(1)}, \ldots, X_i^{(p)}$. Estimer l'espérance conditionnelle en toute généralité à l'aide de ces observations est trop ambitieux, et on est amener à simplifier le problème.

En régression linéaire, on suppose que f est linéaire, ce qui conduit au modèle, pour chaque individu $i = 1, \ldots, n$:

$$Y_i = \beta_1 X_i^{(1)} + \cdots + \beta_p X_i^{(p)} + \epsilon_i \tag{2.1}$$

pour certains paramètres $\beta_1, \ldots, \beta_p$ inconnus et où ϵ_i représente l'erreur individuelle de modélisation. Dans cette écriture, l'une des variables (disons $X^{(1)}$) est souvent la variable constante valant 1, i.e., $X_i^{(1)} = 1$ pour tout i, ce qui conduit en réalité à un modèle affine entre Y et $X^{(2)}, \ldots, X^{(p)}$.

A partir des observations, on souhaite :

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

2.1 Modélisation

On regroupe les observations dans les vecteurs

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X^{(1)} = \begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_n^{(1)} \end{pmatrix}, \quad \cdots \quad , X^{(p)} = \begin{pmatrix} X_1^{(p)} \\ \vdots \\ X_n^{(p)} \end{pmatrix}.$$

Attention, il y a ici un changement dans les notations : Y désigne à présent le vecteur des valeurs $Y_1, \ldots, Y_n$ alors que dans l'introduction il désignait la "variable" Y dont sont issues ces observations, de même pour $X^{(1)}, \ldots, X^{(p)}$.

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

$$X = \left(X^{(1)} | \ldots | X^{(p)}\right) = \begin{pmatrix} X_1^{(1)} & \ldots & X_1^{(p)} \\ \vdots & \vdots & \vdots \\ X_n^{(1)} & \ldots & X_n^{(p)} \end{pmatrix}.$$

On note enfin $\beta \in \mathbb{R}^p$ le vecteur des paramètres et $\epsilon \in \mathbb{R}^n$ celui des erreurs de modélisation de chaque individu :

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

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

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

Nous ferons les **hypothèses** suivantes :

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

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

  où I_n désigne la matrice identité de taille n.
- La matrice de design X est de plein rang ($rg(X) = p$). Cela signifie qu'aucune colonne $X^{(j)}$ n'est combinaison linéaire des autres.

Remarque 2.1.1.

- *La première hypothèse implique que Y est aléatoire avec $\mathbb{E}(Y|X) = X\beta$ et $\mathbb{V}(Y|X) = \sigma^2 I_n$.*
- *Si X n'était pas de plein rang, alors le modèle ne serait pas identifiable, dans le sens où une infinité de paramètres β donneraient le même modèle. $\rightarrow$ Exemple en cours.*
- *L'hypothèse $rg(X) \leq p$ implique que $p \leq n$ en vertu du fait qu'on a toujours $rg(X) \leq \min(n, p)$ pour une matrice de taille (n, p). Le cas $p > n$ fait partie du cadre de la statistique en grande dimension : des méthodes spécifiques d'estimation existent dans ce cas (notamment l'estimation pénalisée).*
- *Si X n'est pas aléatoire, alors on a $\mathbb{E}(\epsilon|X) = \mathbb{E}(\epsilon) = 0$ et $\mathbb{V}(\epsilon|X) = \mathbb{V}(\epsilon) = \sigma^2 I_n$, ce qui implique $\mathbb{E}(Y|X) = \mathbb{E}(Y) = X\beta$ et $\mathbb{V}(Y|X) = \mathbb{V}(Y) = \sigma^2 I_n$.* **Dans la suite de ce chapitre, pour alléger les notations, nous supposerons que X n'est pas aléatoire**. *Si X est aléatoire, tout est identique : il suffit d'ajouter un conditionnement en X dans toutes les espérances et les variances des formules qui suivent.*
- *Dans (2.1), la variable $X^{(1)}$ correspond généralement à $X^{(1)} = \mathbb{1}$, i.e., $X_i^{(1)} = 1$ pour tout i, de telle sorte que le premier coefficient β_1 est simplement la constante (intercept en anglais) du modèle. Certains ouvrages préfèrent introduire un coefficient β_0 pour désigner cette constante, et écrivent $Y_i = \beta_0 + \beta_1 X_i^{(1)} + \cdots + \beta_p X_i^{(p)} + \epsilon_i$ à la place de (2.1). Avec cette écriture $X^{(1)}$ est une "vraie" variable et il y a alors (p+1) coefficients de régression à estimer. C'est un simple choix de notations. Nous adopterons l'écriture (2.1), qui conduit à p coefficients de régression, qu'une constante soit incluse (le cas $X^{(1)} = \mathbb{1}$) ou pas.*

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

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X^{(1)} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad X^{(2)} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

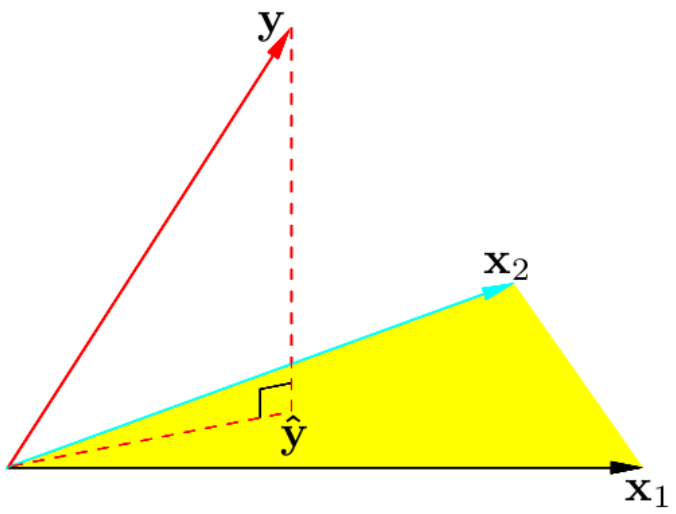

FIGURE 2.1 – Figure extraite de l'ouvrage ESL. Le plan en jaune représente $[X]$ lorsque $p = 2$. Le vecteur Y est projeté sur $[X]$ pour donner $\hat{Y}$.

2.2 Inférence

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

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

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} \left(Y_i - \beta_1 X_i^{(1)} - \cdots - \beta_p X_i^{(p)}\right)^2.$$

Soit $[X]$ l'espace vectoriel engendré par les vecteurs $X^{(1)}, \ldots, X^{(p)}$, i.e.,

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

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

$\rightarrow$ Voir Figure 2.1 et schéma en cours.

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

Démonstration. Cf cours □

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

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

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

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

Si de plus $(X'X)^{-1}$ tend vers 0 lorsque $n \to +\infty$, dans le sens où toutes les valeurs propres de $X'X$ tendent vers l'infini, alors $\hat{\beta}$ converge en moyenne quadratique vers β.

Démonstration. Cf cours □

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

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

Démonstration. Cf cours □

2.2.2 Estimation de σ^2

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

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

Proposition 2.2.5.

- $\hat{\epsilon} = Y - \hat{Y} = Y - P_{[X]}Y = P_{[X]^\perp}Y = P_{[X]^\perp}\epsilon$.
- *Si* $rg(X) = p$, $\mathbb{E}(\epsilon) = 0$ *et* $\mathbb{V}(\epsilon) = \sigma^2 I_n$, *alors*

$$\mathbb{E}(\hat{\epsilon}) = 0 \quad et \quad \mathbb{V}(\hat{\epsilon}) = \sigma^2 P_{[X]^\perp} = \sigma^2(I_n - X(X'X)^{-1}X').$$

- *Si le modèle contient une constante, typiquement* $X_i^{(1)} = 1$ *pour tout* i, *alors* $\bar{\hat{\epsilon}} = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i = 0$, *ou de manière équivalente* $\bar{\hat{Y}} = \bar{Y}$.

Démonstration. Cf cours □

Proposition 2.2.6. *Si* $rg(X) = p$, $\mathbb{E}(\epsilon) = 0$ *et* $\mathbb{V}(\epsilon) = \sigma^2 I_n$, *alors*

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

est un estimateur sans biais de σ^2. *Si de plus les* ϵ_i *sont i.i.d, alors* $\hat{\sigma}^2$ *est un estimateur consistant de* σ^2.

Démonstration. Cf cours □

2.2.3 Cas Gaussien

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

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

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

Démonstration. Cf cours. □

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

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

Démonstration. $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ qui fait partie de la famille des lois exponentielles multivariées. La borne de Cramer-Rao affirme dans ce cas que tout estimateur sans biais de β a une matrice de variance-covariance supérieure à l'inverse de l'information de Fisher (au sens où la différence est semi-définie positive). Cette dernière vaut $\sigma^{-2}X'X$. La matrice de variance-covariance de $\hat{\beta}$ coïncide donc avec la borne minimale : l'estimateur est efficace. □

2.2.4 Tests et intervalles de confiances pour β_j

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

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

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

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

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

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

Démonstration. Cf cours □

Grâce au corollaire 2.2.10, on peut construire des tests et des intervalles de confiance sur chaque paramètre β_j :

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

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

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

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

On vérifie en effet facilement que d'après le corollaire 2.2.10

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

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

2.2.5 Prévision

On suppose qu'on a estimé β et σ^2 par $\hat{\beta}$ et $\hat{\sigma}^2$ à partir des observations des variables Y et $X^{(1)}, \dots, X^{(p)}$ auprès de n individus.

On souhaite prédire Y pour un nouvel individu o, c'est à dire prédire Y_o, connaissant les valeurs prises par cet individu pour les variables explicatives, c'est à dire connaissant les valeurs $X_o^{(1)}, \dots, X_o^{(p)}$.

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

$$x_o = \begin{pmatrix} X_o^{(1)} \\ \vdots \\ X_o^{(p)} \end{pmatrix},$$

cela signifie que

$$Y_o = x_o'\beta + \epsilon_o = \beta_1 X_o^{(1)} + \cdots + \beta_p X_o^{(p)} + \epsilon_o,$$

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

Etant donné ce modèle, la prévision naturelle de Y_o est

$$\hat{Y}_o = x_o'\hat{\beta}.$$

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

L'erreur de prévision vaut $Y_o - \hat{Y}_o$. On a (cf cours) :

- $\mathbb{E}(Y_o - \hat{Y}_o) = 0$,
- $\mathbb{V}(Y_o - \hat{Y}_o) = \sigma^2(x_o'(X'X)^{-1}x_o + 1)$.

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

Intervalles de prévision :
Si l'on suppose que le modèle est Gaussien (c'est à dire que les erreurs suivent une loi Gaussienne comme dans la section 2.2.3), alors $Y_o - \hat{Y}_o \sim \mathcal{N}(0, \sigma^2(x_o'(X'X)^{-1}x_o + 1))$. On en déduit que

$$\frac{Y_o - \hat{Y}_o}{\hat{\sigma}\sqrt{x_o'(X'X)^{-1}x_o + 1}} \sim St(n-p)$$

et on peut donc fournir un intervalle de prévision pour Y_o :

$$IP_{1-\alpha}(Y_o) = \hat{Y}_o \pm t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{x_o'(X'X)^{-1}x_o + 1},$$

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

Attention : si n est grand, $\hat{\beta}$ suit approximativement une loi Gaussienne, même si le modèle n'est pas Gaussien, mais par contre ϵ_o suit sa propre loi, qui n'est pas Gaussienne si le modèle n'est pas Gaussien. Ainsi, $Y_o - \hat{Y}_o$ ne suit pas une loi Gaussienne si le modèle n'est pas Gaussien, même si n est grand. Les intervalles de prévision ne sont donc valables que pour les modèles Gaussiens.

2.3 Validation

2.3.1 Qualité explicative globale

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

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

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

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

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

Dans le cas général (même si $\mathbb{1} \notin [X]$), toujours d'après Pythagore :

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

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

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

Remarque 2.3.2.

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

Le R^2 a un défaut important : il augmente nécessairement lorsqu'on ajoute une variable explicative, même si cette dernière n'est pas significative. En effet, ajouter une variable explicative grossit l'espace $[X]$, ce qui diminue automatiquement la SCR issue de la projection. Utiliser le R^2 pour choisir entre deux modèles possibles conduira donc toujours à prendre le modèle le plus gros. Pour palier ce problème, on introduit le R^2 ajusté.

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

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

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

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

2.3.2 Tests de contraintes linéaires sur les coefficients

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

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

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

Exemples

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

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

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

La statistique de test dans ce cas est souvent écrite

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

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

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

$$Y = \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)} + \epsilon$$

contre le sous-modèle

$$Y = \beta_1 X^{(1)} + \cdots + \beta_{p-q} X^{(p-q)} + \epsilon$$

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

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

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

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

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

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

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

Démonstration. Cf cours □

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

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

où R^2 est le R^2 dans le modèle global, tandis que R_c^2 est le R^2 dans le modèle contraint (cf TD). La formule (2.2) est un cas particulier.

Retour sur le exemples

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

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

2.3.3 Vérification des hypothèses du modèle

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

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

a. Lien linéaire

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

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

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

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

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

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

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

On peut détecter ce phénomène en calculant les VIF (Variance Inflation Factor) pour chaque variable $X^{(j)}$:

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

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

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

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

- Enlever du modèle une des variables dont le VIF est élevé, en recommançant jusqu'à ce que tous les VIF soient faibles.
- Ou faire appel à des méthodes d'estimation robuste comme la régression ridge (cf le module de ML et des cours de 3A), qui évite d'avoir à sélectionner les variables "à la main".

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

c. Analyse des résidus

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

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

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

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

i) Vérification de l'homoscédasticité (cad $\mathbb{V}(\epsilon_i) = \sigma^2$ pour tout i)

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

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

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

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

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

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

De façon plus formelle, on peut appliquer le test de Breusch-Pagan. Ce dernier suppose que dans le modèle de régression linéaire, l'erreur ϵ_i a une variance $\sigma_i^2 = \sigma^2 + z_i'\gamma$ où z_i est un vecteur de k variables à choisir qui pourraient expliquer l'hétéroscédasticité (si elle est présente) et γ est un paramètre inconnu, de dimension k, à estimer. Le choix par défaut sous R est de prendre $z_i' = (X_i^{(1)}, \dots, X_i^{(p)})$, c'est à dire les mêmes variables explicatives que dans le modèle de régression.

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

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

La procédure de test n'est pas détaillée ici.

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

Si on observe un problème d'hétéroscédasticité :

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

ii) Non-corrélation des erreurs (cad $\mathbb{V}(\epsilon)$ est une matrice diagonale)

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

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

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

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

1. Test de Durbin-Watson

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

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

La statistique de test de Durbin-Watson est

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

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

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

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

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

2. Test de Breusch-Godfrey

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

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

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

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

Que faire si une corrélation est détectée ?

Cela n'est pas forcément une mauvaise nouvelle : cela signifie que le modèle peut être enrichi en incluant de l'information contenue dans le passé des variables. Exemple : $Y_i = \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \alpha_1 Y_{i-1} + \alpha_2 X_{i-1}^{(1)} + \epsilon_i$. Ici le passé de Y et de $X^{(1)}$ est intégré au modèle. Avec cette démarche, les nouvelles erreurs ϵ_i du modèle peuvent devenir non-corrélées.

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

iii) Normalité des erreurs

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

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

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

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

On peut également mettre en oeuvre un test de normalité. Le plus utilisé

est le test de Shapiro-Wilk dont l'hypothèse nulle est l'hypothèse de normalité.

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

2.3.4 Analyse des individus atypiques et/ou influents

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

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

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

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

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

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

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

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

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

Cette définition provient du fait que

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

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

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

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

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

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

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

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

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

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

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

2.4 Critères de sélection de modèles

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

2.4.1 Les critères

De nombreux critères existent. Les plus classiques sont :

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

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

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

Soit $Y = X\beta + \epsilon$ un modèle candidat contenant p variables explicatives (qui est potentiellement faux). On note $\hat{\beta}$ l'estimateur des MCO basé sur ce modèle (il dépend donc de Y et de X). Le C_p de Mallows vise à estimer l'erreur $\mathbb{E}(\|\tilde{Y} - X\hat{\beta}\|^2)$ où $\tilde{Y}$ est une copie indépendante de Y (cela signifie que $\tilde{Y} = X^*\beta^* + \tilde{\epsilon}$ où $\tilde{\epsilon}$ est indépendant de ϵ). L'idée est d'évaluer la qualité de prévision du modèle testé sur des nouveaux individus, indépendants de ceux utilisés pour estimer $\hat{\beta}$. Si on avait utilisé Y au lieu de $\tilde{Y}$, c'est à dire si on avait confronté la qualité prédictive aux mêmes individus que ceux utilisés pour estimer le modèle, cela aurait conduit à privilégier automatiquement le plus gros modèle (celui qui passe au plus près des points, donc celui qui minimise SCR), avec le risque d'être peu généralisable, c'est à dire d'expliquer très mal le comportement de nouveaux individus.

L'expression de C_p qui estime l'erreur de prévision précédente est

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

où SCR est la SCR dans le modèle testé et $\hat{\sigma}^2$ est l'estimation de σ^2 dans le plus gros modèle (celui contenant les $p_{\max}$ variables explicatives). La justification de la forme de cet estimateur dépasse le cadre du cours.

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

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

Information Criterion), défini par

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

où SCR est la SCR dans le modèle testé.

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

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

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

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

2.4.2 Lien entre les critères

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

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

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

2.4.3 Aspects théoriques

Supposons que le vrai modèle (inconnu) appartienne aux sous-modèles testés. Sous des hypothèses standards, on a les résultats asymptotiques suivants (cf la partie 9.6 du livre "Le modèle linéaire par l'exemple") :

Pour le critère BIC :

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

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

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

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

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

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

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

2.4.4 Algorithme de sélection automatique

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

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

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

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

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

- Procédure stepwise backward : on part du plus gros modèle contenant $p_{\max}$ variables et on élimine la variable la moins significative (au sens où son retrait optimise le critère choisi, par exemple conduit au BIC le plus faible). On élimine ainsi successivement les variables les unes après les autres, jusqu'à ce que plus aucun retrait n'améliore le modèle (chaque retrait détériore le critère choisi).
- Procédure stepwise forward : on part du plus petit modèle (celui ne contenant que la constante) et on ajoute la meilleure variable (au sens où son ajout optimise le critère choisi). On ajoute ainsi successivement les variables jusqu'à ce que plus aucun ajout n'améliore le modèle.

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

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

Chapitre 3

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

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

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

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

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

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

3.1 Analyse de la variance à 1 facteur

3.1.1 Ecriture du modèle

Notations :

Y : variable à expliquer observée auprès de n individus

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

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

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

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

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

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

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

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

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

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

où les ϵ_k sont centrées, non-corrélées 2 à 2, de variance σ^2, et où $\mathbb{1}_{A_i}$ est la variable indicatrice dont chaque entrée vaut 1 ou 0 selon que l'individu appartient à A_i. De façon matricielle, le modèle s'écrit donc

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

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

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

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

Ce modèle n'est pas de plein rang car la variable $\mathbb{1}$ associée à la constante est une combinaison linéaire des autres variables : $\mathbb{1} = \sum_{i=1}^{I} \mathbb{1}_{A_i}$. Il faut donc ajouter une contrainte pour le rendre identifiable.

Exemples de contraintes

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

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

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

et

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

Démonstration. Cf cours □

3.1.2 Significativité du facteur

On rappelle que d'après la formule d'analyse de la variance :

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

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

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

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

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

d'où la région critique au niveau α :

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

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

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

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

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

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

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

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

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

3.1.3 Analyse post-hoc

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

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

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

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

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

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

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

Cette probabilité est

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

Elle est en générale beaucoup plus grande que α.

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

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

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

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

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

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

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

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

Solution 2 : *Correction de Benjamin Hochberg.* Il s'agit d'une procédure plus puissante que celle de Bonferroni pour gérer les problèmes de tests multiples, mais qui repose sur quelques hypothèses. Elle est très populaire. Elle sera présentée dans un autre cours.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2 Analyse de la variance à 2 facteurs

3.2.1 Modèle

Notations :

Y : variable à expliquer observée auprès de n individus

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

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

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

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

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

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

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

μ_{ij} : espérance de Y dans la classe $A_i \cap B_j$, i.e. $\mu_{ij} = \mathbb{E}(Y | A_i \cap B_j)$.

$\mu_{i.}$, $\mu_{.j}$: espérances marginales dans A_i et B_j, respectivement.

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

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

pour tout $i = 1, \dots, I$, $j = 1, \dots, J$ et $k = 1, \dots, n_{ij}$. Dans cette écriture, les ϵ_{ijk} sont des variables centrées, non-corrélées 2 à 2, et de variance σ^2.

On décompose l'espérance μ_{ij} dans $A_i \cap B_j$, de façon additive pour écrire :

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

Cela permet de mettre en évidence l'effet moyen m de Y (sans tenir compte de A et B) ; l'effet marginal dû à A : $\alpha_i = \mu_{i.} - m$; l'effet marginal dû à B : $\beta_j = \mu_{.j} - m$, et l'effet restant, dû à l'interaction entre A et B : $\gamma_{ij} = \mu_{ij} - m - \alpha_i - \beta_j = \mu_{ij} - \mu_{i.} - \mu_{.j} + m$.

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

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

où les ϵ_k sont centrées, non-corrélées 2 à 2, et de variance σ^2.

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

L'interprétation précédente des coefficients m, α_i, β_j et γ_{ij} en fonction de μ_{ij}, $\mu_{i.}$ et $\mu_{.j}$, est liée au choix de contraintes :

$$\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0, \quad \forall j = 1, \dots, J, \sum_{i=1}^{I} \gamma_{ij} = 0, \quad \forall i = 1, \dots, I, \sum_{j=1}^{J} \gamma_{ij} = 0.$$

Sous R : le modèle complet (avec interaction) se lance avec la commande `lm(Y∼A+B+A:B)` ou de façon équivalente `lm(Y∼A*B)`. Par défaut les contraintes sont $\alpha_1 = 0$, $\beta_1 = 0$, $\gamma_{1j} = 0$ pour tout $j = 1, \dots, J$, et $\gamma_{i1} = 0$ pour tout $i = 1, \dots, I$. Cela en fait bien $1 + I + J$ (la contrainte $\gamma_{11} = 0$ apparaît deux fois). Si l'on souhaite imposer les contraintes précédentes sur les sommes, il faut lancer `lm(Y∼A*B,contrasts=(A=contr.sum,B=contr.sum))`.

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

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

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

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

3.2.2 Tests

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

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

S'il n'y a pas d'interaction, les courbes doivent être plus ou moins parallèles entre elles, comme justifié ci-dessous. C'est la situation que l'on observe

dans les figures 3.1 et 3.2. Inversement, la situation de la figure 3.3 montre une interaction entre A et B.

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

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

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

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

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

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

Le premier terme ne dépend que de i : il correspond à l'effet α_i dans (3.3), le second ne dépend que de j : il correspond à l'effet β_j, et le dernier est constant : il correspond à m. Ainsi dans la relation (3.3), $\gamma_{ij} = 0$. Il n'y a donc pas d'interaction.

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

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

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

où

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

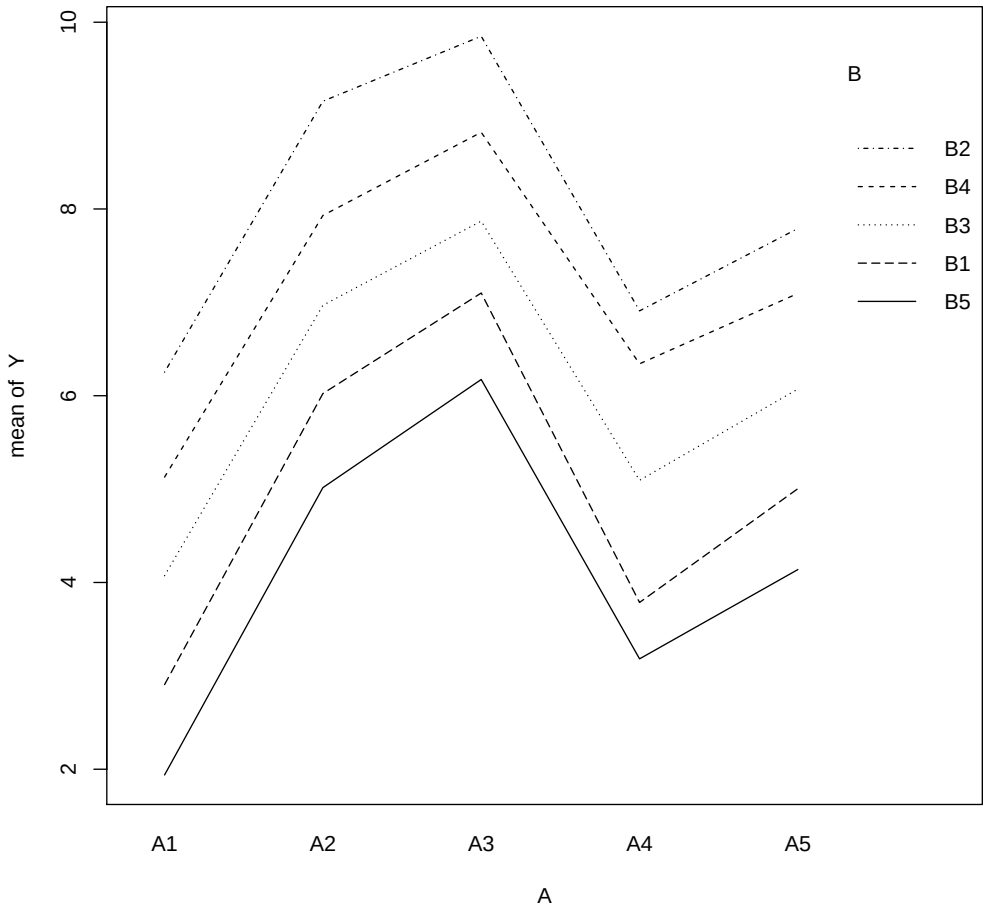

FIGURE 3.1 – Interaction plot : évolution des moyennes de Y selon les modalités croisées de A (en abscisse) et de B (en ordonnée). Les courbes sont à peu près parallèles, ce qui témoigne de l'absence d'interaction entre A et B.

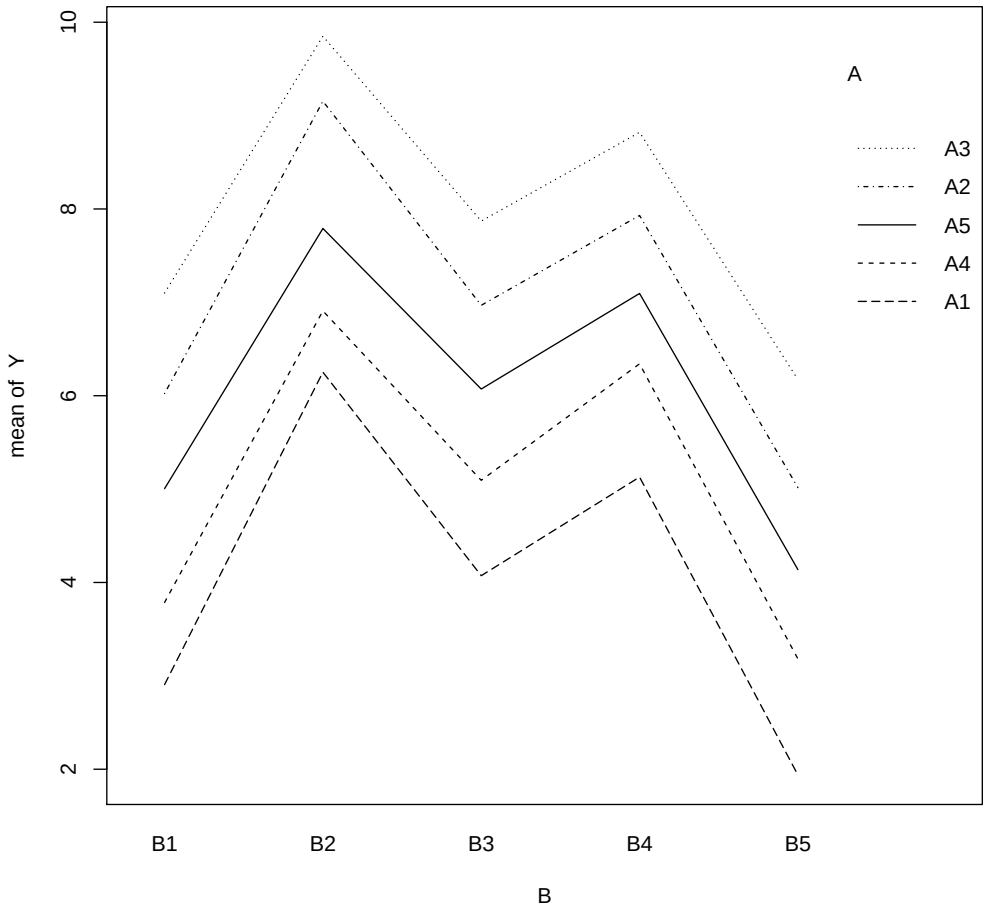

FIGURE 3.2 – Interaction plot : évolution des moyennes de Y selon les modalités croisées de B (en abscisse) et de A (en ordonnée). Il s'agit des mêmes données que dans la figure 3.1.

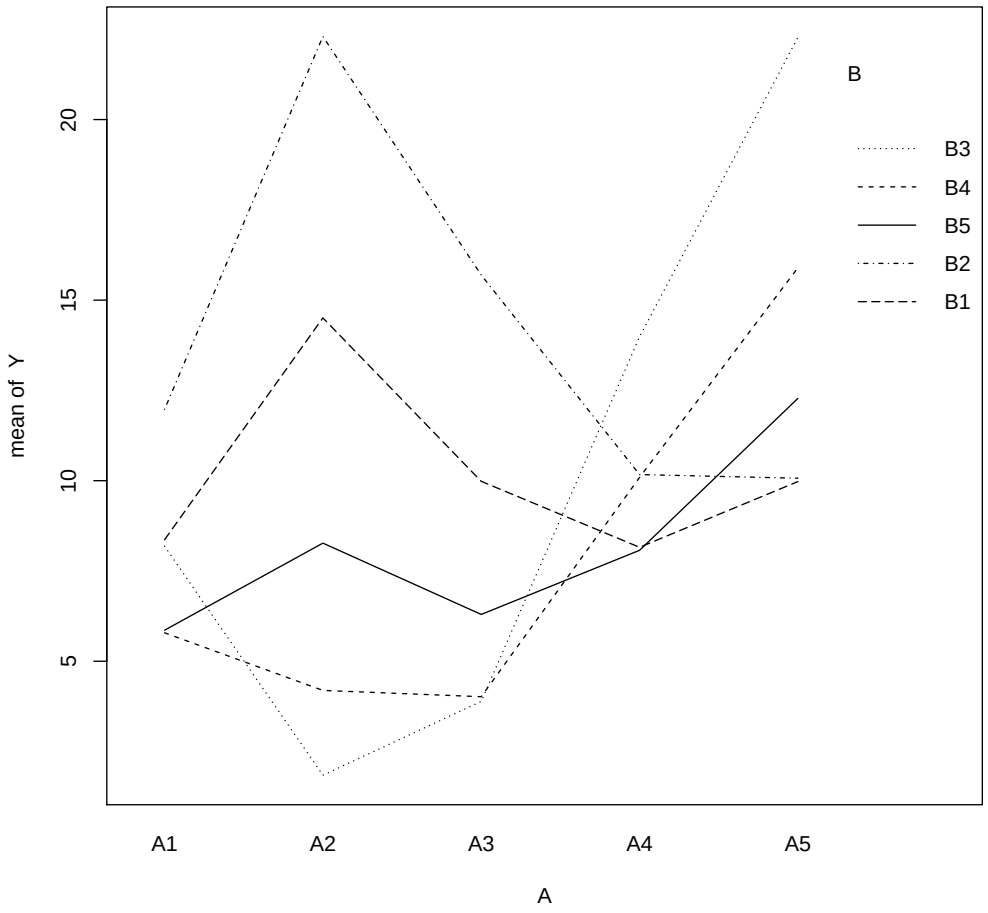

FIGURE 3.3 – Interaction plot, cas d'une présence d'interaction entre A et B : contrairement aux deux figures précédentes, les courbes ne sont pas parallèles.

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

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

Pour cela on utilise la statistique

$$F^{(AB)} = \frac{S^2_{AB}/(I-1)(J-1)}{S^2_R/(n-IJ)}$$

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

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

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

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

on utilise la statistique

$$F^{(A)} = \frac{S^2_A/(I-1)}{S^2_R/(n-IJ)}$$

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

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

on utilise la statistique

$$F^{(B)} = \frac{S^2_B/(J-1)}{S^2_R/(n-IJ)}$$

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

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

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

3.3 Analyse de la variance à k facteurs

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

On peut tester chacun de ces effets par un test d'analyse de la variance comme on l'a présenté ci-dessus dans le cas de 2 facteurs. Néanmoins, si k est grand, cela fait trop de tests à réaliser, et surtout l'effectif dans chaque modalité croisée à k facteurs risque d'être très faible, de l'ordre de 0 ou 1 individu, rendant ces tests inefficaces.

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

3.4 Analyse de la covariance (ANCOVA)

Il s'agit de la situation dans laquelle les variables explicatives incluent à la fois des facteurs et des variables quantitatives. Il s'agit donc d'un mélange de la régression linéaire standard telle que vue dans le chapitre précédent, et de l'ANOVA. Le modèle pourra ainsi inclure : les effets de chaque variable

quantitative (via chaque coefficient de régression β_j associé), les effets des facteurs et des interactions entre les facteurs (comme a l'a vu dans les parties précédentes), mais aussi les effets des interactions entre les facteurs et les variables quantitatives. Par exemple on peut imaginer que le coefficient β_j de la variable quantitative $X^{(j)}$ prend en réalité deux valeurs différentes selon qu'on est dans la première modalité du facteur A ou dans la seconde : il s'agit d'une interaction entre $X^{(j)}$ et A.

Sous R : Si Y est la variable réponse quantitative, X une variable quantitative et A un facteur à I modalités,

- `lm(Y∼X+A)` estime le modèle sans interaction :

$$Y_k = m + \beta X_k + \sum_{i=2}^{I} \alpha_i \mathbb{1}_{A_i}(k) + \epsilon_k$$

  pour chaque individu $k = 1, \ldots, n$. (La contrainte $\alpha_1 = 0$ est adoptée pour rendre le modèle identifiable.)
- `lm(Y∼X+A+X:A)` ou `lm(Y∼X*A)` estime le modèle avec interaction :

$$Y_k = m + \beta X_k + \sum_{i=2}^{I} \beta_i X_k \mathbb{1}_{A_i}(k) + \sum_{i=2}^{I} \alpha_i \mathbb{1}_{A_i}(k) + \epsilon_k.$$

Une vidéo présentant l'ANCOVA et sa mise en oeuvre sous R font l'objet d'une activité disponible dans Moodle.

Chapitre 4

Régression linéaire généralisée

Pour le contenu de ce chapitre, voir les slides associés en support du cours.

4.1 Généralité sur les GLM (generalized linear models)

4.1.1 Limites du modèle linéaire

4.1.2 Vers le modèle linéaire généralisé : 3 cas fondamentaux

4.1.3 Le modèle linéaire généralisé

4.2 Le modèle logistique pour Y binaire

4.2.1 La fonction logit comme fonction de lien

4.2.2 Enjeux du modèle logistique

4.2.3 Interprétation du modèle

4.2.4 Estimation des paramètres

4.2.5 Tests et intervalles de confiance

4.2.6 Déviance, tests et choix de modèles

4.2.7 Classification

4.3 Modèles pour données catégorielles

4.3.1 Modèle logistique nominal

4.3.2 Modèle logistique ordinal

4.4 Modèles pour données de comptage

4.4.1 Le modèle log-linéaire de Poisson

4.4.2 La sur-dispersion

4.4.3 Inflation de zéros

Références

- "Régression avec R", P-A. Cornillon, E. Matzner-Løber
  $\longrightarrow$ *Livre en français, très accessible, en lien avec les 3 premiers chapitres*

- "Le modèle linéaire par l'exemple", J.-M. Azais, J.-M. Bardet.
  $\longrightarrow$ *Livre en français, en lien avec les 3 premiers chapitres : des discussions intéressantes sur l'enjeu des hypothèses, et des résultats théoriques fins.*

- "An introduction to statistical learning with applications in R", G. James, D. Witten, T. Hastie, R. Tibshirani.
  $\longrightarrow$ *Grand classique sur les méthodes de machine learning, y compris les méthodes vues dans ce cours. Exemples avec R.*

- ESL : "The elements of statistical learning", T. Hastie, R. Tibshirani, J. Friedman.
  $\longrightarrow$ *Grand classique également. Version plus théorique (et plus complète) que le précédent.*

- Agresti, A. Foundations of Linear and Generalized Linear Models, Wiley.
  $\longrightarrow$ *Livre classique sur le sujet, en lien avec le chapitre 4*

- Antoniadis, A. Berruyer J. et Carmona R. Régression non linéaire et applications, Economica.
  $\longrightarrow$ *Résultats théoriques complets, en lien avec le chapitre 4*

- Dobson, A.J., Barnett, A.G. An Introduction to Generalized Linear Models, CRC Press.
  $\longrightarrow$ *Des exemples en R, en lien avec le chapitre 4*

- Hosmer, D. et Lemeshow S. Applied Logistic Regression, Wiley.
  $\longrightarrow$ *La régression logistique en applications, en long et en large*