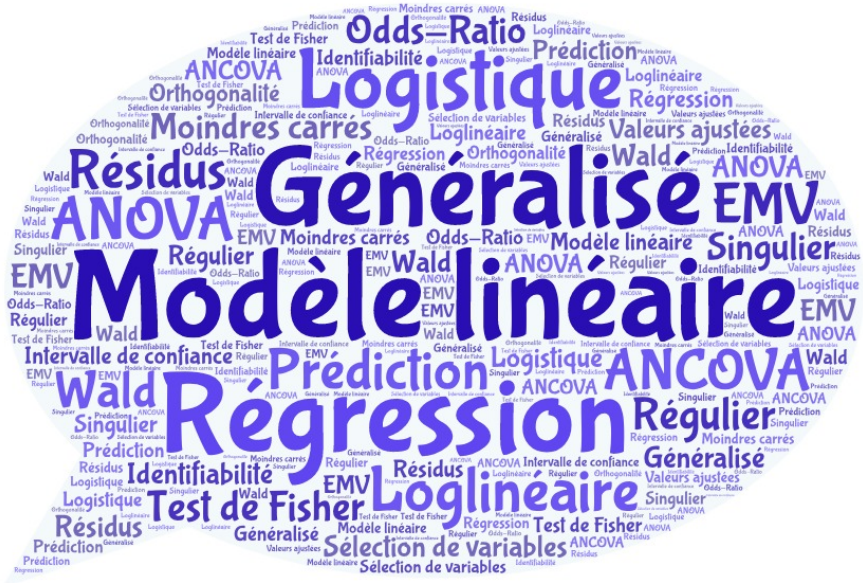
INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Département GMM
4ème année

Modèle linéaire général et modèle linéaire généralisé

Cathy Maugis-Rabusseau

Bureau GMM 116
cathy.maugis@insa-toulouse.fr



Année universitaire 2020-2021

Notations :

- A' est la transposée de la matrice A
- $0_n = (0, \ldots, 0)' \in \mathbb{R}^n$ avec $n \in \mathbb{N}^\star$
- $1_n = (1, \ldots, 1)' \in \mathbb{R}^n$ avec $n \in \mathbb{N}^\star$
- I_n désigne la matrice identité de $\mathcal{M}_n(\mathbb{R})$
- $Var(A)$ est la variance pour une variable aléatoire A
- $Cov(A, B)$ est la covariance entre deux variables aléatoires A et B
- Soit $x = (x_1, \ldots, x_n)$ une série de mesures. On note $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ la moyenne des mesures et $var(x) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$ la variance.
- Soit $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ deux séries de mesures. La covariance est définie par $cov(x, y) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)$.
- On utilisera la même notation pour désigner la matrice et l'application linéaire associée à cette matrice

Illustrations du cours :

Les exemples utilisés pour l'illustrer les chapitres de ce cours sont disponibles sur la page moodle du cours ainsi que les scripts .Rmd. Pour pouvoir compiler ces derniers, les packages R suivants sont nécessaires :
AER, bestglm, boot, corrplot, GGally, ggfortify, ggplot2, gridExtra, ISLR, leaps, MASS, nnet, VGAM

Table des matières

Chapitre 1

Exemples introductifs

Pour illustrer la démarche statistique et les problématiques auxquelles peuvent répondre les modèles linéaires et linéaires généralisés, nous présentons dans cette partie une analyse statistique sur un exemple simple. Cette feuille de bord, constituée de tableaux et de graphiques, a pour objectif de rappeler les principaux outils de statistique descriptive univariée et d'introduire les différents types de modèles linéaires que nous verrons par la suite.

1.1 Jeu de données illustratif

Pour 100 individus, on dispose de leur taille, leur poids, leur âge et leur sexe (75 hommes et 25 femmes). On sait également si ce sont des fumeurs ou non ; s'ils ronflent la nuit ou non. Un extrait des données est présenté ci-dessous :

	age	poids	taille	alcool	sexe	ronfle	tabac
1	47	71	158	0	H	N	O
2	56	58	164	7	H	O	N
3	46	116	208	3	H	N	O
4	70	96	186	3	H	N	O
5	51	91	195	2	H	O	O
6	46	88	188	0	F	N	N

Pour ce chapitre, vous pouvez utiliser le script *ExIntro.Rmd* et le jeu de données *ronfletabac.csv* disponibles sur la page moodle du cours pour reproduire l'étude sous R et l'approfondir.

1.1.1 Description de la population d'étude

Les variables sont analysées différemment selon leur nature : quantitative ou qualitative. Les variables quantitatives sont résumées sous forme d'indicateurs (moyenne, écart-type, ...), comme dans le tableau en Figure 1.1, et sont présentées graphiquement sous forme d'histogramme (quantitative continue), de boîtes à moustache ou de diagramme en bâton (quantitative discrète) (Figure 1.2).

```
> summary(don[,c("age","poids","taille")])
      age            poids            taille
 Min.   :23.00   Min.   : 42.00   Min.   :158.0
 1st Qu.:43.00   1st Qu.: 75.50   1st Qu.:166.0
 Median :52.00   Median : 92.00   Median :186.0
 Mean   :52.27   Mean   : 88.83   Mean   :181.1
 3rd Qu.:62.25   3rd Qu.:104.25   3rd Qu.:194.0
 Max.   :74.00   Max.   :120.00   Max.   :208.0
```

FIGURE 1.1 – Statistiques exploratoires des variables quantitatives

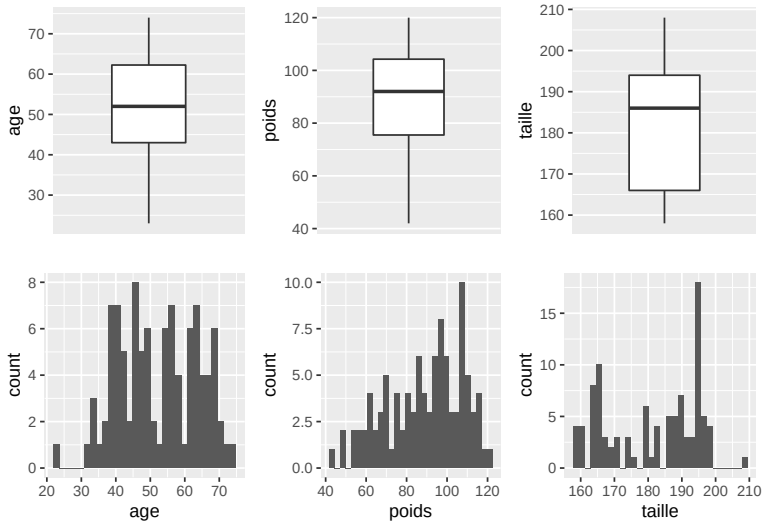


FIGURE 1.2 – Représentations graphiques de la distribution des variables quantitatives : l'âge, le poids et la taille.

Pour les variables qualitatives, on résume les données sous forme de tableau de fréquences (Table 1.1) et on les présente graphiquement par des diagrammes en bâtons (Figure 1.3).

1.2 Modélisation d'une variable quantitative

Dans cette partie, on cherche à évaluer l'effet éventuel des caractéristiques des individus sur leur poids (variable quantitative). Selon la nature des variables, les méthodes d'analyse sont différentes.

1.2.1 Régression linéaire

Pour étudier la relation entre deux variables quantitatives (par exemple, entre le poids et la taille, ou entre le poids et l'âge), on peut tracer un nuage de points (Figure 1.4) et calculer le coefficient de corrélation linéaire entre ces deux variables (Table 1.2).

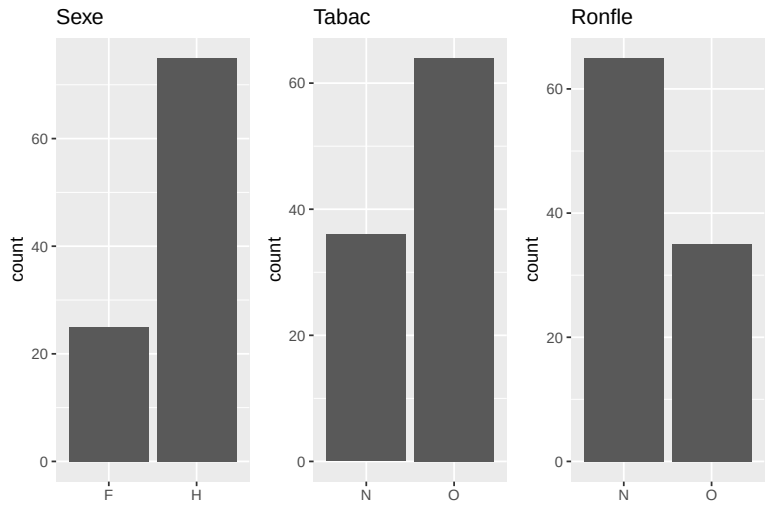


FIGURE 1.3 – Diagrammes en bâtons représentant la distribution des variables qualitatives : sexe, tabac et ronfle.

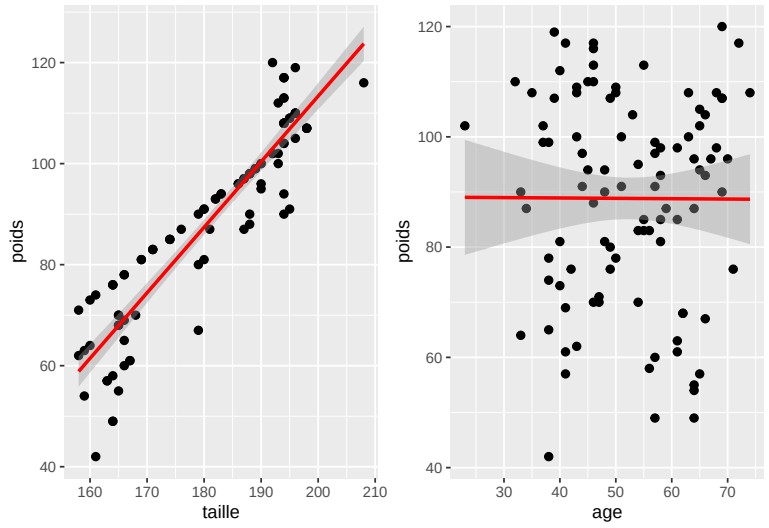


FIGURE 1.4 – Nuage de points représentant la relation entre le poids et la taille (à gauche), entre le poids et l'âge (à droite).

Variable	Modalités	Fréquence en %
Sexe	Féminin	25
	Masculin	75
Tabac	Oui	64
	Non	36
Ronfle	Oui	35
	Non	65

TABLE 1.1 – Tableau de fréquences par sexe, tabac et ronfle

	taille	age
poids	0.92	-0.004
p-valeur	$< 2.2\ 10^{-16}$	0.9687

TABLE 1.2 – Coefficient de corrélation linéaire de Pearson et test de nullité de ce coefficient

Nous remarquons que le coefficient de corrélation linéaire est significativement différent de 0 dans le cas de la régression du poids en fonction de la taille. Ce n'est pas le cas pour la régression du poids en fonction de l'âge.

1.2.1.1 Régression linéaire simple

Le nuage de points peut être résumé par une droite que l'on appellera **la droite de régression linéaire simple**. C'est le cas le plus simple de modèle linéaire, qui permet d'expliquer une variable quantitative en fonction d'une autre variable quantitative. Par exemple, la droite de régression linéaire résumant la relation entre le poids et la taille a pour équation :

$$poids_i = a + b \times taille_i + \varepsilon_i,\ i = 1, \cdots, 100 \tag{1.1}$$

où ε_i désigne l'erreur associée à chaque observation. Généralement, ces erreurs sont supposées être des variables indépendantes gaussiennes centrées de variance constante σ^2 à estimer.

Exercice 1. *Le modèle statistique sous-jacent à l'équation* (1.1) *peut aussi être présenté sous une forme matricielle. En considérant les vecteurs suivants :*

$$poids = (poids_1, \ldots, poids_{100})' \quad \theta = (a, b)' \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_{100})',$$

montrez que le modèle peut s'écrire sous la forme

$$poids = X\theta + \varepsilon. \tag{1.2}$$

où X est une matrice à préciser.

Dans le modèle (1.2), $\theta = (a, b)'$ et σ^2 sont inconnus. Afin d'estimer les paramètres a et b, nous utilisons la **méthode des moindres carrés**. Nous choisissons ainsi le couple $(\hat{a}, \hat{b})$ vérifiant :

$$(\hat{a}, \hat{b}) = \underset{(\alpha,\beta)}{\operatorname{argmin}} \sum_{i=1}^{100} (poids_i - \alpha \ - \beta \ taille_i)^2 = \underset{(\alpha,\beta)}{\operatorname{argmin}} \ ||poids - \alpha \mathbb{1}_{100} - \beta \ taille||^2.$$

Dans le chapitre dédié à la régression linéaire, nous déterminerons l'expression explicite de ces estimateurs et étudierons leurs propriétés. A l'aide de la fonction `lm` sous R, on peut facilement ajuster ce modèle de régression linéaire sur les données :

```
> reg1<-lm(poids~taille,data=don)
> summary(reg1)

Call:
lm(formula = poids ~ taille, data = don)

Residuals:
     Min       1Q   Median       3Q      Max
-20.7482  -3.8787   0.6629   4.1182  17.0261

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -146.16586   10.35384  -14.12   <2e-16 ***
taille         1.29760    0.05702   22.76   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 7.583 on 98 degrees of freedom
Multiple R-squared:  0.8409,        Adjusted R-squared:  0.8393
F-statistic: 517.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

En pratique nous obtenons les estimations suivantes :

- $\left(\widehat{b}\right)^{obs} = 1.298$: estimation de la pente de la droite de régression
  = estimation de la variation moyenne du poids par rapport à la taille
- $(\widehat{a})^{obs} = -146.2$: estimation de l'ordonnée à l'origine de la droite de régression
- $\left(\widehat{\sigma^2}\right)^{obs} = (7.583)^2$

L'estimation de la pente égale à 1.298 est significativement différente de 0, montrant que le poids et la taille sont liés de façon significative.

Ces résultats préliminaires ne donnent qu'une approximation du modèle linéaire sous-jacent. Dans bien des situations, il reste à mener une étude approfondie permettant dans un premier temps de "valider" le modèle, puis d'exploiter ce dernier : construction de tests, intervalles de confiance, ... Nous reviendrons plus en détail sur ces notions dans les chapitres suivants.

1.2.1.2 Régression linéaire multiple

Il peut être également intéressant de modéliser une variable en fonction de plusieurs autres variables quantitatives, par un modèle de **régression linéaire multiple**. Par

exemple, on peut modéliser le poids en fonction de la taille et de l'âge, ce qui donne l'équation suivante :

$$poids_i = a_0 + a_1 \times taille_i + a_2 \times age_i + \varepsilon_i,$$

où les ε_i, $i = 1, \ldots, 100$ désignent des variables indépendantes gaussiennes centrées de variance constante σ^2.

Exercice 2. *En considérant les vecteurs*

$$poids = (poids_1, \ldots, poids_{100})' \quad \theta = (a_0, a_1, a_2)' \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_{100})',$$

montrez que le modèle peut s'écrire sous la forme

$$poids = X\theta + \varepsilon. \tag{1.3}$$

où X est une matrice à préciser.

On peut remarquer en regardant (1.2) et (1.3) que les deux modèles de régression linéaire vus précédemment s'écrivent sous une "même forme" matricielle.

1.2.2 Analyse de la variance (ANOVA)

Il est possible d'étudier la relation entre une variable quantitative et une variable qualitative, par exemple entre le poids et le sexe, ou entre le poids et le tabac. Cette relation est représentée graphiquement par des boxplots parallèles (Figure 1.5).

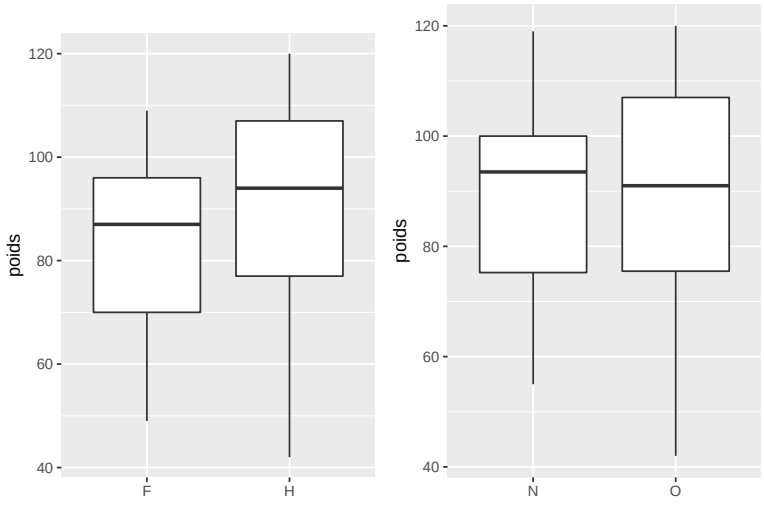


FIGURE 1.5 – Boxplots parallèles représentant la relation entre le poids et le sexe (à gauche) ; entre le poids et le tabac (à droite).

1.2.2.1 ANOVA à un facteur

Intuitivement, pour comparer le poids des hommes et celui des femmes, nous allons calculer le poids moyen pour chaque groupe. Statistiquement, nous modélisons le poids en fonction du sexe en mettant en oeuvre un modèle d'**analyse de variance à un facteur** qui s'écrit sous la forme :

$$poids_i = \mu_1 \mathbb{1}_{sexe_i=F} + \mu_2 \mathbb{1}_{sexe_i=H} + \varepsilon_i,$$

où les ε_i, $i = 1, \ldots, 100$ désignent des variables indépendantes gaussiennes centrées de variance constante σ^2. Dans ce cas, en réordonnant les observations selon le facteur sexe, le modèle peut être écrit sous la forme matricielle suivante :

$$\underbrace{\begin{pmatrix} poids_{11} \\ \vdots \\ poids_{1n_1} \\ poids_{21} \\ \vdots \\ poids_{2n_2} \end{pmatrix}}_{poids} = \underbrace{\begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \end{pmatrix}}_{\varepsilon},$$

où $poids_{ij}$ désigne le poids de l'individu j de sexe $i = F$ ou H avec j variant de 1 à n_i.

En pratique, on utilise la méthode des moindres carrés pour estimer les paramètres inconnus. Toujours à l'aide de la fonction `lm` de R, on obtient les résultats suivants :

```
> anova1<-lm(poids~sexe-1,data=don)
> summary(anova1)

Call:
lm(formula = poids ~ sexe - 1, data = don)

Residuals:
   Min     1Q Median     3Q    Max
-48.77 -13.44   4.00  16.23  29.23

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
sexeF   83.000      3.741   22.19   <2e-16 ***
sexeH   90.773      2.160   42.03   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 18.7 on 98 degrees of freedom
Multiple R-squared:  0.9584,        Adjusted R-squared:  0.9576
F-statistic:  1129 on 2 and 98 DF,  p-value: < 2.2e-16
```

Nous obtenons donc $(\widehat{\mu_1})^{obs} = 83$ et $(\widehat{\mu_2})^{obs} = 90.773$ qui représentent le poids moyen des femmes et celui des hommes respectivement.

1.2.2.2 ANOVA à deux facteurs

Il est également possible d'étudier l'effet conjoint du sexe et du tabac sur le poids. Intuitivement, on peut étudier les moyennes par classe, en croisant les deux variables qualitatives. Pour étudier l'effet combiné du sexe et du tabac sur le poids, nous mettons en oeuvre un modèle d'**analyse de variance à deux facteurs croisés**. Ce modèle s'écrit de la façon suivante :

$$poids_{ijk} = \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

où $poids_{ijk}$ désigne le poids de l'individu k tel que $sexe = i \in \{H, F\}$ et $tabac = j \in \{0, N\}$. Les variables ε_{ijk} sont supposées être des variables indépendantes gaussiennes centrées et de variance constante σ^2. Nous pouvons également écrire ce modèle sous forme matricielle de la forme

$$poids = X\theta + \varepsilon.$$

Ce modèle nous permettra d'étudier l'effet de chaque facteur (sexe et tabac) sur le poids, mais aussi de détecter des combinaisons entre le sexe et le tabac qui donneraient un poids particulièrement différent des autres classes.

1.2.3 Analyse de covariance (ANCOVA)

Sur notre exemple, nous pouvons tenter d'expliquer le poids selon la taille (variable quantitative) et le sexe (variable qualitative). Dans ce cas, nous pouvons représenter deux nuages de points entre le poids et la taille, l'un pour les femmes et l'autre pour les hommes, comme le montre la Figure 1.6.

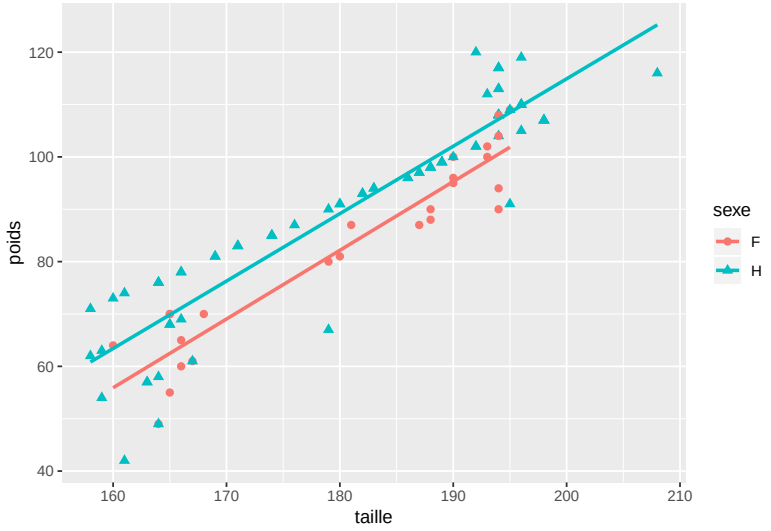


FIGURE 1.6 – Nuages de points représentant la relation entre le poids et la taille selon le sexe.

Le modèle d'analyse de covariance s'écrit de la façon suivante :

$$poids_{ij} = a_i + b_i\ taille_{ij} + \varepsilon_{ij}, i \in \{H, F\} \text{ et } j = 1, \cdots, n_i$$

où $poids_{ij}$ désigne le poids de l'individu j de sexe i et les erreurs ε_{ij} sont supposées gaussiennes centrées indépendantes et de variance σ^2.
Nous pouvons ainsi comparer l'effet de la taille sur le poids, selon le sexe en mettant en oeuvre un modèle d'**analyse de la covariance**. En pratique cela correspond à estimer, pour chaque modalité de la variable sexe, une droite de régression du poids en fonction de la taille.

En conclusion, dans les divers problèmes évoqués dans ce paragraphe, à savoir la régression linéaire, l'analyse de variance et l'analyse de covariance, nous avons utilisé :

- le même type de modélisation matricielle,
- le même type d'hypothèses sur les erreurs,
- l'estimateur des moindres carrés.

En fait, ces différents problèmes ne sont pas si éloignés qu'ils le paraissent a priori car les modèles utilisés font partie d'une même famille de modèles : **le modèle linéaire général**.

1.3 Modélisation d'une variable qualitative

On se place maintenant dans le cas où la variable réponse Y est qualitative et on souhaite expliquer cette variable Y en fonction d'un certain nombre de régresseurs $z^{(1)}, \ldots, z^{(m)}$. Voici quelques exemples illustratifs :

Exemple 1. *Une compagnie d'assurance cherche à détecter les dossiers frauduleux. Elle dispose pour cela d'un panel de n dossiers. À chacun de ces dossiers est associé la valeur* 0 *(pour dossier frauduleux) ou* 1*. Après avoir sélectionné les caractéristiques les plus intéressantes (endettement du foyer, milieu social, lieu de résidence, ...), elle cherche à savoir dans quelle mesure ces dernières variables influencent la probabilité d'existence d'une fraude. Elle espère pouvoir ainsi à l'avenir détecter d'éventuels dossiers "sensibles". On est dans le cas d'une variable réponse Y binaire.*

Exemple 2. *On souhaite expliquer le nombre d'espèces de plantes qui se développent dans différents lieux en fonction de la biomasse de ces différents lieux et du pH du sol. La variable réponse Y prend ici ses valeurs dans* $\mathbb{N}$*.*

Dans le cas d'une variable réponse binaire, on dispose d'un vecteur d'observations $Y = (Y_1, \ldots, Y_n)'$ où $Y_i \sim \mathcal{B}(\pi_i)$ pour tout $i \in \{1, \ldots, n\}$ et de m régresseurs $z^{(1)}, \ldots, z^{(m)}$. Il semblerait assez naturel d'utiliser la modélisation suivante :

$$\mathbb{E}[Y_i] = \pi_i = a_1 z_i^{(1)} + a_2 z_i^{(2)} + \cdots + a_m z_i^{(m)},\ i = 1, \ldots, n.$$

Cependant, comme on cherche à modéliser et prédire des probabilités, cette approche semble peu recommandée dans la mesure où certaines valeurs prédites pourraient ne pas

appartenir à l'intervalle $[0, 1]$. On va donc plutôt chercher à modéliser linéairement une fonction des π. Par exemple dans le cadre de la **régression logistique**, on considère la **fonction de lien** $g :]0, 1[\to \mathbb{R}$ définie par

$$g(t) = \ln\left(\frac{t}{1-t}\right), \ \forall t \in]0, 1[$$

et on modélise

$$g(\pi_i) = a_1 z_i^{(1)} + \cdots + a_m z_i^{(m)}, \ \forall i \in \{1, \ldots, n\}.$$

De manière plus générale, il est possible d'envisager de considérer d'autres distributions pour la variable Y et d'autres fonctions de lien. À ce titre, on pourra remarquer que le modèle de régression abordé au début de ce chapitre correspond à une distribution gaussienne et à une fonction de lien canonique (identité). Nous verrons qu'il est possible d'étudier tous ces modèles à travers un même cheminement : **le modèle linéaire généralisé**.

1.4 Objectifs du cours

Nous verrons que les modèles de régression linéaire (simple ou multiple), d'analyse de variance et d'analyse de covariance peuvent être rassemblés sous un même formalisme : on parle alors de **modèle linéaire général**. Un cran au-dessus, nous pouvons encore rassembler le modèle linéaire général et par exemple la régression logistique sous une même bannière : le **modèle linéaire généralisé**. Les possibilités offertes par ces différentes modélisations ne s'arrêtent pas à une simple écriture commune. C'est en fait tout le traitement et l'exploitation des données qui peuvent être abordés de manière unifiée.

Première partie

Le modèle linéaire général

Dans cette partie, nous allons nous intéresser au modèle linéaire général. Les chapitres 2, 3, 4 et 5 sont des chapitres "théoriques" définissant le cadre du modèle linéaire général, traite le problème de l'estimation des paramètres et du test de sous-modèle de Fisher. Les chapitres 6, 7 et 8 sont dédiés à la régression linéaire, l'ANOVA et l'ANCOVA respectivement. Ils permettront d'illustrer les notions vues précédemment dans ces exemples classiques du modèle linéaire général.

Chapitre 2

Définitions générales

2.1 Modèle linéaire régulier

Définition 2.1. *Une variable Y constituée de n observations Y_i suit un* ***modèle linéaire statistique*** *si Y peut être écrite sous la forme :*

$$Y = X\theta + \varepsilon, \tag{2.1}$$

où

- *X est une matrice réelle à n lignes et k colonnes avec $\mathbf{k < n}$, $X \in \mathcal{M}_{n,k}(\mathbb{R})$,*
- *θ est un vecteur réel inconnu de taille k,*
- *le vecteur $\varepsilon \in \mathbb{R}^n$ représente l'erreur du modèle.*

Cette définition est très générale et dépasse largement le cadre de la régression et de l'analyse de variance. L'hypothèse $k < n$ signifie que le nombre d'observations doit être supérieur au nombre de paramètres à estimer. C'est en quelque sorte une hypothèse d'identifiabilité.

Définition 2.2. *Le modèle linéaire (2.1) est dit* ***régulier*** *si la matrice X est régulière, c'est-à-dire de rang k. Dans le cas contraire où X est de rang $r < k$, on parle de modèle* ***singulier****.*

Proposition 2.1. *Soit $X \in \mathcal{M}_{n,k}(\mathbb{R})$. Les propositions suivantes sont équivalentes :*

- *X est une matrice de rang k.*
- *L'application $X : \mathbb{R}^k \to \mathbb{R}^n$ est injective.*
- *La matrice $X'X$ est inversible.*

Ainsi, si X est régulière alors par injectivité de l'application X, $X\theta = 0_n \Rightarrow \theta = 0_k$ pour tout $\theta \in \mathbb{R}^k$. Cette propriété assure que les colonnes de X sont linéairement indépendantes dans $\mathbb{R}^n$ et garantit l'unicité de θ. Dans certaines situations, la matrice considérée X ne pourra être régulière. Nous verrons cependant (cf section 5.2) qu'il est parfois possible de pallier ce problème en rajoutant des contraintes dites d'identifiabilité sur les paramètres à estimer. À moins que cela ne soit mentionné explicitement, la matrice X sera supposée régulière par la suite.

Proposition 2.2. *Soit $X \in \mathcal{M}_{n,k}(\mathbb{R})$ une matrice régulière. Alors la matrice de projection sur $[X] = Im(X)$ est donnée par $P_{[X]} = X(X'X)^{-1}X'$. Cette matrice $P_{[X]}$, souvent notée H, est appelée la* **matrice chapeau ou Hat Matrix**.

Démonstration. Soit $H := X(X'X)^{-1}X'$ où $X \in \mathcal{M}_{n,k}(\mathbb{R})$ une matrice régulière.
Pour tout $u \in \mathbb{R}^n$, on a $u = Hu + u - Hu$ et $Hu = X(X'X)^{-1}X'u \in [X]$. On va montrer que $u - P_{[X]}u \in [X]^\perp$: $\forall v \in \mathbb{R}^k$,

$$\begin{aligned}(Xv)'(u - P_{[X]}u) &= v'X'(u - X(X'X)^{-1}X'u) \\ &= v'X'u - v'(X'X)(X'X)^{-1}X'u = 0\end{aligned}$$

□

Afin de pouvoir travailler plus simplement et d'aller plus loin dans l'étude de ce modèle, nous allons maintenant imposer quelques restrictions concernant le vecteur ε :

- **Hypothèse H1** : Les erreurs sont centrées : $\mathbb{E}[\varepsilon] = 0_n$.
  Cette hypothèse est relativement importante et assure que le modèle est correctement défini. En effet, s'il s'avérait que $\mathbb{E}[\varepsilon] \neq 0_n$, cela pourrait signifier qu'une partie de l'information n'a pas été prise en compte dans la modélisation de $\mathbb{E}[Y]$.
  En fait cette hypothèse suppose que

  $$\mathbb{E}[Y] = X\theta = \sum_{j=1}^{k} \theta_j x^{(j)}$$

  où $x^{(j)}$ désigne la colonne j de la matrice X. En d'autres termes, l'écriture de ce modèle suppose que l'ensemble des variables $x^{(j)}$ est censé expliquer Y par une relation de cause à effet. Ainsi les variables $x^{(j)}$ sont appelées *variables explicatives* ou *prédicteurs*. Au final, en moyenne Y s'écrit donc comme une combinaison linéaire des $x^{(j)}$: la liaison entre les $x^{(j)}$ et Y est de nature linéaire. C'est la raison pour laquelle ce modèle est appelé **modèle linéaire**.
- **Hypothèse H2** : La variance des erreurs est constante :

  $$\mathbb{E}[\varepsilon_i^2] = \sigma^2,\ \forall i = 1, \ldots, n$$

  où σ^2 est un paramètre inconnu, à estimer. Cette hypothèse impose la caractéristique suivante sur Y : $\forall i = 1, \cdots, n$, $Var(Y_i) = \sigma^2$.
  Il est souvent raisonnable de supposer que **H2** est bien vérifiée. Dans la situation où ce ne serait pas le cas, il est possible de mettre en place un traitement statistique du modèle linéaire... cela nécessite cependant bien plus de travail.
- **Hypothèse H3** : Les variables ε_i sont indépendantes.
  Nous considèrerons que cette hypothèse est vérifiée lorsque chaque donnée correspond à un échantillonnage indépendant ou à une expérience physique menée dans des conditions indépendantes. Il existe un certain nombre de cas où ce postulat ne peut s'appliquer. On pourra par exemple penser aux séries temporelles : l'erreur du passé peut avoir une influence sur l'erreur future. Ces dernières font appel à un traitement statistique particulier (processus ARMA par exemple).

- **Hypothèse H4** : Les données suivent des lois gaussiennes

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2),\ \forall i \in \{1, \cdots, n\}.$$

Cette hypothèse est la moins importante puisque nous pouvons nous en passer quand le nombre de données est important.

Il découle des hypothèses **H1-H4** la normalité de Y :

$$Y \sim \mathcal{N}_n\left(X\theta, \sigma^2 I_n\right).$$

L'hypothèse de normalité des erreurs peut se justifier :

1. par un argument théorique : les erreurs sont caractérisables comme des erreurs de mesure. Ce sont une accumulation de petits aléas non-maîtrisables et indépendants. Par exemple, la mesure du poids d'un animal peut être soumise à des fluctuations dues à des erreurs de mesure à la pesée, à l'état de santé de l'animal, à son bagage génétique, à l'effet individuel de l'animal à prendre plus ou moins du poids. D'après le Théorème Central Limite, si tous ces effets sont indépendants de même moyenne nulle et de même "petite" variance, leur somme tend vers une variable gaussienne. La distribution gaussienne modélise assez bien toutes les situations où le hasard est la résultante de plusieurs causes indépendantes les unes des autres ; les erreurs de mesure suivent généralement assez bien la loi gaussienne.
2. par un argument pratique : il est facile de contrôler si une variable aléatoire suit une loi normale. En étudiant a posteriori la distribution des résidus calculés (erreurs estimées) et en la comparant à la distribution théorique (normale), on constate souvent qu'elle peut être considérée comme s'approchant de la loi gaussienne.

Dans la littérature statistique, un certain nombre de méthodes, souvent graphiques, sont proposées afin de vérifier la satisfaction des hypothèses **H1-H4**. Nous les aborderons à la section 6.6.

2.2 Exemples de modèle linéaire gaussien

2.2.1 Le modèle de régression linéaire

On cherche à modéliser une variable quantitative Y en fonction de variables explicatives quantitatives $x^{(1)}, \cdots, x^{(p)}$. Sous l'hypothèse gaussienne, le modèle de régression linéaire s'écrit :

$$Y_i = \theta_0 + \theta_1 x_i^{(1)} + \cdots + \theta_p x_i^{(p)} + \varepsilon_i,$$

avec $\theta_0, \theta_1, \cdots, \theta_p$ paramètres inconnus et $\varepsilon_1, \cdots, \varepsilon_n$ i.i.d de loi $\mathcal{N}(0, \sigma^2)$ avec σ^2 à estimer. Matriciellement, le modèle peut se réécrire sous la forme

$$Y = X\theta + \varepsilon$$

avec $\theta = (\theta_0, \theta_1, \cdots, \theta_p)'$ et $X = (\mathbb{1}_n, x^{(1)}, \ldots, x^{(p)}) \in \mathcal{M}_{n,(p+1)}(\mathbb{R})$.

Exercice 3. *Quelle est la loi de Y_i ? Quelle est la loi de Y ?*

Le modèle de régression linéaire sera étudié en détail dans le chapitre 6.

2.2.2 Le modèle d'analyse de la variance

On cherche à modéliser une variable quantitative Y en fonction d'une (ou de plusieurs) variable(s) explicative(s) qualitative(s) (appelée facteur). Sous l'hypothèse gaussienne, le modèle à un facteur à I modalités s'écrit :

$$Y_{ij} = \mu_i + \varepsilon_{ij} \text{ pour } i = 1, \cdots, I \,;\, j = 1, \cdots, n_i, \tag{2.2}$$

avec $\mu_1, \cdots, \mu_I$ des paramètres inconnus et $\varepsilon_{11}, \cdots, \varepsilon_{In_I}$ n observations indépendantes de loi $\mathcal{N}(0, \sigma^2)$ avec σ^2 à estimer.

Exercice 4. *Ecriture matricielle de ce modèle*
Afin d'écrire sous forme matricielle ce modèle, les observations sont rangées par modalité du facteur :

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

Soit $n = \sum_{i=1}^{I} n_i$. *Ecrivez le modèle* (2.2) *sous la forme*

$$Y = X\theta + \varepsilon$$

en précisant la matrice de design $X \in \mathcal{M}_{nI}(\mathbb{R})$ *et* $\theta \in \mathbb{R}^I$.
Quelle est la loi de Y_{ij}, *de* $Y_i = (Y_{i1}, \ldots, Y_{in_i})'$ *et de* Y ?

Le modèle d'analyse de la variance sera étudié en détail dans le chapitre 7.

En résumé :

- Modéle linéaire :

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

  avec $Y \in \mathbb{R}^n$, $X \in \mathcal{M}_{n,k}(\mathbb{R})$, $\theta \in \mathbb{R}^k$, $\varepsilon \in \mathbb{R}^n$
- Modèle régulier si $\text{rg}(X) = k$, sinon il est singulier
- Modèle régulier $\Leftrightarrow$ X injective $\Leftrightarrow$ $X'X$ inversible
- Matrice de projection orthogonale sur $[X] = Im(X)$:

  $$P_{[X]} = X(X'X)^{-1}X'$$

Chapitre 3

Estimation des paramètres

Dans ce chapitre, nous allons nous intéresser à l'estimation des paramètres dans un modèle linéaire général régulier :

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

où $X \in \mathcal{M}_{n,k}(\mathbb{R})$, $rg(X) = k$.

3.1 Estimation de θ

Dans cette section, nous nous intéressons à l'estimation du vecteur de paramètres θ. Pour cela, nous allons utiliser la **méthode des moindres carrés**. Il s'agit ici de trouver le vecteur $\hat{\theta}$ qui va minimiser la distance entre l'image de la matrice X et les observations Y. Autrement dit, l'estimateur de θ par la méthode des moindres carrés est défini par :

$$\begin{aligned} \widehat{\theta} &= \arg\min_{\vartheta} \|Y - X\vartheta\|^2 \\ &= \arg\min_{\vartheta} SCR(\vartheta). \end{aligned}$$

La norme $\|.\|$ est issue du produit scalaire usuel dans $\mathbb{R}^n$, i.e.

$$\|u\|^2 = \langle u, u \rangle = \sum_{i=1}^{n} u_i^2 = u'u, \ \forall u \in \mathbb{R}^n.$$

Sous forme matricielle, il est possible d'écrire :

$$\widehat{\theta} = \arg\min_{\vartheta} (Y - X\vartheta)'(Y - X\vartheta).$$

Théorème 3.1. *Soit Y suivant un modèle linéaire régulier. L'estimateur $\widehat{\theta}$ obtenu par la méthode des moindres carrés est*

$$\widehat{\theta} = (X'X)^{-1}X'Y. \tag{3.1}$$

Démonstration. On a :

$$\min_{\vartheta} \|Y - X\vartheta\|^2 = \min_{u \in Im(X)} \|Y - u\|^2 = \|Y - P_{[X]}Y\|^2,$$

où $P_{[X]}$ désigne la projection orthogonale sur $Im(X)$. Ainsi $X\hat{\theta} = P_{[X]}Y = X(X'X)^{-1}X'Y$. X étant supposée régulière, on en déduit que $\hat{\theta} = (X'X)^{-1}X'Y$ par unicité. □

Ce premier théorème nous donne donc une formule explicite pour l'estimateur du vecteur θ par la méthode des moindres carrés. Il est intéressant de noter que cette dernière est purement géométrique et ne demande aucune connaissance de la loi des erreurs. En effet, l'estimateur $\widehat{\theta}$ obtenu par la méthode des moindres carrés vérifie la propriété suivante :

$$X\widehat{\theta} = P_{[X]}Y.$$

Remarque : Dans le cas particulier où les erreurs sont gaussiennes, l'estimateur des moindres carrés $\widehat{\theta}$ correspond exactement à l'estimateur du maximum de vraisemblance. En effet, l'estimation par maximum de vraisemblance est basée sur la vraisemblance du modèle linéaire gaussien :

$$L(\theta, \sigma^2; y) = \prod_{i=1}^{n} f(y_i; \theta)$$

où $f(y_i; \theta)$ est la densité de la loi normale de la variable aléatoire Y_i . Ainsi

$$L(\theta, \sigma^2; (Y_1, \cdots, Y_n)) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{\|Y - X\theta\|^2}{2\sigma^2}\right).$$

Pour obtenir l'estimateur $\widehat{\theta}$ du maximum de vraisemblance, on maximise sa logvraisemblance selon θ. On remarque que par croissance de la fonction exponentielle cela revient à minimiser $\|Y - X\theta\|^2$.

Le résultat suivant explicite les performances de l'estimateur des moindres carrés.

Théorème 3.2. *Soient Y suivant un modèle linéaire régulier et $\hat{\theta}$ l'estimateur par la méthode des moindres carrés défini par* (3.1)*. Alors*

$$\mathbb{E}\left[\widehat{\theta}\right] = \theta \quad \text{et} \quad \text{Var}\left(\widehat{\theta}\right) = \sigma^2(X'X)^{-1}.$$

De plus, si les variables ε_i sont i.i.d, gaussiennes centrées, $\widehat{\theta}$ est le meilleur estimateur parmi tous les estimateurs sans biais de θ, i.e.

$$\text{Var}(C'\widetilde{\theta}) \geq \text{Var}(C'\widehat{\theta}),$$

pour tout $\widetilde{\theta}$ estimateur sans biais de θ et toute combinaison linéaire $C'\theta$, où $C \in \mathbb{R}^k$. Dans ce cas $\widehat{\theta}$ est un vecteur gaussien :

$$\widehat{\theta} \sim \mathcal{N}_k\left(\theta, \sigma^2(X'X)^{-1}\right).$$

Exercice 5.

- *Montrez que* $\mathbb{E}\left[\widehat{\theta}\right] = \theta$ *(rappel :* $\mathbb{E}[Y] = X\theta$*)*
- *Montrez que* $\mathrm{Var}\left(\widehat{\theta}\right) = \sigma^2(X'X)^{-1}$ *(rappel :* $\mathrm{Var}(AY) = A\mathrm{Var}(Y)A'$*)*
- *Pourquoi* $\widehat{\theta}$ *est un vecteur gaussien ?*

3.2 Valeurs ajustées et résidus

Une fois que l'on a estimé θ par $\widehat{\theta}$, on peut définir les **valeurs prédites (ou ajustées)** $\widehat{Y}_i$ par le modèle pour chaque Y_i :

$$\widehat{Y} = (\widehat{Y}_1, \cdots, \widehat{Y}_n)' = X\widehat{\theta} = X(X'X)^{-1}X'Y = P_{[X]}Y = HY.$$

On peut également estimer les erreurs ε_i par **les résidus** :

$$\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)' = Y - \widehat{Y} = (I_n - P_{[X]})Y = (I_n - H)Y.$$

Ayant des réalisations y_i, on obtient alors des valeurs prédites observées $\hat{y}_i = (\hat{Y}_i)^{obs} = (X\widehat{\theta}^{obs})_i$ et des résidus calculés $(\hat{\varepsilon}_i)^{obs} = y_i - \hat{y}_i$.

Proposition 3.1.

1. $\widehat{Y} \sim \mathcal{N}_n(X\theta, \sigma^2 H)$ *où* $H = X(X'X)^{-1}X'$
2. $\widehat{\varepsilon} \sim \mathcal{N}_n(0_n, \sigma^2(I_n - H))$
3. *Les variables aléatoires* $\widehat{Y}$ *et* $\widehat{\varepsilon}$ *sont indépendantes.*
4. *Les variables aléatoires* $\widehat{\theta}$ *et* $\widehat{\varepsilon}$ *sont indépendantes.*

Exercice 6. *Preuve de la proposition 3.1*

1. *Pour démontrer le premier point, utilisez la loi de* $\widehat{\theta}$
2. *Pour démontrer le deuxième point, on peut remarquer que* $\widehat{\varepsilon} = (I_n - H)Y$ *et* $Y \sim \mathcal{N}(X\theta, \sigma^2 I_n)$.
3. *Pour démontrer le troisième point, pensez au théorème de Cochran !*
4. *Pour démontrer le quatrième point, on peut remarquer que* $\widehat{\theta} = (X'X)^{-1}X'X\widehat{\theta}$.

3.3 Estimation de σ^2

Dans cette section, on s'intéresse à l'estimation de la variance des erreurs σ^2, appelée **variance résiduelle**. Par définition du modèle linéaire, la variance résiduelle σ^2 est également donnée comme la variance de Y pour X fixé. Dans le cadre de la régression linéaire, cela s'interprète comme la variance de Y autour de la droite de régression théorique. Cette définition de σ^2 suggère que son estimation est calculée à partir des écarts entre les valeurs observées y_i et les valeurs ajustées $\widehat{y}_i$.

Théorème 3.3. *Soit $\widehat{\theta}$ l'estimateur de θ par la méthode des moindres carrés. Sous les hypothèses H1-H4, et si $X \in \mathcal{M}_{nk}(\mathbb{R})$, alors*

$$\widehat{\sigma^2} = \frac{\|\widehat{\varepsilon}\|^2}{n-k} = \frac{\|Y - \widehat{Y}\|^2}{n-k} = \frac{\|Y - X\widehat{\theta}\|^2}{n-k} = \frac{SCR(\widehat{\theta})}{n-k}$$

est un estimateur sans biais optimal de σ^2, indépendant de $\widehat{\theta}$.
De plus,

$$\widehat{\sigma^2} \sim \frac{\sigma^2}{n-k}\ \chi^2(n-k).$$

Exercice 7. *Preuve du théorème 3.3*

- *Montrez que $SCR(\widehat{\theta}) := \|Y - X\widehat{\theta}\|^2 = \|P_{[X]^\perp}\varepsilon\|^2$*
- *A l'aide du théorème de Cochran, montrez que $SCR(\widehat{\theta}) \sim \sigma^2\chi^2(n-k)$. Déduisez-en que $\widehat{\sigma^2}$ est un estimateur sans biais de σ^2.*
- *Comme $X\widehat{\theta} = X\theta + P_{[X]}\varepsilon$, montrez que $X\widehat{\theta}$ et $\widehat{\sigma^2}$ sont indépendants.*
- *Déduisez-en que $\widehat{\theta}$ et $\widehat{\sigma^2}$ sont indépendants.*

L'estimation de σ^2 est donc

$$(\widehat{\sigma^2})^{obs} = \frac{\|(\widehat{\varepsilon})^{obs}\|^2}{n-k} = \frac{\|y - \widehat{y}\|^2}{n-k}.$$

Le dénominateur $(n-k)$ provient du fait que l'on a déjà estimé k paramètres dans le modèle.

3.4 Erreurs standards de $\widehat{\theta_j}$, $\widehat{Y_i}$, $\widehat{\varepsilon_i}$

La matrice de variance-covariance de $\widehat{\theta}$ notée $\Gamma_{\widehat{\theta}} = \sigma^2(X'X)^{-1}$ est estimée par

$$\widehat{\Gamma_{\widehat{\theta}}} = \widehat{\sigma^2}(X'X)^{-1}.$$

Ainsi $Var\left(\widehat{\theta_j}\right)$ est estimée par $\widehat{\sigma^2}[(X'X)^{-1}]_{jj}$.
Par conséquent, **l'erreur standard** de $\widehat{\theta_j}$, notée se_j, vaut

$$se_j = \sqrt{\widehat{\sigma^2}[(X'X)^{-1}]_{jj}}.$$

La matrice des corrélations de $\widehat{\theta}$ a pour élément j, j' :

$$r(\widehat{\theta_j}, \widehat{\theta_{j'}}) = \frac{\widehat{\sigma^2}[(X'X)^{-1}]_{jj'}}{se_j \times se_{j'}} = \frac{[(X'X)^{-1}]_{jj'}}{\sqrt{[(X'X)^{-1}]_{jj}[(X'X)^{-1}]_{j'j'}}}.$$

La variance $Var(\widehat{Y}) = \sigma^2\ H = \sigma^2 X(X'X)^{-1}X'$ est estimée par $\widehat{\sigma^2}H$.
Par conséquent, $\sqrt{\widehat{\sigma^2}H_{ii}}$ est l'erreur standard de $\widehat{Y_i}$.

De même, $\sqrt{\widehat{\sigma^2}(1-H_{ii})}$ correspond à l'erreur de $\widehat{\varepsilon}_i$.
Ainsi $\widehat{\varepsilon}_i/\sqrt{\widehat{\sigma^2}}$ désigne le **résidu standardisé**
et $\widehat{\varepsilon}_i/\sqrt{\widehat{\sigma^2}(1-H_{ii})}$ désigne le **résidu studentisé**.

3.5 Intervalle de confiance de θ_j, de $(X\theta)_i$ et de $X_0\theta$

3.5.1 Intervalle de confiance de θ_j

Sachant que $\widehat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2(X'X)^{-1})$, on a $\widehat{\theta_j} \sim \mathcal{N}(\theta_j, \sigma^2[(X'X)^{-1}]_{jj})$. Par conséquent

$$\frac{\widehat{\theta_j} - \theta_j}{\sqrt{\sigma^2[(X'X)^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$$

et la variable aléatoire $(n-k)\widehat{\sigma^2}/\sigma^2$ est distribuée selon une loi $\chi^2(n-k)$.
D'après le théorème de Cochran, ces deux variables aléatoires étant indépendantes, on peut donc dire que

$$T = \frac{\widehat{\theta_j} - \theta_j}{\sqrt{\sigma^2[(X'X)^{-1}]_{jj}}} \Big/ \sqrt{\frac{(n-k)\widehat{\sigma^2}}{(n-k)\sigma^2}} = \frac{\widehat{\theta_j} - \theta_j}{\sqrt{\widehat{\sigma^2}[(X'X)^{-1}]_{jj}}} \sim \mathcal{T}(n-k).$$

Si on note $t_{n-k,1-\frac{\alpha}{2}}$ le $(1-\alpha/2)$-quantile de la loi de Student à $(n-k)$ ddl, alors l'intervalle de confiance du paramètre θ_j de sécurité $1-\alpha$ est défini par :

$$IC_{1-\alpha}(\theta_j) = \left[\widehat{\theta_j} \pm t_{n-k,1-\frac{\alpha}{2}} \sqrt{\widehat{\sigma^2}[(X'X)^{-1}]_{jj}}\right] = \left[\widehat{\theta_j} \pm t_{n-k,1-\frac{\alpha}{2}}\ se_j\right].$$

3.5.2 Intervalle de confiance de $(X\theta)_i$

Soit $\mathbb{E}[Y_i] = (X\theta)_i$ la réponse moyenne de Y_i. On l'estime par $\widehat{Y_i} = (X\widehat{\theta})_i$. Puisque $\widehat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2(X'X)^{-1})$, d'après les propriétés des vecteurs gaussiens (Théorème A.1), la loi de $\widehat{Y_i}$ est $\mathcal{N}\left((X\theta)_i, \sigma^2[X(X'X)^{-1}X']_{ii}\right)$. De plus, $(n-k)\hat{\sigma}^2 \sim \sigma^2\chi^2(n-k)$ et $\hat{\theta}$ et $\hat{\sigma}^2$ sont indépendants. On obtient donc que

$$\frac{\widehat{Y_i} - (X\theta)_i}{\sqrt{\widehat{\sigma}^2[X(X'X)^{-1}X']_{ii}}} \sim \mathcal{T}(n-k).$$

L'intervalle de confiance de $(X\theta)_i$ au niveau de confiance de $1-\alpha$ est donc donné par :

$$IC_{1-\alpha}((X\theta)_i) = \left[\widehat{Y_i} \pm t_{n-k,1-\alpha/2} \times \sqrt{\widehat{\sigma}^2[X(X'X)^{-1}X']_{ii}}\right].$$

3.5.3 Intervalle de confiance de $X_0\theta$

On considère des nouvelles valeurs pour les variables explicatives, rassemblées dans le vecteur ligne $X_0 \in \mathcal{M}_{1k}(\mathbb{R})$. La réponse moyenne est alors $X_0\theta$. L'estimateur de $X_0\theta$ est $\widehat{Y_0} = X_0\widehat{\theta}$. Puisque $\widehat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2(X'X)^{-1})$, d'après les propriétés des vecteurs gaussiens (Théorème A.1), la loi de cet estimateur est

$$\widehat{Y_0} = X_0\widehat{\theta} \sim \mathcal{N}(X_0\theta, \sigma^2 X_0(X'X)^{-1}X_0').$$

Ainsi l'intervalle de confiance de $X_0\theta$ au niveau de confiance de $1-\alpha$ s'écrit :

$$IC_{1-\alpha}(X_0\theta) = \left[\widehat{Y_0} \pm t_{n-k,1-\alpha/2} \times \sqrt{\widehat{\sigma^2} X_0(X'X)^{-1}X_0'}\right].$$

3.6 Intervalles de prédiction

Avant toute chose, il est important de comprendre la différence entre l'intervalle de confiance de $X_0\theta$ et l'intervalle de prédiction. Dans les deux cas, on suppose un nouveau jeu de valeurs données des variables explicatives. Dans le premier cas, on veut prédire une réponse moyenne correspondant à ces variables explicatives alors que dans le second cas, on cherche à prédire une nouvelle valeur "individuelle". Par exemple, si on étudie la liaison entre le poids et l'âge d'un animal, on peut prédire la valeur du poids à 20 jours soit comme le poids moyen d'animaux à 20 jours, soit comme le poids à 20 jours d'un nouvel animal. Pour le nouvel animal, on doit prendre en compte la variabilité individuelle, ce qui augmente la variance de l'estimateur et donc la largeur de l'intervalle.

Si on veut prédire dans quel intervalle se trouvera le résultat d'un nouvel essai $X_0 \in \mathcal{M}_{1k}(\mathbb{R})$, on doit tenir compte de deux facteurs d'incertitude :

- l'incertitude sur l'estimation du résultat moyen de l'essai $X_0\theta$,
- l'incertitude sur le terme d'erreur ε_0.

Le vecteur de paramètres θ est estimé par

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

où $Y = (Y_1, \ldots, Y_n)'$. Une nouvelle observation Y_0, correspondant à X_0, s'écrit :

$$Y_0 = X_0\theta + \varepsilon_0,$$

où ε_0 est supposé indépendant des ε_i, $1 \leq i \leq n$ et $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$.
Le modèle linéaire prédit la valeur

$$\widehat{Y_0} = X_0\widehat{\theta} \sim \mathcal{N}(X_0\theta, \sigma^2 X_0(X'X)^{-1}X_0').$$

D'après les hypothèses sur ε_0, on a que $Y_0 \sim \mathcal{N}(X_0\theta, \sigma^2)$ et Y_0 est indépendant de $\hat{Y}_0$. On a donc

$$Y_0 - \widehat{Y_0} \sim \mathcal{N}(0, \sigma^2\left(1 + X_0(X'X)^{-1}X_0'\right)).$$

Par ailleurs, d'après le Théorème 3.3

$$\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - X\widehat{\theta})^2 \sim \frac{\sigma^2}{n-k}\chi^2(n-k)$$

et comme $\widehat{\sigma}^2$ est indépendant de $\widehat{\theta}$ et de ε_0 (car ε_0 indépendant des ε_i), la variable aléatoire

$$\frac{Y_0 - \widehat{Y_0}}{\widehat{\sigma}\sqrt{1 + X_0(X'X)^{-1}X_0'}} \sim \mathcal{T}(n-k).$$

Au final, en notant $t_{n-k,1-\alpha/2}$ le $1-\alpha/2$ quantile d'une loi de Student à $n-k$ degrés de liberté, on obtient

$$\mathbb{P}\left(Y_0 \in \left[\widehat{Y_0} \pm t_{n-k,1-\alpha/2}\widehat{\sigma}\sqrt{1 + X_0(X'X)^{-1}X_0'}\right]\right) = 1-\alpha.$$

Par conséquent, l'intervalle de prédiction de la variable Y pour une nouvelle observation au point X_0 est défini par

$$IC_{1-\alpha}(Y_0) = \left[\widehat{Y_0} \pm t_{n-k,1-\alpha/2} \times \widehat{\sigma}\sqrt{1 + X_0(X'X)^{-1}X_0'}\right].$$

Notez bien la différence entre $IC_{1-\alpha}(Y_0)$ et

$$IC_{1-\alpha}(X_0\theta) = \left[\widehat{Y_0} \pm t_{n-k,1-\alpha/2} \times \widehat{\sigma}\sqrt{X_0(X'X)^{-1}X_0'}\right].$$

3.7 Décomposition de la variance

La mise en oeuvre d'un modèle linéaire a pour objectif d'expliquer la variabilité d'une variable Y par d'autres variables. On note :

- $SCT = \|Y - \bar{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = n\ var(Y)$ la **variabilité totale** de Y.
- $SCE = \|\widehat{Y} - \bar{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(\widehat{Y_i} - \overline{Y})^2 = n\ var(\widehat{Y})$ la **variabilité expliquée** par le modèle, c'est-à-dire par les prédicteurs.
- $SCR = \|Y - \widehat{Y}\|^2 = \sum_{i=1}^{n}(\widehat{\varepsilon_i})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2 = n\ var(\widehat{\varepsilon})$ la **variabilité résiduelle** non expliquée par le modèle.

La variance totale de Y admet alors la décomposition suivante :

$$var(Y) = var(\widehat{Y}) + var(\widehat{\varepsilon})$$

c'est-à-dire

$$SCT = SCE + SCR.$$

Exercice 8. *Démontrez ce résultat.*

On verra par la suite que selon le modèle étudié, cette décomposition amène à des définitions spécifiques à chaque modèle.

D'après le critère des moindres carrés utilisé pour estimer les paramètres, on cherche à minimiser la Somme des Carrés des Résidus SCR et donc à maximiser la Somme des Carrés Expliquée par le modèle SCE. Pour juger de la qualité d'ajustement du modèle aux données, on définit le critère R^2 qui représente la part de variance de Y expliquée par le modèle :

$$R^2 = \frac{SCE}{SCT} = \frac{var(\widehat{Y})}{var(Y)} \in [0,1]$$

Plus R^2 est proche de 1, plus le modèle s'ajuste aux données. Nous discuterons de l'efficacité de ce critère dans les chapitres suivants.

En résumé :

Dans le cadre d'un modèle linéaire régulier,

- $\widehat{\theta} = (X'X)^{-1}X'Y \sim \mathcal{N}_k(\theta, \sigma^2(X'X)^{-1})$
- $\widehat{\sigma^2} = \frac{\|Y - X\widehat{\theta}\|^2}{n-k} \sim \frac{\sigma^2}{n-k}\chi^2(n-k)$
- $\widehat{\theta}$ et $\widehat{\sigma^2}$ sont indépendants
- Connaitre les définitions de valeurs ajustées $\widehat{Y} = X\widehat{\theta} = P_{[X]}Y$ et de résidus $\widehat{\varepsilon} = Y - \widehat{Y}$
- Savoir refaire la construction
  - d'un IC pour un paramètre
  - d'un IC pour une réponse moyenne
  - d'un intervalle de prédiction pour une nouvelle réponse

  Surtout, ne pas apprendre par coeur les formules !
- Décomposition de la variance

$$\underbrace{\|Y - \bar{Y}\mathbb{1}_n\|^2}_{SCT} = \underbrace{\|Y - \widehat{Y}\|^2}_{SCR} + \underbrace{\|\widehat{Y} - \bar{Y}\mathbb{1}_n\|^2}_{SCE}$$

  et $R^2 = \frac{SCE}{SCT}$.

Chapitre 4

Test de Fisher-Snedecor

Nous allons nous intéresser dans ce chapitre à un certain nombre de tests pouvant être mis en oeuvre sur le modèle linéaire. Nous supposerons pendant toute cette partie que les hypothèses H1-H4 sont vérifiées. Les tests présentés ci-dessous ne peuvent être utilisés si ces contraintes ne sont pas satisfaites.

4.1 Hypothèses testées

On considère un modèle linéaire gaussien

$$Y = X\theta + \varepsilon \text{ avec } \varepsilon \sim \mathcal{N}_n\left(0_n, \sigma^2 I_n\right) \tag{4.1}$$

et on s'intéresse à examiner la nullité de certaines composantes du paramètre θ ou de certaines combinaisons linéaires des composantes de θ, par exemple : $\theta_j = 0$; $\theta_j = \theta_k = 0$ ou $\theta_j = \theta_k$. Ces hypothèses reposent sur la notion de modèles emboîtés : deux modèles sont dits **emboîtés** si l'un peut être considéré comme un cas particulier de l'autre. Cela revient à comparer un modèle de référence à un modèle réduit ou contraint. Cette approche vise donc à déterminer si le modèle utilisé peut être oui ou non simplifié. Voici deux exemples de sous-modèles :

Modèle général de la régression linéaire simple :	$Y_i = a + bX_i + \varepsilon_i$
Sous-modèle avec nullité de la pente :	$Y_i = a + \varepsilon_i$
Modèle général de l'analyse de variance à 1 facteur :	$Y_{ij} = \mu_i + \varepsilon_{ij}$
Sous-modèle avec égalité des groupes :	$Y_{ij} = \mu + \varepsilon_{ij}$

Par la suite, nous allons considérer deux écritures équivalentes de l'hypothèse nulle $\mathcal{H}_0$, la première est plus pratique et la deuxième est plus théorique.

Ecriture 1 : Pour spécifier la nullité de certaines composantes du paramètre θ, on introduit la matrice $C \in \mathcal{M}_{qk}(\mathbb{R})$ où k désigne le nombre de paramètres du modèle de référence et q le nombre de contraintes testées ($1 \leq q \leq k$) telle que :

$$\mathcal{H}_0 : C \in \mathcal{M}_{qk}(\mathbb{R}) \text{ telle que } C\theta = 0_q.$$

La matrice C sera supposée être de rang q.

Exercice 9. *On suppose un modèle à $k = 3$ paramètres. Dans les trois cas suivants, précisez la matrice C :*

- *Tester l'hypothèse $\mathcal{H}_0 : \theta_2 = 0$*
- *Tester l'hypothèse $\mathcal{H}_0 : \theta_3 = \theta_2$*
- *Tester l'hypothèse $\mathcal{H}_0 : \theta_3 = \theta_2 = 0$*

Ecriture 2 : Plaçons-nous dans le cadre général du modèle linéaire. Soit le modèle (4.1) et soit X_0 une matrice telle que $Im(X_0) \subset Im(X)$ et $k_0 = dim(Im(X_0)) < k = dim(Im(X))$. Le modèle défini par

$$Y = X_0\beta + \varepsilon, \tag{4.2}$$

est appelé **sous-modèle** issu du modèle linéaire défini en (4.1). Le plus souvent, X_0 est la matrice constituée de k_0 vecteurs colonnes de X avec $k_0 < k$ et β est un vecteur de longueur k_0. Nous notons alors SCR_0 la somme des carrés des résidus de ce sous-modèle, associée à $n - k_0$ degrés de liberté et définie de la façon suivante

$$SCR_0 = \|Y - X_0\widehat{\beta}\|^2,$$

où $\widehat{\beta}$ est l'estimateur des moindres carrés issus du modèle (4.2) pour β. Dans la mesure où $Im(X_0) \subset Im(X)$ et par définition des estimateurs des moindre carrés, nous pouvons remarquer que $SCR_0 \geq SCR$.

Il peut être parfois intéressant d'essayer de savoir si les observations sont issues du modèle (4.1) ou (4.2). Soit le modèle défini par :

$$Y = R + \varepsilon.$$

Tester la présence d'un sous-modèle revient donc à tester :

$$\mathcal{H}_0 : R \in Im(X_0) \text{ contre } \mathcal{H}_1 : R \in Im(X) \backslash Im(X_0).$$

4.2 Le test de Fisher-Snedecor

4.2.1 Principe

Le test de Fisher-Snedecor est la règle de décision qui permet de décider si on rejette ou ne rejette pas $\mathcal{H}_0 : C\theta = 0_q$, c'est-à-dire $\mathcal{H}_0 : R \in Im(X_0)$:

- Rejeter $\mathcal{H}_0$, c'est décider que $C\theta \neq 0_q$, c'est-à-dire que certaines composantes de $C\theta$ ne sont pas nulles. Nous n'avons donc pas confiance dans le sous-modèle et nous préfèrerons continuer à travailler avec le modèle de référence.
- Ne pas rejeter $\mathcal{H}_0$, c'est ne pas exclure que toutes les composantes de $C\theta$ sont nulles. Dans ce cas, il n'est pas nécessaire de conserver un modèle trop compliqué et nous préfèrerons conserver le modèle contraint pour expliquer les données.

4.2.2 La statistique de test

Théorème 4.1. *Dans le cadre du modèle linéaire général (4.1) avec les hypothèses H1-H4 et les notations précédentes, sous l'hypothèse nulle $\mathcal{H}_0$ (le sous-modèle (4.2) est vrai), la variable*

$$F = \frac{(SCR_0 - SCR)/(k - k_0)}{SCR/(n-k)} = \frac{\|\widehat{Y} - \widehat{Y_0}\|^2/(k-k_0)}{\|Y - \widehat{Y}\|^2/(n-k)},$$

suit une loi de Fisher de paramètres $(k - k_0, n - k)$. De plus, F est indépendante de $\widehat{Y_0} = X_0\widehat{\beta}$ (calculé sous l'hypothèse $\mathcal{H}_0$).

Exercice 10.

- *Montrez que $SCR = \|P_{[X]^\perp}\varepsilon\|^2 \sim \sigma^2\chi^2(n-k)$*
- *Soit A un sous-espace vectoriel de $Im(X) = [X]$ tel que $A \overset{\perp}{\oplus} Im(X_0) = Im(X)$, $dim(A) = k - k_0$. Montrez que $SCR_0 - SCR = \|P_A\varepsilon\|^2 \underset{\mathcal{H}_0}{\sim} \sigma^2\chi^2(k-k_0)$*
- *Déduisez-en que $F \underset{\mathcal{H}_0}{\sim} \mathcal{F}(k-k_0, n-k)$.*
- *Montrez que F est indépendante de $\widehat{Y_0}$ et $\widehat{\beta}$.*

Cette statistique de test peut s'écrire sous une autre forme donnée dans la proposition suivante :

Proposition 4.1. *La statistique de test de Fisher-Snedecor peut également s'écrire sous la forme suivante :*

$$F = \frac{[C\widehat{\theta}]' \left[C(X'X)^{-1}C'\right]^{-1} [C\widehat{\theta}]}{q\widehat{\sigma^2}} \text{ avec } q = k - k_0.$$

Démonstration. La preuve de cette proposition est donnée en annexe B.1. □

Cette dernière expression a l'avantage de ne pas nécessiter l'estimation du modèle contraint pour tester $\mathcal{H}_0 : C\theta = 0_q$ contre $\mathcal{H}_1 : C\theta \neq 0_q$.

Par la suite, on notera F^{obs} la valeur observée de la variable aléatoire F.

4.2.3 Règle de décision

La quantité d'importance dans notre construction du test de Fisher est $SCR_0 - SCR$. Intuitivement, si la valeur observée de $SCR_0 - SCR$ est très grande, il y a peu de chance que les observations Y soient "issues" du sous-modèle. À l'opposé, si la valeur observée $SCR_0 - SCR$ est petite, il est fort possible que le modèle initial puisse être simplifié : le sous-modèle explique aussi bien les observations dans la mesure où SCR_0 est comparable à SCR. Par conséquent, la zone de rejet avec un risque de première espèce α s'écrit

$$\mathcal{R}_\alpha = \{F > f_{q,n-k,1-\alpha}\}$$

où $f_{q,n-k,1-\alpha}$ est le $(1-\alpha)$-quantile de la distribution de Fisher de degrés de liberté $q = k - k_0$ et $n - k$.

4.2.4 Cas particulier où $q = 1$: Test de Student

Dans le cas particulier où l'on teste la nullité d'une seule combinaison linéaire des composantes du paramètre θ, i.e. $q = 1$ et $C \in \mathcal{M}_{1,k}(\mathbb{R})$, alors l'hypothèse nulle s'écrit :

$$\mathcal{H}_0 : C\theta = 0.$$

On a donc $C(X'X)^{-1}C' \in \mathbb{R}$ et la variable aléatoire F s'écrit alors de la façon suivante :

$$F = \frac{(C\widehat{\theta})^2}{\widehat{\sigma^2}C(X'X)^{-1}C'}.$$

F suit une loi de Fisher à 1 et $n - k$ degrés de liberté. Or une propriété de la distribution de Fisher-Snedecor est qu'une distribution de Fisher-Snedecor à 1 et m_2 degrés de liberté est le carré d'une distribution de Student à m_2 degrés de liberté. Par conséquent, on obtient l'égalité suivante : si $A \sim \mathcal{F}(1, n-k)$ et $T \sim \mathcal{T}(n-k)$,

$$\mathbb{P}\left[A \geq f_{1,n-k,1-\alpha}\right] = \alpha = \mathbb{P}\left[T^2 \geq f_{1,n-k,1-\alpha}\right].$$

On en déduit donc la propriété suivante sur les quantiles :

$$f_{1,n-k,1-\alpha} = t^2_{n-k,1-\alpha/2}.$$

Selon le test de Fisher, on rejette l'hypothèse $\mathcal{H}_0$ si $F \geq f_{1,n-k,1-\alpha}$. Or on a les équivalences suivantes :

$$\begin{aligned} F \leq f_{1,n-k,1-\alpha} &\iff |C\widehat{\theta}| \leq t_{n-k,1-\alpha/2}\sqrt{\widehat{\sigma^2}C(X'X)^{-1}C'} \\ &\iff -t_{n-k,1-\alpha/2}\sqrt{\widehat{\sigma^2}C(X'X)^{-1}C'} \leq C\widehat{\theta} \leq t_{n-k,1-\alpha/2}\sqrt{\widehat{\sigma^2}C(X'X)^{-1}C'}. \end{aligned}$$

Ainsi l'intervalle de confiance au niveau de sécurité $1-\alpha$ de $C\theta$ est

$$\left[C\widehat{\theta} \pm t_{n-k,1-\alpha/2}\sqrt{\widehat{\sigma^2}C(X'X)^{-1}C'}\right].$$

Au final, le test consiste donc à rejeter l'hypothèse nulle si et seulement si 0 n'appartient pas à l'intervalle de confiance de $C\theta$.

Exercice 11. *Construisez directement le test de Student de nullité du paramètre θ_j au niveau α.*

4.3 Intervalle (région) de confiance pour $C\theta$

4.3.1 IC pour $C\theta \in \mathbb{R}$

Commençons par l'intervalle de confiance pour une combinaison linéaire $C\theta \in \mathbb{R}$. Nous reprenons les notations de la section 4.2.4. Comme $\widehat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2(X'X)^{-1})$, on a $C\widehat{\theta} \sim \mathcal{N}(C\theta, \sigma^2\Delta)$ avec $\Delta = C(X'X)^{-1}C' \in \mathbb{R}$. De plus $(n-k)\widehat{\sigma^2}/\sigma^2 \sim \chi^2(n-k)$ et $\widehat{\theta}$ et $\widehat{\sigma^2}$ sont indépendantes. Ainsi,

$$\frac{C\widehat{\theta} - C\theta}{\widehat{\sigma}\sqrt{\Delta}} \sim \mathcal{T}(n-k).$$

On obtient ainsi l'intervalle de confiance suivant au niveau de confiance $1-\alpha$:

$$IC_{1-\alpha}(C\theta) = \left[C\widehat{\theta} \pm t_{n-k,1-\alpha/2}\sqrt{\widehat{\sigma^2}C(X'X)^{-1}C'}\right].$$

Rappelons le lien entre test et intervalle de confiance : l'ensemble des c_0 acceptés pour un test

$$\mathcal{H}_0 : C\theta = c_0 \text{ contre } \mathcal{H}_1 : C\theta \neq c_0$$

au niveau α, définit un intervalle de confiance au niveau de confiance $1-\alpha$.

4.3.2 Région de confiance pour $C\theta \in \mathbb{R}^q$

Si maintenant, comme dans la partie 4.2.2, $C\theta$ est de dimension $q > 1$ et si c_0 est une valeur particulière appartenant à $\mathbb{R}^q$, nous pouvons généraliser la construction de l'intervalle de confiance. Dans ce cas, $C\widehat{\theta} - C\theta \sim \mathcal{N}_q(0_q, \sigma^2\Delta)$ avec $\Delta = C(X'X)^{-1}C' \in \mathcal{M}_q(\mathbb{R})$. Ainsi

$$\frac{[C\widehat{\theta} - C\theta]'\Delta^{-1}[C\widehat{\theta} - C\theta]}{\sigma^2} \sim \chi^2(q).$$

On a aussi $(n-k)\widehat{\sigma^2}/\sigma^2 \sim \chi^2(n-k)$ et les deux statistiques sont indépendantes. On en déduit donc que

$$A := \frac{[C\widehat{\theta} - C\theta]'\Delta^{-1}[C\widehat{\theta} - C\theta]}{q\widehat{\sigma^2}} \sim \mathcal{F}(q, n-k).$$

Finalement,

$$\begin{aligned} & \mathbb{P}(A \leq f_{q,n-k,1-\alpha}) = 1-\alpha \\ \Leftrightarrow \quad & \mathbb{P}([C\widehat{\theta} - C\theta]'\Delta^{-1}[C\widehat{\theta} - C\theta] \leq q\widehat{\sigma^2} f_{q,n-k,1-\alpha}) = 1-\alpha \\ \Leftrightarrow \quad & \mathbb{P}(C\theta \in RC) = 1-\alpha \end{aligned}$$

où RC est l'ellipsoïde de confiance défini par :

$$RC = \left\{ u \in \mathbb{R}^q;\ (C\widehat{\theta} - u)'[C(X'X)^{-1}C']^{-1}(C\widehat{\theta} - u) \leq q\widehat{\sigma^2} f_{q,n-k,1-\alpha} \right\}.$$

L'ensemble des $c_0 \in \mathbb{R}^q$ acceptés par le test

$$\mathcal{H}_0 : C\theta = c_0 \text{ contre } \mathcal{H}_1 : C\theta \neq c_0$$

au niveau α forme l'ellipsoïde de confiance RC défini ci-dessus.

En résumé :

- Savoir écrire les hypothèses d'un test de Fisher de sous-modèle
- Savoir justifier qu'un modèle est sous-modèle d'un autre
- Connaitre la forme de la statistique du test de Fisher, sa loi sous $\mathcal{H}_0$ et savoir définir les quantités qui la composent selon le contexte (Théorème 4.1)
- Savoir mener la construction d'un test de Fisher de sous-modèle
- Savoir mener la construction d'un test de Student quand $q = 1$
- Savoir mener la construction d'un intervalle de confiance pour $C\theta$. Ne pas apprendre la formule !

Chapitre 5

Modèles singuliers, orthogonalité et importance des hypothèses sur les erreurs

5.1 Quand H1-H4 ne sont pas respectées...

L'hypothèse de gaussianité des erreurs est la plus difficile à vérifier en pratique. Les tests classiques de normalité (test de Kolmogorov-Smirnov, Cramer-Von Mises, Anderson-Darling ou de Shapiro-Wilks) demanderaient l'observation des erreurs ε_i elles-mêmes ; ils perdent beaucoup de leur puissance quand ils sont appliqués sur les résidus $\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$, notamment en raison du fait que ces résidus ne sont pas, en général, indépendants. Nous pouvons cependant toujours faire des droites de Henri ou des QQ-plots pour mettre en évidence des écarts évidents. Il n'en reste pas moins que l'hypothèse de gaussianité sera le plus souvent un credo que nous ne pourrons pas vraiment vérifier expérimentalement. Fort heureusement, il existe une théorie asymptotique (donc de grands échantillons) du modèle linéaire qui n'a pas besoin de cette hypothèse. Comme il est dit dans la section 2.1, c'est dans cette optique là qu'il faut réellement penser le modèle linéaire.

5.1.1 Propriétés de l'estimateur des moindres carrés $\widehat{\theta}$

Proposition 5.1. *Soit $\widehat{\theta} = (X'X)^{-1}X'Y$.*

- *$\widehat{\theta}$ reste sans biais, $\mathbb{E}\left[\widehat{\theta}\right] = \theta$, sous l'hypothèse H1.*
- *la matrice de variance-covariance de $\widehat{\theta}$ reste égale à $\sigma^2(X'X)^{-1}$ sous les hypothèses H2 et H3, mais si H1 n'est pas vraie cette propriété a peu d'intérêt.*
- *$\widehat{\theta}$ n'est plus un estimateur optimal parmi les estimateurs sans biais, mais il le reste parmi les estimateurs linéaires sans biais sous H1-H3.*
- *$\widehat{\theta}$ est gaussien sous H3 et H4. Si H4 n'est pas vraie, alors il tend à être gaussien pour de grands échantillons. On dit qu'il est asymptotiquement gaussien.*

5.1.2 Propriétés de l'estimateur des moindres carrés $\widehat{\sigma^2}$

Cette étude n'a bien sûr d'intérêt que si σ^2 est bien définie ce qui nécessite l'hypothèse H2. Nous considérons

$$\widehat{\sigma^2} = \frac{1}{n-k}\|Y - X\widehat{\theta}\|^2 \text{ avec } \widehat{\theta} = (X'X)^{-1}X'Y.$$

Proposition 5.2.

- *Sous les hypothèses H1-H3, $\widehat{\sigma^2}$ reste un estimateur sans biais de σ^2 même si l'hypothèse H4 n'est pas vérifiée : $\mathbb{E}\left[\widehat{\sigma^2}\right] = \sigma^2$.*
- *Il est clair que $(n-k)\widehat{\sigma^2}$ ne suit plus une loi $\sigma^2\chi^2(n-k)$ dès que l'hypothèse H4 n'est pas vérifiée.*
- *Nous montrons facilement que sous les hypothèses H1-H3, $\widehat{\sigma^2}$ converge en probabilité vers σ^2 quand le nombre d'observations devient grand, même si l'hypothèse H4 n'est pas vérifiée.*
- *Enfin, sous les seules hypothèses H1-H3, dès que la loi de ε_i admet un moment d'ordre 4, alors $\widehat{\sigma^2}$ converge à la vitesse $\sqrt{n}$ vers σ^2 mais sa vitesse exacte de convergence dépend du type de loi, plus précisément du coefficient de Kurtosis[5].*

5.1.3 Propriétés des statistiques de test T et F

Dans cette partie, nous considérons simplement le cas du modèle linéaire non-gaussien, sans l'hypothèse H4. Le résultat général est la validité asymptotique (pour de grands effectifs) des tests évoqués (sous certaines conditions peu restrictives), voir [5] pour les détails et les explications théoriques d'un tel résultat.

Pour illustrer cette validité dans un cas simple, prenons l'exemple de l'analyse de variance à un facteur. Nous avons vu que l'estimateur $\widehat{\mu_i}$ de la valeur moyenne d'une classe est une simple moyenne empirique : $\widehat{\mu_i} = Y_{i.} = \frac{1}{n}\sum_{j=1}^{n_i} Y_{ij}$. Pour "mesurer" la vitesse de convergence de $\widehat{\mu_i}$ vers μ_i, nous utilisons le Théorème de la Limite Centrale. Pour revenir au problème précédent d'analyse de la variance à un facteur, une conséquence de ce théorème est que pour n grand, $\widehat{\mu_i} = Y_{i.}$ suit approximativement une loi gaussienne et une conséquence de la Loi des Grands Nombres est que $\widehat{\sigma^2}$ est très proche de σ^2. Ceci implique par exemple que si nous voulons tester l'hypothèse "$\mu_i = p$", pour un i donné et avec un réel p connu, alors comme précédemment nous considèrerons la statistique de test :

$$\widehat{T} = \frac{\widehat{\mu_i} - p}{\sqrt{\widehat{\sigma^2}/n}}.$$

Pour n grand, $\widehat{T}$ suit approximativement une loi gaussienne centrée réduite qui n'est autre que la limite d'une loi de Student $\mathcal{T}(n)$ dont le nombre de degrés de liberté tend vers l'infini, voir [5].

Ce résultat théorique peut être complété par une étude de simulation. Dans un mémoire, Bonnet et Lansiaux [**?**] ont étudié le comportement du test de Fisher en analyse

de variance à un facteur à 2, 5 ou 10 niveaux, avec des indices de répétition de 2, 4 ou 8. Nous avons donc de 4 à 80 données dans chaque expérience. La validité du test est appréciée par le niveau réel du test pour un niveau nominal de 10%, 5% ou 1%.
Divers types de loi non normales sont utilisées. Nous apercevons un écart au comportement nominal du test seulement dans le cas où les éléments suivants sont réunis : dispositifs déséquilibrés, petits échantillons, loi dissymétrique.

Dans les autres cas, tout se passe comme si les données étaient gaussiennes. Ainsi cette étude confirme, que sauf cas extrêmes, le test de Fisher n'a pas besoin de l'hypothèse de gaussiannité pour être approximativement exact.

5.1.4 Modèles avec corrélations

Il est possible de modéliser des corrélations entre erreurs, par exemple en supposant que ces erreurs sont issues d'un processus ARMA, ce qui permet de ne plus avoir besoin de l'hypothèse H3, voir Guyon [14].
Il est également possible de modéliser les liaisons par des modèles à effets aléatoires et poser un modèle mixte, voir Pinheiro et Bates [20].

5.2 Modèles singuliers

Nous nous sommes jusqu'à présent cantonnés à l'étude des modèles linéaires réguliers. Or certains modèles ne peuvent être paramétrés de façon régulière : ils sont naturellement sur-paramétrés. Un exemple simple est celui du modèle additif en analyse de la variance à 2 facteurs. Considérons le cas où les 2 facteurs ont chacun 2 niveaux et que les 4 combinaisons sont observées une fois et une seule. On a donc, avec les notations vues précédemment :

$$Y_{i,j} = \mu + a - i + b_j + \varepsilon_{i,j},\ i \in \{1,2\},\ j \in \{1,2\}.$$

Le vecteur $\theta = (\mu, a_1, a_2, b_1, b_2)'$ et la matrice X du modèle vaut :

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Nous remarquons que tout vecteur de la forme $(\alpha+\beta, -\alpha, -\alpha, -\beta, -\beta)$ donne le vecteur nul lorsqu'il est multiplié par la matrice X. Les valeurs μ, a_i, b_i pour $i = 1$ ou 2 ne sont donc pas identifiables de manière unique. Le modèle est en fait **sur-paramétré** : nous avons 5 paramètres inconnus pour seulement 4 observations.

Définition 5.1. *Le modèle est dit **singulier** ou **non régulier** quand la matrice X est non injective, c'est-à-dire s'il existe $\theta \neq 0_k$ tel que $X\theta = 0_n$.*

Nous rappelons que $Ker(X) = \{u \in \mathbb{R}^k;\ Xu = 0_n\}$ désigne le noyau de X. Nous pouvons faire deux remarques :

- $X\widehat{\theta}$ reste unique, puisque c'est la projection de Y sur $Im(X)$.
- $\widehat{\theta}$ ne peut être unique puisque si $\widehat{\theta}$ est solution et si $u \in Ker(X)$ alors $\widehat{\theta} + u$ est encore solution.

Si X n'est pas régulière, alors la matrice $X'X$ n'est pas inversible. Pour contourner ce problème, nous définissons alors un inverse généralisé de $(X'X)$.

Définition 5.2. *Soit M une matrice. Alors la matrice M^- est* **une matrice inverse généralisée** *de M si $MM^-M = M$.*

Cette construction est toujours possible. En effet, $(X'X)$ définit une application bijective de $Ker(X)^\perp$ sur lui-même. Il suffit donc simplement de négliger la partie contenue dans le noyau : on prend l'inverse sur $Ker(X)^\perp$, complété arbitrairement sur $Ker(X)$. La définition de $(X'X)^-$ est donc loin d'être unique ! Il est alors possible de généraliser les résultats du cas régulier.

Proposition 5.3. *Si $(X'X)^-$ est une matrice inverse généralisée de $X'X$ alors $\widehat{\theta} = (X'X)^-X'Y$ est une solution des équations normales :*

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

Démonstration. On commence par remarquer que

$$\forall \omega \in \mathbb{R}^k,\ < X\omega, P_{[X]^\perp}Y >=< \omega, X'P_{[X]^\perp}Y >= 0$$

donc

$$X'Y = X'P_{[X]}Y + X'P_{[X]^\perp}Y = X'P_{[X]}Y.$$

Ainsi, $\exists u \in \mathbb{R}^k,\ X'Y = X'Xu$. Finalement,

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

□

Remarque : Cet estimateur n'est pas unique et dépend de la définition choisie pour $(X'X)^-$. Par contre, le vecteur $X\widehat{\theta}$ reste unique, même si la matrice X est singulière. Ce vecteur correspond en effet à la projection orthogonale de Y sur $Im(X)$.

En règle générale, nous préférons lever l'indétermination sur $\hat{\theta}$ en fixant des contraintes, souvent afin de donner un sens plus intuitif à θ.

5.2.1 Contraintes d'identifiabilité

Proposition 5.4. *Supposons la matrice X singulière de rang $r < k$ de sorte qu'il y ait $k-r$ paramètres redondants. Soit M une matrice à $k-r$ lignes et k colonnes, supposée de rang $k-r$ et telle que :*

$$Ker(M) \cap Ker(X) = \{0_k\}.$$

Alors,

- *la matrice $(X'X + M'M)$ est inversible et son inverse est une matrice inverse généralisée de $X'X$;*
- *le vecteur $\widehat{\theta} = (X'X+M'M)^{-1}X'Y$ est l'unique solution du système* $\begin{cases} X'X\alpha = X'Y \\ M\alpha = 0_{k-r}. \end{cases}$

Exercice 12.

1. *Pour montrer que $X'X + M'M$ est inversible : montrez que la matrice*

$$A = \begin{pmatrix} X \\ M \end{pmatrix} \in \mathcal{M}_{n+k-r,k}(\mathbb{R})$$

est injective et donc $A'A$ est inversible.

2. *Considérez le problème de minimisation suivante :*

$$g : \alpha \mapsto \|Y - X\alpha\|^2 + \|M\alpha\|^2.$$

Ecrivez $g(\alpha)$ sous la forme $g(\alpha) = \|\tilde{Y} - A\alpha\|^2$ avec $\tilde{Y}$ à préciser. Déduisez-en que $\widehat{\theta}$ est solution du système $\begin{cases} X'X\alpha = X'Y \\ M\alpha = 0_{k-r}. \end{cases}$

3. *Montrez l'unicité de la solution*

Le choix de la contrainte n'est pas toujours évident. Par ailleurs, pour chaque contrainte H, nous aurons un estimateur correspondant ce qui est parfois gênant.

Exemple 3. *Prenons l'exemple de l'analyse de variance à un facteur avec effet différentiel : nous supposons pour simplifier que $I = 4$. Le modèle s'écrit donc de la façon suivante :*

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{ij} \text{ pour } i = 1, \cdots, 4 \text{ et } j = 1.$$

La matrice X associée au modèle est :

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Il nous faut poser une contrainte (dite d'identifiabilité) sur le vecteur θ au travers du choix d'une matrice à 1 ligne et k colonnes. Une possibilité est de considérer $M = (0\ 1\ 1\ 1\ 1)$. La contrainte correspondante est :

$$M\theta = 0 \Leftrightarrow \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0.$$

Nous imposons donc que la somme des effets différentiels est nulle. Nous pouvons vérifier que les conditions de la proposition précédente sont bien satisfaites : l'estimateur suggéré peut alors être construit.

5.2.2 Fonctions estimables et contrastes

En présence de matrice singulière, il est donc toujours possible de construire un estimateur. Qu'en est-il des tests ? En particulier, ces contraintes sont-elles systématiquement nécessaires ?

La plupart des quantités que nous avons voulu tester sont des fonctions de θ qui ne dépendent pas de la solution particulière des équations normales, c'est-à-dire du type de contraintes d'identifiabilité choisi. Ces fonctions sont appelées estimables car elles sont intrinsèques.

Définition 5.3. *Une combinaison linéaire $C\theta$ est dite **fonction estimable** (de paramètre θ) si elle ne dépend pas du choix particulier d'une solution des équations normales. On caractérise ces fonctions comme étant celles qui s'écrivent $C\theta = DX\theta$ où D est une matrice de plein rang.*

Définition 5.4. *On appelle **contraste** une fonction estimable $C\theta$ telle que $C\ \mathbb{1} = 0$, où $\mathbb{1}$ désigne le vecteur unité.*

En analyse de variance, la plupart des combinaisons linéaires que l'on teste sont en fait des contrastes (cf chapitre 7). Dans l'exemple précédent, $\alpha_1 - \alpha_2$ est un contraste.

5.3 Orthogonalité

5.3.1 Orthogonalité pour les modèles réguliers

L'orthogonalité est une notion qui peut notablement simplifier la résolution et la compréhension d'un modèle linéaire. Un modèle linéaire admet le plus souvent une décomposition naturelle des paramètres θ (cf exemple ci-dessous) et conséquemment une décomposition de la matrice X associée au modèle. On va s'intéresser ici à l'orthogonalité éventuelle des différents espaces associés à cette décomposition (l'orthogonalité sera toujours comprise par la suite au sens d'orthogonalité liée au produit scalaire euclidien usuel). Le problème sera plus ou moins délicat suivant que le modèle est régulier ou non. En premier lieu, illustrons par deux exemples ce que l'on entend par décomposition des paramètres.

Exemple 4. *Soit le modèle de régression linéaire multiple sur trois variables* $x^{(1)}$, $x^{(2)}$ *et* $x^{(3)}$ *:*

$$Y_i = \mu + \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + \theta_3 x_i^{(3)} + \varepsilon_i, i = 1, \cdots, n > 4.$$

Le vecteur θ *comprend 4 coordonnées :* μ, θ_1, θ_2, θ_3 *et la matrice* X *quatre colonnes. Assez naturellement ici, on peut considérer la décomposition, plus précisément on parlera par la suite de partition en quatre éléments. La partition de la matrice revient alors à l'écrire comme concaténation de 4 vecteurs colonnes. L'orthogonalité de la partition correspondra alors strictement à l'orthogonalité des 4 droites vectorielles :* $[\mathbb{1}]$, $[x^{(1)}]$, $[x^{(2)}]$ *et* $[x^{(3)}]$.

Exemple 5. *Soit le modèle de régression quadratique sur* $x^{(1)}$ *et* $x^{(2)}$ *:*

$$Y_i = \mu + \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + \gamma_1 \left(x_i^{(1)}\right)^2 + \gamma_2 \left(x_i^{(2)}\right)^2 + \delta x_i^{(1)} x_i^{(2)} + \varepsilon_i, i = 1, \cdots, n > 6.$$

Ici plutôt que de demander comme précédemment l'orthogonalité de chacun des régresseurs (ce qui serait beaucoup demander), on peut définir la partition naturelle correspondant à :

- *la constante* μ *;*
- *les effets linéaires* θ_1, θ_2 *;*
- *les effets carrés* γ_1, γ_2 *;*
- *l'effet produit* δ.

L'orthogonalité de la partition est alors définie comme l'orthogonalité des sous-espaces vectoriels : $[\mathbb{1}]$, $[(x^{(1)}, x^{(2)})]$, $\left[\left((x^{(1)})^2, (x^{(2)})^2\right)\right]$ *et* $[x^{(1)}x^{(2)}]$.

En conséquence, on voit bien, à partir de ces deux exemples, qu'il faudra parler de modèle avec partition orthogonale plutôt que de modèle orthogonal.

Formalisons ces exemples dans une définition.

Définition 5.5. *Soit un modèle linéaire général régulier* $Y = X\theta + \varepsilon$. *Considérons une partition en* m *termes de* X *et de* θ, *soit*

$$Y = X_1\theta_1 + \cdots + X_m\theta_m + \varepsilon,$$

où la matrice X_j *est une matrice de taille* (n, k_j) *et* $\theta_j \in \mathbb{R}^{k_j}$ *avec* $k_j \in \{1, \cdots, k\}$ *pour* $j = 1, \cdots, m$ *et avec* $\sum_{j=1}^m k_j = k$. *On dit que cette partition est orthogonale si les sous-espaces vectoriels de* $\mathbb{R}^n$, $[X_1], \cdots, [X_m]$, *sont orthogonaux.*

Une conséquence simple de l'orthogonalité d'un modèle linéaire est que la matrice d'information $X'X$ a une structure bloc diagonale, chaque bloc étant associé à chaque élément de la partition.

Le plus souvent, la partition du vecteur de paramètres θ en différents effets vient

- en régression, des différentes variables ;
- en analyse de la variance, des décompositions en interactions.

L'orthogonalité donne aux modèles statistiques les deux propriétés suivantes :

Proposition 5.5. *Soit un modèle linéaire régulier muni d'une partition orthogonale :*

$$Y = X_1\theta_1 + \cdots + X_m\theta_m + \varepsilon.$$

Alors

- *les estimateurs des moindres carrés des différents effets $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$ sont non-corrélés et indépendants sous l'hypothèse gaussienne.*
- *pour $l = 1, \cdots, m$, l'expression de l'estimateur $\widehat{\theta}_l$ ne dépend pas de la présence ou non des autres termes θ_j dans le modèle.*

L'orthogonalité apporte une simplification des calculs : elle permet d'obtenir facilement une expression explicite des estimateurs. Par ailleurs, elle donne une indépendance approximative entre les tests des différents effets. Les tests portant sur des effets orthogonaux ne sont liés que par l'estimation du σ^2.

5.3.2 Orthogonalité pour les modèles non-réguliers

Lorsque le modèle est singulier, il est nécessaire de rajouter des contraintes. Il est alors raisonnable d'effectuer cette démarche en tenant compte de la partition, i.e. $C_j\theta_j = 0$ où $X_j|_{Ker(C_j)}$ sont injectives.

Définition 5.6. *Soit la partition suivante d'un modèle linéaire*

$$Y = X_1\theta_1 + \cdots + X_m\theta_m + \varepsilon.$$

Soit un système de contraintes $C_1\theta_1 = 0, \cdots, C_m\theta_m = 0$ qui rendent le modèle identifiable. On dit que ces contraintes rendent la partition orthogonale si les sous-espaces vectoriels

$$V_j = \{X_j\theta_j;\ \theta_j \in Ker(C_j)\},\ j = 1, \cdots, m$$

sont orthogonaux.

Cette notion est proche du cas régulier. Cependant, la notion d'orthogonalité dépend des contraintes choisies. L'idée sera en général de choisir des contraintes qui rendent le modèle orthogonal. On verra que cette définition prend tout son sens avec l'exemple incontournable du modèle d'analyse de la variance à deux facteurs croisés (cf chapitre 7).

En résumé

Dans ce chapitre, il est attendu que vous ayez compris

- la problématique de l'estimation des paramètres pour un modèle linéaire singulier
- l'intérêt d'avoir l'orthogonalité dans un modèle linéaire

Les résultats énoncés dans ce chapitre ne sont pas à connaitre. Il faudra savoir les mettre en application dans le cadre de l'ANOVA (voir Chapitre 7) et de l'ANCOVA (voir Chapitre 8).

Chapitre 6

La régression linéaire

6.1 Introduction

6.1.1 Exemple illustratif

Pour illustrer les notions abordées dans ce chapitre, nous allons considérer l'exemple suivant : On s'intéresse au lien éventuel entre le poids d'un homme et diverses caractéristiques physiques. Pour 22 hommes en bonne santé âgés de 16 à 30 ans, on dispose :

- du poids en kg (variable Y),
- de la circonférence maximale de l'avant-bras en cm (variable X1),
- de la circonférence maximale du biceps en cm (variable X2),
- de la distance autour de la poitrine directement sous les aisselles en cm (variable X3),
- de la distance autour du cou, à peu près à mi-hauteur, en cm (variable X4),
- de la distance autour des épaules, mesurées autour de la pointe des omoplates, en cm (variable X5),
- de la distance autour de la taille au niveau de la ligne de pantalon, en cm (variable X6),
- de la hauteur de la tête aux pieds en cm (variable X7),
- de la circonférence maximum du mollet en cm (variable X8),
- de la circonférence de la cuisse, mesurée à mi-chemin entre le genou et le haut de la jambe, en cm (variable X9),
- de la circonférence de la tête en cm (variable X10).

Le jeu de données **Data-ExRegMultiple.txt** ainsi que le fichier **ExRegressionLineaire.Rmd** contenant le script R illustrant ce chapitre sont disponibles sur la page moodle du cours.

Quelques statistiques descriptives sont codées dans le script, les boxplots et les corrélations deux à deux entre les variables quantitatives sont représentés sur la Figure 6.1.

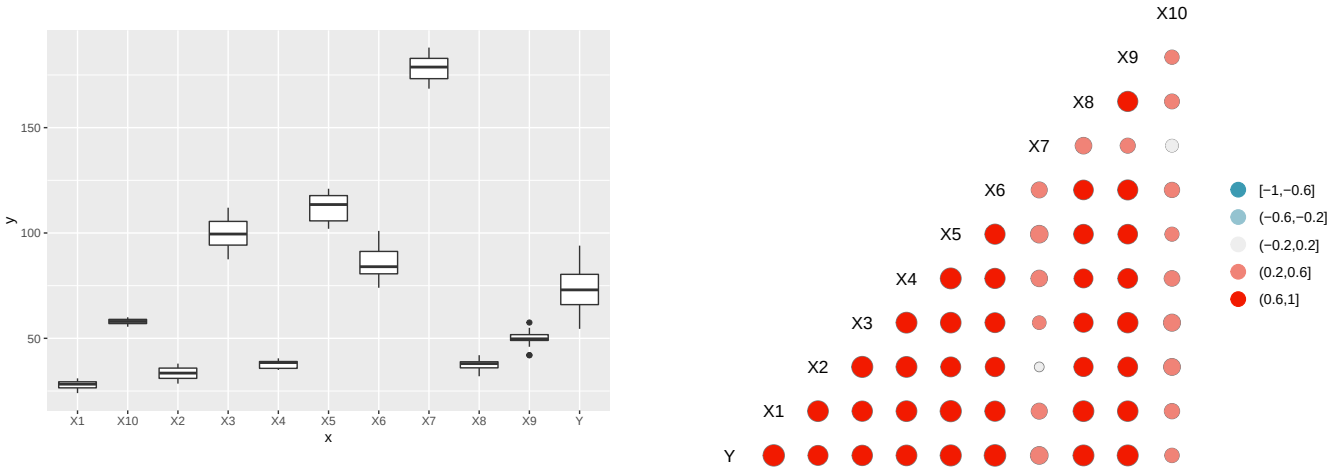


FIGURE 6.1 – Description des données. A gauche, boxplot des différentes variables quantitatives. A droite, représentation graphique des corrélations deux à deux des variables quantitatives.

6.1.2 Problématique

La régression est une des méthodes les plus connues et les plus appliquées en statistique pour l'analyse des données quantitatives. Elle est utilisée pour établir une liaison entre une variable quantitative, et une ou plusieurs autres variables quantitatives, sous la forme d'un modèle. Si on s'intéresse à la relation entre deux variables (par exemple, le poids en fonction de la circonférence maximale de l'avant-bras X1), on parlera de **régression simple** en exprimant une variable en fonction de l'autre. Si la relation porte entre une variable et plusieurs autres variables (par exemple, le poids en fonction de toutes les autres variables quantitatives), on parlera de **régression multiple**. La mise en oeuvre d'une régression impose l'existence d'une relation de cause à effet entre les variables prises en compte dans le modèle.

Cette méthode peut être mise en place sur des données quantitatives observées sur n individus et présentées sous la forme :

- une variable quantitative Y prenant la valeur Y_i pour l'individu $i, i = 1, \cdots, n$ appelée **variable à expliquer** ou **variable réponse**,
- p variables quantitatives $z^{(1)}, z^{(2)}, \cdots, z^{(p)}$ prenant respectivement les valeurs $z_i^{(1)}, z_i^{(2)}, \cdots, z_i^{(p)}$ pour l'individu i, appelées **variables explicatives** ou **prédicteurs**. Si $p = 1$, on est dans le cas de la régression simple. Lorsque les valeurs prises par une variable explicative sont choisies par l'expérimentateur, on dit que la variable explicative est *contrôlée*.

Dans notre exemple, $n = 22$, Y est la variable poids et $p = 10$.

Considérons un couple de variables aléatoires quantitatives (Z, Y). S'il existe une liaison entre ces deux variables, la connaissance de la valeur prise par Z change notre incertitude concernant la réalisation de Y. Si l'on admet qu'il existe une relation de cause à effet entre Z et Y, le phénomène aléatoire représenté par Z peut donc servir à

prédire celui représenté par Y et la liaison s'écrit sous la forme $y = f(z)$. On dit que l'on fait de la régression de Y sur Z.
Dans le cas les plus fréquents, on choisit l'ensemble des fonctions affines (du type $f(z) = \theta_0 + \theta_1 z$ ou $f(z^{(1)}, z^{(2)}, \cdots, z^{(p)}) = \theta_0 + \theta_1 z^{(1)} + \theta_2 z^{(2)} + \cdots + \theta_p z^{(p)}$) et on parle de **régression linéaire**.

6.1.3 Le modèle de régression linéaire simple

Soit un échantillon de n individus. Pour un individu i ($i = 1, \cdots, n$), on a observé
- Y_i la valeur de la variable quantitative Y (ex : le poids),
- z_i la valeur de la variable quantitative z (ex : la circonférence maximale de l'avant-bras)

On veut étudier la relation entre ces deux variables, et en particulier, l'effet de z (*variable explicative*) sur Y (*variable réponse*). Dans un premier temps, on peut représenter graphiquement cette relation en traçant le nuage des n points de coordonnées $(z_i, Y_i)_{1 \leq i \leq n}$ (cf Figure 6.2). Dans le cas où le nuage de points est de forme "linéaire", on cherchera à ajuster ce nuage de points par une droite. La relation entre Y_i et z_i s'écrit alors sous la forme d'un modèle de régression linéaire simple :

$$\begin{cases} Y_i = \theta_0 + \theta_1 z_i + \varepsilon_i, \, \forall i = 1, \cdots, n, \\ \\ \varepsilon_1, \ldots, \varepsilon_n \text{ i.i.d de loi } \mathcal{N}(0, \sigma^2) \end{cases} \tag{6.1}$$

La première partie du modèle $\theta_0 + \theta_1 z_i$ représente la moyenne de Y_i sachant z_i et la seconde partie ε_i, la différence entre cette moyenne et la valeur Y_i. Le nuage de points est résumé par la droite d'équation $y = \theta_0 + \theta_1 z$.

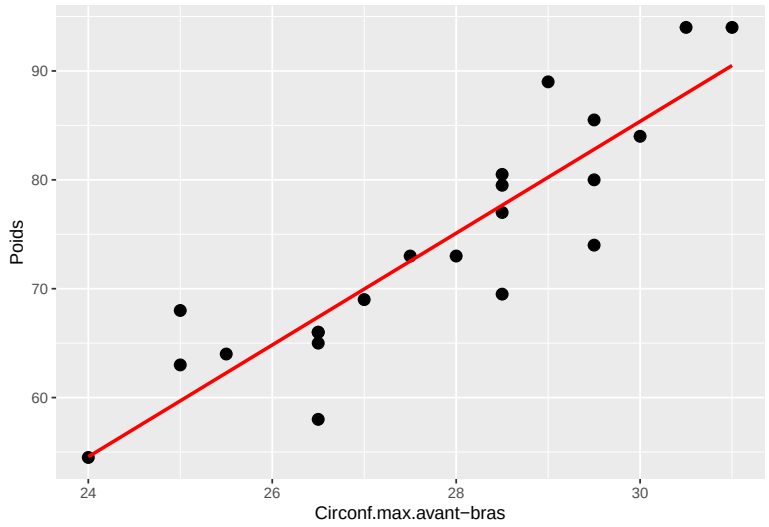


FIGURE 6.2 – Représentation du poids en fonction de la circonférence maximale de l'avant-bras. En rouge, la droite de régression linéaire simple ajustée.

6.1.4 Le modèle de régression linéaire multiple

On dispose d'un échantillon de n individus pour lesquels on a observé
- Y_i la valeur de la variable réponse Y quantitative (ex : le poids),
- $z_i^{(1)}, \cdots, z_i^{(p)}$ les valeurs de p autres variables quantitatives $z^{(1)}, \cdots, z^{(p)}$.

On veut expliquer la variable quantitative Y par les p variables quantitatives $z^{(1)}, \cdots, z^{(p)}$.
Le modèle s'écrit

$$\begin{cases} Y_i = \theta_0 + \theta_1 z_i^{(1)} + \cdots, +\theta_p z_i^{(p)} + \varepsilon_i, \forall i = 1, \cdots, n, \\ \\ \varepsilon_1, \ldots, \varepsilon_n \text{ i.i.d de loi } \mathcal{N}(0, \sigma^2) \end{cases} \tag{6.2}$$

6.2 Estimation

6.2.1 Résultats généraux

Le modèle (6.2) peut se réécrire sous la forme matricielle

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & z_1^{(1)} & z_1^{(2)} & \ldots & z_1^{(p)} \\ 1 & z_2^{(1)} & z_2^{(2)} & \ldots & z_2^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & z_n^{(1)} & z_n^{(2)} & \ldots & z_n^{(p)} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}$$

où $X \in \mathcal{M}_{n,p+1}(\mathbb{R})$ (ici, $k = p+1$). Si le modèle est régulier, on peut alors estimer le vecteur des paramètres θ par la méthode des moindres carrés d'où

$$\widehat{\theta} = (X'X)^{-1}X'Y \sim \mathcal{N}_{p+1}(\theta, \sigma^2(X'X)^{-1}).$$

On en déduit alors $\widehat{Y_i} = (X\widehat{\theta})_i = \widehat{\theta_0} + \sum_{j=1}^{p} \widehat{\theta_j} z_i^{(j)}$ la valeur ajustée de Y_i et le résidu $\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$, de valeur observée $(\widehat{\varepsilon_i})^{obs} = y_i - \widehat{y_i}$.
La variance σ^2 est estimée par

$$\widehat{\sigma^2} = \frac{\|Y - X\widehat{\theta}\|^2}{n-(p+1)} = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \left(\widehat{\varepsilon_i}\right)^2.$$

Les erreurs standards des estimateurs $\widehat{\theta_0}, \cdots, \widehat{\theta_p}$, des valeurs ajustées et des résidus calculés valent :
- erreur standard de $\widehat{\theta_j}$ vaut $se(\widehat{\theta_j}) = \sqrt{\widehat{\sigma^2}[(X'X)^{-1}]_{j+1,j+1}}$
- erreur standard de $\widehat{Y_i}$ vaut $se(\widehat{Y_i}) = \sqrt{\widehat{\sigma^2}[X(X'X)^{-1}X']_{ii}} = \sqrt{\widehat{\sigma^2}H_{ii}}$
- erreur standard de $\widehat{\varepsilon_i}$ vaut $se(\widehat{\varepsilon_i}) = \sqrt{\widehat{\sigma^2}(1 - H_{ii})}$.

Exercice 13. *On se place dans le cadre de la régression linéaire simple d'équation* (6.1). *Montrez que les estimateurs de* θ_0 *et* θ_1 *par la méthode des moindres carrés sont donnés par :*

$$\begin{cases} \widehat{\theta_1} = \frac{cov(Y,z)}{var(z)} = \frac{\sum_{i=1}^{n}(z_i-\overline{z})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(z_i-\overline{z})^2}, \\ \widehat{\theta_0} = \overline{Y} - \widehat{\theta_1}\,\overline{z}, \end{cases}$$

où $\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ *et* $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
Pour cela, on cherche à minimiser la fonction des moindres carrés

$$(a,b) \mapsto \sum_{i=1}^{n}(Y_i - a - bz_i)^2.$$

Les résultats obtenus avec la commande `lm` pour l'exemple de la régression linéaire simple sont en Figure 6.3. On a en particulier $(\widehat{\theta_0})^{obs} = -68.644$ et $(\widehat{\theta_1})^{obs} = 5.134$ ainsi que leur erreur standard dans la colonne suivante.

```
> reg.simple<-lm(Y~X1,data=Data)
> summary(reg.simple)

Call:
lm(formula = Y ~ X1, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3981 -1.9234 -0.3646  2.8012  8.7678

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -68.644     15.589  -4.403 0.000274 ***
X1             5.134      0.560   9.167 1.34e-08 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 4.926 on 20 degrees of freedom
Multiple R-squared:  0.8078,        Adjusted R-squared:  0.7981
F-statistic: 84.03 on 1 and 20 DF,  p-value: 1.338e-08
```

FIGURE 6.3 – Résultats pour l'exemple de la régression linéaire simple

Les résultats obtenus avec la commande `lm` pour l'exemple de la régression linéaire multiple sont en Figure 6.4. Les deux premières colonnes correspondent aux estimations et aux erreurs standards respectivement pour chaque paramètre.

6.2.2 Propriétés en régression linéaire simple

On se place dans cette section dans le cadre de la régression linéaire simple (cf Equation (6.1)). La proposition suivante donne des propriétés entre les résidus et les valeurs prédites par le modèle.

```
> reg<-lm(Y~.,data=Data)
> summary(reg)

Call:
lm(formula = Y ~ ., data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5523 -0.9965  0.0461  1.0499  4.1719

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -69.51714   29.03739  -2.394 0.035605 *
X1            1.78182    0.85473   2.085 0.061204 .
X2            0.15509    0.48530   0.320 0.755275
X3            0.18914    0.22583   0.838 0.420132
X4           -0.48184    0.72067  -0.669 0.517537
X5           -0.02931    0.23943  -0.122 0.904769
X6            0.66144    0.11648   5.679 0.000143 ***
X7            0.31785    0.13037   2.438 0.032935 *
X8            0.44589    0.41251   1.081 0.302865
X9            0.29721    0.30510   0.974 0.350917
X10          -0.91956    0.52009  -1.768 0.104735
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.287 on 11 degrees of freedom
Multiple R-squared:  0.9772,        Adjusted R-squared:  0.9565
F-statistic: 47.17 on 10 and 11 DF,  p-value: 1.408e-07
```

FIGURE 6.4 – Résultats pour l'exemple de la régression linéaire multiple

Proposition 6.1.

1. $\sum_{i=1}^{n} \widehat{\varepsilon_i} = 0$, $\sum_{i=1}^{n} \widehat{Y_i} = \sum_{i=1}^{n} Y_i$
2. *La droite de régression passe par le point de coordonnées* $(\overline{z}, \overline{Y})$.
3. *Le vecteur des résidus n'est pas corrélé avec la variable explicative :* $cov(z, \widehat{\varepsilon}) = 0$.
4. *Le vecteur des résidus n'est pas corrélé avec la variable ajustée :* $cov(\widehat{Y}, \widehat{\varepsilon}) = 0$.
5. *La variance de* Y *admet la décomposition :*

$$var(Y) = var(\widehat{Y}) + var(\widehat{\varepsilon}). \tag{6.3}$$

6. *Le carré du coefficient de corrélation de* z *et de* Y *s'écrit sous les formes suivantes :*

$$r^2(z, Y) = \frac{var(\widehat{Y})}{var(Y)} = 1 - \frac{var(\widehat{\varepsilon})}{var(Y)}.$$

On en déduit que la variance empirique de Y *se décompose en somme d'une part de variance expliquée* $(var(\widehat{Y}))$ *et d'une variance résiduelle* $(var(\widehat{\varepsilon}))$*, et que* $r^2(z, Y)$ *est le rapport de la variance expliquée sur la variance de la variable à expliquer.*

Preuve : En utilisant que $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$, $\widehat{Y}_i = \widehat{\theta}_0 + \widehat{\theta}_1 z_i$ et $\widehat{\theta}_0 = \bar{Y} - \widehat{\theta}_1 \bar{z}$, on a

1. $\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^n \left\{Y_i - [\widehat{\theta}_0 + \widehat{\theta}_1 z_i]\right\} = \bar{Y} - \widehat{\theta}_0 - \widehat{\theta}_1 \bar{z} = 0$ par définition de $\widehat{\theta}_0$.
2. $\widehat{\theta}_0 + \widehat{\theta}_1 \bar{z} = \left[\bar{Y} - \widehat{\theta}_1 \bar{z}\right] + \widehat{\theta}_1 \bar{z} = \bar{Y}$
3. On a

$$\begin{aligned} n\ cov(z, \widehat{\varepsilon}) &= \sum_{i=1}^n \widehat{\varepsilon}_i (z_i - \bar{z}) \\ &= \sum_{i=1}^n [Y_i - \bar{Y} - \widehat{\theta}_1 (z_i - \bar{z})][z_i - \bar{z}] \\ &= n \left\{cov(Y, z) - \widehat{\theta}_1 var(z)\right\} = 0 \end{aligned}$$

   par définition de $\widehat{\theta}_1$.
4. $n\ cov(\widehat{Y}, \widehat{\varepsilon}) = \sum_{i=1}^n \widehat{\varepsilon}_i (\widehat{Y}_i - \bar{Y}) = \sum_{i=1}^n \widehat{\varepsilon}_i \widehat{\theta}_1 (z_i - \bar{z}) = n\widehat{\theta}_1 cov(z, \widehat{\varepsilon}) = 0$.
5. $n\, var(Y) = \sum_{i=1}^n (Y_i - \widehat{Y}_i + \widehat{Y}_i - \bar{Y})^2 = n\, var(\widehat{\varepsilon}) + n\, var(\widehat{Y}) + 2n\, cov(\widehat{\varepsilon}, \widehat{Y})$.
6. On a $r^2(z, Y) = \frac{cov(z,Y)^2}{var(z) var(Y)}$ et

$$n\, cov(z, Y) = \sum_{i=1}^n (Y_i - \widehat{Y}_i + \widehat{Y}_i - \bar{Y})(z_i - \bar{z}) = n\, cov(\widehat{\varepsilon}, z) + n\, cov(\widehat{Y}, z) = n\, cov(\widehat{Y}, z).$$

   Ainsi,

$$r^2(z, Y) = \frac{cov(\widehat{Y}, z)^2}{var(z) var(\widehat{Y})} \frac{var(\widehat{Y})}{var(Y)} = cor(\widehat{Y}, z)^2 \frac{var(\widehat{Y})}{var(Y)} = \frac{var(\widehat{Y})}{var(Y)}$$

   car $\widehat{Y}_i = \widehat{\theta}_0 + \widehat{\theta}_1 z_i, \forall i$ (relation linéaire).

6.2.3 Le coefficient R^2

6.2.3.1 Définition

Le coefficient R^2, défini comme le carré du coefficient de corrélation de z et Y est une mesure de qualité de l'ajustement, égale au rapport de la variance effectivement expliquée sur la variance à expliquer :

$$R^2 = r^2(z, Y) = \frac{var(\widehat{Y})}{var(Y)}.$$

Ainsi $R^2 \in [0, 1]$ et s'interprète comme la *proportion de variance expliquée par la régression.*

La plupart des logiciels n'utilise pas la décomposition (6.3), mais plutôt la décomposition obtenue en multipliant cette expression par n :

$$SCT = SCE + SCR$$

où

1. $SCT = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ est la somme totale des carrés corrigés de Y,
2. $SCE = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$ est la somme des carrés expliquée par le modèle,
3. $SCR = \sum_{i=1}^{n}(\widehat{\varepsilon}_i)^2$ est la somme des carrés des résidus.

Ainsi, pour calculer le R^2, on utilise également l'expression

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}.$$

Dans l'exemple de la régression linéaire simple, la valeur du R^2 vaut 0.8078 (cf Figure 6.3). Pour retrouver les valeurs de SCT, SCR et SCE, on peut utiliser la commande `anova` (cf Figure 6.5).

```
> anova(reg.simple)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1 2038.88 2038.88  84.032 1.338e-08 ***
Residuals 20  485.27   24.26
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
> var(Yfitted)*(n-1)  # SCE
[1] 2038.882
> var(residus)*(n-1)  # SCR
[1] 485.2655
> var(Y)*(n-1)
[1] 2524.148
```

FIGURE 6.5 – Résultats avec la commande anova pour l'exemple de la régression linéaire simple

Dans le cas d'une régression multiple de Y par $z^{(1)}, \cdots, z^{(p)}$, le *coefficient de corrélation multiple* noté $r(Y, z^{(1)}, \cdots, z^{(p)})$ est défini comme le coefficient de corrélation linéaire empirique de Y par $\widehat{Y}$:

$$r(Y, z^{(1)}, \cdots, z^{(p)}) = r(Y, \widehat{Y}).$$

Ainsi le coefficient R^2 de la régression multiple est égal au carré du coefficient de corrélation linéaire multiple empirique $r(Y, z^{(1)}, \cdots, z^{(p)})$. Dans l'exemple de la régression linéaire multiple, la valeur du R^2 vaut 0.9772 (cf Figure 6.4).

6.2.3.2 Augmentation mécanique du R^2

Lorsque l'on ajoute une variable explicative à un modèle, la somme des carrés des résidus diminue ou au moins reste stable. En effet, si on considère un modèle à $p-1$ variables :

$$Y_i = \theta_0 + \theta_1 z_i^{(1)} + \cdots + \theta_{p-1} z_i^{(p-1)} + \varepsilon_i$$

alors les coefficients $(\widehat{\theta}_0, \widehat{\theta}_1, \cdots, \widehat{\theta}_{p-1})$ estimés minimisent

$$\phi(\theta_0, \theta_1, \cdots, \theta_{p-1}) = \sum_{i=1}^{n} \left[Y_i - (\theta_0 + \theta_1 z_i^{(1)} + \cdots + \theta_{p-1} z_i^{(p-1)}) \right]^2.$$

Si on rajoute une nouvelle variable explicative $z^{(p)}$ au modèle, on obtient

$$Y_i = \theta_0 + \theta_1 z_i^{(1)} + \cdots + \theta_{p-1} z_i^{(p-1)} + \theta_p z_i^{(p)} + \varepsilon_i,$$

et les coefficients estimés, notés $(\widetilde{\theta}_0, \widetilde{\theta}_1, \cdots, \widetilde{\theta}_p)$ minimisent la fonction :

$$\tilde{\psi}(\theta_0, \theta_1, \cdots, \theta_p) = \sum_{i=1}^{n} \left[Y_i - (\theta_0 + \theta_1 z_i^{(1)} + \cdots + \theta_p z_i^{(p)}) \right]^2$$

qui, par construction, vérifie l'égalité :

$$\tilde{\psi}(\theta_0, \theta_1, \cdots, \theta_{p-1}, 0) = \phi(\theta_0, \theta_1, \cdots, \theta_{p-1}).$$

D'où l'inégalité :

$$\tilde{\psi}(\widetilde{\theta}_0, \widetilde{\theta}_1, \cdots, \widetilde{\theta}_p) \leq \tilde{\psi}(\widehat{\theta}_0, \widehat{\theta}_1, \cdots, \widehat{\theta}_{p-1}, 0) = \phi(\widehat{\theta}_0, \widehat{\theta}_1, \cdots, \widehat{\theta}_{p-1}).$$

Ceci prouve l'augmentation "mécanique" du R^2 sans pour autant améliorer le modèle, comme nous le verrons par la suite.

6.3 Tests et intervalles de confiance

6.3.1 Test de nullité d'un paramètre du modèle

En testant l'hypothèse nulle $\mathcal{H}_0^{(j)} : \theta_j = 0$ où θ_j est le paramètre associé à la variable explicative $z^{(j)}$, on étudie l'effet de la présence de la variable explicative $z^{(j)}$. Pour tester $\mathcal{H}_0^{(j)} : \theta_j = 0$ contre $\mathcal{H}_1^{(j)} : \theta_j \neq 0$, on met en place un test classique de Student.

Exercice 14. *Construisez le test statistique de Student pour tester* $\mathcal{H}_0^{(j)} : \theta_j = 0$ *contre* $\mathcal{H}_1^{(j)} : \theta_j \neq 0$ *au niveau* α.

Dans les exemples de la régression linéaire simple et la régression linéaire multiple, la pvaleur associée au test de nullité de chacun des coefficients θ_j est donnée dans la dernière colonne (la valeur de la statistique de test est donnée dans l'avant dernière colonne). D'après les résultats reportés dans la Figure 6.3, on rejette fortement la nullité de chacun des coefficients dans le modèle de régression simple pour un test à 5%. D'après les résultats en Figure 6.4, on rejette la nullité des coefficients θ_0, θ_6 et θ_7 dans l'exemple de régression linéaire multiple pour un test à 5%. Chaque test de nullité est fait séparément, attention aux conclusions trop rapides !

6.3.2 Test de nullité de quelques paramètres du modèle

Soit un modèle de référence à p variables explicatives. On veut étudier l'influence de q variables explicatives (avec $q \leq p$) sur la variable à expliquer. Cela revient à tester l'hypothèse de **nullité de q paramètres du modèle** :

$$\mathcal{H}_0 : \theta_1 = \theta_2 = \cdots = \theta_q = 0, \text{ avec } q \leq p.$$

Sous l'hypothèse alternative, au moins un des paramètres $\theta_1, \cdots, \theta_q$ est non nul.

Ce test peut être formulé comme la comparaison de deux modèles emboîtés, l'un à $p+1$ paramètres et l'autre à $p+1-q$ paramètres :

$$\begin{array}{ll} (M1) & Y_i = \theta_0 + \theta_1 z_i^{(1)} + \cdots + \theta_p z_i^{(p)} + \varepsilon_i \qquad \text{sous } \mathcal{H}_1 \\ \text{versus} & \\ (M0) & Y_i = \theta_0 + \theta_{q+1} z_i^{(q+1)} + \cdots + \theta_p z_i^{(p)} + \varepsilon_i \quad \text{sous } \mathcal{H}_0. \end{array}$$

L'hypothèse $\mathcal{H}_0$ peut être testée au moyen de la statistique de Fisher :

$$F = \frac{(SCR_0 - SCR_1)/q}{SCR_1/(n-(p+1))} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(q, n-(p+1))$$

où SCR_0 désigne la somme des carrés des résidus du modèle "réduit" sous $\mathcal{H}_0$ et SCR_1 correspond à la somme des carrés des résidus du modèle de référence.
On compare F au quantile $f_{q,n-p-1,1-\alpha}$: si $F \geq f_{q,n-p-1,1-\alpha}$, alors on rejette $\mathcal{H}_0$.
On remarque que dans le cas où $q = 1$, on teste la nullité d'un seul paramètre du modèle et on retrouve les mêmes conclusions qu'avec le test précédent de Student.

Exercice 15. *En écrivant les modèles $(M0)$ et $(M1)$ sous la forme $Y = Z\beta + \varepsilon$ et $Y = X\theta + \varepsilon$ respectivement, donnez l'expression de SCR_0 et SCR_1 en fonction de $\widehat{\theta}$ et $\widehat{\beta}$ dans ce test.*

Dans notre exemple en régression linéaire multiple, on souhaite tester le sous-modèle composé uniquement des variables X_1, X_6 et X_7. A l'aide de la fonction `anova`, on va faire un test de Fisher entre ce sous-modèle et le modèle complet :

```
> reg0<-lm(Y~X1+X6+X7,data=Data)
> anova(reg0,reg)
Analysis of Variance Table

Model 1: Y ~ X1 + X6 + X7
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     18 112.149
2     11  57.524  7    54.625 1.4922 0.2653
```

La pvaleur valant 0.2653, on accepte le sous-modèle M_0.

Exercice 16. *Dans la sortie R de* `anova(reg0,reg)` *ci-dessus, à quoi corresponde chacune des valeurs numériques ?*

6.3.3 Test de nullité de tous les paramètres du modèle

Dans cette section, on souhaite tester l'hypothèse de nullité de tous les paramètres du modèle (associés aux variables explicatives) :

$$\mathcal{H}_0 : \theta_1 = \cdots = \theta_p = 0.$$

Ce test revient à comparer la qualité d'ajustement du modèle de référence à celle du "modèle blanc". Cette hypothèse composée de p contraintes signifie que les p paramètres associés aux p variables explicatives sont nuls, c'est-à-dire qu'aucune variable explicative présente dans le modèle ne permet d'expliquer la variable Y.
Sous $\mathcal{H}_0$, le modèle s'écrit :

$$Y_i = \theta_0 + \varepsilon_i \text{ avec } \widehat{\theta_0} = \overline{Y}$$

et la somme des carrés des résidus (SCR_0) est égale à la somme des carrés totales (SCT).

Exercice 17. *Montrez que la statistique de test de Fisher dans ce cas s'écrit :*

$$F = \frac{SCE_1/p}{SCR_1/n-(p+1)} = \frac{R^2}{1-R^2} \times \frac{n-p-1}{p} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(p, n-p-1)$$

où SCE_1 désigne la somme des carrés du modèle de référence avec $SCT = SCE_1 + SCR_1$ et R^2 est le critère d'ajustement du modèle de référence.
On compare F au quantile $f_{p,n-(p+1),1-\alpha}$: si $F \geq f_{p,n-(p+1),1-\alpha}$, alors on rejette $\mathcal{H}_0$ et on conclut qu'il existe au moins un paramètre non nul dans le modèle.

Dans l'exemple de régression linéaire multiple, on peut mettre en place ce test avec la fonction `anova`. On peut aussi remarquer que le résultat de ce test est donné directement dans le summary(reg) (cf Figure 6.4).

```
> regblanc<-lm(Y~1,data=Data)
> anova(regblanc,reg)
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     21 2524.15
2     11   57.52 10    2466.6 47.168 1.408e-07 ***
```

Ici, la pvaleur vaut $1.408e^{-07}$, on rejette donc l'hypothèse que tous les coefficients sont nuls.

6.3.4 Intervalle de confiance de θ_j, de $(X\theta)_i$ et de $X_0\theta$

6.3.4.1 Intervalle de confiance de θ_j

On reprend ici la construction générale faite en section 3.5.1, ici $k = 1 + p$. En utilisant que

- $\hat{\theta}_j \sim \mathcal{N}(\theta_j, \sigma^2[(X'X)^{-1}]_{j+1,j+1})$
- $(n-(p+1))\hat{\sigma}^2 \sim \sigma^2\chi(n-(p+1))$
- $\hat{\theta}_j$ et $\hat{\sigma}^2$ indépendants

on obtient que

$$\frac{\widehat{\theta_j} - \theta_j}{\sqrt{\widehat{\sigma}^2[(X'X)^{-1}]_{j+1,j+1}}} \sim \mathcal{T}(n-(p+1)).$$

On construit alors l'intervalle de confiance suivant pour le paramètre θ_j au niveau de confiance $1-\alpha$:

$$IC_{1-\alpha}(\theta_j) = \left[\widehat{\theta_j} \pm t_{n-(p+1),1-\alpha/2} \times \sqrt{\widehat{\sigma}^2[(X'X)^{-1}]_{j+1,j+1}}\right].$$

Dans les deux exemples de ce chapitre, on peut facilement obtenir les intervalles de confiance pour les coefficient θ_j à l'aide de la fonction `confint`.

```
> confint(reg.simple,level=0.9)
                   5 %       95 %
(Intercept) -95.530374 -41.757816
X1            4.167783   6.099549
> confint(reg)
                    2.5 %     97.5 %
(Intercept) -133.42800877 -5.6062615
X1            -0.09943047  3.6630678
X2            -0.91303791  1.2232187
X3            -0.30791607  0.6861870
X4            -2.06801804  1.1043439
X5            -0.55628825  0.4976636
X6             0.40506844  0.9178140
X7             0.03090785  0.6047850
X8            -0.46204518  1.3538255
X9            -0.37430495  0.9687296
X10           -2.06427506  0.2251497
```

6.3.4.2 Intervalle de confiance de $(X\theta)_i$

En reprenant la construction faite en section 3.5.2, l'intervalle de confiance de $(X\theta)_i$ au niveau de confiance de $1-\alpha$ est donc donné par :

$$IC_{1-\alpha}((X\theta)_i) = \left[\widehat{Y}_i \pm t_{n-(p+1),1-\alpha/2} \times \sqrt{\widehat{\sigma}^2 [X(X'X)^{-1}X']_{ii}}\right].$$

Pour l'exemple de la régression linéaire simple, ces intervalles de confiance sont représentés en Figure 6.6.

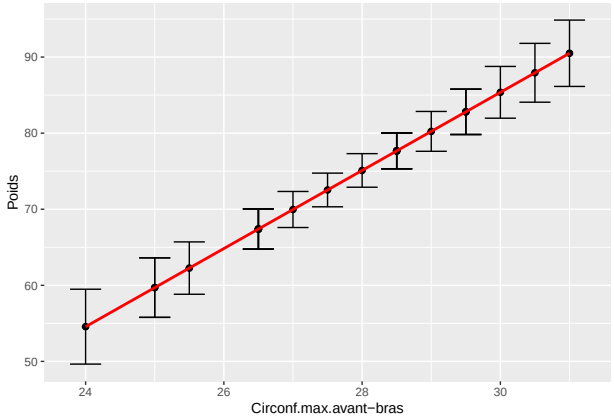


FIGURE 6.6 – Intervalles de confiance au niveau de confiance 95% pour les $(X\theta)_i$

6.3.4.3 Intervalle de confiance de $X_0\theta$

Pour des nouvelles données $z_0^{(1)}, \cdots, z_0^{(p)}$ des variables explicatives, on définit $X_0 = (1, z_0^{(1)}, \cdots, z_0^{(p)}) \in \mathcal{M}_{1,(p+1)}(\mathbb{R})$. La réponse moyenne est alors :

$$X_0\theta = \theta_0 + \sum_{j=1}^{p} \theta_j z_0^{(j)}.$$

En reprenant la construction faite en section 3.5.3, on obtient que l'intervalle de confiance de $X_0\theta$ au niveau de confiance de $1-\alpha$ s'écrit :

$$IC_{1-\alpha}(X_0\theta) = \left[X_0\widehat{\theta} \pm t_{n-(p+1),1-\alpha/2} \times \sqrt{\widehat{\sigma}^2 X_0(X'X)^{-1}X_0'}\right].$$

Dans l'exemple de la régression linéaire simple, voir Figure 6.7.

6.3.5 Intervalle de prédiction

On veut prédire dans quel intervalle se trouvera le résultat d'un nouvel essai $(z_0^{(1)}, \cdots, z_0^{(p)})$. On veut donc construire un intervalle de prédiction pour une nouvelle observation Y_0, correspondant à $X_0 = (1, z_0^{(1)}, z_0^{(2)}, \cdots, z_0^{(p)})$:

$$Y_0 = X_0\theta + \varepsilon_0,$$

où ε_0 est indépendant des ε_i, $1 \leq i \leq n$ et où $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$. En reprenant la construction faite en section 3.6, on obtient que l'intervalle de prédiction de la variable Y pour une nouvelle observation au point X_0 est défini par

$$IC_{1-\alpha}(Y_0) = \left[X_0\widehat{\theta} \pm t_{n-(p+1),1-\alpha/2}\widehat{\sigma}\sqrt{1 + X_0(X'X)^{-1}X_0'}\right].$$

Notez bien la différence entre $IC_{1-\alpha}(Y_0)$ et

$$IC_{1-\alpha}(X_0\theta) = \left[X_0\widehat{\theta} \pm t_{n-(p+1),1-\alpha/2} \times \widehat{\sigma}\sqrt{X_0(X'X)^{-1}X_0'}\right].$$

Dans l'exemple de la régression linéaire simple, les intervalles de confiance $IC_{1-\alpha}(X_0\theta)$ et $IC_{1-\alpha}(Y_0)$ sont représentés sur la Figure 6.7

Remarque : Pour faire de la prédiction à l'aide de ce modèle de régression linéaire, il est recommandé de n'utiliser ce modèle que dans le domaine couvert par les données. En effet, le phénomène étudié peut être linéaire dans le domaine observé et avoir un comportement différent dans un autre domaine.

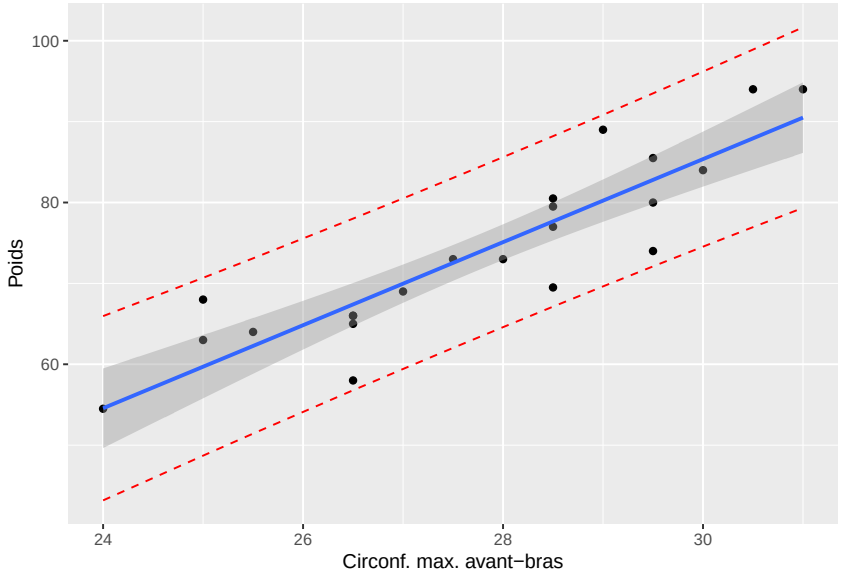


FIGURE 6.7 – Intervalle de confiance de $X_0\theta$ en gris foncé et intervalle de prédiction de Y_0 en rouge pointillé

6.4 Sélection des variables explicatives

En présence de p variables explicatives dont on ignore celles qui sont réellement influentes, on doit rechercher un modèle d'explication de Y à la fois performant (résidus les plus petits possibles) et économique (le moins possible de variables explicatives). Nous allons donc maintenant nous concentrer sur l'étude de la matrice X autrement dit sur les variables explicatives elles-mêmes. Dans cette partie, nous allons voir comment choisir le modèle le plus en adéquation avec nos données et éliminer certaines variables peu explicatives pour gagner en interprétation. Ce problème de sélection de variables est en fait un problème de sélection de modèles.

6.4.1 Cadre général de sélection de modèles

Par soucis de simplicité, on présente ce problème dans le cadre de la régression linéaire multiple. Les outils présentés ici peuvent être bien sûr utilisés dans un cadre plus général (bien souvent sans travail supplémentaire).

On se donne une famille de modèles $\mathcal{M}$ représentant formellement une famille de sous-ensembles de $\{1,\dots,p\}$. Ce choix est fait a priori et peut ne pas être exhaustif. Par exemple, on peut considérer

- famille exhaustive : $\mathcal{M} = \mathcal{P}(\{1,\dots,p\})$ i.e. la famille de tous les sous-ensembles de $\{1,\dots,p\}$,
- famille croissante : $\mathcal{M} = (\{1,\dots,m\})_{m=1,\dots,p}$.

Par la suite, pour $m \in \mathcal{M}$, on notera $|m|$ le cardinal de m et $X_{(m)}$ représente la matrice constituée des vecteurs $\mathbf{z}^{(j)}$ pour $j \in m$. On supposera également que pour tout $m \in \mathcal{M}$, la matrice $X_{(m)}$ est régulière, i.e. de rang $|m|+1$. Il faut noter que le "+1" vient de la constante (de l'intercept) qui est supposée être présente systématiquement dans tous les modèles.

Hypothèses sur le vrai modèle : On suppose qu'il existe $m^\star \in \mathcal{M}$, inconnu, tel que le vrai modèle s'écrive :

$$Y = \mu^\star + \varepsilon^\star = X_{(m^\star)}\theta_{(m^\star)} + \varepsilon^\star, \text{ avec } \varepsilon^\star \sim \mathcal{N}(0_n, {\sigma^\star}^2 I_n),$$

le vecteur $\theta_{(m^\star)} \in \mathbb{R}^{|m^\star|+1}$ ayant toutes ses coordonnées non nulles.

Modèles d'analyse : Pour modéliser l'expérience et essayer d'identifier le vrai modèle on utilise la famille de modèles suivante, qui est en correspondance avec $\mathcal{M}$, i.e.

$$Y = \mu + \varepsilon = X_{(m)}\theta_{(m)} + \varepsilon, \text{ avec } \varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n).$$

Pour préciser la modélisation, nous utiliserons le vocabulaire suivant :

Définition 6.1. *On suppose que le modèle d'analyse est $m \in \mathcal{M}$. Alors*

- *si $m = m_p = \{1,\cdots,p\}$, on dit que le modèle est **complet**, i.e. que toutes les variables explicatives disponibles sont significatives ;*
- *si $m^\star \subset m$ avec $m \neq m^\star$, on dit que le modèle est **sur-ajusté** ;*
- *si $|m \cap m^\star| < |m^\star|$, on dit que le modèle est **faux** ;*

- *si $m \subset m^\star$ avec $m \neq m^\star$, on dit que le modèle est* ***sous-ajusté****.*

Rappelons que chaque modèle correspond à un choix parmi l'ensemble des variables explicatives, et qu'il y a donc potentiellement des variables explicatives superflues. En cas de sur-ajustement, i.e. s'il y a des variables superflues, un **modèle sur-ajusté** est un modèle contenant toutes les variables du vrai modèle plus un certain nombre de variables superflues. Un **faux modèle** est typiquement un modèle où les variables du vrai modèle n'ont pas toutes été choisies et où certaines variables superflues ont pu être choisies. Un cas particulier est celui du **sous-ajustement** correspondant à un faux modèle ne contenant aucune variable superflue.

Nous allons voir dans la suite diverses approches permettant, non pas de retrouver $m^\star$, mais au moins de s'en approcher. Ceci correspond aux bases de la **sélection de modèle**.

6.4.2 Quelques critères pour sélectionner un modèle

6.4.2.1 Les coefficients d'ajustement

Dans la situation où seul un petit nombre de régresseurs est en jeu, il existe déjà un certain nombre d'approches s'inspirant plus ou moins directement des outils étudiés précédemment. Pour "tester" la validité d'un sous-modèle m par rapport à un modèle plus grand, il existe deux indices (ou coefficients) dont le calcul et l'interprétation sont assez immédiats.

Une première possibilité consiste à s'intéresser au coefficient de détermination :

$$R^2_m = \frac{SCT - SCR(m)}{SCT} = 1 - \frac{\|Y - X_{(m)}\hat{\theta}_{(m)}\|^2}{\|Y - \overline{Y}\mathbb{1}_n\|^2}.$$

Cet indice compare donc les valeurs prédites de Y aux valeurs observées par l'intermédiaire de $\|\widehat{Y}_{(m)} - Y\|^2$, le dénominateur correspondant en quelque sorte à une renormalisation. Plus le coefficient R^2_m sera proche de 1, plus l'adéquation du modèle retenu aux données sera importante. Si on est amené à choisir entre deux modèles explicatifs, on est donc facilement tenté de retenir celui possédant le coefficient de détermination le plus important.

Il est cependant important d'apporter un petit bémol à ce type de raisonnement. En effet la maximisation de ce critère R^2_m revenant à maximiser $\|Y - \widehat{Y}_{(m)}\|^2$, il est clair que la quantité $\|Y - \widehat{Y}_{(m)}\|^2 = \|P_{[X_{(m)}]^\perp} Y\|^2$ décroît pour une suite emboîtée de modèles. Par conséquent, la maximisation de R^2_m conduit à coup sûr à choisir le modèle complet m_k. Utiliser ce type de critère favorise ainsi la sélection de modèles très paramétrés. En revanche, pour des modèles de même cardinal $|m|$, ce coefficient peut être utilisé pour choisir un modèle optimal.

Il est possible d'améliorer le coefficient R^2 pour permettre de sélectionner des modèles comportant un nombre différent de variables explicatives en définissant **le coefficient de détermination ajusté** $\widetilde{R}^2_m$. Ce coefficient permet de tenir compte du

nombre de régresseurs retenus et propose donc un compromis entre l'adéquation et le paramétrage du modèle. Cet indice est défini par :

$$\widetilde{R^2}_m = 1 - \frac{n-1}{n-|m|-1}.\frac{SCR(m)}{SCT} = 1 - \frac{n-1}{n-|m|-1}.\frac{\|Y - X_m\hat{\theta}_{(m)}\|^2}{\|Y - \overline{Y}\mathbb{1}_n\|^2}.$$

L'interprétation est similaire à celle du R^2.

6.4.2.2 Les stratégies de sélections ascendantes et descendantes par le test de Fisher

Le coefficient d'ajustement peut être utilisé en présence d'un petit nombre de modèles. Dans le cas contraire, on peut utiliser une stratégie dite de **régression descendante** faisant appel au test de Fisher sur la présence d'un sous-modèle. La méthodologie est la suivante : on part du modèle utilisant tous les régresseurs possibles. À chaque étape, on calcule la statistique de Fisher correspondant au retrait de chacune des variables encore présentes. On retire alors la variable possédant la plus petite valeur, i.e. la plus grande p-valeur. En fait à chaque étape, on retire la variable la moins significative au sens du test de Fisher. On réitère ensuite ce processus jusqu'à ce que toutes les statistiques soient supérieures à un seuil pré-déterminé, i.e. lorsque toutes les p-valeurs sont toutes plus petites qu'un seuil fixé au préalable, par exemple 5%. Attention, cette stratégie peut être extrêmement lourde à mettre en place suivant le nombre de variables en question (on peut aller jusqu'à $|m|!$ tests de Fisher).

Initialisation : on se donne un seuil s et $m_{[0]} = \{1, \dots, p\}$
Itération t :
Etape 1 : Pour tout $j \in m_{[t]}$, on calcule la p-valeur p_j du test de Fisher de sous-modèle de

$$(M_0) : m_{[t]} \setminus \{j\} \text{ contre } (M_1) : m_{[t]}$$

Etape 2 : $\hat{\jmath} = \arg\max_{j \in m_{[t]}} p_j$
Etape 3 :
- Si $p_{\hat{\jmath}} > s$, $m_{[t+1]} = m_{[t]} \setminus \{\hat{\jmath}\}$ et on retourne à l'étape 1
- Sinon stop.

La sélection de modèle par régression ascendante reprend exactement les mêmes arguments, sauf que l'on part du modèle vide (sans régresseur, uniquement l'intercept) et l'on rajoute au fur et à mesure les variables les plus significatives (au sens du test de Fisher), jusqu'au dépassement par les p-valeurs d'un seuil fixé préalablement.

6.4.2.3 Le critère C_p de Mallows

Le risque quadratique est un critère usuel pour mesurer l'écart entre le vrai modèle $m^\star$ et un modèle d'analyse $m \in \mathcal{M}$.

Définition 6.2. *Soit $m \in \mathcal{M}$.* **Le risque quadratique** *entre les modèles m et $m^\star$ est défini par :*

$$\mathcal{R}(m, m^\star) = \mathbb{E}\left[\left\|\mu^\star - \widehat{Y}_{(m)}\right\|^2\right] = \mathbb{E}\left[\left\|X_{(m^\star)}\theta_{(m^\star)} - X_{(m)}\hat{\theta}_{(m)}\right\|^2\right],$$

où $\mu^\star = X_{(m^\star)}\theta_{(m^\star)}$ et $\widehat{Y}_{(m)} = X_{(m)}\hat{\theta}_{(m)}$.

Par la suite, pour tout $m \in \mathcal{M}$, on définit $\mu^\star_{(m)} = P_{[X_{(m)}]}\mu^\star$, le projeté orthogonal de $\mu^\star$ sur l'espace vectoriel $Im(X_{(m)})$. Il est alors possible de calculer explicitement ce risque quadratique.

Proposition 6.2. *Pour tout $m \in \mathcal{M}$, on a :*

$$\mathcal{R}(m, m^\star) = \sigma^{\star\,2}(|m| + 1) + \|\mu^\star_{(m)} - \mu^\star\|^2. \tag{6.4}$$

La preuve de la proposition 6.2 est donnée en annexe B.3.

Afin de minimiser la distance entre m et $m^\star$, il y a donc un compromis à trouver. Si $|m|$ est petit, il en sera de même pour le terme de variance $\sigma^{\star\,2}(|m| + 1)$, au dépend du terme de biais $\|\mu^\star_{(m)} - \mu^\star\|^2$. Au contraire, pour de grandes valeurs de $|m|$, on peut espérer avoir un petit biais, mais au risque d'avoir une erreur plus importante, ce qui se traduit par une augmentation du terme $\sigma^{\star\,2}(|m| + 1)$. Ce compromis biais-variance est très classique dans ce cadre de sélection de modèle et se retrouve dans un grand nombre de thématiques.

Remarque : À partir du moment où $m^\star \subset m$, on a $\|\mu^\star_{(m)} - \mu^\star\|^2 = 0$, puisque $\mu^\star_{(m)}$ correspond au projeté orthogonal de $\mu^\star$ sur $[X_{(m)}]$.

La question qui se pose à présent est : comment approcher le modèle qui va minimiser le risque quadratique ? Clairement, trouver le meilleur modèle possible nécessite la connaissance de $\mu^\star$... que l'on cherche justement à estimer ! L'idée proposée par Mallows [17] consiste à estimer le risque quadratique à partir des données elles-mêmes et de prendre ensuite une décision à partir de cette estimation. Le modèle $\hat{m}_{CP}$ retenu vérifie :

$$\hat{m}_{CP} = \arg\min_{m \in \mathcal{M}} C_p(m)$$

où le critère C_p de Mallows est défini par

$$C_p(m) = \|Y - \widehat{Y}_{(m)}\|^2 + 2|m|\sigma^2$$

si la variance est connue. Dans le cas où la variance est inconnue, on utilisera l'estimateur $\widehat{\sigma^2} = \widehat{\sigma^2}_{(m_p)}$ où $m_p = \{1, \ldots, p\}$ est le modèle prenant en compte tous les régresseurs. La construction de ce critère est présentée en annexe B.5.

6.4.2.4 Les critères AIC et BIC

Le critère C_p de Mallows est basé sur une volonté de minimiser la distance entre m et le vrai modèle au sens du risque quadratique. Les critères AIC (Akaike Information

Criterion) [2, 3] et BIC (Bayesian Information Criterion) [22] sont eux construits pour minimiser la dissemblance de Kullback entre les 2 modèles.

À chaque modèle d'analyse qui, en général, est un faux modèle, on peut faire correspondre la mesure de probabilité de Y en procédant comme si ce modèle d'analyse était réellement le vrai modèle. On fait donc correspondre la loi de $X_{(m)}\hat{\theta}_{(m)} + \hat{\varepsilon}$. On peut ainsi mesurer l'écart entre la loi du vrai modèle (loi paramétrée par des paramètres inconnus) et la loi engendrée par le modèle d'analyse. Pour mesurer cet écart, un outil souvent utilisé est la dissemblance de Kullback-Leibler.

Définition 6.3. *Soient $\mathbb{P}$ et $\mathbb{P}^\star$ deux mesures de probabilité dominées par une même mesure (dans notre cas la mesure de Lebesgue). La dissemblance de Kullback entre ces deux mesures est donnée par :*

$$KL(\mathbb{P}^\star, \mathbb{P}) = \mathbb{E}_{\mathbb{P}^\star}\left[\log \frac{d\mathbb{P}^\star}{d\mathbb{P}}\right].$$

Si $f = \dfrac{d\mathbb{P}}{d\nu}$ et si $f^\star = \dfrac{d\mathbb{P}^\star}{d\nu}$, alors $KL(\mathbb{P}^\star, \mathbb{P}) = \begin{cases} \int f^\star \log \frac{f^\star}{f} d\nu \text{ si } \mathbb{P}^\star \ll \mathbb{P}, \\ +\infty \text{ sinon.} \end{cases}$

Remarquons, en premier lieu la non symétrie de $KL(.,.)$. C'est pour cette raison que l'on préfèrera parler de dissemblance plutôt que de distance. Cependant, cette dissemblance vérifie comme toute distance "classique" les propriétés suivantes :

- $KL(\mathbb{P}^\star, \mathbb{P}) \geq 0$ pour toutes mesures $\mathbb{P}^\star$ et $\mathbb{P}$;
- $KL(\mathbb{P}^\star, \mathbb{P}) = 0$ si et seulement si $\mathbb{P} = \mathbb{P}^\star$.

Ces propriétés peuvent être démontrées par des arguments de convexité.

Dans le cas où les erreurs sont gaussiennes, ce que nous avons supposé jusqu'à présent, il est possible d'obtenir une expression relativement simple de la dissemblance de Kullback.

Proposition 6.3. *Soit $m \in \mathcal{M}$ fixé. On a alors :*

$$KL(m^\star, m) = \frac{n}{2}\left[\log\left(\frac{\sigma^2_{(m)}}{\sigma^{\star\,2}}\right) + \frac{\sigma^{\star\,2}}{\sigma^2_{(m)}} - 1\right] + \frac{1}{2\sigma^2_{(m)}}\|\mu^\star - \mu^\star_{(m)}\|^2,$$

où $KL(m^\star, m)$ désigne la dissemblance de Kullback entre les deux modèles $m^\star$ et m.

La preuve de la proposition 6.3 en annexe B.4.

Le critère AIC consiste à sélection le modèle vérifiant

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \mathrm{AIC}(m)$$

avec

$$\mathrm{AIC}(m) = -2\text{logvraisemblance au maximum de vraisemblance} + 2D_m$$

où D_m est la dimension du modèle m (i.e le nombre de paramètres pour le modèle m). Nous n'allons pas ici présenter la construction théorique de ce critère AIC. Une preuve est disponible dans [5].

Dans le cas gaussien, la logvraisemblance au maximum de vraisemblance vaut

$$\ln\left[(2\pi\tilde{\sigma}^2_{(m)})^{-n/2}\exp\left(-\frac{1}{2\tilde{\sigma}^2_{(m)}}\|Y-\hat{Y}_{(m)}\|^2\right)\right] = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\tilde{\sigma}^2_{(m)}) - \frac{n}{2}$$

car $\tilde{\sigma}^2_{(m)} = \frac{1}{n}\|Y-\hat{Y}_{(m)}\|^2$. Ainsi la sélection de modèle par le critère AIC peut se réécrire sous la forme

$$\hat{m} = \arg\min_{m\in\mathcal{M}} n\ln(\tilde{\sigma}^2_{(m)}) + 2(|m|+2).$$

Ce type de critère fonctionne plutôt bien pour de petites collections de modèles. Des simulations numériques montrent toutefois que la qualité d'estimation a tendance à se dégrader lorsque m augmente.
Afin de pallier ce problème, il est possible d'utiliser le critère AIC corrigé :

$$\mathrm{AIC}_c(m) = n\ln\left(\tilde{\sigma}^2_{(m)}\right) + n\frac{n+|m|-1}{n-|m|-3}.$$

Le critère BIC (Bayesian Information Criterion) introduit en 1978 par Schwarz [22], est une extension de l'écriture générale du critère d'AIC et utilise le point de vue bayésien. On ne considère plus le paramètre inconnu θ comme un vecteur de $\mathbb{R}^{p+1}$ mais plutôt comme une variable aléatoire à valeurs dans $\mathbb{R}^{p+1}$. Une loi a priori est alors placée sur le 'paramètre' à estimer. La démarche consiste ensuite à essayer d'exploiter cette information pour l'estimation. Ce type d'approche apporte en théorie plus de richesse puisque l'on étend l'éventail des solutions possibles.

Cette approche conduit au critère BIC défini par :

$$BIC(m) = n\log(\hat{\sigma}^2_{(m)}) + \log n \times |m|. \tag{6.5}$$

Le modèle correspondant $\hat{m}_{BIC}$ est obtenu en posant :

$$\hat{m}_{BIC} = \arg\min_{m\in\mathcal{M}} BIC(m). \tag{6.6}$$

Nous ne nous étendrons pas sur les détails permettant d'arriver à la construction de ce critère.

6.4.3 Algorithmes de sélection de variables

En pratique, une fois un critère de sélection de modèles choisi, la détermination du "meilleur" modèle par une recherche exhaustive est impossible en raison du nombre de modèles à explorer. On a donc recourt à des méthodes pas à pas :

1. Les méthodes descendantes :
   On part du modèle en utilisant les p variables explicatives et on cherche, à chaque étape de l'algorithme, la variable la plus pertinente à retirer selon le critère choisi. On itère ainsi l'algorithme jusqu'à atteindre l'ensemble vide. Parmi les ensembles de variables visités pendant l'algorithme, on retient le meilleur au vu du critère. Certains algorithmes s'arrêtent dès lors qu'un seuil donné est atteint.

   Initialisation : $m_{[0]} = \{1, \ldots, p\}$
   Itération t :
   Etape 1 Pour tout $j \in m_{[t]}$, on calcule $c_j = \text{CRIT}(m_{[t]} \backslash \{j\})$.
   Etape 2 $\hat{\jmath} = \arg\max_{j \in m_{[t]}} c_j$
   Etape 3 $m_{[t+1]} = m_{[t]} \setminus \{\hat{\jmath}\}$
   - Si $m_{[t+1]} \neq \emptyset$, on retourne à l'étape 1
   - Sinon stop.

2. Les méthodes ascendantes :
   On part de l'ensemble vide de variables et on cherche, à chaque étape de l'algorithme, la variable la plus pertinente à ajouter selon le critère choisi. On itère ainsi l'algorithme jusqu'à intégrer toutes les variables. Parmi les ensembles de variables visités pendant l'algorithme, on retient le meilleur au vu du critère. Certains algorithmes s'arrêtent dès lors qu'un seuil donné est atteint.

   Initialisation : $m_{[0]} = \emptyset$
   Itération t :
   Etape 1 Pour tout $j \in \{1, \ldots, p\} \setminus m_{[t]}$,
   on calcule $c_j = \text{CRIT}(m_{[t]} \cup \{j\})$.
   Etape 2 $\hat{\jmath} = \arg\min_{j} c_j$
   Etape 3 $m_{[t+1]} = m_{[t]} \cup \{\hat{\jmath}\}$
   - Si $m_{[t+1]} \neq \{1, \ldots, p\}$, on retourne à l'étape 1
   - Sinon stop.

3. Les méthodes stepwise :
   Partant d'un modèle donné, on opère une sélection d'une nouvelle variable (comme avec une méthode ascendante), puis on cherche si on peut éliminer une des variables du modèle (comme pour une méthode descendante) et ainsi de suite. Il faut définir pour une telle méthode un critère d'entrée et un critère de sortie.

4. On peut citer la méthode des "s best subsets" (ou "s meilleurs sous-ensembles") :
   On cherche de façon exhaustive parmi tous les sous-ensembles de s variables, les s meilleures, au sens du critère considéré.

6.4.4 Illustration sur l'exemple

Dans cette section, nous allons illustrer sur notre exemple quelques stratégies de sélection de variables. Grâce à la fonction `regsubsets`, on peut mettre en place une

méthode ascendante, descendante ou séquentielle. On peut également choisir un critère parmi le Cp de Mallows, le R^2 ajusté et le critère BIC. On peut aussi utiliser la fonction `stepAIC`.

```
> library(leaps)
> choixb<-regsubsets(Y~.,data=Data,nbest=1,nvmax=10,method="backward")
> summary(choixb)
> choixf<-regsubsets(Y~.,data=Data,nbest=1,nvmax=10,method="forward")

> plot(choixb,scale="Cp")
> plot(choixb,scale="adjr2")
> plot(choixb,scale="bic")

> reg.fin<-lm(Y~X1 + X6 + X7 + X9 + X10,data=Data)
> anova(reg.fin,reg)
Analysis of Variance Table

Model 1: Y ~ X1 + X6 + X7 + X9 + X10
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     16 73.955
2     11 57.524  5    16.432 0.6284 0.6822

> library(MASS)
> modselect_aic=stepAIC(reg,trace=TRUE,direction=c("backward"))
> modselect_bic=stepAIC(reg,trace=TRUE,direction=c("backward"),k=log(n))
```

Par exemple, avec les critères BIC et le Cp de Mallows (Figures 6.8 et 6.9), on retient le modèle composé des variables X1, X6, X7, X9 et X10. Le test de sous-modèle `anova(reg.fin,reg)` confirme que le sous-modèle suffit pour expliquer la variable Y. Avec le R^2 ajusté, le modèle retenu contient plus de variables (X1, X3, X6, X7, X8, X9 et X10).

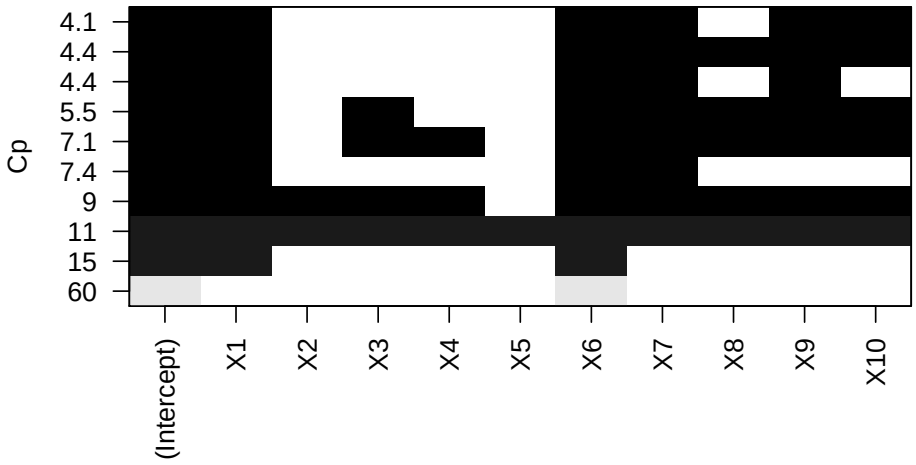


FIGURE 6.8 – Résultat du processus de sélection de variables avec le critère Cp de Mallows.

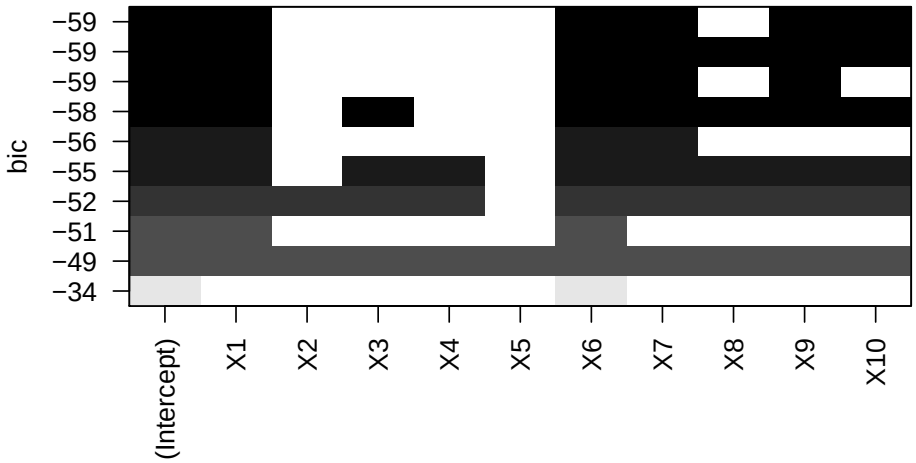


FIGURE 6.9 – Résultat du processus de sélection de variables avec le critère BIC.

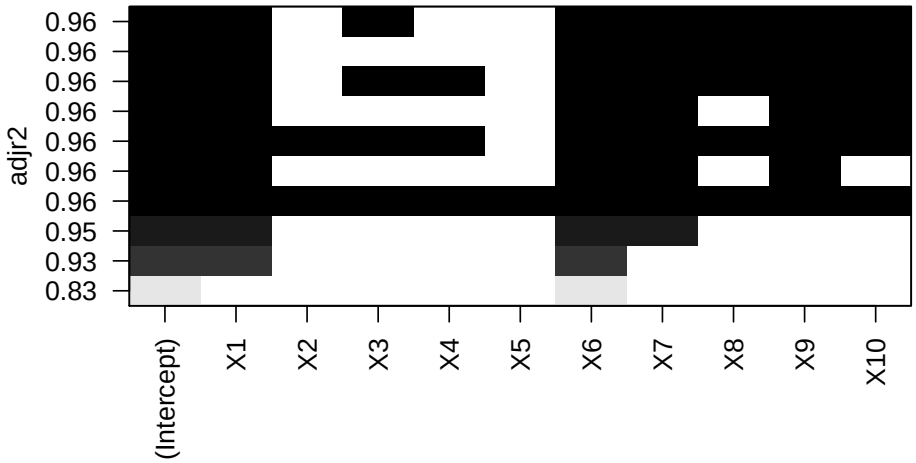


FIGURE 6.10 – Résultat du processus de sélection de variables avec le R^2 ajusté.

6.5 Régression linéaire régularisée

Quand on se retrouve avec un modèle singulier, $\text{rg}(X) < k$, la matrice $X'X$ n'est plus inversible. Ce cas se présente quand

- le nombre de variables explicatives est supérieur au nombre d'observations ($n < p$)
- $n > p$ mais des variables sont linéairement redondantes (la famille $\{X^{(1)}, \ldots, X^{(p)}\}$ est liée)

Dans cette situation, on a vu précédemment que l'estimateur des moindres carrés $\widehat{\theta}$ n'existe pas. La projection $\widehat{Y} = P_{[X]}Y$ de la réponse Y sur $Im(X) = [X]$ n'a pas une décomposition unique sur les colonnes de X (le modèle est non identifiable, voir Chapitre 5). De plus, comme la matrice de variance-covariance de $\widehat{\theta}$ vaut $\sigma^2(X'X)^{-1}$, la précision de l'estimateur $\widehat{\theta}$ diminue quand $X'X$ se rapproche d'une matrice non inversible.

Du point de vue de la prédiction, si $x^\star$ est un nouveau vecteur de valeurs des variables explicatives, on sait que la qualité (au sens écart quadratique) de la prédiction $\hat{Y}^\star$ de la vraie réponse $Y^\star$ se décompose en le biais^2 + variance. Donc pour améliorer la prédiction, on peut préférer une augmentation légère du biais pour avoir une diminution de la variance.

On va donc chercher dans ce contexte à utiliser des méthodes de régression dites régularisées pour pallier ces difficultés. Elles ont pour formalisme commun l'optimisation d'un critère de la forme

$$\underset{\theta \in \mathbb{R}^k}{\text{argmin}} \ \|Y - X\theta\|^2 + \tau \ \text{pen}(\theta)$$

où $\tau > 0$ est une quantité à choisir. Elles se distinguent par la forme de la fonction de pénalité $\text{pen}(\theta)$ qui fera intervenir le contrôle d'une norme de θ.

En pratique on commence par centrer et réduire les variables explicatives $z^{(j)}$ pour ne pas pénaliser ou favoriser un coefficient de θ car les pénalisations que nous allons considérer portent sur une norme de θ. Il est donc préférable que chaque coefficient soit affecté de façon "semblable". La matrice des variables explicatives centrées-réduites est notée $\tilde{X}$. De plus, l'intercept θ_0 étant un coefficient qui a un rôle particulier assurant au modèle de se positionner autour du comportement moyen de Y, il n'a pas à intervenir dans la contrainte sur la norme de θ. Aussi, on centre le vecteur réponse Y, $\tilde{Y} = Y - \bar{Y}\mathbb{1}_n$, et on peut potentiellement le réduire. A noter que le modèle est alors de la forme $\tilde{Y} = \tilde{X}\theta + \varepsilon$ avec $\theta = (\theta_1, \ldots, \theta_p)'$ (donc $k = p$ et sans intercept).

Ainsi, après transformation initiale des données, nous allons ici nous intéresser à des méthodes de régression régularisées qui cherchent à minimiser le risque empirique régularisé (pour la perte quadratique) :

$$\underset{\theta \in \mathbb{R}^k}{\text{argmin}} \ \left\{ \|\tilde{Y} - \tilde{X}\theta\|^2 + \tau \|\theta\|_q^q \right\} \ \text{où} \ \|\theta\|_q^q = \sum_{j=1}^{p} (\theta_j)^q.$$

On parle de régression ridge quand $q = 2$, de régression Lasso quand $q = 1$. Nous allons détailler ces deux méthodes et la régression Elasticnet qui combine les deux premières. Pour illustrer cette section, nous reprenons le jeu de données **Data-ExRegMultiple.txt** auquel on a ajouté 10 variables de bruit (simulation selon une loi $\mathcal{N}(0, 1)$).

6.5.1 Régression ridge

Dans le contexte présenté précédemment, la difficulté vient de l'inversibilité de $\tilde{X}'\tilde{X} \in \mathcal{M}_p(\mathbb{R})$. Cette matrice $\tilde{X}'\tilde{X}$ est une matrice semi-définie positive donc ses valeurs propres sont positives et on les ordonne $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Si $\tilde{X}'\tilde{X}$ n'est pas inversible, c'est qu'au moins l'une de ses valeurs propres est nulle.

Proposition 6.4. *Soit $\tau > 0$. Les matrices $\tilde{X}'\tilde{X}$ et $\tilde{X}'\tilde{X} + \tau I_p$ ont les mêmes vecteurs propres mais leur valeurs propres sont $\{\lambda_j\}_{j\in[|1,p|]}$ et $\{\lambda_j + \tau\}_{j\in[|1,p|]}$ respectivement. Ainsi, $det(\tilde{X}'\tilde{X}+\tau I_p) > det(\tilde{X}'\tilde{X})$, donc $\tilde{X}'\tilde{X}+\tau I_p$ a "plus de chance" d'être inversible que $\tilde{X}'\tilde{X}$.*

En exploitant la Proposition 6.4, l'idée consiste à remplacer $(\tilde{X}'\tilde{X})^{-1}$ dans l'expression de l'estimateur des moindres carrés $\widehat{\theta}$ par $(\tilde{X}'\tilde{X} + \tau I_p)^{-1}$. Ainsi l'estimateur ridge est donné par

$$\widehat{\theta}_{\text{ridge}}(\tau) = (\tilde{X}'\tilde{X} + \tau I_p)^{-1}\tilde{X}'\tilde{Y}.$$

Cet estimateur ridge est solution du problème optimisation suivant

$$\widehat{\theta}_{\text{ridge}}(\tau) \in \underset{\theta\in\mathbb{R}^p}{\text{argmin}} \ \|\tilde{Y} - \tilde{X}\theta\|_2^2 + \tau\|\theta\|_2^2,$$

qui peut être reformulé en le problème de minimisation sous contrainte suivant :

$$\|\tilde{Y} - \tilde{X}\theta\|_2^2 \text{ sous la contrainte } \|\theta\|_2^2 \leq r(\tau)$$

où $r(.)$ est bijective. La régression ridge conserve toutes les variables mais avec la contrainte $\|\theta\|_2^2 \leq r(\tau)$, elle empêche les estimateurs de prendre de trop grandes valeurs et limite ainsi la variance des prédictions. On parle de "shrinkage" car on rétrécit l'étendue des valeurs possibles des paramètres estimés.

Proposition 6.5. *Soit l'estimateur ridge $\widehat{\theta}_{ridge}(\tau) = (\tilde{X}'\tilde{X} + \tau I_p)^{-1}\tilde{X}'\tilde{Y}$. On a*

- $\mathbb{E}[\widehat{\theta}_{ridge}(\tau)] = \theta - \tau(\tilde{X}'\tilde{X} + \tau I_p)^{-1}\theta$ *donc il est biaisé.*
- $Var(\widehat{\theta}_{ridge}(\tau)) = \sigma^2(\tilde{X}'\tilde{X} + \tau I_p)^{-1}(\tilde{X}'\tilde{X})(\tilde{X}'\tilde{X} + \tau I_p)^{-1} \leq \sigma^2(\tilde{X}'\tilde{X})^{-1} = Var(\widehat{\theta})$.
- *Les valeurs ajustées pour Y sont*

$$\widehat{Y}_{ridge}(\tau) = \tilde{X}\widehat{\theta}_{ridge}(\tau) + \bar{Y}\mathbb{1}_n$$

- *Quand $\tau \to +\infty$, $\widehat{\theta}_{ridge}(\tau) \to 0$*
- *Quand $\tau \to 0$, $\widehat{\theta}_{ridge}(\tau) \to \widehat{\theta}$*

L'estimateur $\widehat{\theta}_{\text{ridge}}(\tau)$ dépend du choix de τ qui est un point délicat. C'est pratiquement impossible de pouvoir faire ce choix a priori. On peut tracer le *chemin de régularisation* de la régression ridge qui est l'ensemble des fonctions $\tau \mapsto (\widehat{\theta}_{\text{ridge}}(\tau))_j$ pour $j = 1, \ldots, p$ (voir Figure 6.11). On constate que le chemin de régularisation de la régression ridge est continu, ne permettant pas un ajustement aisé de τ. On peut également suivre les recommandations proposées dans la littérature, voir par exemple [15, 16, 17, 19]. En pratique, on passe par une procédure de validation croisée pour calibrer τ (Figure 6.12) :
On commence par séparer les données en un jeu d'apprentissage (Y_a, X_a) et un jeu de test (Y_v, X_v). On estime alors la régression ridge sur le jeu d'apprentissage pour chaque valeur de τ dans une grille de valeurs choisie et on prédit la réponse sur le jeu de test pour chaque valeur de τ : $\widehat{Y}_{\text{ridge},v}(\tau)$. La qualité du modèle est alors obtenue en comparant les vraies données Y_v et les valeurs prédites $\widehat{Y}_{\text{ridge},v}(\tau)$. Par exemple, on peut utiliser le critère PRESS

$$PRESS(\tau) = \|Y_v - \widehat{Y}_{\text{ridge},v}(\tau)\|^2.$$

Finalement on choisit la valeur de τ qui minime ce critère.
Le principe de la validation croisée est de répéter plusieurs fois le découpage entre test et apprentissage et de considérer la moyenne des valeurs du critère pour chaque valeur de τ.

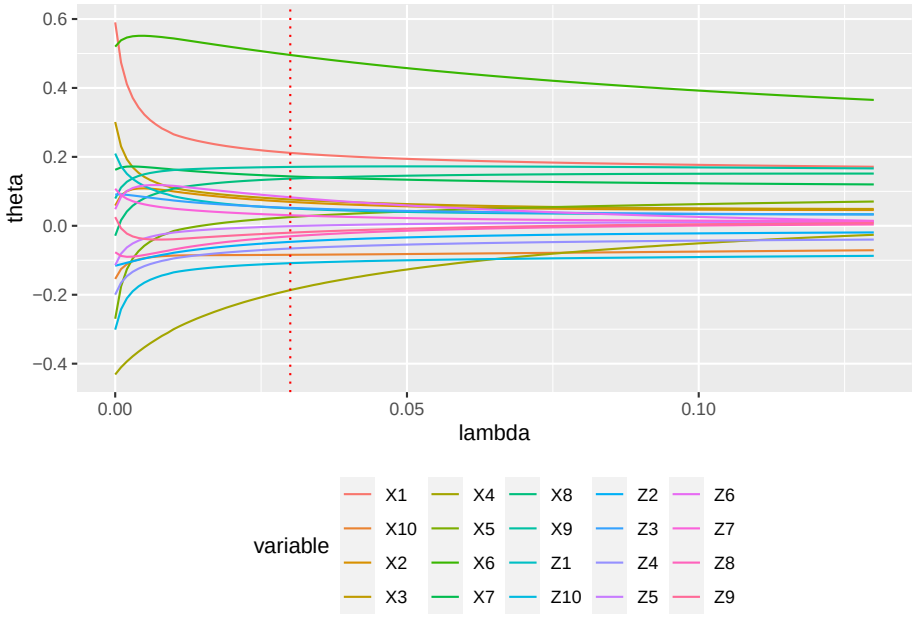


FIGURE 6.11 – Chemins de regularisation pour la régression ridge sur notre exemple.

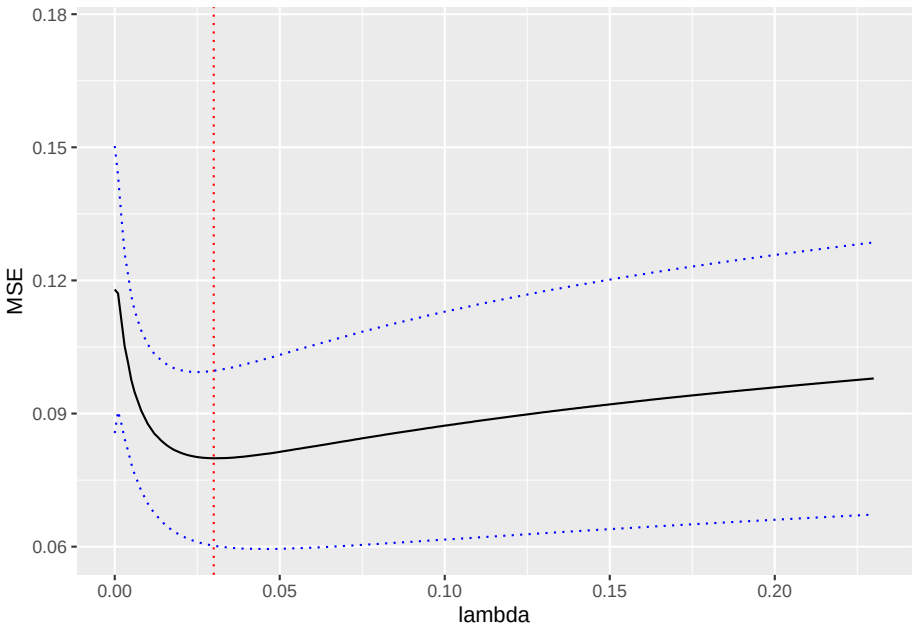


FIGURE 6.12 – Sélection de τ par validation croisée pour la régression ridge sur notre exemple.

6.5.2 Régression Lasso

L'idée de la régression LASSO (Least Absolute Selection and Shrinkage Operator) proposée par Tibshirani [23] est d'essayer d'annuler des coefficients du vecteur θ afin d'avoir un estimateur parcimonieux (sparse en anglais). Cela induit une sélection de variables rendant le modèle plus interprétable et une matrice des variables explicatives avec de meilleures propriétés que $X'X$. Pour forcer à annuler des coordonnées de θ, on contraint la norme ℓ_1 : $\|\theta\|_1 = \sum_{j=1}^{p} |\theta_j|$. Comme pour la régression ridge, on commence par centrer-réduire les variables explicatives $(\tilde{X})$ et au moins centrer le vecteur des réponses $(\tilde{Y})$.

L'estimateur LASSO est défini pour $\tau > 0$ par

$$\widehat{\theta}_{\text{lasso}}(\tau) \in \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \ \|\tilde{Y} - \tilde{X}\theta\|_2^2 + \tau\|\theta\|_1. \tag{6.7}$$

Ce problème de minimisation est équivalent à minimiser $\|\tilde{Y} - \tilde{X}\theta\|_2^2$ sous la contrainte $\|\theta\|_1 \leq r(\tau)$ avec $r(.)$ bijective. La solution du problème (6.7) peut ne pas être unique mais le vecteur des valeurs ajustées en résultant $\tilde{X}\widehat{\theta}_{\text{lasso}}(\tau)$ est lui toujours unique. Quand $\tau = 0$, $\widehat{\theta}_{\text{lasso}}(0) = \widehat{\theta}$; quand $\tau \to +\infty$, $\widehat{\theta}_{\text{lasso}}(+\infty) = 0$.

Comme pour la régression ridge, le choix de τ est délicat, il est impossible de faire ce choix a priori. On peut tracer le *chemin de régularisation* de la régression Lasso c'est-à-dire l'ensemble des fonctions $\tau \mapsto \widehat{\theta}_{\text{lasso}}(\tau)_j$ pour $j = 1, \ldots, p$ (voir Figure 6.13). Comme pour la régression ridge, on passe par une procédure de validation croisée pour stabiliser le choix de τ (voir Figure 6.14).

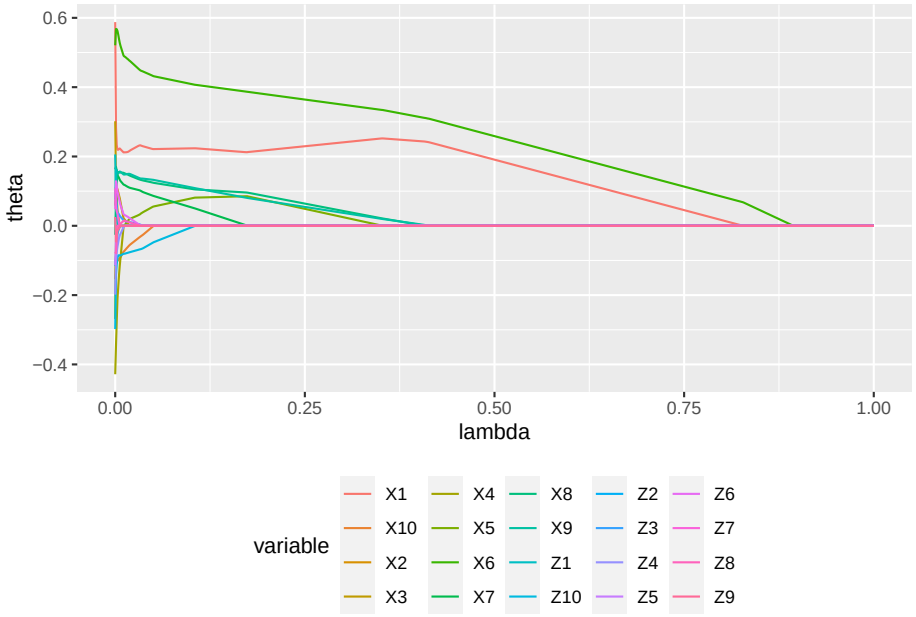


FIGURE 6.13 – Chemins de régularisation pour la régression Lasso sur notre exemple

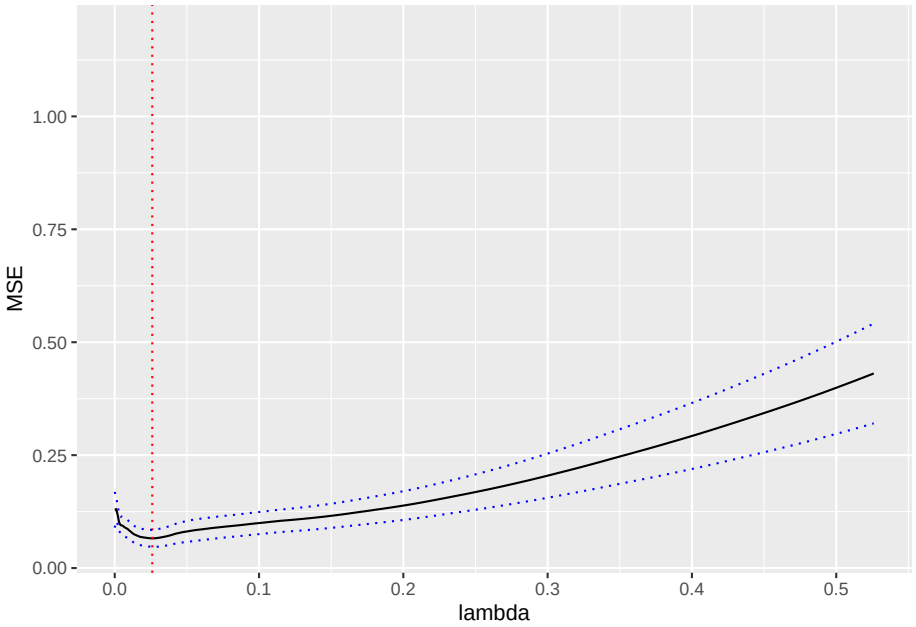


FIGURE 6.14 – Sélection de τ par validation croisée pour la régression lasso sur notre exemple.

6.5.3 Régression Elastic-Net

La régression Elastic-Net combine les avantages de la régression ridge et de la régression Lasso. En particulier, elle pallie le défaut de l'estimation Lasso lorsque les $x^{(j)}$ sont fortement corrélées. L'estimateur Elastic-Net [24] est défini pour $\tau > 0$ et $\alpha > 0$ par

$$\widehat{\theta}_{\text{net}}(\tau, \alpha) \in \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \ \|\tilde{Y} - \tilde{X}\theta\|_2^2 + \tau\{\alpha\|\theta\|_1 + (1-\alpha)\|\theta\|_2^2\}$$

ce qui peut se reformuler en minimiser $\|\tilde{Y} - \tilde{X}\theta\|_2^2$ sous la contrainte $\alpha\|\theta\|_1 + (1-\alpha)\|\theta\|_2^2 \leq r(\tau)$. Il faut alors utiliser des algorithmes d'optimisation pour déterminer $\widehat{\theta}_{\text{net}}(\tau, \alpha)$ et la calibration des seuils τ et α est souvent faite par validation croisée en pratique. La Figure 6.15 illustre les différences sur les chemins de régularisation des trois méthodes.

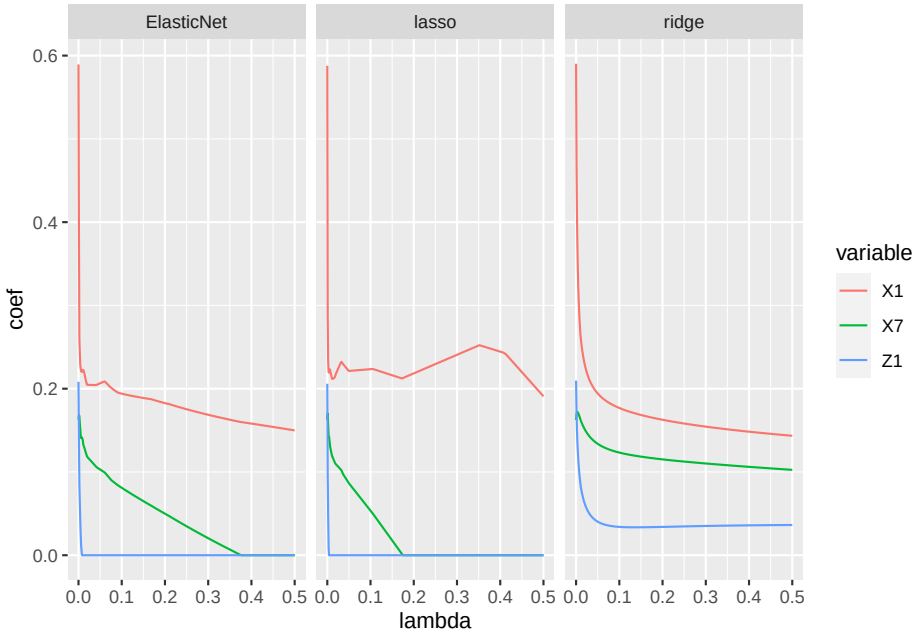


FIGURE 6.15 – Chemins de régularisation pour 3 variables du jeu de données pour la régression Lasso ($\alpha = 1$), régression ridge ($\alpha = 0$) et la régression Elastic Net (ici $\alpha = 0.5$)

6.6 Validation du modèle

6.6.1 Contrôle graphique a posteriori

Une fois le modèle mis en oeuvre, on doit vérifier *a posteriori* le "bien-fondé statistique" de ce modèle du point de vue de la normalité des résidus et de l'adéquation de la valeur ajustée $\widehat{Y}_i$ à la valeur observée Y_i et de l'absence de données aberrantes. Il

est alors indispensable de commencer par s'entourer de "protections" graphiques pour vérifier empiriquement les 4 postulats de base (au moins les hypothèses H1-H3, puisque l'hypothèse H4 n'est pas vraiment important dès que l'on dispose de suffisamment de données).

- En régression linéaire simple, la confrontation graphique entre le nuage de points (z_i, y_i) et la droite de régression de Y par z par moindres carrés ordinaires donne une information quasi exhaustive (cf Figure 6.2).
  Sur ce graphique, si nous voyons une courbure de la "vraie" courbe de régression de Y, nous pouvons alors penser que le modèle est inadéquat et que l'hypothèse H1 n'est pas vérifiée.
- Dans le cas de la régression multiple, ce type de graphique n'est pas utilisable car il y a plusieurs régresseurs. Les différentes hypothèses sont donc à vérifier sur les termes des erreurs ε_i qui sont malheureusement inobservables. Nous utilisons alors leurs prédicteurs naturels, les résidus $\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$.
- Le graphe des n points $(y_i, \widehat{y_i})$ est également très informatif. Il suffit alors de vérifier si les points sont alignés selon la première bissectrice (cf. Figure 6.16).

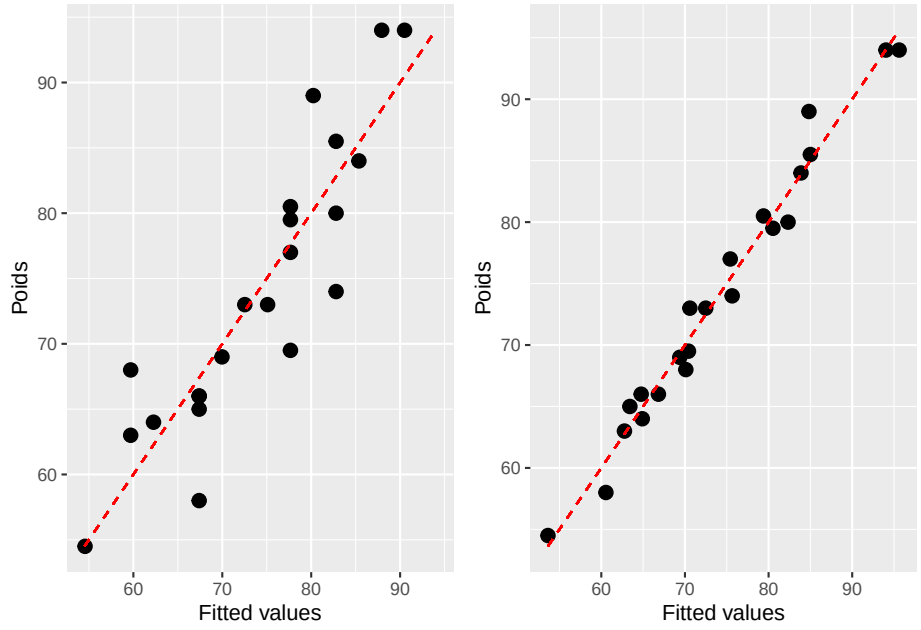


FIGURE 6.16 – Graphique des points $(y_i, \widehat{y_i})$ pour l'exemple en régression linéaire simple (à gauche) et multiple (à droite)

Voici maintenant plusieurs démarches permettant de s'assurer de la légitimité des conclusions, démarches à effectuer pour toute régression linéaire multiple.

6.6.2 Pour vérifier les hypothèses H1 et H2 : adéquation et homoscédasticité

Le graphique le plus classique consiste à représenter les résidus $(\widehat{\varepsilon_i})_i$ en fonction des valeurs prédites $(\widehat{Y_i})_i$ (cf graphique en haut à gauche Figures 6.17 et 6.18). Ce graphique doit être fait pratiquement systématiquement. Cela revient encore à tracer les coordonnées du vecteur $P_{[X]^\perp}Y$ en fonction de celles de $P_{[X]}Y$. L'intérêt d'un tel graphique réside dans le fait que si les 4 hypothèses H1-H4 sont bien respectées, il y a indépendance entre ces 2 vecteurs qui sont centrés et gaussiens (d'après le théorème de Cochran). Cependant, à partir de ce graphe, nous ne pourrons nous apercevoir que de la possible déficience des hypothèses H1 et H2. Concrètement, si on ne voit rien de notable sur le graphique, i.e. si l'on observe un nuage de points centrés et alignés quelconque, c'est très bon signe : les résidus ne semblent alors n'avoir aucune propriété intéressante et c'est bien ce que l'on demande à l'erreur.

Voyons maintenant 2 types de graphes résidus/valeurs prédites "pathologiques" (Figure 6.19) :

1. Type 1 "forme banane" :
   Dans ce cas, on peut penser que le modèle n'est pas adapté aux données. En effet, il ne semble pas y avoir indépendance entre les $\widehat{\varepsilon_i}$ et les $\widehat{Y_i}$, puisque, par exemple, les $\widehat{\varepsilon_i}$ ont tendance à décroître lorsque les $\widehat{Y_i}$ sont dans un certain intervalle et croissent. Il faut donc améliorer l'analyse du problème pour proposer d'autres régresseurs pertinents ou transformer les régresseurs $z^{(j)}$ par une fonction de type (log, sin).

2. Type 2 "forme trompette"
   Dans ce cas la variance des résidus semble inhomogène, puisque les $\widehat{\varepsilon_i}$ ont une dispersion de plus en plus importante au fur et à mesure que les $\widehat{Y_i}$ croissent. Un changement de variable pour Y pourrait être une solution envisageable afin de "rendre" constante la variance du bruit (cf pararaphe suivant).

En cas de comportement inadéquat, les modifications possibles à apporter au modèle sont :

- On peut librement transformer les régresseurs $z^{(1)}, \cdots, z^{(p)}$ par toutes les transformations algébriques ou analytiques connues (fonctions puissances, exponentielles, logarithmiques...), pourvu que le nouveau modèle reste interprétable. Cela peut permettre d'améliorer l'adéquation du modèle ou diminuer son nombre de termes si on utilise ensuite une procédure de choix de modèles.
- En revanche, on ne peut envisager de transformer Y que si les graphiques font suspecter une hétéroscédasticité. Dans ce cas, cette transformation doit obéir à des règles précises basées sur la relation suspectée entre l'écart-type résiduel σ et la réponse Y : c'est ce que précise le Tableau 6.1.

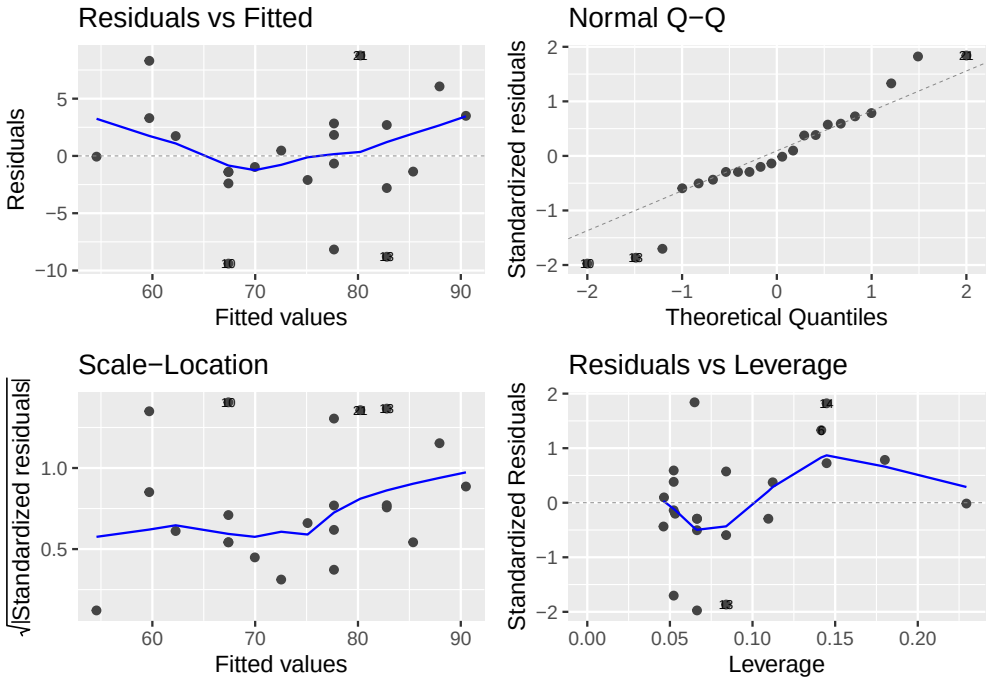


FIGURE 6.17 – Graphiques pour l'étude des résidus pour l'exemple en régression linéaire simple

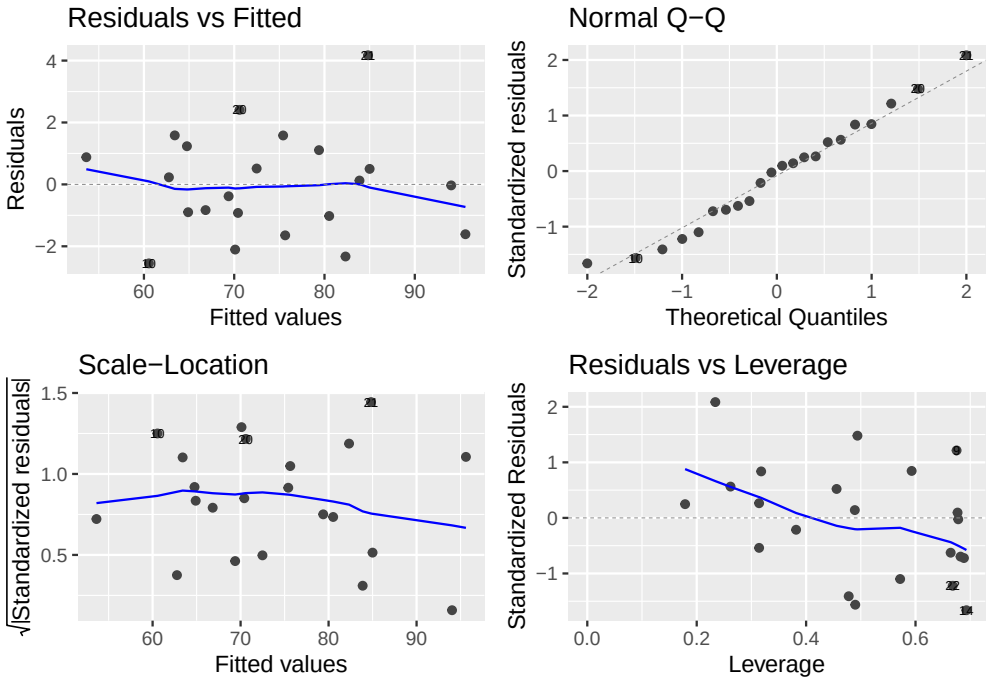


FIGURE 6.18 – Graphiques pour l'étude des résidus pour l'exemple en régression linéaire multiple

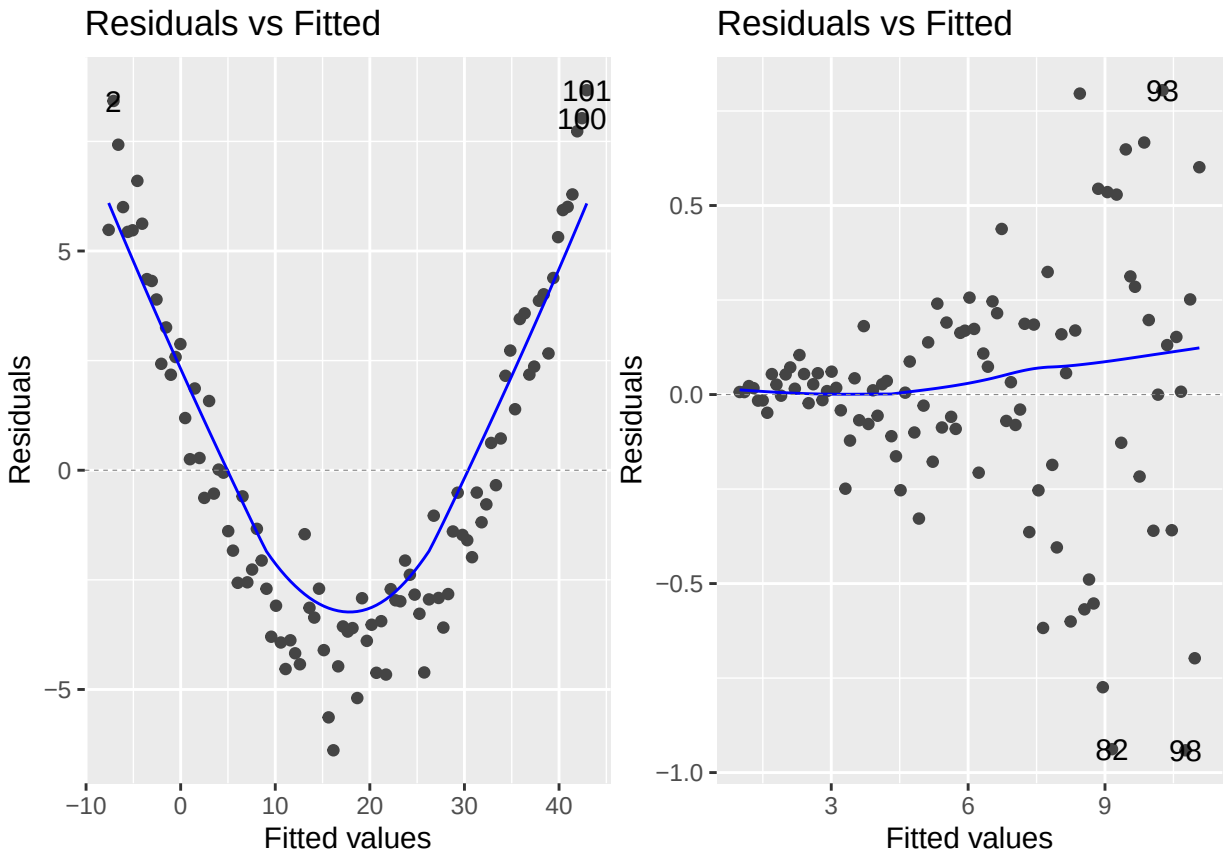


Figure 6.19 – Exemple du type "forme banane" (gauche) et "forme trompette" (droite)

6.6.3 Pour vérifier l'hypothèse H3 : indépendance

Un graphe pertinent pour s'assurer de l'indépendance des erreurs entre elles est celui des résidus $\widehat{\varepsilon_i}$ en fonction de l'ordre des données (lorsque celui-ci a un sens, en particulier s'il représente le temps). Un tel graphique est potentiellement suspect si les résidus ont tendance à rester par paquets lorsqu'ils se trouvent d'un côté ou de l'autre de 0. On pourra confirmer ces doutes en effectuant un test de runs (cf [12], p. 157). Ce test est basé sur le nombre de runs, i.e. sur le nombre de paquets de résidus consécutifs de même signe.

Par ailleurs, si les erreurs sont corrélées suivant certaines conditions (par exemple si ce sont des processus ARMA), il est tout d'abord possible d'obtenir encore des résultats quant à l'estimation des paramètres. Mais il existe également des méthodes de correction telles que les estimations par moindres carrés généralisés ou pseudo-généralisés, cf [14] ou d'autres.

Nature de la relation	Domaine pour Y	Transformation
$\sigma = (cste)Y^k$, $k \neq 1$	$\mathbb{R}^{+*}$	$Y \mapsto Y^{1-k}$
$\sigma = (cste)\sqrt{Y}$	$\mathbb{R}^{+*}$	$Y \mapsto \sqrt{Y}$
$\sigma = (cste)Y$	$\mathbb{R}^{+*}$	$Y \mapsto \log(Y)$
$\sigma = (cste)Y^2$	$\mathbb{R}^{+*}$	$Y \mapsto Y^{-1}$
$\sigma = (cste)\sqrt{Y(1-Y)}$	$[0;1]$	$Y \mapsto \arcsin\sqrt{Y}$
$\sigma = (cste)\sqrt{1-Y}Y^{-1}$	$[0;1]$	$Y \mapsto (1-Y)^{\frac{1}{2}} - \frac{1}{3}(1-Y)^{\frac{3}{2}}$
$\sigma = (cste)(1-Y)^{-2}$	$[-1;1]$	$Y \mapsto \log(1+Y) - \log(1-Y)$

TABLE 6.1 – Table des changements de variable pour la variable à expliquer afin de stabiliser la variance de Y

6.6.4 Pour vérifier l'hypothèse H4 : gaussianité

Notamment pour que les tests de Fisher et de Student aient un sens, il peut être intéressant de vérifier si l'hypothèse de gaussianité est acceptable. Pour cela, nous déconseillons fortement les tests d'adéquation classiques de Kolmogorov-Smirnov, Cramer-Von Mises,..., du fait qu'on les appliquera sur les résidus $\widehat{\varepsilon_i}$, qui ne sont (quasiment) jamais indépendants. On préfèrera se "contenter" d'une vérification graphique à partir du tracé d'une droite de Henri, dite encore graphique QQ-plot (cf graphiques en haut à droite Figures 6.17 et 6.18). Celle-ci relie les points de $\mathbb{R}^2$ formés par les quantiles empiriques des résidus studentisés (i.e. le $\widehat{\varepsilon_i}$ divisés par leur écart-type empirique) en fonction des quantiles théoriques (pour les probabilités $k/(n+1)$ où $k = 1, \cdots, n$, n étant le nombre de données) d'une loi normale centrée réduite. La loi de Student "ressemblant" fortement à une loi gaussienne dès que le paramètre dépasse la dizaine, si les erreurs (ε_i) sont gaussiennes, i.e. sous H4, alors la droite de Henri est une bissectrice du plan. Ce type de tracé permet surtout de voir si une loi à "queue de distribution lourde" ne pourrait pas être plus adéquate (dans ce cas, les points s'éloignent de la droite de Henri en ses extrémités).

6.6.5 Détection de données aberrantes

Nous allons ici décrire deux méthodes permettant de détecter des données "aberrantes".

Effet levier avec la matrice H

On reprend la matrice chapeau $H = X(X'X)^{-1}X'$. La prédiction pour le ième individu est donné par

$$\hat{Y}_i = (X\hat{\theta})_i = (HY)_i = H_{ii}Y_i + \sum_{j \neq i} H_{ij}Y_j.$$

Si $H_{ii} = 1$, $\hat{Y}_i$ est entièrement déterminée par la ième observation alors que, si $H_{ii} = 0$, la ième observation n'a aucune influence sur $\hat{Y}_i$. Ainsi, pour mesurer l'influence d'une

observation sur sa propre estimation, on peut examiner le diagramme en batons des termes diagonaux de H (cf Figure 6.20). En pratique, on déclare la ième observation comme **levier** si H_{ii} dépasse $2k/n$ ou $3k/n$.

Distances de Cook

Les points influents sont les points tels que, si on les retire de l'étude, l'estimation des coefficients du modèle sera fortement modifiée. La mesure la plus classique d'influence est la distance de Cook. C'est une distance entre le coefficient estimé avec toutes les observations et celui estimé en enlevant une observation. La distance de Cook pour la ième observation est définie par

$$DC_i = (\widehat{\theta} - \widehat{\theta}^{(-i)})' T' T (\widehat{\theta} - \widehat{\theta}^{(-i)})$$

où T est le vecteur des résidus studentisés et $\widehat{\theta}^{(-i)}$ l'EMV sans l'observation i. On peut là encore tracer le diagramme en bâtons des DC_i (cf Figure 6.20). Si une distance se révèle grande par rapport aux autres alors ce point sera considéré comme influent. Il faut alors chercher à comprendre pourquoi il est influent : il est levier, aberrant, les deux, ...

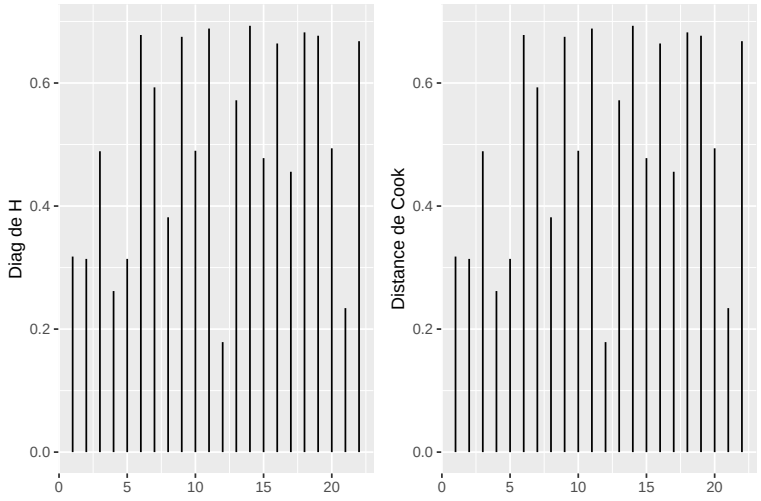


Figure 6.20 – Diagramme en bâtons des termes diagonaux de la matrice chapeau H (à gauche) et des distances de Cook (à droite).

En résumé :

- Savoir écrire un modèle de régression linéaire (pour chaque individu et matriciellement)
- Savoir déterminer les estimateurs de la régression linéaire et leur loi.
- Savoir construire en régression linéaire un intervalle de confiance pour un paramètre et un intervalle de prédiction
- Savoir construire un test de nullité d'un (ou des) paramètre(s)
- Comprendre la problématique de la sélection de variables, savoir proposer des stratégies en pratique et savoir interpréter mes sorties de R
- Comprendre le principe des méthodes de régression régularisées et savoir lire les sorties de R associées.
- Savoir interpréter les graphiques de contrôle pour valider les hypothèses du modèle linéaire.

Chapitre 7

Analyse de variance (ANOVA)

7.1 Vocabulaire

On se place ici dans le cas où l'on souhaite expliquer une variable quantitative à l'aide d'une ou plusieurs variables *qualitatives* explicatives, appelées **facteurs**. Un facteur est dit **contrôlé** si ses valeurs ne sont pas observées mais fixées par l'expérimentateur. Les modalités d'une variable qualitative explicative sont appelées **niveaux** du facteur.

Un plan d'expérience répertorie l'ensemble des combinaisons des différents facteurs considérés par l'expérimentateur. Nous donnons ici qu'un peu de vocabulaire sur les plans d'expérience pour la suite, nous n'aborderons pas la théorie de la planification expérimentale dans ce cours.

Définition 7.1.

- *On appelle* ***cellule*** *d'un plan d'expérience une case du tableau, associée à une combinaison des facteurs contrôlés.*
- *Un plan est dit* ***complet*** *s'il a au moins une observation dans chaque cellule.*
- *Un plan est dit* ***répété*** *s'il a plus d'une observation par cellule.*
- *Un plan est dit* ***équilibré*** *si chaque cellule comporte le même nombre d'observations.*
- *Un plan équilibré et répété est dit* ***equirépété****.*

7.2 Analyse de variance à un facteur

7.2.1 Exemple et notations

On dispose d'une variable quantitative Y à expliquer et d'un seul facteur explicatif. On note

- i l'indice du niveau (ou de la "cellule") pour le facteur explicatif,

- I le nombre de niveaux ($i = 1, \cdots, I$),
- n_i le nombre d'expériences dans le niveau i,
- $j = 1, \cdots, n_i$ l'indice de l'expérience dans le niveau i,
- $n = \sum_{i=1}^{I} n_i$ le nombre total d'expériences.

Une expérience (ou encore un "individu") est repérée par deux indices : le numéro de la cellule (i) et le numéro de l'observation dans la cellule (j). Ainsi on note Y_{ij} la valeur théorique de la réponse quantitative pour l'expérience j du niveau i.

Dans cette section, nous allons illustrer les notions abordées avec l'exemple suivant : On s'intéresse aux notes obtenues par des étudiants à un oral. On s'interroge sur un effet potentiel de l'examinateur sur la note obtenue.

Examinateur (i)	A	B	C
Notes (Y_{ij})	10,11,11	8,11,11,13	10,13,14,14
	12,13,15	14,15,16,16	15,16,16
Effectifs (n_i)	6	8	7
Moyenne ($Y_{i.}$)	12	13	14

TABLE 7.1 – Données pour illustrer l'anova à un facteur

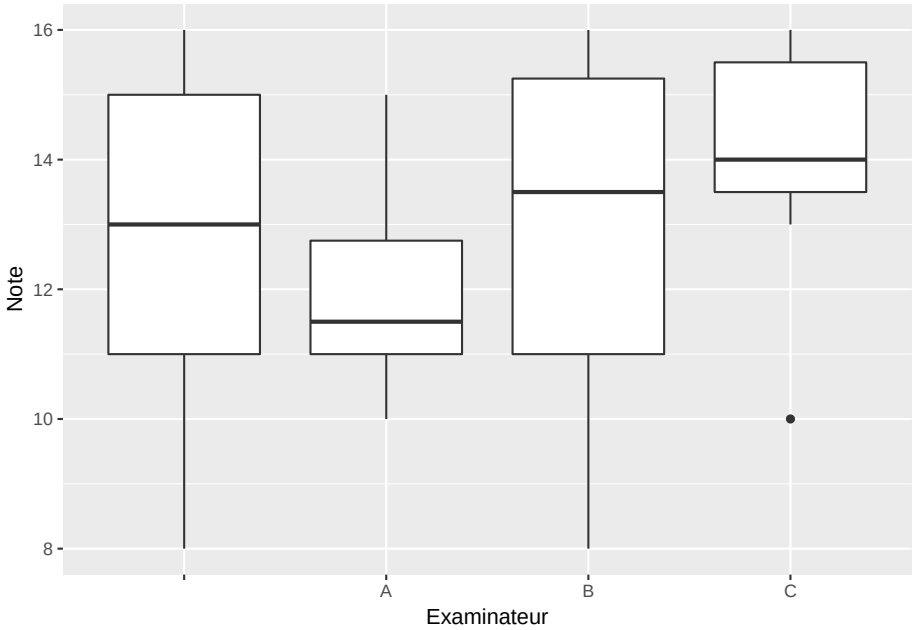


FIGURE 7.1 – Boxplot des notes globalement (à gauche) puis par examinateur

7.2.2 Modèle d'ANOVA à un facteur

On modélise une variable quantitative en fonction d'un facteur à I niveaux. Le modèle s'écrit :

$$\begin{cases} Y_{ij} = m_i + \varepsilon_{ij},\ \forall i = 1, \cdots I,\ \forall j = 1, \cdots, n_i \\ \\ \varepsilon_{ij} \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases} \tag{7.1}$$

Le modèle peut se réécrire sous la forme matricielle suivante :

$$Y = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbb{1}_{n_1} & 0_{n_1} & \dots & 0_{n_1} \\ 0_{n_2} & \mathbb{1}_{n_2} & \dots & 0_{n_2} \\ \vdots & \vdots & \vdots & \vdots \\ 0_{n_I} & 0_{n_I} & \dots & \mathbb{1}_{n_I} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_I \end{pmatrix}}_{\theta} + \varepsilon \qquad \text{avec } \varepsilon \sim \mathcal{N}_n\left(0_n, \sigma^2 I_n\right).$$

Avec le logiciel R, il suffit d'utiliser la commande `an1<-lm(Notes~Exam -1)`. On peut vérifier la matrice de design X par la commande `model.matrix(Notes~Exam -1)`.

Pour des raisons d'interprétation, on peut s'intéresser à un changement de paramétrage. Il s'agit d'un changement de variables dans la fonction à minimiser dont les variables sont les paramètres du modèle. Soulignons que les nouvelles équations que nous allons définir ci-après correspondent toujours à celles d'un modèle à un facteur. Si on veut comparer les effets des niveaux du facteur, on peut prendre comme référence un effet moyen et examiner les écarts des effets des différents niveaux à cet effet moyen. Le modèle initial (7.1) peut s'écrire sous la forme :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{7.2}$$

où μ est l'effet moyen et $\alpha_i = m_i - \mu$ l'effet différentiel (centré) du niveau i. Mais ce modèle est alors surparamétré (cf Chapitre 5). Pour le rendre régulier, il faut imposer une contrainte entre les paramètres. Généralement, on considère le modèle (7.2) sous la contrainte $\sum_{i=1}^{I} n_i \alpha_i = 0$ (on l'appellera par la suite la contrainte "naturelle") car elle rend le modèle orthogonal. Attention sous R, la commande "par défaut" `an2<-lm(Notes~Exam)` correspond à la modélisation (7.2) sous la contrainte $\alpha_1 = 0$.

Exercice 18. *Quel est le lien entre les paramètres des trois modélisations* (7.1), (7.2) *sous la contrainte naturelle et* (7.2) *sous la contrainte* $\alpha_1 = 0$ *?*

7.2.3 Estimation

Proposition 7.1.

1. *Dans la modélisation* (7.1)*, les* m_i *sont estimés par*

$$\widehat{m_i} = Y_{i.} := \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

On les appelle les ***effets principaux des facteurs****. Leur variance vaut* σ^2/n_i*.*

2. *Dans la modélisation* (7.2) *sous la contrainte "naturelle",* μ *et les* α_i *sont estimés par :*

$$\widehat{\mu} = Y_{..} := \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} \text{ et } \widehat{\alpha_i} = Y_{i.} - Y_{..}$$

3. *Dans la modélisation* (7.2) *sous la contrainte* $\alpha_1 = 0$*,* μ *et les* α_i *sont estimés par :*

$$\widehat{\mu} = Y_{1.} \text{ et } \widehat{\alpha_i} = Y_{i.} - Y_{1.}, \forall i > 1.$$

Les valeurs ajustées $\widehat{Y_{ij}}$ *dans la cellule* i *sont constantes et sont égales à la moyenne* $Y_{i.}$ *des observations dans la cellule* i *:*

$$\widehat{Y_{ij}} = Y_{i.},$$

dont on déduit les résidus :

$$\widehat{\varepsilon_{ij}} = Y_{ij} - \widehat{Y_{ij}}.$$

L'estimateur de σ^2 *est donné par :*

$$\widehat{\sigma^2} = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - Y_{i.})^2.$$

Exercice 19. *Preuve de la proposition 7.1*

1. *La modélisation* (7.1) *étant régulière, vous pouvez utiliser la formule générale* $\widehat{\theta} = (X'X)^{-1}X'Y$ *ou minimiser la fonction des moindres carrés*

$$h(m_1, \ldots, m_I) = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - m_i)^2$$

2. *Pour la modélisation* (7.2) *sous la contrainte "naturelle", vous pouvez minimiser la fonction des moindres carrés*

$$h(\mu, \alpha_1, \ldots, \alpha_I) = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \mu - \alpha_i)^2$$

sous la contrainte $\sum_{i=1}^{I} n_i\alpha_i = 0$*.*

3. *Pour la modélisation* (7.2) *sous la contrainte* $\alpha_1 = 0$*, faire comme dans la question précédente en adaptant la contrainte.*

Pour notre exemple, les résultats obtenus avec R sont reportés en Figure 7.2.

```
> an1<-lm(Notes~Exam -1)
> summary(an1)

Call:
lm(formula = Notes ~ Exam - 1)

Residuals:
   Min     1Q Median     3Q    Max
    -5     -1      0      2      3

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
ExamA  12.0000     0.9526   12.60 2.30e-10 ***
ExamB  13.0000     0.8250   15.76 5.63e-12 ***
ExamC  14.0000     0.8819   15.88 4.98e-12 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.333 on 18 degrees of freedom
Multiple R-squared:  0.9734,        Adjusted R-squared:  0.969
F-statistic: 219.7 on 3 and 18 DF,  p-value: 2.311e-14

> an2<-lm(Notes~Exam)
> summary(an2)

Call:
lm(formula = Notes ~ Exam)

Residuals:
   Min     1Q Median     3Q    Max
    -5     -1      0      2      3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.0000     0.9526  12.597  2.3e-10 ***
ExamB         1.0000     1.2601   0.794    0.438
ExamC         2.0000     1.2981   1.541    0.141
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.333 on 18 degrees of freedom
Multiple R-squared:  0.1167,        Adjusted R-squared:  0.0186
F-statistic:  1.19 on 2 and 18 DF,  p-value: 0.3272
```

Figure 7.2 – Résultats pour l'ANOVA à un facteur

7.2.4 Propriétés

On a les propriétés suivantes analogues à celles de la régression linéaire :

Proposition 7.2.

- *La moyenne des résidus par cellule est nulle :* $\forall i = 1, \cdots, I$, $\frac{1}{n_i}\sum_{j=1}^{n_i} \widehat{\varepsilon}_{ij} = 0$.
- *La moyenne générale des résidus est nulle :* $\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} \widehat{\varepsilon}_{ij} = 0$.
- *La moyenne des valeurs ajustées est égale à la moyenne des valeurs observées :* $\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} \widehat{Y_{ij}} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}$.
- $cov(\widehat{\varepsilon}, \widehat{Y}) = 0$.
- $var(Y) = var(\widehat{Y}) + var(\widehat{\varepsilon})$.

Exercice 20. *Démontrez la proposition 7.2. Vous pouvez vous aider de la preuve de la proposition 6.1.*

La dernière propriété nous amène à définir les quantités suivantes :

- On appelle **variance inter-groupe** la quantité $var(\widehat{Y})$:

$$var(\widehat{Y}) = \frac{1}{n}\sum_{i=1}^{I} n_i \left(Y_{i.} - Y_{..}\right)^2,$$

  variance des moyennes par cellule, pondérées par les poids des cellules n_i/n.
- On appelle **variance intra-groupe**, ou variance résiduelle, la quantité $var(\widehat{\varepsilon})$:

$$var(\widehat{\varepsilon}) = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - Y_{i.})^2 = \frac{1}{n}\sum_{i=1}^{I} n_i var_i(Y)$$

  où $var_i(Y)$ est la variance des valeurs observées dans le niveau i :
  $var_i(Y) = \frac{1}{n_i}\sum_{j=1}^{n_i}(Y_{ij} - Y_{i.})^2$.
  Par conséquent, $var(\widehat{\varepsilon})$ est la moyenne des variances des observations dans les cellules.

La relation $var(Y) = var(\widehat{Y}) + var(\widehat{\varepsilon})$ s'écrit ici :

Variance totale = Variance inter + Variance intra.

On définit également le coefficient R^2 comme le rapport de la variance inter-groupe sur la variance totale :

$$R^2 = \frac{var(\widehat{Y})}{var(Y)} = 1 - \frac{var(\widehat{\varepsilon})}{var(Y)}.$$

On l'appelle rapport de corrélation empirique entre la variable quantitative Y et le facteur considéré. C'est une mesure de liaison entre une variable quantitative et une variable qualitative. On peut mentionner les deux cas particuliers suivants :

1. $R^2 = 1 \leftrightarrow \widehat{\varepsilon} = 0_n \leftrightarrow \forall j = 1, \cdots, n_i, Y_{ij} = Y_{i.}$ i.e. Y est constant dans chaque cellule.
2. $R^2 = 0 \leftrightarrow var(\widehat{Y}) = 0 \leftrightarrow \forall i = 1, \cdots, I, Y_{i.} = Y_{..}$, i.e. la moyenne de Y est la même dans chaque cellule.

7.2.5 Intervalle de confiance et test sur l'effet facteur

7.2.5.1 Intervalle de confiance pour les m_i

Dans le cadre général du modèle gaussien, on a montré que les estimateurs des paramètres du modèle sont distribués selon une loi gaussienne. Cette propriété peut s'appliquer au modèle à un facteur pour lequel on a posé l'hypothèse de normalité et d'indépendance des erreurs. Pour construire un intervalle de confiance pour les m_i, il suffit donc de construire un intervalle de confiance de Student en utilisant que

$$\widehat{m_i} \sim \mathcal{N}(m_i, \sigma^2/n_i) \text{ et } (n-I)\widehat{\sigma}^2 \sim \sigma^2\chi^2(n-I).$$

On obtient donc

$$IC_{1-\alpha}(m_i) = \left[\widehat{m_i} \pm t_{n-I,1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2}{n_i}} \right].$$

Sous R, on les obtient de la manière suivante :

```
> an1<-lm(Notes~Exam -1)
> confint(an1)
          2.5 %   97.5 %
ExamA  9.998705 14.00129
ExamB 11.266828 14.73317
ExamC 12.147161 15.85284
```

Exercice 21. *Construisez les intervalles de confiance donnés par les commandes suivantes :*

```
> an2<-lm(Notes~Exam)
> confint(an2)
                2.5 %    97.5 %
(Intercept)  9.9987051 14.001295
ExamB       -1.6474644  3.647464
ExamC       -0.7273053  4.727305
```

7.2.5.2 Test d'effet du facteur

On peut étudier l'effet du facteur sur la variable Y en posant l'hypothèse d'égalité de tous les paramètres du modèle :

$$\mathcal{H}_0 : m_1 = m_2 = \cdots = m_I = m \Longleftrightarrow \forall i = 1, \cdots, I,\ \alpha_i = 0$$

versus

$$\mathcal{H}_1 : \exists (i, i') \text{ tel que } m_i \neq m_{i'}.$$

Sous $\mathcal{H}_0$, tous les paramètres m_i sont égaux et le modèle s'écrit :

$$Y_{ij} = m + \varepsilon_{ij} \text{ avec } \widehat{m} = Y_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}.$$

On teste l'hypothèse d'égalité des paramètres m_i du modèle à partir de la statistique de Fisher :

$$F = \frac{\displaystyle\sum_{i=1}^{I} n_i(Y_{i\cdot} - Y_{\cdot\cdot})^2/(I-1)}{\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - Y_{i\cdot})^2/(n-I)} = \frac{SCE/(I-1)}{SCR/(n-I)} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(I-1, n-I),$$

où SCE désigne la sommes des carrés inter-groupes et SCR est la somme des carrés intra-groupes. On rejette $\mathcal{H}_0$ si $F > f_{1-\alpha,I-1,n-I}$.

```
> anmequal<-lm(Notes~1)
> anova(anmequal,an1)
Analysis of Variance Table

Model 1: Notes ~ 1
Model 2: Notes ~ Exam - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     20 110.95
2     18  98.00  2    12.952 1.1895 0.3272
```

Exercice 22. *Dans les deux sorties suivantes, quelles sont les hypothèses testées ? Construisez le test de Fisher associé. Notez bien la différence entre les deux procédures.*

```
> anova(an1)
Analysis of Variance Table

Response: Notes
          Df Sum Sq Mean Sq F value    Pr(>F)
Exam       3   3588 1196.00  219.67 2.311e-14 ***
Residuals 18     98    5.44

> anova(an2)
Analysis of Variance Table

Response: Notes
          Df Sum Sq Mean Sq F value Pr(>F)
Exam       2 12.952  6.4762  1.1895 0.3272
Residuals 18 98.000  5.4444
```

7.2.5.3 Tableau d'analyse de la variance à un facteur

Toutes ces estimations peuvent être présentées sous la forme d'un tableau d'analyse de la variance à un facteur :

Source	ddl	Somme des Carrés	Moyenne des Carrés	F	$f_{1-\alpha}$
Facteur	$I-1$	$SCE = \sum_{i=1}^{I} n_i (Y_{i.} - Y_{..})^2$	$\frac{SCE}{I-1} = MCE$	$\frac{MCE}{\widehat{\sigma}^2}$	$f_{1-\alpha,I-1,n-I}$
Résiduel	$n-I$	$SCR = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i.})^2$	$\frac{SCR}{n-I} = \widehat{\sigma}^2$		
Total	$n-1$	$SCT = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - Y_{..})^2$			

7.3 Analyse de variance à deux facteurs

7.3.1 Notations et exemple

Soit Y la variable réponse quantitative que l'on veut expliquer ici par rapport à deux variables qualitatives, i.e deux facteurs. Le premier facteur (**facteur ligne**), dit "A", admet I niveaux, le deuxième (**facteur colonne**), dit "B", admet J niveaux. On considère le tableau croisé entre les I modalités du facteur A et les J modalités du facteur B. On appelle **cellule** une case du tableau. Par la suite, on note :

$i = 1, \cdots I$	les indices des niveaux du facteur ligne A
$j = 1, \cdots J$	les indices des niveaux du facteur colonne B
n_{ij}	le nombre d'observations pour le niveau i du facteur A et pour le niveau j du facteur B (nombre d'observations dans la cellule (i,j))
$\ell = 1, \cdots, n_{ij}$	les indices des observations de la cellule (i,j)
$Y_{ij\ell}$	la ℓ-ième observation dans la cellule (i,j)
$Y_{ij.}$	la moyenne des observations dans la cellule (i,j) : $Y_{ij.} = \frac{1}{n_{ij}} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell}$.

On utilisera également les notations suivantes :

$$Y_{i..} = \frac{1}{n_{i+}} \sum_{j=1}^{J} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell} \text{ avec } n_{i+} = \sum_{j=1}^{J} n_{ij}$$

$$Y_{.j.} = \frac{1}{n_{+j}} \sum_{i=1}^{I} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell} \text{ avec } n_{+j} = \sum_{i=1}^{I} n_{ij}$$

$$Y_{...} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell} \text{ avec } n = \sum_{i=1}^{I} n_{i+} = \sum_{j=1}^{J} n_{+j}$$

Exemple :

Cet exemple est issu du livre de Husson et Pagès (2013). Au cours d'une étude sur les facteurs influençant le rendement de blé, on a comparé trois variétés de blé (facteur

B de modalités L, N et NF) et deux apports d'azote (facteur A apport normal, dose 1 ; apport intensif, dose 2). Trois répétitions pour chaque couple (variété, dose) ont été effectuées ($n_{ij} = 3$, plan équilibré) et le rendement (en q/ha) a été mesuré ($Y_{ij\ell}$). On s'intéresse aux différences qui pourraient exister d'une variété à l'autre, et aux interactions éventuelles des variétés avec les apports azotés.

```
> summary(Ble)
 Dose  Variete   Rendement
 1:9   L :6    Min.   :55.65
 2:9   N :6    1st Qu.:62.82
       NF:6    Median :65.50
               Mean   :66.58
               3rd Qu.:69.75
               Max.   :79.83
```

7.3.2 Modélisation

7.3.2.1 ANOVA à deux facteurs croisés

Le modèle général à deux facteurs croisés s'écrit sous la forme :

$$Y_{ij\ell} = m_{ij} + \varepsilon_{ij\ell} \text{ avec } i = 1, \cdots, I,\ j = 1, \cdots, J,\ \ell = 1, \cdots, n_{ij} \tag{7.3}$$

où $\varepsilon_{ij\ell} \sim \mathcal{N}(0, \sigma^2)$, n variables aléatoires indépendantes.

Cette paramétrisation ne permet pas de distinguer les effets de chaque facteur et de leur interaction. On considère donc la paramétrisation centrée qui décompose m_{ij} par rapport à un effet moyen général et permet de mesurer des "effets séparés" des deux facteurs et les "effets conjoints". Le modèle complet s'écrit sous la forme :

$$Y_{ij\ell} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\ell}. \tag{7.4}$$

Mais on se retrouve avec un modèle défini avec $1 + I + J + IJ$ paramètres. On doit donc introduire $1 + I + J$ contraintes pour estimer ces paramètres. On a vu dans le chapitre 5 qu'il est intéressant de considérer des contraintes dans le cadre d'un dispositif orthogonal. Dans le cas de l'analyse de la variance à deux facteurs croisés, le dispositif orthogonal est caractérisé par la propriété suivante :

Proposition 7.3. *Dans le modèle d'analyse de variance à deux facteurs croisés il existe des contraintes qui rendent la partition μ, α, β, γ orthogonale si et seulement si*

$$n_{ij} = \frac{n_{i+} n_{+j}}{n}. \tag{7.5}$$

Dans ce cas, les contraintes (dites de type I) sont

$$\sum_{i=1}^{I} n_{i+}\alpha_i = 0;\ \sum_{j=1}^{J} n_{+j}\beta_j = 0;\ \forall i,\ \sum_{j=1}^{J} n_{ij}\gamma_{ij} = 0;\ \forall j,\ \sum_{i=1}^{I} n_{ij}\gamma_{ij} = 0. \tag{7.6}$$

Exercice 23. *Travaillez la preuve de la proposition 7.3 donnée en annexe B.2.*

En pratique, les contraintes utilisées sont souvent celles dites de type III :

$$\sum_i \alpha_i = 0, \sum_j \beta_j = 0, \forall i, \sum_j \gamma_{ij} = 0 \text{ et } \forall j, \sum_j \gamma_{ij} = 0$$

Avec ce système de contraintes, il n'y a possibilité d'orthogonalité que si le modèle est **équilibré**, i.e. $n_{ij} = cte$, d'après la proposition 7.3. Attention, les contraintes 7.6 ne sont pas les contraintes par défaut sous R (cf `model.matrix(Rendement~Dose * Variete )`). Sous R, les contraintes sont $\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0$ mais on peut les modifier (cf section suivante).

Les IJ paramètres m_{ij} sont donc redéfinis en fonction de :
- μ un paramètre de centrage général,
- α_i, $I-1$ paramètres qui caractérisent les effets principaux du facteur A,
- β_j, $J-1$ paramètres qui caractérisent les effets principaux du facteur B,
- γ_{ij}, $(I-1)(J-1)$ paramètres qui prennent en compte les effets d'interaction.

Dans la suite, on va se placer dans le cadre d'un **dispositif orthogonal**.

7.3.2.2 ANOVA à deux facteurs additifs

Le modèle d'ANOVA à deux facteurs **additif** est un modèle où on suppose qu'il n'y a pas d'effet d'interaction entre les deux facteurs. Le modèle s'écrit donc sous la forme

$$Y_{ij\ell} = \mu + \alpha_i + \beta_j + \varepsilon_{ij\ell}. \tag{7.7}$$

Le modèle additif est un sous-modèle du modèle complet avec interaction.

Exercice 24. *A l'aide de l'exercice 23, déterminez sous quelle condition il existe des contraintes rendant le modèle additif ci-dessus orthogonal.*

7.3.3 Estimation des paramètres

Proposition 7.4. *Dans le cadre de la paramétrisation générale $Y_{ij\ell} = m_{ij} + \varepsilon_{ij\ell}$, m_{ij} est estimé par*

$$\widehat{m}_{ij} = \frac{1}{n_{ij}} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell} = Y_{ij.} \sim \mathcal{N}\left(m_{ij}, \frac{\sigma^2}{n_{ij}}\right).$$

Les valeurs ajustées et les résidus sont alors donnés par :

$$\widehat{Y}_{ij\ell} = \widehat{m}_{ij} = Y_{ij.} \text{ et } \widehat{\varepsilon}_{ij\ell} = Y_{ij\ell} - Y_{ij.}.$$

La variance est estimée par

$$\widehat{\sigma}^2 = \frac{1}{n - IJ} \sum_{ij\ell} (\widehat{\varepsilon}_{ij\ell})^2 = \frac{1}{n - IJ} \sum_{ij\ell} (Y_{ij\ell} - Y_{ij.})^2.$$

Exercice 25. *Démontrez cette proposition en utilisant les propriétés du modèle linéaire régulier.*

Proposition 7.5. *Les paramètres du modèle complet d'équation* (7.4) *sous les contraintes* (7.6) *sont estimés par*

$$\begin{cases} \widehat{\mu} = Y_{...} \\ \widehat{\alpha}_i = Y_{i..} - Y_{...} \\ \widehat{\beta}_j = Y_{.j.} - Y_{...} \\ \widehat{\gamma}_{ij} = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...} \end{cases}$$

Exercice 26. *Pour démontrer cette proposition, vous pouvez minimiser la fonction des moindres carrés sous les contraintes* (7.6).

Pour notre exemple, les résultats obtenus avec R sont reportés en Figure 7.3.

7.3.4 Décomposition de la variabilité

Comme dans l'analyse de variance à un facteur, la variabilité totale de Y se décompose en une variabilité inter-cellule expliquée par le modèle (notée SCE) et une variabilité intra-cellule non expliquée par le modèle (notée SCR) :

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{\ell=1}^{n_{ij}}(Y_{ij\ell} - Y_{...})^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(Y_{ij.} - Y_{...})^2}_{\text{SCE}} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} var_{ij}(Y)}_{\text{SCR}}$$

avec $var_{ij}(Y) = \frac{1}{n_{ij}}\sum_{\ell=1}^{n_{ij}}(Y_{ij\ell} - Y_{ij.})^2$.

Dans le cas du modèle à deux facteurs croisés, la variance inter-cellule SCE peut être décomposée en une variance expliquée par le premier facteur, une variance expliquée par le second facteur et une variance expliquée par les interactions entre les deux facteurs. Dans le cas d'un plan orthogonal à deux facteurs, on définit les quantités suivantes :

- SCA, la somme des carrés corrigés de l'effet différentiel du facteur A :

$$SCA = \sum_{i=1}^{I} n_{i+}(Y_{i..} - Y_{...})^2 = \sum_{i=1}^{I} n_{i+}(\widehat{\alpha}_i)^2.$$

- SCB, la somme des carrés corrigés de l'effet différentiel du facteur B :

$$SCB = \sum_{j=1}^{J} n_{+j}(Y_{.j.} - Y_{...})^2 = \sum_{j=1}^{J} n_{+j}(\widehat{\beta}_j)^2.$$

- SCI, la somme des carrés corrigés de l'effet d'interaction entre les deux facteurs :

$$SCI = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(\widehat{\gamma}_{ij})^2.$$

On peut montrer que :

$$SCE = SCA + SCB + SCI.$$

```
> aov2C<-lm(Rendement~Dose * Variete ,data=Ble)
> summary(aov2C)

Call:
lm(formula = Rendement ~ Dose * Variete, data = Ble)

Residuals:
   Min     1Q Median     3Q    Max
-7.667 -2.296 -0.325  2.623  8.573

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       71.257      2.536  28.101 2.55e-12 ***
Dose2              2.500      3.586   0.697  0.49899
VarieteN         -12.223      3.586  -3.409  0.00519 **
VarieteNF         -4.453      3.586  -1.242  0.23801
Dose2:VarieteN    -0.200      5.071  -0.039  0.96919
Dose2:VarieteNF   -2.007      5.071  -0.396  0.69928
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 4.392 on 12 degrees of freedom
Multiple R-squared:  0.6725,        Adjusted R-squared:  0.536
F-statistic: 4.928 on 5 and 12 DF,  p-value: 0.01105


> aov2Cbis<-lm(Rendement~C(Dose,sum) + C(Variete,sum) +  C(Dose,sum):C(Variete,sum),data=Ble)
> #lm(Rendement ~ Dose * Variete,data=Ble,contrasts=list(Dose="contr.sum",Variete="contr.sum"))
> summary(aov2Cbis)

Call:
lm(formula = Rendement ~ C(Dose, sum) + C(Variete, sum) + C(Dose,
    sum):C(Variete, sum), data = Ble)

Residuals:
   Min     1Q Median     3Q    Max
-7.667 -2.296 -0.325  2.623  8.573

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    66.5800     1.0352  64.316  < 2e-16 ***
C(Dose, sum)1                  -0.8822     1.0352  -0.852 0.410775
C(Variete, sum)1                5.9267     1.4640   4.048 0.001615 **
C(Variete, sum)2               -6.3967     1.4640  -4.369 0.000913 ***
C(Dose, sum)1:C(Variete, sum)1 -0.3678     1.4640  -0.251 0.805897
C(Dose, sum)1:C(Variete, sum)2 -0.2678     1.4640  -0.183 0.857923
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 4.392 on 12 degrees of freedom
Multiple R-squared:  0.6725,        Adjusted R-squared:  0.536
F-statistic: 4.928 on 5 and 12 DF,  p-value: 0.01105
```

Figure 7.3 – Résultats pour l'étude de l'ANOVA à deux facteurs

7.3.5 Le diagramme d'interactions

Le diagramme d'interactions permet de visualiser graphiquement la présence ou l'absence d'interactions. Pour chaque j fixé, on représente dans un repère orthogonal les points (i, j) de coordonnées $(i, \widehat{m}_{ij} = Y_{ij.})$. Puis on trace les segments joignants les couples de points $((i-1, j), (i, j))$. On obtient ainsi pour chaque j fixé une ligne brisée.

Proposition 7.6. *Si l'hypothèse de non-interaction est vraie, alors les lignes brisées dans le diagramme d'interaction sont parallèles.*

Preuve : La ligne brisée associée au niveau j joint les points $(1, \widehat{m}_{1j}), (2, \widehat{m}_{2j}), \cdots, (I, \widehat{m}_{Ij})$. S'il n'y a pas d'interaction, alors ces points ont pour coordonnées $(1, \widehat{\alpha}_1 + \widehat{\beta}_j), (2, \widehat{\alpha}_2 + \widehat{\beta}_j), \cdots, (I, \widehat{\alpha}_I + \widehat{\beta}_j)$. Par conséquent, les lignes brisées associées aux niveaux j et j' se correspondent par une translation verticale d'amplitude $\widehat{\beta}_j - \widehat{\beta}_{j'}$.

On lit sur ce graphique l'effet principal des modalités j (le niveau moyen d'une ligne brisée), l'effet principal des modalités i (la moyenne des ordonnées des points à abscisse fixée). En ce qui concerne les interactions, on obtiendra rarement des lignes brisées strictement parallèles. Le problème sera alors de savoir si leur non-parallélisme traduit une interaction significative. Un test est donc nécessaire.

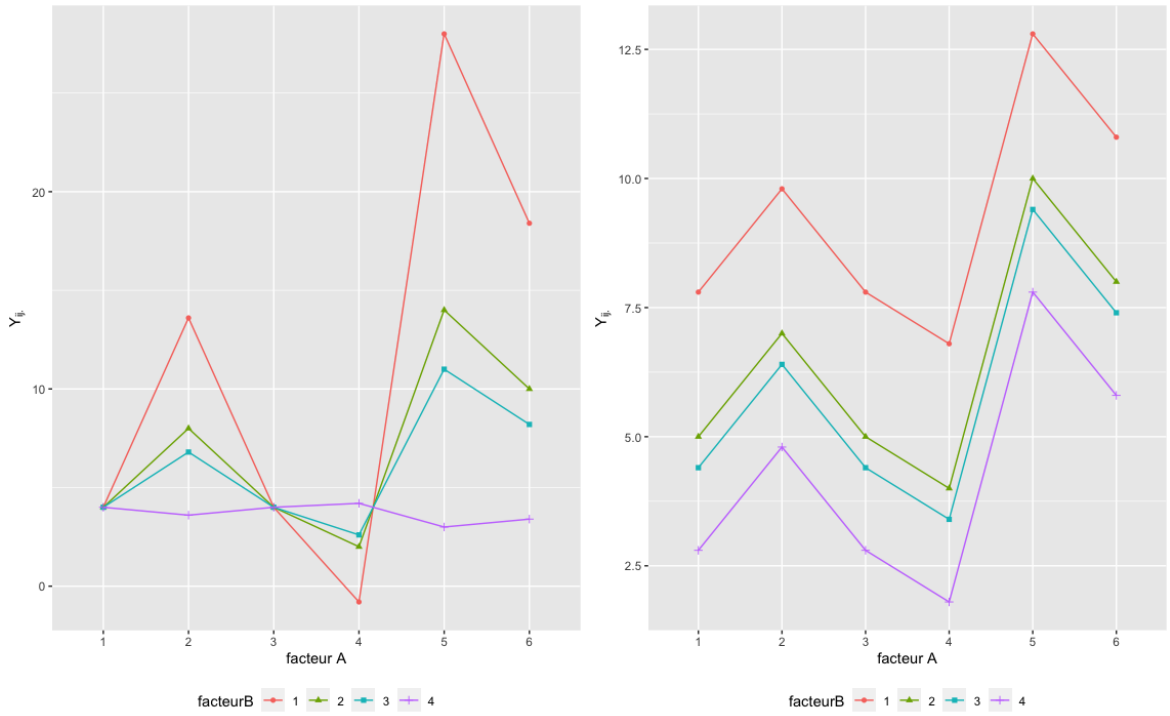


FIGURE 7.4 – Moyennes de la variable Y pour chaque niveau d'un facteur en fonction des niveaux de l'autre facteur : avec interaction à gauche ; sans interaction à droite.

La Figure 7.4 illustre les comportements des moyennes des cellules de modèles avec ou sans interaction (additif). Chaque ligne est appelée un **profil**, et la présence d'interactions se caractérise par le croisement de ces profils tandis que le parallélisme indique l'absence d'interactions. La question est évidemment de tester si des croisements observés sont jugés significatifs.

Attention, un manque de parallélisme peut aussi être dû à la présence d'une relation non-linéaire entre la variable Y et l'un des facteurs.

7.3.6 Tests d'hypothèses

Trois hypothèses sont couramment considérées :

- L'hypothèse d'absence d'interactions entre les deux facteurs ou hypothèse d'additivité des 2 facteurs :

$$\mathcal{H}_I : \forall i = 1, \cdots, I, \forall j = 1, \cdots, J, \gamma_{ij} = 0.$$

  Cette hypothèse impose $(I-1)(J-1)$ contraintes.
- L'hypothèse d'absence d'effet du facteur A :

$$\mathcal{H}_A : \forall i = 1, \cdots, I, \alpha_i = 0.$$

  Cette hypothèse impose $(I-1)$ contraintes.
- L'hypothèse d'absence d'effet du facteur B :

$$\mathcal{H}_B : \forall j = 1, \cdots, J, \beta_j = 0.$$

  Cette hypothèse impose $(J-1)$ contraintes.

Une remarque très importante porte sur la démarche de ces tests d'hypothèses. **S'il existe des interactions entre les deux facteurs, alors les deux facteurs qui constituent cette interaction doivent impérativement être introduits dans le modèle ; dans ce cas, il est donc inutile de tester l'effet de chacun des deux facteurs**. En effet, la présence d'interactions entre les deux facteurs signifie qu'il y a un effet combiné des deux facteurs et donc un effet de chaque facteur.

Exercice 27.

- **Test de non-interaction entre les deux facteurs** *:*
  Ce test consiste à comparer le modèle complet avec interactions (7.4) *et le modèle additif* (7.7)*. Montrez que la statistique de Fisher pour ce test vaut :*

$$F = \frac{SCI/(I-1)(J-1)}{SCR/(n-IJ)} \underset{\mathcal{H}_0}{\sim} \mathcal{F}((I-1)(J-1), n-IJ).$$

- **Test d'absence d'effet du facteur** A *:*
  Ce test est intéressant que si le test précédent a permis de montrer l'absence d'interaction. En effet, si les termes d'interaction sont introduits dans le modèle,

les facteurs qui constituent cette interaction doivent également apparaître dans le modèle. Pour étudier l'effet du facteur A, on compare le modèle additif

$$Y_{ij\ell} = \mu + \alpha_i + \beta_j + \varepsilon_{ij\ell}$$

et le modèle à un facteur B à J paramètres

$$Y_{ij\ell} = \mu + \beta_j + \varepsilon_{ij\ell}.$$

Montrez que le test est basé sur la statistique de Fisher suivante :

$$F = \frac{SCA/(I-1)}{SCRAB/(n-(I+J-1))} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(I-1, n-(I+J-1)),$$

où SCA désigne la somme des carrés du facteur A et $SCRAB$ correspond à la somme des carrés des résidus du modèle additif.

- **Test d'absence d'effet du facteur B :**
  On compare le modèle additif au modèle à un facteur A à I paramètres :

$$Y_{ij\ell} = \mu + \alpha_i + \varepsilon_{ij\ell}.$$

Montrez que le test est basé sur la statistique de Fisher suivante :

$$F = \frac{(SCB)/(J-1)}{SCRAB/(n-(I+J-1))} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(J-1, n-(I+J-1)),$$

où SCB désigne la somme des carrés du facteur B et $SCRAB$ correspond à la somme des carrés des résidus du modèle additif.

Pour chacun de ces tests, étudiez les résultats obtenus dans notre exemple.

```
> aov2SI<-lm(Rendement~Variete + Dose,data=Ble)
> anova(aov2SI,aov2Cbis)
Analysis of Variance Table

Model 1: Rendement ~ Variete + Dose
Model 2: Rendement ~ C(Dose, sum) + C(Variete, sum) + C(Dose, sum):C(Variete, sum)
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     14 235.14
2     12 231.47  2    3.6654 0.095   0.91

> aovVariete<-lm(Rendement~Variete,data=Ble)
> anova(aovVariete,aov2SI)
Analysis of Variance Table

Model 1: Rendement ~ Variete
Model 2: Rendement ~ Variete + Dose
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     15 249.15
2     14 235.14  1     14.01 0.8341 0.3765

> aovDose<-lm(Rendement~Dose,data=Ble)
> anova(aovDose,aov2SI)
Analysis of Variance Table

Model 1: Rendement ~ Dose
Model 2: Rendement ~ Variete + Dose
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     16 692.72
2     14 235.14  2    457.58 13.622 0.0005192 ***
```

7.3.7 Tableau d'analyse de variance à deux facteurs croisés dans le cas d'un plan orthogonal

Dans le cas du modèle à deux facteurs croisés avec dispositif orthogonal, on rappelle que la variabilité totale peut être décomposée en

$$SCT = SCE + SCR = SCA + SCB + SCI + SCR.$$

On peut ainsi dresser le tableau d'analyse de variance d'un plan orthogonal à deux facteurs croisés :

Source	ddl	Somme des Carrés	Moyenne des Carrés	F	$f_{1-\alpha}$
Ligne	$I-1$	SCA	$MCA = SCA/(I-1)$	$MCA/\widehat{\sigma}^2$	$f_{1-\alpha,I-1,n-IJ}$
Colonne	$J-1$	SCB	$MCB = SCB/(J-1)$	$MCB/\widehat{\sigma}^2$	$f_{1-\alpha,J-1,n-IJ}$
Interactions	$(I-1)(J-1)$	SCI	$MCI = SCI/(I-1)(J-1)$	$MCI/\widehat{\sigma}^2$	$f_{1-\alpha,(I-1)(J-1),n-IJ}$
Résiduel	$n-IJ$	SCR	$SCR/(n-IJ) = \widehat{\sigma}^2$		
Total	$n-1$	SCT			

En résumé :

- Savoir écrire un modèle d'ANOVA à un et deux facteurs (individuellement et matriciellement), régulier et singulier
- Savoir distinguer un modèle régulier d'un modèle singulier
- Savoir estimer les paramètres du modèle d'ANOVA dans le cas régulier et dans le cas singulier (en s'adaptant à la / les contrainte(s) choisie(s))
- Savoir construire un intervalle de confiance pour un paramètre du modèle d'ANOVA
- Savoir construire un test pour tester l'effet d'un facteur, l'effet d'interaction entre facteurs, ... et savoir organiser ces différents tests
- Savoir interpréter un diagramme d'interaction
- Savoir manipuler SCA, SCB, SCI, SCE, SCR dans le cas d'un plan orthogonal.

Chapitre 8

Analyse de covariance (ANCOVA)

8.1 Les données

Dans ce chapitre, nous allons présenter le modèle d'analyse de covariance (ANCOVA) seulement dans le cadre simple où on observe deux variables quantitatives z et Y, et une variable qualitative T sur un échantillon de n individus : la variable quantitative Y est la variable réponse que l'on cherche à expliquer en fonction de la variable quantitative z (appelée **covariable**) et du facteur T à I niveaux. Les notions peuvent être généralisées pour plusieurs covariables et plusieurs facteurs.

Chaque individu de l'échantillon est repéré par un double indice (i, j), i représente le niveau du facteur T auquel appartient l'individu et j correspond à l'indice de l'individu dans le niveau i. Pour chaque individu (i, j), on dispose d'une valeur z_{ij} de la variable z et d'une valeur Y_{ij} de la variable Y. Pour chaque niveau i de T (avec $i = 1, \cdots, I$), on observe n_i valeurs $z_{(i)} = (z_{i1}, \cdots, z_{in_i})'$ de z et n_i valeurs $Y_{(i)} = (Y_{i1}, \cdots, Y_{in_i})'$ de Y. Au final $n = \sum_{i=1}^{I} n_i$ est le nombre d'observations disponibles.

Dans tout ce chapitre, nous allons illustrer les notions abordées à l'aide de l'exemple suivant : On cherche à savoir si des conditions de température et d'oxygénation influencent l'évolution du poids des huîtres. On dispose de $n = 20$ sacs de 10 huitres. On place, pendant un mois, ces 20 sacs de façon aléatoire dans $I = 5$ emplacements différents d'un canal de refroidissement d'une centrale électrique à raison de $n_i = 4$ sacs par emplacement. Ces emplacements se différencient par leurs températures et oxygénations. Pour chaque sac, on a

- son poids avant l'expérience (variable Pds Init),
- son poids après l'expérience (variable Pds Final),
- l'emplacement (variable Traitement) codé de 1 à 5.

Les données peuvent être représentées conjointement sur un même graphique permettant de visualiser la relation éventuelle entre Y, z et T. Il s'agit de tracer un nuage de points de coordonnées (z_{ij}, Y_{ij}), où tous les points du niveau i, $i = 1, \cdots, I$, sont représentés par le même symbole (Figure 8.1). On peut également tracer un boxplot du poids initial et du poids final pour chaque emplacement (Figure 8.2).

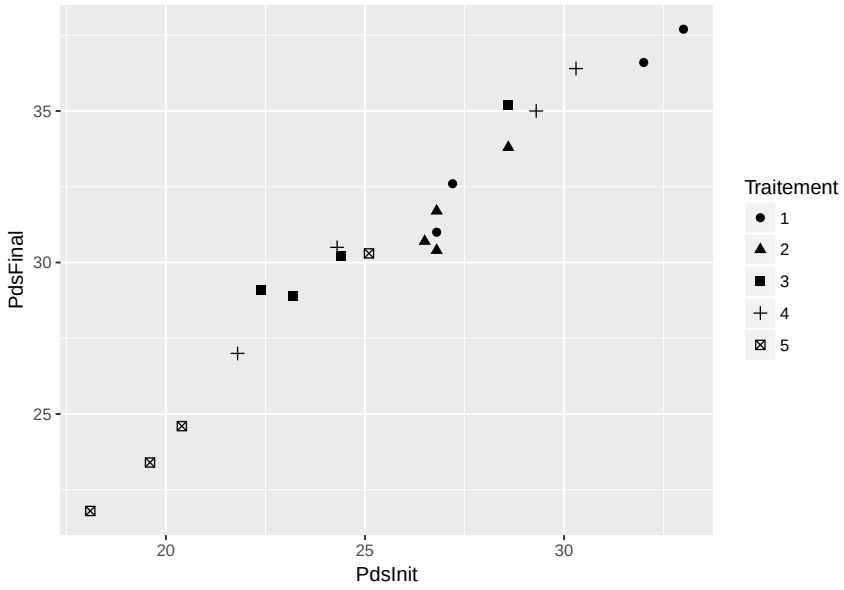


FIGURE 8.1 – Graphique des poids finaux par rapport aux poids initiaux selon chaque emplacement.

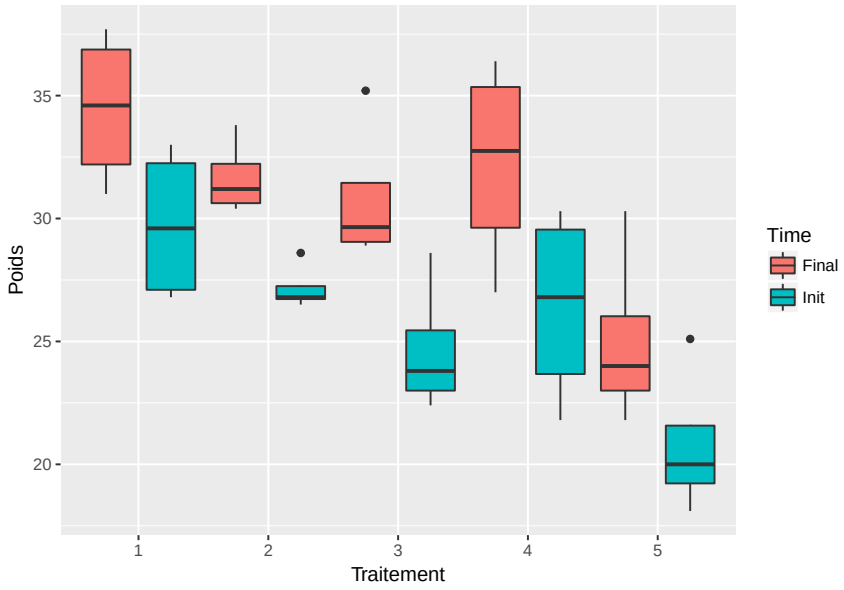


FIGURE 8.2 – Evolution des poids initiaux et poids finaux pour chaque traitement

8.2 Modélisation

8.2.1 Modélisation régulière

Dans le cadre d'une ANCOVA simple, le modèle régulier s'écrit sous la forme :

$$(MR) : Y_{ij} = a_i + b_i\ z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I,\ j = 1, \cdots, n_i$$

où $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, n variables indépendantes. Cela revient à estimer une droite de régression linéaire de Y sur z pour chaque niveau i du facteur T. Pour le niveau i, on estime les paramètres a_i, constantes à l'origine des droites de régression et b_i, pentes des droites de régression.

Matriciellement, le modèle s'écrit sous la forme

$$\underbrace{\begin{pmatrix} Y_{(1)} \\ \vdots \\ \vdots \\ Y_{(I)} \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} X_{(1)} & & & \\ & X_{(2)} & & \\ & & \ddots & \\ & & & X_{(I)} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} a_1 \\ b_1 \\ \vdots \\ a_I \\ b_I \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_{(1)} \\ \vdots \\ \vdots \\ \varepsilon_{(I)} \end{pmatrix}}_{\varepsilon}$$

avec

$$X_{(i)} = \begin{pmatrix} 1 & z_{i1} \\ \vdots & \vdots \\ 1 & z_{i,n_i} \end{pmatrix}$$

8.2.2 Modélisation singulière

Comme pour les modèles factoriels, il existe une reparamétrisation faisant apparaître des effets différentiels par rapport à un niveau de référence. Le modèle associé à cette nouvelle paramétrisation s'écrit :

$$(MS) : Y_{ij} = (\mu + \alpha_i) + (\beta + \gamma_i) z_{ij} + \varepsilon_{ij}, \quad \forall i = 1, \cdots, I,\ j = 1, \cdots, n_i.$$

Cette paramétrisation permet de faire apparaître :

- un effet d'interaction entre la covariable z et le facteur T : γ_i ;
- un effet différentiel du facteur T sur la variable Y : α_i ;
- un effet différentiel de la covariable z sur la variable Y : β.

Exercice 28. *Considérons le premier niveau comme référence dans cette paramétrisation (MS) pour rendre le modèle identifiable (ce qui est fait sous R par défaut). Les contraintes d'identifiabilité sont donc $\alpha_1 = \gamma_1 = 0$. Dans ce cas, donnez le lien entre les paramètres μ, α_i, β, γ_i et les paramètres a_i et b_i de la modélisation (MR). Donnez une interprétation "graphique" des paramètres μ, α_i, β et γ_i.*

8.3 Estimation des paramètres

Exercice 29. *Estimation pour le modèle régulier*
Dans le cas du modèle régulier (MR), on peut utiliser la formule générale $\hat{\theta} = (X'X)^{-1}X'Y$. *En utilisant le fait que la matrice* X *est diagonale par bloc,* $X = diag(X_{(1)}, \ldots, X_{(I)})$, *montrez que*

$$\hat{\theta} = \begin{pmatrix} (X'_{(1)}X_{(1)})^{-1}X'_{(1)}Y_{(1)} \\ \vdots \\ X'_{(I)}X_{(I)})^{-1}X'_{(I)}Y_{(I)} \end{pmatrix}$$

Déduisez-en les estimateurs $\widehat{b_i}$ *et* $\widehat{a_i}$. *Que remarquez-vous ?*

Dans notre exemple, on obtient les estimations suivantes :

```
      a1        b1        a2       b2       a3       b3       a4       b4         a5       b5
5.241259 0.9826468 -9.149322 1.501355 4.817959 1.056067 4.295756 1.056925 -0.4318298 1.223886
```

et on peut graphiquement observer l'ajustement des droites de régression aux données (Figure 8.3).

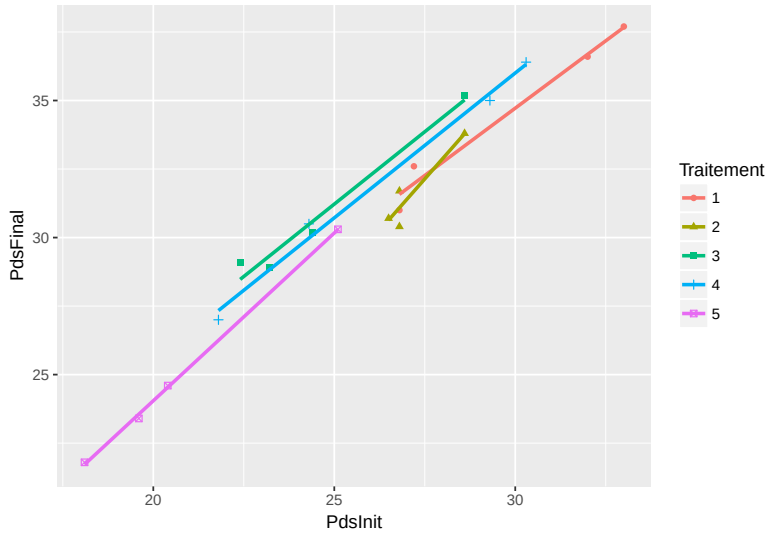


FIGURE 8.3 – Ajustement des droites de régression aux données

Exercice 30. *Estimation pour le modèle singulier*
On considère le modèle singulier (MS) avec les contraintes d'identifiabilité $\alpha_1 = \gamma_1 = 0$. *A l'aide des estimateurs du modèle (MR), déduisez-en les estimateurs pour ce modèle singulier (MS).*

Dans notre exemple, les résultats obtenus sous R sont reportés en Figure 8.4. On peut ensuite s'intéresser à construire des intervalles de confiance pour ces paramètres, faire des tests de nullité pour chacun des paramètres, ...

```
> complet<-lm(PdsFinal~PdsInit * Traitement)
> summary(complet)

Call:
lm(formula = PdsFinal ~ PdsInit * Traitement)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68699 -0.28193  0.02184  0.10425  0.63075

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.24126    2.86473   1.830   0.0972 .
PdsInit               0.98265    0.09588  10.249 1.27e-06 ***
Traitement2         -14.39058    9.15971  -1.571   0.1472
Traitement3          -0.42330    3.97747  -0.106   0.9174
Traitement4          -0.94550    3.50725  -0.270   0.7930
Traitement5          -5.67309    3.57150  -1.588   0.1433
PdsInit:Traitement2   0.51871    0.33406   1.553   0.1515
PdsInit:Traitement3   0.07342    0.14699   0.499   0.6282
PdsInit:Traitement4   0.07428    0.12229   0.607   0.5571
PdsInit:Traitement5   0.24124    0.13980   1.726   0.1151
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.5324 on 10 degrees of freedom
Multiple R-squared:  0.9921,        Adjusted R-squared:  0.985
F-statistic: 139.5 on 9 and 10 DF,  p-value: 2.572e-09
```

FIGURE 8.4 – Résultats du modèle (MS) pour le jeu de données des Huitres.

Exercice 31. *Dans la sortie de R en Figure 8.4, à quel test correspond la pvaleur* $1.27e-06$ *? Construisez le test statistique permettant de faire ce test au niveau* 5%.

8.4 Tests d'hypothèses

On peut tout d'abord commencer par tester l'absence de tout effet, aussi bien de la covariable z que du facteur T. Pour cela, on veut comparer le modèle "blanc"

$$(M0) : Y_{ij} = \mu + \varepsilon_{ij},\ \forall i = 1, \cdots, I,\ \forall j = 1, \cdots, n_i$$

contre le modèle complet (MS)

$$(MS) :\ Y_{ij} = \mu + \alpha_i + \beta z_{ij} + \gamma_i z_{ij} + \varepsilon_{ij},\ \forall i = 1, \cdots, I,\ \forall j = 1, \cdots, n_i.$$

Exercice 32. *Ecrivez le test de Fisher permettant de tester* $(M0)$ *contre* (MS).
A l'aide des résultats en Figure 8.9, répondez à ce test.

Le modèle $(M0)$ consiste à ajuster une droite horizontale aux données (Figure 8.5). Si on rejette le modèle (M0), on peut poursuivre l'étude mais il est important de suivre une démarche logique dans la mise en place des tests d'hypothèses, comme dans le cadre de l'analyse de la variance.

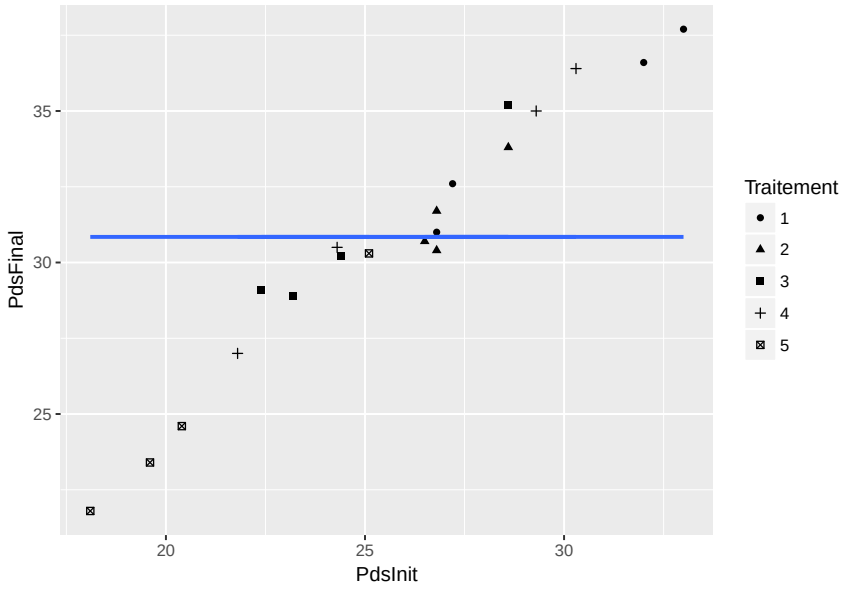


FIGURE 8.5 – Ajustement du modèle (M0) aux données

On commence par tester l'hypothèse de non-interaction entre le facteur T et la covariable z. On souhaite donc tester l'hypothèse nulle suivante :

$$\mathcal{H}_0^{(M1)} : b_1 = b_2 = \cdots = b_I \Longleftrightarrow \gamma_1 = \gamma_2 = \cdots = \gamma_I = 0.$$

Ce test revient à comparer le modèle complet :

$$(MS) : \ Y_{ij} = \mu + \alpha_i + (\beta + \gamma_i) z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i$$

avec le sous-modèle sans interaction :

$$(M1) : \ Y_{ij} = \mu + \alpha_i + \beta z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i.$$

Exercice 33. *Ecrivez le test de Fisher permettant de tester* $(M1)$ *contre* (MS)*. A l'aide des résultats en Figure 8.9, répondez à ce test.*

Le modèle $(M1)$ consiste graphiquement à ajuster des droites parallèles pour chaque traitement (Figure 8.6). Si on rejette l'hypothèse $\mathcal{H}_0^{(M1)}$, on conclut à la présence d'interactions dans le modèle. Il est alors inutile de tester l'absence d'effet du facteur T ou de la covariable z sur Y, car toute variable constituant une interaction doit apparaître dans le modèle.

En revanche, si le test montre que l'hypothèse $\mathcal{H}_0^{(M1)}$ est vraisemblable (i.e. les I droites de régression partagent la même pente de régression), on peut alors évaluer l'effet de la covariable z sur Y et celui du facteur T sur Y. On peut tester deux hypothèses en comparant le modèle sans interaction $(M1)$ à chacun des modèles réduits suivants :

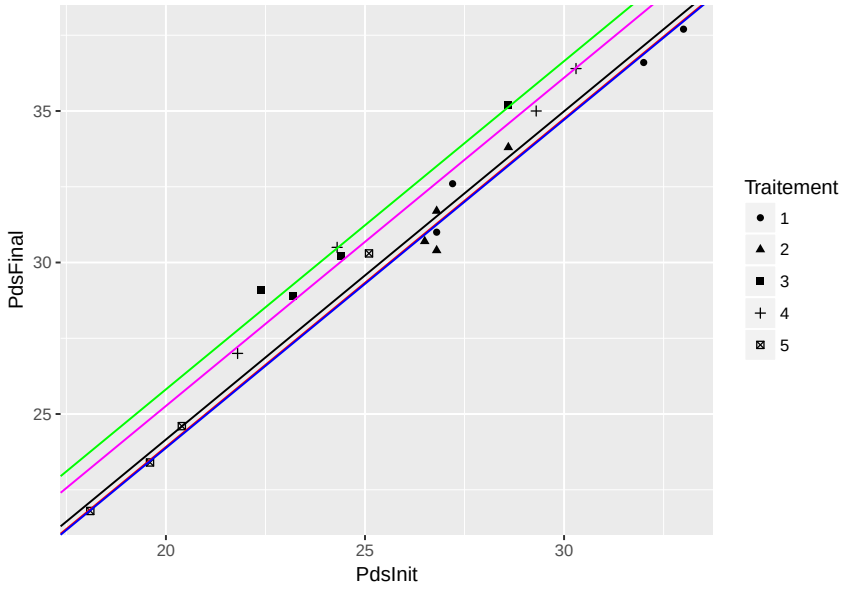


FIGURE 8.6 – Ajustement du modèle sans interaction (M1) aux données

- Le modèle défini par l'équation :

$$(M2): \ Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \ \forall i = 1, \cdots, I, \ \forall j = 1, \cdots, n_i.$$

Ce modèle correspond à l'hypothèse d'absence d'effet de la covariable z sur Y :

$$\mathcal{H}_0^{(M2)} : b_1 = b_2 = \cdots = b_I = 0.$$

Seul le facteur T explique Y. On met donc en place un modèle d'analyse de la variance à un facteur (Figure 8.7).

- Le modèle défini par l'équation :

$$(M3): \ Y_{ij} = \mu + \beta z_{ij} + \varepsilon_{ij}, \ \forall i = 1, \cdots, I, \ \forall j = 1, \cdots, n_i$$

Ce modèle correspond à l'hypothèse d'absence d'effet du facteur T sur Y.

$$\mathcal{H}_0^{(M3)} : a_1 = a_2 = \ldots = a_I \Longleftrightarrow \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0.$$

Les I droites de régression partagent la même constante à l'origine, seule la covariable z explique Y. On met alors en place un modèle de régression linéaire simple (Figure 8.8)

Exercice 34. *Ecrivez les tests de Fisher permettant de tester* $(M2)$ *contre* $(M1)$ *et* $(M3)$ *contre* $(M1)$. *A l'aide des résultats en Figure 8.9, répondez à ces deux tests.*

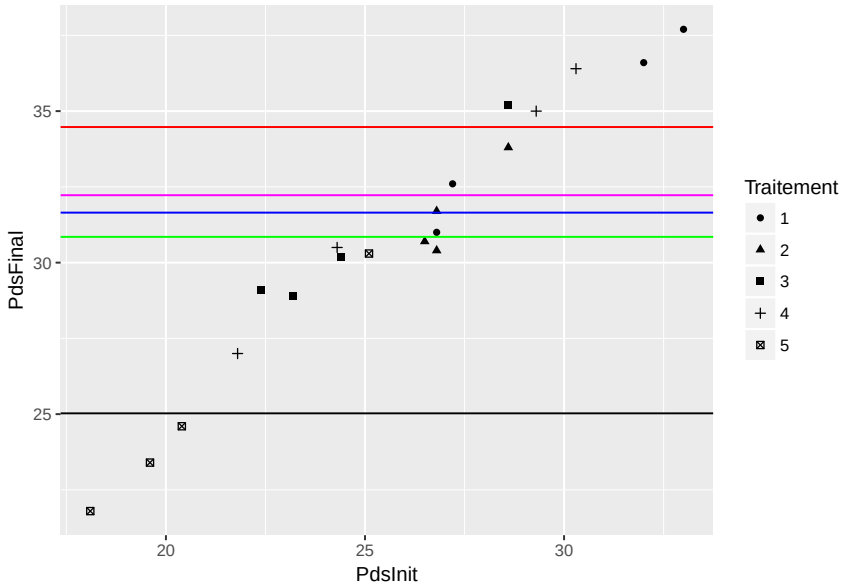


FIGURE 8.7 – Ajustement du modèle d'analyse de la variance à un facteur (M2) aux données

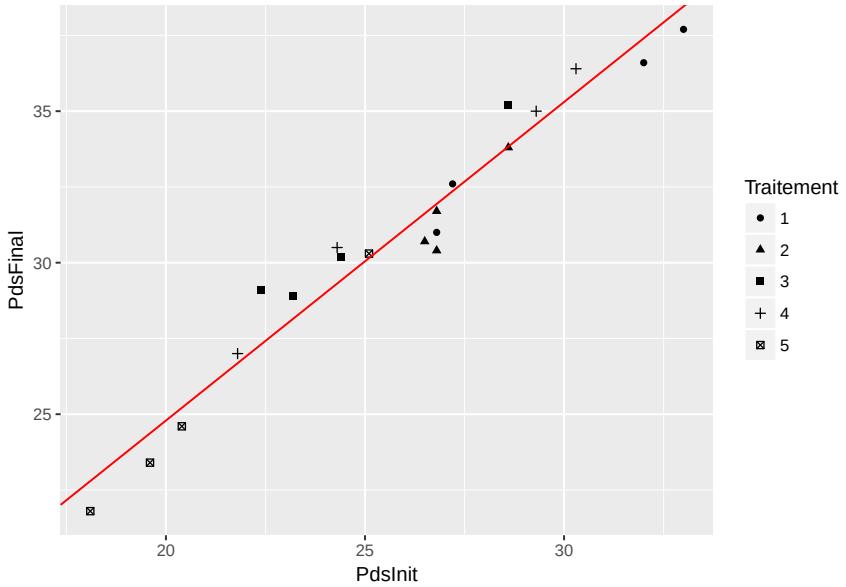


FIGURE 8.8 – Ajustement du modèle de régression linéaire simple (M3) aux données

```
> M0<-lm(PdsFinal~1)
> anova(M0,complet)
Analysis of Variance Table

Model 1: PdsFinal ~ 1
Model 2: PdsFinal ~ PdsInit * Traitement
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     19 358.67
2     10   2.83  9    355.84 139.51 2.572e-09 ***

> nonI<-lm(PdsFinal~PdsInit+Traitement)
> anova(nonI,complet)
Analysis of Variance Table

Model 1: PdsFinal ~ PdsInit + Traitement
Model 2: PdsFinal ~ PdsInit * Traitement
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     14 4.2223
2     10 2.8340  4    1.3883 1.2247 0.3602

> M2<-lm(PdsFinal~Traitement)
> anova(M2,nonI)
Analysis of Variance Table

Model 1: PdsFinal ~ Traitement
Model 2: PdsFinal ~ PdsInit + Traitement
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     15 160.263
2     14   4.222  1    156.04 517.38 1.867e-12 ***

> M3<-lm(PdsFinal~PdsInit)
> anova(M3,nonI)
Analysis of Variance Table

Model 1: PdsFinal ~ PdsInit
Model 2: PdsFinal ~ PdsInit + Traitement
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     18 16.3117
2     14  4.2223  4    12.089 10.021 0.0004819 ***

> summary(nonI)

Call:
lm(formula = PdsFinal ~ PdsInit + Traitement)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8438 -0.3154 -0.2171  0.4863  0.8871

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.25040    1.44308   1.559 0.141205
PdsInit      1.08318    0.04762  22.746 1.87e-12 ***
Traitement2 -0.03581    0.40723  -0.088 0.931169
Traitement3  1.89922    0.45802   4.147 0.000988 ***
Traitement4  1.35157    0.41937   3.223 0.006135 **
Traitement5  0.24446    0.57658   0.424 0.678022
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.5492 on 14 degrees of freedom
Multiple R-squared:  0.9882,        Adjusted R-squared:  0.984
F-statistic:   235 on 5 and 14 DF,  p-value: 5.493e-13
```

Figure 8.9 – Résultats des différents tests effectués avec la commande anova

En résumé :

- Savoir écrire un modèle d'ANCOVA (individuellement et matriciellement), régulier et singulier
- Savoir distinguer un modèle régulier d'un modèle singulier
- Savoir estimer les paramètres du modèle d'ANCOVA dans le cas régulier et dans le cas singulier (en s'adaptant à la / les contrainte(s) choisie(s))
- Savoir construire un intervalle de confiance pour un paramètre du modèle d'ANCOVA
- Savoir construire un test pour tester l'effet du facteur, l'effet d'interaction, ... et savoir organiser ces différents tests
- Savoir associer une représentation graphique à un sous-modèle d'ANCOVA

Deuxième partie

Le modèle linéaire généralisé

Chapitre 9

Principe du modèle linéaire généralisé

9.1 Introduction

Nous observons un vecteur Y de taille n, réalisation d'une variable aléatoire de moyenne μ et dont les composants sont indépendants. Dans le cadre du modèle linéaire, on a $\mu = X\theta$ où X est une matrice $n \times k$: le design. Le vecteur θ est inconnu et modélise l'influence des régresseurs sur la réponse Y.

Le modèle linéaire tel que nous l'avons vu peut donc être caractérisé de la manière suivante :

1. une **composante aléatoire** : Y est un vecteur aléatoire de moyenne μ,
2. une **composante "systématique"**, les régresseurs $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(p)}$ définissent un prédicteur linéaire : $\eta = X\theta (= \theta_0 \mathbb{1}_n + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)})$ $(k = p + 1)$
3. **la relation liant** μ et η : $\mu = \eta$ pour le modèle linéaire.

Imposer une dépendance linéaire entre les régresseurs et $\mathbb{E}[Y]$ permet une étude approfondie mais peut être parfois trop restrictive. Une généralisation possible du modèle linéaire consiste donc à supposer que la relation liant μ à η n'est pas l'identité, mais plutôt un lien du type :

$$\eta_i = g(\mu_i), \text{ pour } \eta = (\eta_1, \ldots, \eta_n)' \text{ et } \mu = (\mu_1, \ldots, \mu_n)'.$$

La fonction g modélise donc le lien entre ces deux vecteurs. Cette formulation permet de modéliser un panel plus riche d'expériences.

Exemple 6. *Dans une expérience clinique, on cherche à comparer deux modes opératoires pour une opération chirurgicale donnée. L'expérience est menée sur deux hôpitaux différents. On dispose donc ici de deux facteurs à deux modalités : mode opératoire et hôpital. La variable réponse correspond pour chaque patient au succès ou à l'échec de l'intervention : il s'agit d'une variable binaire.*

Exemple 7. *Un assureur s'intéresse au nombre de sinistres automobile déclarés pendant ces dix dernières années. Il souhaite étudier si ce nombre de sinistres est lié à l'âge*

du conducteur, la taille de la voiture, ... Le nombre de sinistres peut être modélisé par une loi de Poisson.

Dans le cas particulier où la fonction de lien est de type canonique (i.e. $g(x) = x$), rien n'interdit d'utiliser la méthode des moindres carrés introduite dans la Partie I. Cette dernière est en effet purement géométrique et peut donc tout à fait s'appliquer à des réponses de type "binaire". Cependant, la partie inférentielle traitée dans ce cours nécessite quant à elle des hypothèses très fortes sur la distribution des observations. Pour des modèles alternatifs, il faut donc complètement repenser la construction des tests et des intervalles de confiance. Par ailleurs, une relation de type canonique est relativement restrictive (cf Figure 9.1). Il convient donc de se placer dans un cadre plus général afin de pouvoir faire face à des problèmes plus variés.

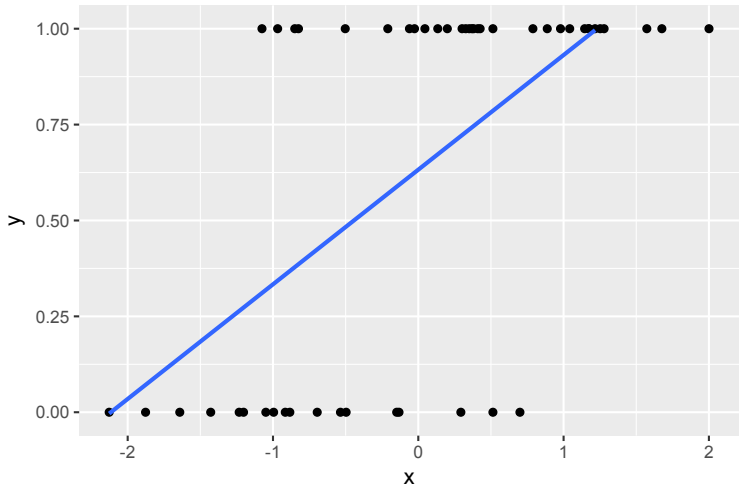


FIGURE 9.1 – Exemple d'observations pour un modèle de type binaire

De manière plus générale, la méthode des moindres carrés introduite dans la Partie I ne peut être implémentée. Bien souvent, le problème d'optimisation associé n'est en effet pas convexe. Une première "parade" consiste à utiliser l'estimateur du maximum de vraisemblance... mais dans la plupart des cas, ce dernier n'est pas calculable analytiquement. Il est cependant possible d'utiliser un algorithme itératif inspiré de la méthode de Newton-Raphson permettant d'approcher le maximum de vraisemblance. Sous certaines conditions, cet algorithme propose des résultats tout à fait satisfaisants.

9.2 Caractérisation d'un modèle linéaire généralisé

L'objet de cette section est d'introduire le cadre théorique global permettant de regrouper tous les modèles (linéaire gaussien, logit, log-linéaire) de ce cours qui cherchent à modéliser l'espérance d'une variable réponse Y en fonction d'une combinaison linéaire

de variables explicatives. Le **modèle linéaire généralisé** développé initialement en 1972 par Nelder et Wedderburn et dont on trouvera des exposés détaillés dans Nelder et Mc Cullagh [18], Agresti [1] ou Antoniadis et al. [4], n'est ici qu'esquissé afin de définir les concepts communs à ces modèles : famille exponentielle, estimation par maximum de vraisemblance, tests, diagnostics, résidus.

Le modèle linéaire généralisé est caractérisé par trois quantités :

1. La variable réponse Y, composante aléatoire à laquelle est associée une loi de probabilité
2. les variables explicatives $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)}$ (prédicteurs)
3. le lien qui décrit la relation fonctionnelle entre la combinaison linéaire des $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)}$ et l'espérance de la variable réponse Y.

Nous allons par la suite détailler ces différentes quantités.

9.2.1 Loi de la variable réponse Y

La composante aléatoire identifie la distribution de probabilité de la variable à expliquer Y. On suppose que l'échantillon statistique est constitué de n variables aléatoires $(Y_i)_{i=1,\cdots,n}$ indépendantes admettant des distributions issues d'une structure de **famille exponentielle**.

Définition 9.1. *Soit Y une variable aléatoire unidimensionnelle. On dit que la loi de Y appartient à une **famille exponentielle** si la loi de Y est dominée par une mesure dite de référence et si la vraisemblance de Y calculée en y par rapport à cette mesure s'écrit de la façon suivante :*

$$f_Y(y, \omega, \phi) = \exp\left[\frac{y\omega - b(\omega)}{\gamma(\phi)} + c(y, \phi)\right]. \tag{9.1}$$

Cette formulation inclut la plupart des lois usuelles comportant un ou deux paramètres : gaussienne, gaussienne inverse, gamma, Poisson, binomiale (cf Table 9.1). Le paramètre ω est appelé le **paramètre naturel** de la famille exponentielle.

Attention, la mesure de référence change d'une structure exponentielle à l'autre : la mesure de Lebesgue pour une loi continue, une mesure discrète combinaison de Dirac pour une loi discrète. Consulter Antoniadis et al [4] pour une présentation générale de la famille exponentielle et des propriétés asymptotiques des estimateurs de leurs paramètres.

Proposition 9.1. *Soit Y une variable aléatoire dont la loi de probabilité appartient à la famille exponentielle alors*

$$\mathbb{E}[Y] = b'(\omega)$$

et

$$Var(Y) = b''(\omega)\gamma(\phi).$$

Exercice 35.

1. *Pour évaluer* $\mathbb{E}[Y]$*, calculer* $\frac{\partial}{\partial\omega} f_Y(y,\omega,\phi)$ *et intégrer par rapport à* y
2. *Pour évaluer* $Var(Y)$*, calculer* $\frac{\partial^2}{\partial\omega^2} f_Y(y,\omega,\phi)$ *et intégrer par rapport à* y

Pour certaines lois, la fonction γ est de la forme : $\gamma(\phi) = \phi$. Dans ce cas, ϕ est appelé **paramètre de dispersion**, c'est un paramètre de nuisance intervenant par exemple lorsque les variances des lois gaussiennes sont inconnues, mais égal à 1 pour les lois à un paramètre (Poisson, Bernoulli). L'expression de la structure exponentielle se met alors sous la forme canonique :

$$f(y,\omega) = a(\omega)d(y)\exp[yQ(\omega)] \tag{9.2}$$

avec $Q(\omega) = \frac{\omega}{\phi}$, $a(\omega) = \exp\left(-\frac{b(\omega)}{\phi}\right)$ et $d(y) = \exp[c(y,\phi)]$.

Exercice 36. *Exemples dans la famille exponentielle :*

1. *Loi gaussienne :*
   Montrez que la loi $\mathcal{N}(\mu,\sigma^2)$ *est dans la famille exponentielle de paramètre de dispersion* $\phi = \sigma^2$ *et de paramètre naturel* $\omega = \mu$*.*
2. *Loi de Bernoulli :*
   Montrez que la loi de Bernoulli $\mathcal{B}(\pi)$ *est dans la famille exponentielle, de paramètre naturel* $\omega = \ln(\frac{\pi}{1-\pi})$*.*
   La loi binomiale conduit à des résultats identiques en considérant la somme de n *(connu) variables de Bernoulli.*
3. *Loi de Poisson :*
   Montrez que la loi de Poisson de paramètre λ *est dans la famille exponentielle, de paramètre naturel* $\omega = \ln(\lambda)$*.*

Distribution	ω	$b(\omega)$	$\gamma(\phi)$	$\mathbb{E}[Y] = b'(\omega)$	$\mathrm{Var}(Y) = b''(\omega)\gamma(\phi)$
Gaussienne $\mathcal{N}(\mu,\sigma^2)$	μ	$\frac{\omega^2}{2}$	$\phi = \sigma^2$	$\mu = \omega$	σ^2
Bernoulli $\mathcal{B}(p)$	$\ln(p/1-p)$	$\ln(1+e^\omega)$	1	$p = \frac{e^\omega}{1+e^\omega}$	$p(1-p)$
Poisson $\mathcal{P}(\lambda)$	$\ln(\lambda)$	$\lambda = e^\omega$	1	$\lambda = e^\omega$	$\lambda = e^\omega$
Gamma $\mathcal{G}(\mu,\nu)$	$-\frac{1}{\mu}$	$-\ln(-\omega)$	$\frac{1}{\nu}$	$\mu = -\frac{1}{\omega}$	$\frac{\mu^2}{\nu}$
Inverse Gamma $IG(\mu,\sigma^2)$	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\omega}$	σ^2	$\mu = (\sqrt{-2\omega})^{-1}$	$\mu^3\sigma^2$

TABLE 9.1 – Exemples de lois de probabilité appartenant à la famille exponentielle

9.2.2 Prédicteur linéaire

Les observations planifiées des variables explicatives sont organisées dans la matrice $\mathbf{X}$ de planification d'expérience (design matrix). Soit θ un vecteur de k $(= p + 1)$ paramètres. Le **prédicteur linéaire**, composante déterministe du modèle est le vecteur à n composantes défini par

$$\eta = \mathbf{X}\theta.$$

9.2.3 Fonction de lien

Cette troisième quantité exprime une *relation fonctionnelle* entre la composante aléatoire et le prédicteur linéaire. Soit $\mu_i = \mathbb{E}[Y_i]\,;\, i = 1, \cdots, n$. On pose

$$\forall i = 1, \cdots, n,\ \eta_i = g(\mu_i)$$

où g, appelée **fonction de lien**, est supposée monotone et différentiable. Ceci revient donc à écrire un modèle dans lequel une fonction de la moyenne appartient au sous-espace vectoriel engendré par les variables explicatives :

$$\forall i = 1, \cdots, n,\ g(\mu_i) = \mathbf{x}_i\theta.$$

La fonction de lien qui associe la moyenne μ_i au paramètre naturel ω_i est appelée **fonction de lien canonique**. Dans ce cas,

$$\forall i = 1, \cdots, n,\ g(\mu_i) = \omega_i = \mathbf{x}_i\theta.$$

Exemples :
La fonction de lien canonique pour

- la loi gaussienne est l'identité : $\omega_i = \mu_i$
- la loi de Poisson est le logarithme $\omega_i = \ln(\mu_i)$
- la loi de Bernoulli est la fonction logit $\omega_i = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$

Dans le cas d'une variable réponse binaire Y, on peut aussi considérer la fonction de lien **probit** :

$$\eta_i = g(\pi_i) = \Phi^{-1}(\pi_i)$$

où $\Phi(.)$ est la fonction de répartition de la loi normale $\mathcal{N}(0, 1)$.

Dans le cadre de l'étude d'une variable réponse Y suivant une loi binomiale et considérant la fonction de lien logit, le modèle linéaire généralisé est appelé **régression logistique**. Dans le cadre de l'étude d'une variable réponse Y suivant une loi de Poisson et considérant la fonction de lien logarithme, le modèle linéaire généralisé est appelé **modèle log-linéaire**.

9.3 Estimation

Le modèle étant posé, on souhaite maintenant estimer le vecteur des paramètres $\theta = (\theta_0, \ldots, \theta_p)'$ et le paramètre de dispersion ϕ. Comme ce dernier paramètre n'apparait pas dans l'espérance, ce n'est pas le paramètre d'intérêt. Pour simplifier, on va supposer par la suite que ϕ est fixé (ou estimé préalablement), seul θ reste à estimer.

9.3.1 Estimation par maximum de vraisemblance

La méthode des moindres carrés n'est pas applicable dans un grand nombre de situations pour le modèle linéaire généralisé (excepté pour des fonctions de lien canonique, i.e. identité). Pour ce problème d'estimation, on utilise donc la méthode d'estimation du maximum de vraisemblance (EMV). On va donc maximiser la log-vraisemblance du modèle linéaire généralisé.

Par indépendance des observations, la vraisemblance du n-échantillon $\underline{Y} = (Y_1, \ldots, Y_n)$ s'écrit $\theta \mapsto L(\underline{Y}; \theta)$ telle que

$$\theta \mapsto L(\underline{y}; \theta) = \prod_{i=1}^{n} f_{Y_i}(y_i; \omega_i)$$

et la log-vraisemblance vaut $\theta \mapsto l(Y; \theta)$ avec

$$\theta \mapsto l(\underline{y}; \theta) = \sum_{i=1}^{n} \ln[f_{Y_i}(y_i; \omega_i)],$$

où θ, η, μ et ω sont liés par le modèle. L'EMV associé vérifie donc

$$\widehat{\theta}_{MV} \in \arg\max_{\theta} L(\underline{Y}; \theta) = \arg\max_{\theta} l(\underline{Y}; \theta).$$

En particulier, si la fonction de lien g est celle du lien canonique, on a $\omega_i = \mathbf{x}_i\theta$ et donc

$$l(\underline{y}; \theta) = \sum_{i=1}^{n} \frac{y_i \mathbf{x}_i\theta - b(\mathbf{x}_i\theta)}{\gamma(\phi)} + c(y_i, \phi).$$

Afin d'obtenir une expression de l'EMV, on s'intéresse au score

$$S(\underline{Y}; \theta) = \left(\frac{\partial}{\partial\theta_0} l(\underline{Y}; \theta), \ldots, \frac{\partial}{\partial\theta_p} l(\underline{Y}; \theta) \right)'.$$

L'estimateur du maximum de vraisemblance vérifie donc

$$S(\underline{Y}; \widehat{\theta}_{MV}) = 0_k. \tag{9.3}$$

Dans le cas particulier où g est le lien canonique, on a

$$\forall j = 0, \ldots, p, \ \frac{\partial}{\partial\theta_j} l(\underline{y}; \theta) = \sum_{i=1}^{n} \frac{1}{\gamma(\phi)} x_i^{(j)} [y_i - b'(\mathbf{x}_i\theta)] = 0 \Leftrightarrow \sum_{i=1}^{n} [y_i - b'(\mathbf{x}_i\theta)] \frac{\mathbf{x}_i}{\gamma(\phi)} = 0_k$$

On peut constater que ce système n'est linéaire que si $b'(a) = a$, c'est-à-dire si on est dans le cas du modèle linéaire. Pour tous les autres modèles linéaires généralisés, (9.3) est un système non linéaire en θ et il n'existe pas de formule analytique pour cet estimateur. Il est cependant possible de montrer que le problème associé à la détermination de $\hat{\theta}_{MV}$ est un problème d'optimisation convexe qui peut donc être traité par un algorithme de type Newton-Raphson, adapté à un cadre statistique, cf l'Annexe des rappels A.3 pour les détails de cet algorithme.

9.3.2 Algorithmes de Newton-Raphson et Fisher-scoring

L'algorithme de Newton-Raphson est un algorithme itératif basé sur le développement de Taylor à l'ordre 1 du score (cf Figure 9.2). Il fait donc intervenir la matrice Hessienne de la log-vraisemblance

$$\mathcal{J}_{j\ell} = \frac{\partial^2 l(\underline{y};\theta)}{\partial\theta_j \partial\theta_\ell}.$$

Il faut que $\mathcal{J}$ soit inversible et comme elle dépend de θ, il convient de mettre à jour cette matrice à chaque étape de cet algorithme itératif. Cet algorithme est implémenté dans la plupart des logiciels statistiques.

- Initialisation : $u^{(0)}$.
- Pour tout entier h

$$u^{(h)} = u^{(h-1)} - [\mathcal{J}^{(h-1)}]^{-1} S(\underline{Y}; u^{(h-1)}). \tag{9.4}$$

- Arrêt quand

$$|u^{(h)} - u^{(h-1)}| \leq \Delta.$$

- on pose $\hat{\theta}_{MV} = u^{(h)}$.

FIGURE 9.2 – Principe de l'algorithme de Newton-Raphson

Parfois, au lieu d'utiliser la matrice hessienne, on utilise la matrice d'information de Fisher

$$\mathcal{I}_n(\theta)_{j,\ell} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_\ell} l(\underline{Y};\theta)\right].$$

C'est l'algorithme de Fisher-scoring. Ici aussi, on a besoin que $\mathcal{I}_n(\theta)$ soit inversible, quitte à imposer des contraintes sur θ. Cette solution peut permettre d'éviter des problèmes de non inversibilité de la hessienne.

9.3.3 Equations de vraisemblance

Les algorithmes de type Newton-Raphson précédents nécessitent d'évaluer le score et la matrice d'information de Fisher.

Proposition 9.2. *Soit le score* $S(\underline{Y};\theta) = (S_0,\dots,S_p)'$ *avec* $S_j = \frac{\partial}{\partial\theta_j} l(\underline{Y};\theta)$. *Alors pour* $j \in \{0,\dots,p\}$,

$$S_j = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_i^{(j)}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \tag{9.5}$$

et $\mathbb{E}[S_j] = 0$.

Exercice 37. *Comprenez la preuve donnée en annexe B.6.*

Proposition 9.3. *La matrice d'information de Fisher s'écrit*

$$\mathcal{I}_n(\theta) = \mathbf{X}'\mathbf{W}\mathbf{X}$$

où $\mathbf{W}$ *est la matrice diagonale de "pondération" :*

$$[\mathbf{W}]_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Exercice 38. *On rappelle que* $\mathcal{I}_n(\theta)$ *est la matrice de variance-covariance de* $S(\underline{Y};\theta)$ *donc* $(\mathcal{I}_n(\theta))_{j\ell} = \mathbb{E}[S_j S_\ell]$. *En utilisant* (9.5), *démontrez la Proposition 9.3.*

Corollaire 9.1. *Dans le cas particulier où la fonction lien est le lien canonique associé à la structure exponentielle alors* $\eta_i = \omega_i = \mathbf{x}_i\boldsymbol{\theta}$. *On obtient donc les simplifications suivantes :*

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \omega_i} = b''(\omega_i) = \frac{Var(Y_i)}{\gamma(\phi)}.$$

Ainsi,

$$S_j = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\gamma(\phi)} x_i^{(j)} \text{ et } W_{ii} = \frac{Var(Y_i)}{\gamma(\phi)^2}.$$

En particulier, comme $\mathcal{I}_n(\theta)$ *ne dépend plus de* Y_i, *la hessienne est égale à la matrice d'information de Fisher et donc les méthodes de résolution du score de Fisher et de Newton-Raphson coïncident.*

Si de plus $\gamma(\phi)$ *est une constante pour les observations,*

$$S_j = \frac{1}{\gamma(\phi)} \sum_{i=1}^{n} (Y_i - \mu_i)x_i^{(j)} = 0 \ \forall j \iff X'Y = X'\mu.$$

Dans le cas gaussien, comme $\mu = X\theta$ *avec la fonction de lien canonique identité, on retrouve la solution* $(X'X)^{-1}X'Y = \theta$ *qui coincide avec celle obtenue par minimisation des moindres carrés.*

9.4 Loi asymptotique de l'EMV et inférence

De part la complexité du modèle linéaire généralisé, l'obtention d'un intervalle de confiance va nécessiter un peu plus de travail que dans un cadre de statistique paramétrique usuel.
Le théorème suivant donne des propriétés sur l'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$.

Théorème 9.1. *Sous certaines conditions de régularité de la densité de probabilité, l'EMV vérifie les propriétés suivantes :*

- $\hat{\theta}_{MV}$ *converge en probabilité vers* $\theta \in \mathbb{R}^k$
- $\hat{\theta}_{MV}$ *converge en loi vers une loi gaussienne :*

$$\mathcal{I}_n(\theta)^{1/2}(\hat{\theta}_{MV} - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0_k, I_k)$$

- *La* ***statistique de Wald*** $\mathcal{W}$ *vérifie*

$$\mathcal{W} := (\hat{\theta}_{MV} - \theta)'\mathcal{I}_n(\theta)(\hat{\theta}_{MV} - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(k).$$

Remarque : Dans le cas particulier où la distribution des Y_i est gaussienne et la fonction de lien est canonique, il est possible de montrer que l'estimateur du maximum de vraisemblance est lui aussi gaussien et ce sans avoir recours à l'approximation $\hat{\theta}_{MV} - \theta \overset{\mathcal{L}}{\simeq} \mathcal{N}(0_k, \mathcal{I}_n(\theta)^{-1})$ quand $n \to +\infty$. Si maintenant les erreurs ne sont pas gaussiennes, le résultat précédent propose une alternative intéressante aux tests de Fisher. Ce théorème permet déjà de répondre à des problèmes intéressants comme la construction d'intervalles de confiance pour les θ_j, tests sur des valeurs de θ,.... D'autres approches complémentaires sont disponibles pour ce type de modèle, la plus connue étant basée sur le test du rapport de vraisemblance.

A noter qu'un tel résultat n'est pas utilisable tel quel puisque la matrice $\mathcal{I}_n(\theta)$ est inconnue. Mais en remplaçant $\mathcal{I}_n(\theta)$ par $\mathcal{I}_n(\hat{\theta}_{MV})$ avec $\hat{\theta}_{MV}$ converge en probabilité vers θ, on obtient que

$$\mathcal{I}_n(\hat{\theta}_{MV})^{1/2}(\hat{\theta}_{MV} - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0_k, I_k)$$

et

$$(\hat{\theta}_{MV} - \theta)'\mathcal{I}_n(\hat{\theta}_{MV})(\hat{\theta}_{MV} - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(k).$$

9.5 Tests d'hypothèses

Contrairement au cas du modèle linéaire, la loi de l'estimateur du maximum de vraisemblance dans le cadre du modèle linéaire généralisé n'est connue qu'asymptotiquement. Aussi les procédures de test vont être menées dans un cadre asymptotique. Nous allons dans la suite considérer plusieurs problèmes de test qui permettent d'examiner la qualités du modèle, de déterminer si les différentes variables explicatives du modèles sont pertinentes ou pas, ...

9.5.1 Test de modèles emboîtés

Le test de comparaison des modèles emboîtés permet de déterminer si un sous-ensemble de variables explicatives est suffisant pour expliquer la réponse Y comme dans le cas du modèle linéaire.

On considère deux modèles emboîtés M_1 et M_0, définis par $g(\mu) = X_1\theta_1$ et $g(\mu) = X_0\theta_0$ respectivement, avec M_0 sous-modèle de M_1. On veut tester $\mathcal{H}_0 : M_0$ contre $\mathcal{H}_1 : M_1$.

9.5.1.1 Test du rapport de vraisemblance

On considère le **test du rapport de vraisemblance** dont la statistique de test est donnée par

$$T = -2\ln\left[\frac{L(\underline{Y};\hat{\theta}_0)}{L(\underline{Y};\hat{\theta}_1)}\right] = -2\left[l(\underline{Y};\hat{\theta}_0) - l(\underline{Y};\hat{\theta}_1)\right]$$

où $\hat{\theta}_0$ et $\hat{\theta}_1$ sont les EMV de θ dans le modèle M_0 et M_1 respectivement. Sous certaines conditions, on peut montrer que

$$T \underset{n\to+\infty}{\overset{\mathcal{L}}{\to}} \chi^2(k_1 - k_0)$$

où k_0 et k_1 sont les dimensions des sous-espaces engendrés par les colonnes de X_0 et X_1 respectivement. La zone de rejet est alors définie par

$$\mathcal{R}_\alpha = \{T > v_{1-\alpha,k_1-k_0}\}$$

où $v_{1-\alpha,k_1-k_0}$ est le $(1-\alpha)$- quantile de la loi du χ^2 à $k_1 - k_0$ degrés de liberté.

Ce test est parfois présenté de façon un peu différente en faisant intervenir la **déviance**, qui est l'écart entre la log-vraisemblance du modèle d'intérêt M et celle du modèle le plus complet possible M_{sat}, appelé modèle saturé. Le modèle saturé est le modèle comportant n paramètres, c'est à dire autant que d'observations. La déviance de M est définie par :

$$\mathcal{D}(M) = -2\left[l(\underline{Y};\hat{\theta}) - l(\underline{Y};\hat{\theta}_{sat})\right].$$

La statistique de test T peut donc se réécrire avec la déviance :

$$T = \mathcal{D}(M_0) - \mathcal{D}(M_1).$$

Le test global consiste à tester $\mathcal{H}_0 : g(\mu_i) = a$ contre $\mathcal{H}_1 : g(\mu_i) = \mathbf{x}_i\theta$ avec le test du rapport de vraisemblance. Il consiste donc à tester si toutes les variables sont inutiles pour expliquer la réponse Y.

9.5.1.2 Test de Wald

Le **test de Wald** est basé sur la forme quadratique faisant intervenir la matrice de covariance des paramètres, l'inverse de la matrice d'information observée $(X'WX)^{-1}$. Cette matrice généralise la matrice $(X'X)^{-1}$ utilisée dans le cas du modèle linéaire gaussien en faisant intervenir une matrice W de pondération. Ainsi, les test de Wald et test de Fisher sont équivalents dans le cas particulier du modèle gaussien.

Si la matrice $C \in \mathcal{M}_{qk}(\mathbb{R})$ définit l'ensemble $\mathcal{H}_0$ des hypothèses à tester sur les paramètres correspondant à q contraintes : $C\,\theta = 0_q$, on montre que

$$(C\widehat{\theta}_{MV})'[C(X'WX)^{-1}C']^{-1}(C\widehat{\theta}_{MV}) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(q).$$

Attention, le test de Wald peut ne pas être précis si le nombre d'observations est faible.

9.5.2 Test d'un paramètre θ_j

Si la réponse au test global est de rejeter $\mathcal{H}_0$, on s'intéresse ensuite à tester quelles sont les variables qui ont une influence. On revient au problème général de vouloir tester $\mathcal{H}_0 : \theta_j = a$ contre $\mathcal{H}_0 : \theta_j \neq a$, où a est une valeur définie a priori (souvent $a = 0$). D'après le théorème 9.1, on peut faire l'approximation de loi suivante sous $\mathcal{H}_0$:

$$(\hat{\theta}_{MV})_j - a \underset{\mathcal{H}_0}{\overset{\mathcal{L}}{\simeq}} \mathcal{N}\left(0, [\mathcal{I}_n(\hat{\theta}_{MV})^{-1}]_{jj}\right).$$

On va donc rejeter $\mathcal{H}_0$ si

$$T_j := \left|(\hat{\theta}_{MV})_j - a\right| / \sqrt{[\mathcal{I}_n(\hat{\theta}_{MV})^{-1}]_{jj}} > z_{1-\alpha/2}$$

où $z_{1-\alpha/2}$ est le $1-\alpha/2$ quantile de la loi $\mathcal{N}(0,1)$. Ce test est appelé le **Z-test**.

Ce test est équivalent au test de Wald : on rejette $\mathcal{H}_0$ si

$$\left[(\hat{\theta}_{MV})_j - a\right]^2 / [\mathcal{I}_n(\hat{\theta}_{MV})^{-1}]_{jj} > v_{1-\alpha,1}$$

où $v_{1-\alpha,1}$ est le $1-\alpha$ quantile de la loi $\chi^2(1)$.

9.5.3 Test de $C\theta = 0_q$

Comme dans le modèle linéaire, on peut vouloir tester

$$\mathcal{H}_0 : C\theta = 0_q \text{ contre } \mathcal{H}_1 : C\theta \neq 0_q$$

où $C \in \mathcal{M}_{qk}(\mathbb{R})$. Connaissant la loi asymptotique de $\hat{\theta}_{MV}$, on obtient que

$$\left[C\mathcal{I}_n(\hat{\theta}_{MV})^{-1}C'\right]^{-1/2}(C\hat{\theta}_{MV} - C\theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0_q, I_q)$$

et

$$(C\hat{\theta}_{MV} - C\theta)' \left[C\mathcal{I}_n(\hat{\theta}_{MV})^{-1}C'\right]^{-1} (C\hat{\theta}_{MV} - C\theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(q).$$

On considère donc la zone de rejet

$$\mathcal{R}_\alpha = \{(C\hat{\theta}_{MV})' \left[C\mathcal{I}_n(\hat{\theta}_{MV})^{-1}C'\right]^{-1} (C\hat{\theta}_{MV}) > v_{1-\alpha,q}\}$$

où $v_{1-\alpha,q}$ est le $(1-\alpha)$ quantile d'un $\chi^2(q)$. Quand $q=1$, on peut aussi considèrer

$$\mathcal{R}_\alpha = \left\{ \left| \left[C\mathcal{I}_n(\hat{\theta}_{MV})^{-1}C'\right]^{-1/2} C\hat{\theta}_{MV} \right| > z_{1-\alpha/2} \right\}$$

où $z_{1-\alpha/2}$ est le $1-\alpha/2$ quantile d'une loi $\mathcal{N}(0,1)$.

9.6 Intervalle de confiance pour θ_j

9.6.1 Par Wald

D'après le théorème 9.1, on peut faire l'approximation de loi suivante :

$$\left[(\hat{\theta}_{MV})_j - \theta_j\right] / \sqrt{[\mathcal{I}_n(\hat{\theta}_{MV})^{-1}]_{jj}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

On peut donc construire l'intervalle de confiance asymptotique pour θ_j au niveau de confiance $1-\alpha$ suivant :

$$IC_{1-\alpha}(\theta_j) = \left[(\hat{\theta}_{MV})_j \pm z_{1-\alpha/2}\sqrt{[\mathcal{I}_n(\hat{\theta}_{MV})^{-1}]_{jj}}\right]$$

où $z_{1-\alpha/2}$ est le $1-\alpha/2$ quantile de la loi $\mathcal{N}(0,1)$.

9.6.2 Fondé sur le rapport de vraisemblances

La **fonction de vraisemblance profil** de θ_j est définie par

$$l^\star(\underline{Y};\theta_j) = \max_{\tilde{\theta}} l(\underline{Y};\tilde{\theta})$$

où $\tilde{\theta}$ est le vecteur θ avec le $j$ème élément fixé à θ_j. Si θ_j est la vraie valeur du paramètre alors

$$2\left[l(\underline{Y};\hat{\theta}_{MV}) - l^\star(\underline{Y};\theta_j)\right] \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(1).$$

Ainsi, si on considère l'ensemble

$$\mathcal{G} = \left\{u; 2\left[l(\underline{Y};\hat{\theta}_{MV}) - l^\star(\underline{Y};u)\right] \leq v_{1-\alpha,1}\right\},$$

on obtient que $\mathbb{P}(\theta_j \in \mathcal{G}) \to 1-\alpha$. Ainsi $\mathcal{G}$ est un intervalle de confiance asymptotique pour θ_j au niveau de confiance $1-\alpha$.

9.7 Qualité d'ajustement

9.7.1 Le pseudo R^2

Par analogie avec le R^2 utilisé dans le cadre du modèle linéaire, on définit le pseudo-R^2 en fonction de la déviance $\mathcal{D}(M_0)$ du modèle M_0 (le modèle nul) et de la différence de déviance $\mathcal{D}(M_0) - \mathcal{D}(M)$:

$$pseudoR^2 = \frac{\mathcal{D}(M_0) - \mathcal{D}(M)}{\mathcal{D}(M_0)}.$$

Ce pseudo-R^2 varie entre 0 et 1. Plus il est proche de 1, meilleur est l'ajustement du modèle.

9.7.2 Le χ^2 de Pearson généralisé

Le χ^2 de Pearson généralisé est la statistique définie par

$$Z^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\mathrm{Var}_{\hat{\mu}_i}(Y_i)}$$

où $\hat{\mu}_i = g^{-1}(\mathbf{x}_i\hat{\theta}_{MV})$ et $\mathrm{Var}_{\hat{\mu}_i}(Y_i) = \mathrm{Var}_{\mu}(Y_i)|_{\mu=\hat{\mu}_i}$ (la variance théorique de Y_i évaluée en $\hat{\mu}_i$).
Sous l'hypothèse que le modèle étudié est le bon modèle et si l'approximation asymptotique est valable, alors la loi de Z^2 est approchée par $\chi^2(n-k)$. On rejette donc le modèle étudié au niveau α si la valeur observée de Z^2 est supérieure au $(1-\alpha)$ quantile de la loi $\chi^2(n-k)$.

9.8 Diagnostic, résidus

Dans le modèle linéaire généralisé, la définition la plus naturelle pour le résidu consiste à quantifier l'écart entre Y_i et sa prédiction par le modèle $\hat{\mu}_i$. On définit ainsi les résidus bruts $\varepsilon_i = Y_i - \hat{\mu}_i$. Mais ces résidus n'ayant pas toujours la même variance, il est difficile de les comparer à un comportement type attendu. Par exemple dans le cas d'un modèle de Poisson, l'écart-type d'un effectif est $\sqrt{\hat{\mu}_i}$, de grosses différences tendent à apparaitre quand μ_i prend des valeurs élevées. En normalisant les résidus bruts par une variance estimée, on obtient les résidus "standardisés" de Pearson :

$$r_{Pi} = \frac{Y_i - \hat{\mu}_i}{\sqrt{Var_{\hat{\mu}_i}(Y_i)}}.$$

On remarque que la somme des carrés des r_{Pi} correspond au χ^2 de Pearson généralisé.

On peut également étudier les résidus déviance définis par :

$$r_{Di} = \sqrt{d_i}\ sgn(Y_i - \hat{\mu}_i)$$

où d_i représente la contribution de l'observation i à la déviance $\mathcal{D}$.

Du fait que les résidus sont calculés sur les données de l'échantillon qui ont permis de construire le modèle, on risque de sous-estimer les résidus. En notant h_i le levier associé à l'observation i, $i$ème terme diagonal de la matrice $H = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$, on définit alors :

- le résidu de Pearson normalisé :

$$r^{\star}_{Pi} = \frac{Y_i - \hat{\mu}_i}{\sqrt{Var_{\hat{\mu}_i}(Y_i)(1-h_i)}}$$

- le résidu déviance normalisé :

$$r^{\star}_{Di} = \frac{\sqrt{d_i}\ sgn(Y_i - \hat{\mu}_i)}{\phi(1-h_i)}$$

- le résidu vraisemblance normalisé :

$$r_{Gi} = sgn(Y_i - \hat{\mu}_i)\sqrt{(1-h_i){r^{\star}_{Di}}^2 + h_i {r^{\star}_{Pi}}^2}$$

Chapitre 10

Régression logistique

Dans ce chapitre, on s'intéresse au cas où la variable réponse Y est binaire. Nous allons illustrer les différents points abordés dans ce chapitre avec l'exemple suivant :

Exemple : Problème de défaut bancaire
On s'intéresse à la variable réponse binaire `default` qui indique si des clients sont en défaut sur leur dette de carte de crédit (default=1 si le client fait défaut sur sa dette, 0 sinon). On considère ici un échantillon de $n = 10000$ clients et l'on souhaite expliquer la variable `default` à l'aide des 3 variables suivantes :

- `student` : 1 si le client est étudiant, 0 sinon
- `balance` : montant moyen mensuel d'utilisation de la carte de crédit
- `income` : revenu du client

On considère donc ici des variables explicatives quantitatives et qualitatives. Le comportement de ces variables est résumé sur la Figure 10.1.

L'ensemble des méthodes de modélisation disponible pour apporter des réponses à ce type de problème est désigné par le terme de **régression logistique**.

10.1 Pourquoi des modèles particuliers ?

Dans la suite, on note $Y = (Y_1, \ldots, Y_n)' \in \{0, 1\}^n$ le vecteur des réponses, et $\mathbf{x}_i$ le vecteur ligne des variables explicatives considérées pour l'individu i dans $\{1 \ldots, n\}$. La variable réponse à expliquer $Y_i|\mathbf{x}_i \sim \mathcal{B}(\pi(\mathbf{x}_i))$ vérifie

$$\mathbb{P}(Y_i = 1|\mathbf{x}_i) = \pi(\mathbf{x}_i).$$

L'objectif est de construire un modèle pour reconstituer $\pi(\mathbf{x}_i) = \mathbb{E}[Y_i|\mathbf{x}_i]$ en fonction des variables explicatives. Si on utilise le modèle de régression usuel $Y_i = \mathbf{x}_i\theta + \varepsilon_i$ pour une variable binaire, le résidu serait distribué selon la loi

$$\varepsilon_i = \begin{cases} 1 - \mathbf{x}_i\theta & \text{avec probabilité } \pi(\mathbf{x}_i), \\ -\mathbf{x}_i\theta & \text{avec probabilité } 1 - \pi(\mathbf{x}_i). \end{cases}$$

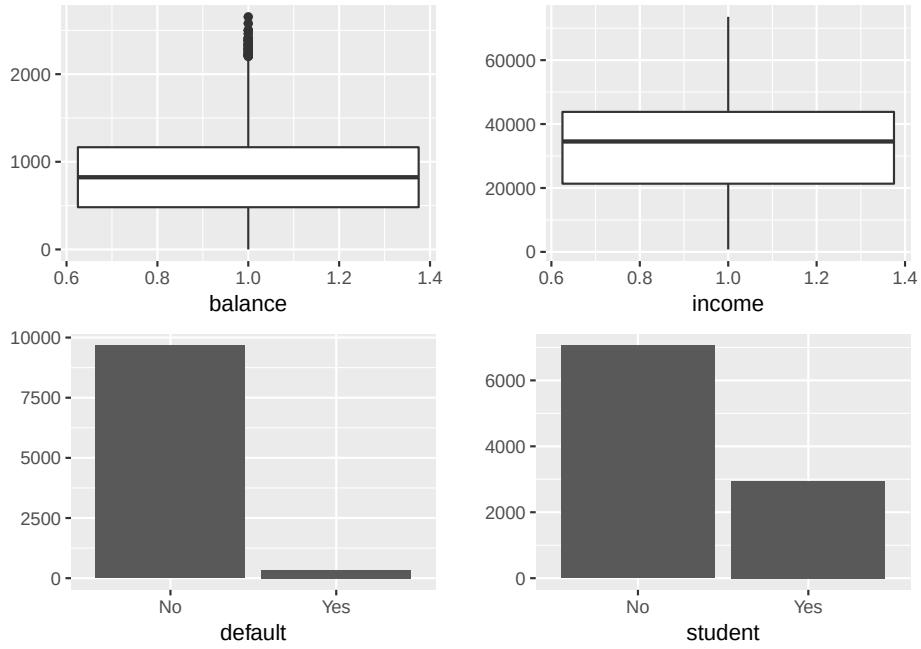


FIGURE 10.1 – Résumé des 4 variables de l'exemple de défaut bancaire.

ce qui est trop éloigné des hypothèses usuelles de normalité des résidus. De plus, la régression linéaire implique que $\mathbb{E}[Y_i|\mathbf{x}_i] = \mathbf{x}_i\theta$. Or $Y_i|\mathbf{x}_i \sim \mathcal{B}(\pi(\mathbf{x}_i))$ donc $\pi(\mathbf{x}_i) = \mathbf{x}_i\theta$. Cependant, rien n'indique que $\mathbf{x}_i\theta \in [0,1]$.

Les méthodes proposées partent du principe que le phénomène étudié est l'observation de Y_i (binaire), qui est la manifestation visible d'une variable Z_i latente (non observée) continue : $Y_i = \mathbb{1}_{Z_i>0}$. On considère un modèle linéaire entre Z_i et $\mathbf{x}_i$:

$$Z_i = \mathbf{x}_i\theta + \varepsilon_i.$$

On peut alors remarquer que

$$\pi(\mathbf{x}_i) = \mathbb{P}(Y_i = 1|\mathbf{x}_i) = \mathbb{P}(Z_i > 0|\mathbf{x}_i) = \mathbb{P}(-\varepsilon_i < \mathbf{x}_i\theta) = F(\mathbf{x}_i\theta),$$

où F est la fonction de répartition de $-\varepsilon_i$, qui correspond à l'inverse de la fonction de lien g.

Le choix du modèle porte donc sur le choix de cette fonction de répartition F ou de façon équivalente à la fonction de lien g. Dans le cadre binaire, les fonctions les plus usuellement utilisées sont (Figure 10.2) :

- **la fonction logistique** :

$$F(u) = \frac{e^u}{1+e^u} \quad \Longleftrightarrow \quad g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi).$$

Cette fonction est bien adaptée à la modélisation de probabilité car elle prend ses valeurs dans $[0,1]$. Dans ce cas, on parle de **modèle logistique**.

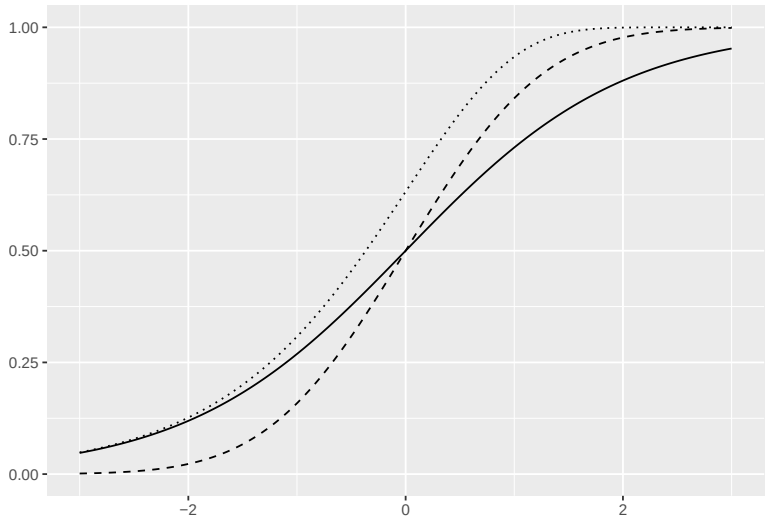


FIGURE 10.2 – Comparaison des fonctions de répartition logistique (ligne pleine), probit (tirets) et Gompit(pointillés).

- **la fonction probit** :
  F est la fonction de répartition de la loi $\mathcal{N}(0,1)$ et donc $g = F^{-1}$ est la fonction probit. Dans ce cas, on parle de **modèle probit**.
- **la fonction Gompit ou log-log** :
  F est la fonction de répartition de la loi de Gompertz

$$F(u) = 1 - \exp\left(-e^u\right) \quad \Longleftrightarrow \quad g(\pi) = \ln[-\ln(1-\pi)],$$

  mais cette fonction est dissymétrique. Dans ce cas, on parle de **modèle log-log**.

10.2 Odds et odds ratio

Il est souvent difficile d'interpréter directement les coefficients θ, l'interprétation se fait plutôt via les **odds ratio**. Ces odds ratio servent à mesurer l'effet d'une variable quantitative ou le contraste entre les effets d'une variable qualitative. L'idée générale est de raisonner en termes de probabilités ou de rapport de "chances" (odds).

Définition 10.1. *L'***odds** *(chance) pour un individu* $\mathbf{x}$ *d'obtenir la réponse* $Y = 1$ *est défini par :*

$$odds(\mathbf{x}) = \frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}, \qquad avec\ \pi(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x}).$$

L'odds ratio (rapport de chances) entre deux individus $\mathbf{x}$ *et* $\tilde{\mathbf{x}}$ *est*

$$OR(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{odds(\mathbf{x})}{odds(\tilde{\mathbf{x}})}.$$

Par exemple, si un joueur $\mathbf{x}$ a une probabilité $\pi(\mathbf{x}) = 1/4 = 0.25$ de gagner à un jeu, l'odds de succès vaut $0.25/0.75 = 1/3$. On dira que les chances de succès sont de 1 contre 3. Remarquons que, l'odds d'échec est de $0.75/0.25 = 3$ et on dira que les chances d'échec sont de 3 contre 1.

Les odds ratio peuvent être utilisés de plusieurs manières :

- Comparaison de probabilités de succès entre deux individus :

$$\begin{cases} \mathrm{OR}(\mathbf{x}, \tilde{\mathbf{x}}) > 1 & \Leftrightarrow & \pi(\mathbf{x}) > \pi(\tilde{\mathbf{x}}) \\ \mathrm{OR}(\mathbf{x}, \tilde{\mathbf{x}}) = 1 & \Leftrightarrow & \pi(\mathbf{x}) = \pi(\tilde{\mathbf{x}}) \\ \mathrm{OR}(\mathbf{x}, \tilde{\mathbf{x}}) < 1 & \Leftrightarrow & \pi(\mathbf{x}) < \pi(\tilde{\mathbf{x}}) \end{cases}$$

- Mesure de l'impact d'une variable : pour le modèle logistique avec intercept,

$$\mathrm{logit}[\pi(\mathbf{x})] = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)},$$

il est facile de vérifier que

$$\mathrm{OR}(\mathbf{x}, \tilde{\mathbf{x}}) = \prod_{j=1}^{p} \exp\left[\theta_j (x^{(j)} - \tilde{x}^{(j)})\right].$$

Pour mesurer l'influence d'une variable sur l'odds ratio, il suffit de considérer deux individus qui diffèrent uniquement sur la $j$ème variable. On obtient alors

$$\mathrm{OR}(\mathbf{x}, \tilde{\mathbf{x}}) = \exp\left[\theta_j (x^{(j)} - \tilde{x}^{(j)})\right].$$

Ainsi une variation de la $j$ème variable d'une unité correspond à un odds ratio $\exp(\theta_j)$ qui est uniquement fonction du coefficient θ_j. Le coefficient θ_j permet de mesurer l'influence de la $j$ème variable sur le rapport $\pi(\mathbf{x})/[1-\pi(\mathbf{x})]$ lorsque $x^{(j)}$ varie d'une unité, et ce, indépendamment de la valeur $x^{(j)}$. Une telle analyse peut se révéler intéressante pour étudier l'influence d'un changement d'état d'une variable qualitative.

- Interprétation d'un risque relatif : si $\pi(\mathbf{x})$ et $\pi(\tilde{\mathbf{x}})$ sont petits par rapport à 1 (ex pour une maladie rare), l'odds ratio peut être approché par $\pi(\mathbf{x})/\pi(\tilde{\mathbf{x}})$. On fait alors une interprétation simple : si $\mathrm{OR}(\mathbf{x}, \tilde{\mathbf{x}}) = 4$, la réponse $Y = 1$ est 4 fois plus probable en $\mathbf{x}$ qu'en $\tilde{\mathbf{x}}$.

10.3 Régression logistique simple

Dans cette section, on cherche à expliquer la variable réponse binaire Y par une seule variable explicative. Nous allons distinguer deux cas : celui où la variable explicative est quantitative et celui où elle est qualitative.

10.3.1 Avec une variable explicative quantitative

Dans notre exemple, nous allons chercher à expliquer la variable `default` à l'aide de la variable explicative `balance`. La Figure 10.3 montre qu'il est difficile de modéliser les données brutes mais si on regroupe les clients par classe de valeurs de `balance`, la liaison entre `balance` et `default` devient plus claire. Il apparait que lorsque le montant mensuel d'utilisation de la carte de crédit augmente, la proportion de clients en défaut augmente. Au vu de la forme de la courbe de liaison, une modélisation avec le lien logit semble "naturelle". On va donc chercher à modéliser l'espérance conditionnelle de Y_i sachant $\mathbf{x}_i = (1, x_i)$, par $\mathbb{E}[Y_i|\mathbf{x}_i] = \pi_\theta(\mathbf{x}_i)$, où

$$\pi_\theta(\mathbf{x}_i) = F(\theta_0 + \theta_1 x_i) = \frac{e^{\theta_0+\theta_1 x_i}}{1 + e^{\theta_0+\theta_1 x_i}} \quad \Longleftrightarrow \quad \ln\left(\frac{\pi_\theta(\mathbf{x_i})}{1 - \pi_\theta(\mathbf{x_i})}\right) = \theta_0 + \theta_1 x_i.$$

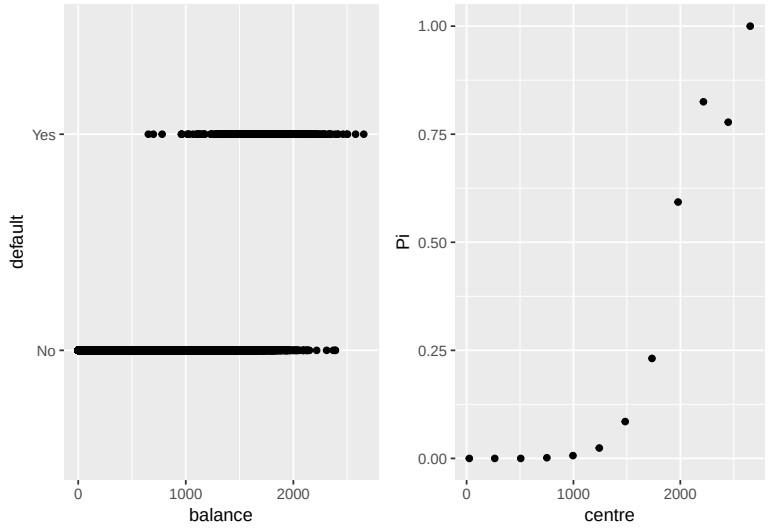


FIGURE 10.3 – A gauche, représentation de default en fonction de balance. A droite, proportion de clients en défaut par classe de valeurs pour balance.

10.3.1.1 Estimation des paramètres

Les paramètres $\theta = (\theta_0, \theta_1)$ sont estimés par la méthode du maximum de vraisemblance. La vraisemblance des données $\underline{Y} = (Y_1, \ldots, Y_n)$ est définie par :

$$L(\underline{Y}; \theta) = \prod_{i=1}^{n} \pi_\theta(\mathbf{x}_i)^{Y_i}[1 - \pi_\theta(\mathbf{x}_i)]^{1-Y_i},$$

et la log-vraisemblance par :

$$\begin{aligned} l(\underline{Y};\theta) &= \sum_{i=1}^{n} \left\{ Y_i \ln[\pi_\theta(\mathbf{x}_i)] + (1-Y_i)\ln[1-\pi_\theta(\mathbf{x}_i)] \right\} \\ &= \sum_{i=1}^{n} \left\{ Y_i \ln\left[F(\theta_0+\theta_1 x_i)\right] + (1-Y_i)\ln\left[1-F(\theta_0+\theta_1 x_i)\right] \right\} \end{aligned}$$

On cherche alors à annuler les dérivées partielles. On commence par remarquer que si $F(u) = e^u/(1+e^u)$, on a $F'(u) = F(u)\left[1-F(u)\right]$. D'où

$$\begin{aligned} \frac{\partial l(\underline{Y};\theta)}{\partial \theta_0} &= \sum_{i=1}^{n} \left[Y_i \frac{F'(\theta_0+\theta_1 x_i)}{F(\theta_0+\theta_1 x_i)} - (1-Y_i)\frac{F'(\theta_0+\theta_1 x_i)}{1-F(\theta_0+\theta_1 x_i)} \right] \\ &= \sum_{i=1}^{n} \left[Y_i \left[1-F(\theta_0+\theta_1 x_i)\right] - (1-Y_i)F(\theta_0+\theta_1 x_i) \right], \end{aligned}$$

et

$$\begin{aligned} \frac{\partial l(\underline{Y};\theta)}{\partial \theta_1} &= \sum_{i=1}^{n} \left[Y_i x_i \frac{F'(\theta_0+\theta_1 x_i)}{F(\theta_0+\theta_1 x_i)} - (1-Y_i)x_i\frac{F'(\theta_0+\theta_1 x_i)}{1-F(\theta_0+\theta_1 x_i)} \right] \\ &= \sum_{i=1}^{n} \left[x_i \left[Y_i - F(\theta_0+\theta_1 x_i)\right] \right]. \end{aligned}$$

On obtient donc le système suivant :

$$\left\{ \begin{array}{l} \sum_{i=1}^{n} \left[Y_i - F(\theta_0+\theta_1 x_i)\right] = 0 \\ \sum_{i=1}^{n} x_i \left[Y_i - F(\theta_0+\theta_1 x_i)\right] = 0 \end{array} \right. \iff \left\{ \begin{array}{l} \sum_{i=1}^{n} \left[Y_i - \pi_\theta(\mathbf{x}_i)\right] = 0 \\ \sum_{i=1}^{n} x_i \left[Y_i - \pi_\theta(\mathbf{x}_i)\right] = 0 \end{array} \right.$$

Pour mettre ensuite en place un algorithme de type Newton-Raphson ou de Fisher-scoring, on a besoin d'évaluer la matrice hessienne ou la matrice d'information de Fisher. Pour cela, on évalue les dérivées secondes :

$$\left\{ \begin{array}{l} \dfrac{\partial^2 l(\underline{Y};\theta)}{\partial \theta_0^2} = -\sum_{i=1}^{n} F'(\theta_0+\theta_1 x_i) = -\sum_{i=1}^{n} F(\theta_0+\theta_1 x_i)[1-F(\theta_0+\theta_1 x_i)] \\ \dfrac{\partial^2 l(\underline{Y};\theta)}{\partial \theta_1^2} = -\sum_{i=1}^{n} x_i^2 F'(\theta_0+\theta_1 x_i) = -\sum_{i=1}^{n} x_i^2 F(\theta_0+\theta_1 x_i)[1-F(\theta_0+\theta_1 x_i)] \\ \dfrac{\partial^2 l(\underline{Y};\theta)}{\partial \theta_0 \partial \theta_1} = -\sum_{i=1}^{n} x_i F'(\theta_0+\theta_1 x_i) = -\sum_{i=1}^{n} x_i F(\theta_0+\theta_1 x_i)[1-F(\theta_0+\theta_1 x_i)] \end{array} \right.$$

$$\mathcal{I}_n(\theta) = \begin{pmatrix} \sum_{i=1}^{n} \pi_\theta(\mathbf{x}_i)(1-\pi_\theta(\mathbf{x}_i)) & \sum_{i=1}^{n} x_i \pi_\theta(\mathbf{x}_i)(1-\pi_\theta(\mathbf{x}_i)) \\ \sum_{i=1}^{n} x_i \pi_\theta(\mathbf{x}_i)(1-\pi_\theta(\mathbf{x}_i)) & \sum_{i=1}^{n} x_i^2 \pi_\theta(\mathbf{x}_i)(1-\pi_\theta(\mathbf{x}_i)) \end{pmatrix} = (X'WX),$$

avec

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \text{ et } W = \text{diag}\big[\pi_\theta(\mathbf{x}_1)(1-\pi_\theta(\mathbf{x}_1)) \ , \ \ldots \ , \ \pi_\theta(\mathbf{x}_n)(1-\pi_\theta(\mathbf{x}_n))\big].$$

Sur notre exemple, on ajuste ce modèle entre la variable `default` et la variable `balance` avec la fonction `glm` de R :

```
> glm.balance <-glm(default~balance,data=Default,family=binomial(link="logit"))
> glm.balance$coefficients
  (Intercept)       balance
-10.651330614   0.005498917
```

10.3.1.2 Prédiction

Une fois le modèle ajusté, on obtient une estimation pour chaque prédicteur linéaire $\eta_i = \theta_0 + x_i\theta_1$ par $\hat{\eta}_i = \hat{\theta}_0 + x_i\hat{\theta}_1$ et pour chaque paramètre

$$\hat{\pi}(\mathbf{x}_i) = F(\hat{\eta}_i) = F(\hat{\theta}_0 + \hat{\theta}_1 x_i) = \pi_{\hat{\theta}}(\mathbf{x}_i).$$

En appliquant ensuite la règle de Bayes sur les $\hat{\pi}(\mathbf{x}_i)$, on récupère les valeurs ajustées $\widehat{Y}_i$ pour les Y_i :

$$\widehat{Y}_i = \begin{cases} 1 & \text{si } \hat{\pi}(\mathbf{x}_i) > 0.5 \\ 0 & \text{sinon.} \end{cases}$$

On peut alors comparer par une table de contingence les valeurs prédites par le modèle avec les valeurs observées des réponses. Dans notre exemple, on obtient

```
> hatpi <- glm.balance$fitted.values
> hatY <- (hatpi > 0.5)
> table(default,hatY)
       hatY
default FALSE TRUE
    No   9625   42
    Yes   233  100
```

On constate que l'on retrouve assez bien les clients n'ayant pas de défaut, mais mal ceux avec un défaut.

Si on se donne maintenant un nouvel individu décrit par $\mathbf{x}_0 = (1, x_0)$ alors le modèle ajusté permet de prédire une proportion $\hat{\pi}(\mathbf{x}_0) = F(\hat{\theta}_0 + \hat{\theta}_1 x_0)$ et une réponse prédite $\widehat{Y}_0 = \mathbb{1}_{\hat{\pi}(\mathbf{x}_0)>0.5}$.

La Figure 10.4 représente l'ajustement des proportions estimées par le modèle logistique et par le modèle probit avec les proportions observées.

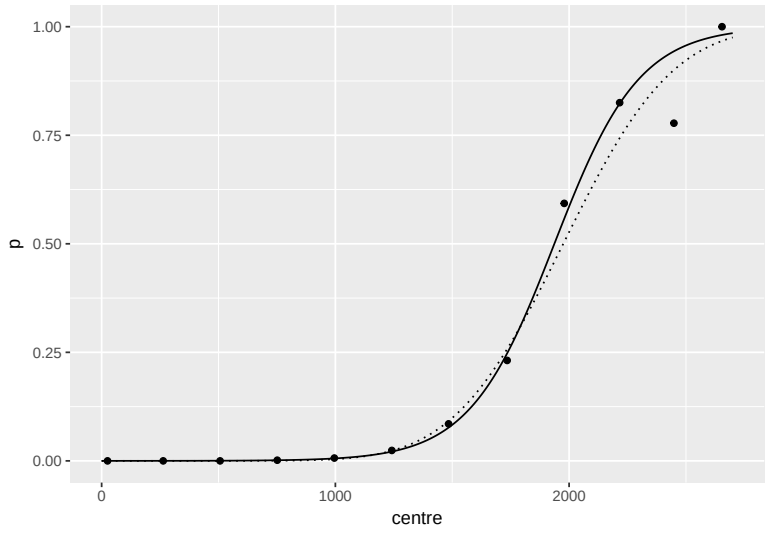


FIGURE 10.4 – Représentation des proportions prédites par le modèle logistique (en trait plein) et par le modèle probit (en pointillé).

10.3.1.3 Intervalle de confiance

Pour obtenir un intervalle de confiance pour chaque θ_j, on peut utiliser la méthode de Wald ou la méthode fondée sur le rapport de vraisemblance présentées en section 9.6. Sous R, le premier est obtenu avec la commande `confint.default`, le second avec la commande `confint`. Sur notre exemple, on obtient :

```
> confint(glm.balance)

                    2.5 %       97.5 %
(Intercept) -11.383288936 -9.966565064
balance       0.005078926  0.005943365
> confint.default(glm.balance)

                    2.5 %       97.5 %
(Intercept) -11.359186056 -9.943475172
balance       0.005066999  0.005930835
```

10.3.1.4 Test de nullité des paramètres

Si l'on souhaite tester $\mathcal{H}_0 : \theta_j = 0$ contre $\mathcal{H}_1 : \theta_j \neq 0$, il suffit de remarquer que sous $\mathcal{H}_0$,

$$\frac{\hat{\theta}_j}{\hat{\sigma}_j} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

avec $\hat{\sigma}_j = \sqrt{[\mathcal{I}(\hat{\theta}_{MV})^{-1}]_{jj}}$. On peut alors construire un test asymptotique de niveau α de zone de rejet

$$\mathcal{R}_\alpha = \left\{ \left| \hat{\theta}_j / \hat{\sigma}_j \right| > z_{1-\alpha/2} \right\},$$

qui correspond au Z-test. Pour notre exemple, on obtient

```
> summary(glm.balance)

[...]

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8
```

Les p-valeurs étant $< 2e-16$, on rejette la nullité pour les deux coefficients.

On peut aussi voir le problème comme un test de sous-modèle et le résoudre avec un test de Wald ou un test du rapport du maximum de vraisemblance. Par exemple, pour tester la nullité de θ_1 :

```
> anova(glm(default~1,data=Default,family=binomial(link="logit")), glm.balance,
+       test="Chisq")
Analysis of Deviance Table

Model 1: default ~ 1
Model 2: default ~ balance
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      9999     2920.7
2      9998     1596.5  1   1324.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

10.3.2 Avec une variable explicative qualitative

On va chercher ici à expliquer la variable `default` par rapport à la variable `student` (binaire). On peut considérer le modèle suivant :

$$\text{logit}(\pi(\mathbf{x}_i)) = \ln\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \theta_0 + \theta_1 \mathbb{1}_{x_i=1} + \theta_2 \mathbb{1}_{x_i=0}.$$

Mais on peut remarquer qu'il est possible d'écrire le modèle également sous la forme

$$\text{logit}(\pi(\mathbf{x}_i)) = \ln\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = (\theta_0 + \theta_2) + (\theta_1 - \theta_2)\mathbb{1}_{x_i=1} + 0\ \mathbb{1}_{x_i=0}.$$

On se retrouve comme pour l'analyse de variance avec un modèle non identifiable. Il faut donc imposer une contrainte sur les paramètres pour le rendre identifiable. Par exemple si on suppose que $\theta_2 = 0$, le modèle devient $\text{logit}(\pi(\mathbf{x}_i)) = \theta_0$ si $x_i = 0$ et $\text{logit}(\pi(\mathbf{x}_i)) = \theta_0 + \theta_1$ si $x_i = 1$. Il faut donc adapter l'interprétation des paramètres selon la contrainte considérée.

On peut alors reprendre les mêmes raisonnements et les mêmes calculs que dans la section 10.3.1 pour estimer les paramètres, construire un intervalle de confiance pour chacun des paramètres, tester la nullité de chacun des paramètres, etc.

Dans notre exemple traité sous R, le modèle considéré est celui avec la contrainte $\theta_2 = 0$:

```
> table(default,student)
       student
default   No  Yes
    No  6850 2817
    Yes  206  127

> glm.student = glm(default~student, data=Default, family=binomial)
> glm.student$coefficients
(Intercept)  studentYes
 -3.5041278   0.4048871
```

Exercice 39. *Dans cet exemple montrez que* $odds(x_i = 0) = e^{\theta_0}$, $odds(x_i = 1) = e^{\theta_0+\theta_1}$ *et* $OR(x_i = 1, x_i = 0) = e^{\theta_1}$. *Contrôlez ce résultat sur la sortie de R suivante :*

```
> new.data=data.frame(student=factor(c("No","Yes")))
> inv.logit(predict(glm.student,new.data)) # nécessite la librairie boot
         1          2
0.02919501 0.04313859

> exp(c(coef(glm.student),sum=sum(coef(glm.student))))
(Intercept)  studentYes         sum
 0.03007299  1.49913321  0.04508342
```

Si on n'est pas étudiant, $\text{logit}(\hat{\pi}) = \hat{\theta}_0$ donc $\hat{\pi} = 0.029$ alors que si l'on est étudiant, $\text{logit}(\hat{\pi}) = \hat{\theta}_0 + \hat{\theta}_1$ d'où $\hat{\pi} = 0.043$. Ainsi,

$$\begin{cases} \text{odds("étudiant")} = 0.045, \\ \text{odds("non étudiant")} = 0.030, \\ \text{OR("étudiant", "non étudiant")} = 1.5. \end{cases}$$

Ainsi un étudiant a 1.5 fois plus de chance d'être en défaut qu'une personne non étudiante.

On peut également faire un test de sous-modèle :

```
> anova(glm(default~1, data=Default, family=binomial), glm.student, test="Chisq")
Analysis of Deviance Table

Model 1: default ~ 1
Model 2: default ~ student
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      9999     2920.7
2      9998     2908.7  1   11.967 0.0005416 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On en déduit que le modèle prenant en compte la variable `student` est meilleur que le modèle nul.

10.4 Régression logistique multiple

Dans cette section, nous allons considérer le cas plus général où l'on souhaite expliquer la variable réponse binaire Y par rapport à p régresseurs $x^{(1)}, \ldots, x^{(p)}$. Dans l'exemple, on a $p = 3$ régresseurs, une variable qualitative (`student` $= x^{(1)}$) et deux variables quantitatives (`balance` $= x^{(2)}$, `income`$=x^{(3)}$). Nous allons aborder plusieurs questions au travers de l'étude de cet exemple.

10.4.1 Modèle sans interaction

Dans un premier temps, on considère le modèle complet sans interaction suivant :

$$\text{logit}(\pi(\mathbf{x}_i)) = \theta_0 + \theta_1 \, \mathbb{1}_{x_i^{(1)}=1} + \theta_2 \, x_i^{(2)} + \theta_3 \, x_i^{(3)}.$$

Le code suivant sous R permet d'ajuster ce modèle sans interaction :

```
> glm.additif<-glm(default~.,data=Default,family=binomial(link="logit"))
> summary(glm.additif)

Call:
glm(formula = default ~ ., family = binomial(link = "logit"),
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

On obtient les estimations pour chacun des paramètres. On constate que si on teste la nullité individuellement de chaque paramètre par le test de Wald ou le Z-test, on rejette la nullité de $\theta_0, \theta_1, \theta_2$ et on accepte la nullité de θ_3.

10.4.1.1 Tests successifs de modèles emboîtés

```
> anova(glm.additif,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: default

Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     9999     2920.7
student  1    11.97      9998     2908.7 0.0005416 ***
balance  1  1337.00      9997     1571.7 < 2.2e-16 ***
income   1     0.14      9996     1571.5 0.7115139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Exercice 40. *A quoi correspondent les différentes quantités dans la sortie de R ci-dessus ?*

Remarquons que l'on retrouve le test du rapport de vraisemblance de $\mathcal{H}_0 : \theta_3 = 0$ contre $\mathcal{H}_1 : \theta_3 \neq 0$, sur la dernière ligne, que l'on aurait pu directement obtenir en faisant :

```
> glm.sansincome<-glm(default~student+balance,data=Default,family=binomial(link="logit"))
> anova(glm.sansincome,glm.additif,test="Chisq")
Analysis of Deviance Table

Model 1: default ~ student + balance
Model 2: default ~ student + balance + income
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      9997     1571.7
2      9996     1571.5  1  0.13677   0.7115
```

10.4.1.2 Test de nullité de plusieurs coefficients simultanément.

Testons la nullité de θ_2 et θ_3 simultanément. On peut le voir comme un test de sous-modèle comparant le modèle complet sans interaction avec le modèle (introduit en section 10.3.2)

$$\text{logit}(\pi(\mathbf{x}_i)) = \theta_0 + \theta_1 \mathbb{1}_{x_i^{(1)}=1}.$$

```
> anova(glm.student,glm.additif,test="Chisq")
Analysis of Deviance Table

Model 1: default ~ student
Model 2: default ~ student + balance + income
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      9998     2908.7
2      9996     1571.5  2   1337.1 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On rejette la nullité simultanée de ces deux paramètres. On peut aussi voir l'hypothèse nulle sous la forme

$$\mathcal{H}_0 : C\theta = 0_2 \text{ avec } C = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

```
> hattheta <- glm.additif$coefficients
> hatpi <- glm.additif$fitted.values
> W <- diag(hatpi*(1-hatpi))
> X <- cbind(rep(1,10000),student,balance,income)
> In <- t(X) %*% W %*% X
> C <- matrix(c(0,0,1,0,0,0,0,1),nrow=4)
> t(t(C)%*%hattheta) %*% solve(t(C)%*%solve(In)%*%C) %*% (t(C)%*%hattheta) > qchisq(0.95,df=2)
     [,1]
[1,] TRUE
```

10.4.1.3 Sélection de variables

Plus généralement, on peut envisager une procédure de sélection de variables. On choisit ici de faire une sélection de variable descendante avec le critère AIC :

```
> step.backward <- step(glm.additif)
Start:  AIC=1579.54
default ~ student + balance + income

          Df Deviance    AIC
- income   1   1571.7 1577.7
<none>         1571.5 1579.5
- student  1   1579.0 1585.0
- balance  1   2907.5 2913.5

Step:  AIC=1577.68
default ~ student + balance

          Df Deviance    AIC
<none>         1571.7 1577.7
- student  1   1596.5 1600.5
- balance  1   2908.7 2912.7
```

On peut également utiliser la fonction `stepAIC` de la librairie `MASS` avec le critère AIC (option "p=2") ou BIC (option "p=log(n))

```
library(MASS)
stepAIC(glm.additif, direction=c("backward"),p=2) # AIC
stepAIC(glm.additif, direction=c("backward"),p=log(nrow(Default))) # BIC
```

Pour les trois procédures, le modèle sélectionné est également celui sans la variable `income`.

10.4.2 Modèle avec interactions

On va considérer ici le modèle complet avec toutes les interactions (d'ordre 2) entre variables et on met en place une procédure de sélection de variables pour déterminer un modèle plus simple pour expliquer la variable réponse `default`. Le modèle s'écrit :

$$\text{logit}(\pi(\mathbf{x}_i)) = \theta_0 + \theta_2 x_i^{(2)} + \theta_3 x_i^{(3)} + \theta_{23} x_i^{(2)} x_i^{(3)} + (\beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)}) \mathbb{1}_{x_i^{(1)}=1}.$$

On commence par ajuster le modèle complet avec interactions avec le code suivant :

```
> glm.complet<-glm(default~.^2,data=Default,family="binomial")
> summary(glm.complet)

Call:
glm(formula = default ~ .^2, family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4848  -0.1417  -0.0554  -0.0202   3.7579

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.104e+01  1.866e+00  -5.914 3.33e-09 ***
studentYes         -5.201e-01  1.344e+00  -0.387    0.699
balance             5.882e-03  1.180e-03   4.983 6.27e-07 ***
income              4.050e-06  4.459e-05   0.091    0.928
studentYes:balance -2.551e-04  7.905e-04  -0.323    0.747
studentYes:income   1.447e-05  2.779e-05   0.521    0.602
balance:income     -1.579e-09  2.815e-08  -0.056    0.955
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.1  on 9993  degrees of freedom
AIC: 1585.1

Number of Fisher Scoring iterations: 8
```

On peut remarquer que le test de nullité de chaque paramètre individuellement est accepté pour plusieurs des paramètres. On poursuit notre étude en cherchant à simplifier le modèle par sélection de variable. La Figure 10.5 donne les résultats en utilisant le critère AIC. Une fois encore c'est le modèle additif avec seulement les variables `student` et `balance` qui est retenu.

10.4.3 Etude complémentaire du modèle retenu

On commence par ajuster le modèle retenu

$$\text{logit}(\pi(\mathbf{x}_i)) = \theta_0 + \theta_1 \mathbb{1}_{x_i^{(1)}=1} + \theta_2 x_i^{(2)}$$

à l'aide du code en Figure 10.6

On peut vérifier graphiquement la non-interaction entre les variables `student` et `balance`. Sur la Figure 10.7, on constate que les droites représentant $\text{logit}(\pi_\theta(\cdot))$ en fonction de la variable `balance` en distinguant selon la valeur de la variable `student` sont parallèles.

Afin de comparer les valeurs de la variable réponse et les valeurs prédites par le modèle, on forme la table de contingence :

```
> hatpi <- glm.final$fitted.values
> table(default,hatpi>0.5)

default FALSE TRUE
    No   9628   39
    Yes   228  105
```

```
> step(glm.complet)
Start:  AIC=1585.07
default ~ (student + balance + income)^2

                  Df Deviance    AIC
- balance:income   1   1571.1 1583.1
- student:balance  1   1571.2 1583.2
- student:income   1   1571.3 1583.3
<none>                 1571.1 1585.1

Step:  AIC=1583.07
default ~ student + balance + income + student:balance + student:income

                  Df Deviance    AIC
- student:balance  1   1571.3 1581.3
- student:income   1   1571.3 1581.3
<none>                 1571.1 1583.1

Step:  AIC=1581.28
default ~ student + balance + income + student:income

                 Df Deviance    AIC
- student:income  1   1571.5 1579.5
<none>                1571.3 1581.3
- balance         1   2907.3 2915.3

Step:  AIC=1579.54
default ~ student + balance + income

          Df Deviance    AIC
- income   1   1571.7 1577.7
<none>         1571.5 1579.5
- student  1   1579.0 1585.0
- balance  1   2907.5 2913.5

Step:  AIC=1577.68
default ~ student + balance

          Df Deviance    AIC
<none>         1571.7 1577.7
- student  1   1596.5 1600.5
- balance  1   2908.7 2912.7

Call:  glm(formula = default ~ student + balance, family = "binomial",
    data = Default)

Coefficients:
(Intercept)   studentYes      balance
 -10.749496    -0.714878     0.005738

Degrees of Freedom: 9999 Total (i.e. Null);  9997 Residual
Null Deviance:            2921
Residual Deviance: 1572         AIC: 1578
```

FIGURE 10.5 – Résultats de la sélection de variables avec AIC sur le modèle complet

```
> glm.final = glm(default ~ student + balance, data=Default, family=binomial)
> summary(glm.final)

Call:
glm(formula = default ~ student + balance, family = binomial,
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4578  -0.1422  -0.0559  -0.0203   3.7435

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.7  on 9997  degrees of freedom
AIC: 1577.7

Number of Fisher Scoring iterations: 8
```

FIGURE 10.6 – Résultats pour le modèle retenu

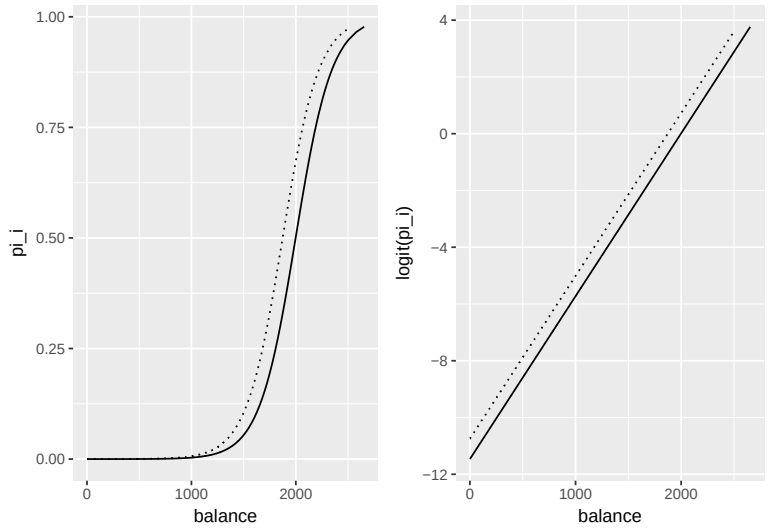


FIGURE 10.7 – Tracé de $\pi_{\hat{\theta}}(\cdot)$ (à gauche) et $\text{logit}(\pi_{\hat{\theta}}(\cdot))$ (à droite) en fonction de la variable balance en distinguant selon la valeur de la variable student (student = 1 en ligne pleine et = 0 en pointillé).

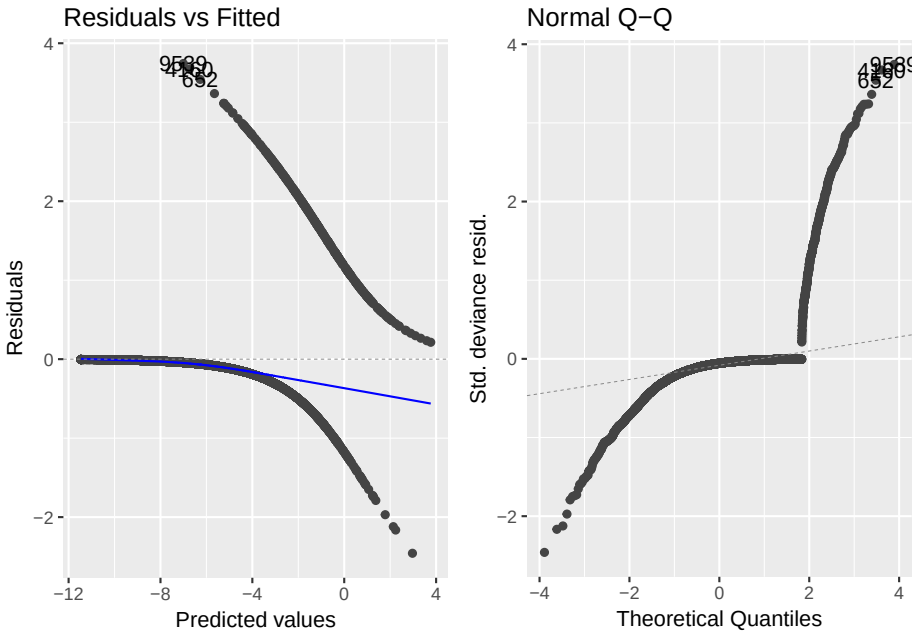


FIGURE 10.8 – Représentation des résidus.

On constate que l'on retrouve assez bien les clients sans défaut sur leur dette, par contre la détection des clients étant en défaut sur leur dette sont très mal prédits.

Concernant les résidus du modèle ajusté, on peut calculer les différents résidus introduits dans le chapitre général (section 9.8). La Figure 10.8 représente un diagnostic graphique des résidus dans notre exemple.

```
library(boot)
#residus y_i - \hat\mu_i
res<-residuals(glmfinal,type="response")
#residus de deviance
res_dev<-residuals(glmfinal)
#residus de Pearson
res_pear<-residuals(glmfinal,type="pearson")
#residus de deviance standardisés
res_dev_stand<-rstandard(glmfinal)
    res_dev_stand<-glm.diag(glmfinal)$rd
# residus de Pearson standardisés
H<-influence(glmfinal)$hat
res_pear_stand<-res_pear/sqrt(1-H)
  res_pear_stand<-glm.diag(glmfinal)$rp
# residus de Jackknife
res_Jackknife<-glm.diag(glmfinal)$res
```

10.5 Régression polytomique

On souhaite étendre la régression logistique au cas de l'étude d'une variable réponse qualitative Y pouvant prendre M modalités $u_1, \ldots, u_M$ avec $M > 2$. On note $x^{(1)}, \ldots, x^{(p)}$ les variables explicatives.

10.5.1 Régression multinomiale ou polytomique non-ordonnée

Dans cette section, on considère l'exemple suivant :
On considère $n = 735$ clients. On souhaite étudier pour un produit le choix du client entre la marque à petit prix, la marque de l'enseigne ou la marque de référence du marché. On souhaite savoir si l'âge et/ou le genre du client a une influence sur son choix.

```
> D <- read.table("MarqueSexe.csv",header=T,sep=";")
> D[,"femme"] <- as.factor(D[,"femme"])
> D[,1] <- factor(D[,1], levels = c("Reference","Enseigne","PetitPrix"))
> summary(D)
       brand     femme        age
 Reference :221   0:269   Min.   :24.0
 Enseigne  :307   1:466   1st Qu.:32.0
 PetitPrix :207           Median :32.0
                          Mean   :32.9
                          3rd Qu.:34.0
                          Max.   :38.0
```

La variable réponse Y est ici nominale, c'est-à-dire que les M modalités n'ont pas de lien hiérarchique, pas d'ordre. On parle alors de **régression multinomiale (ou polytomique non-ordonnée)**. Comme dans le cas binaire, on cherche à modéliser

$$\pi_m(\mathbf{x}) = \mathbb{P}(Y = u_m|\mathbf{x}),\ \forall m \in \{1, \ldots, M\}.$$

Comme $\sum_{m=1}^{M} \pi_m(\mathbf{x}) = 1$, il suffit de modéliser $(M-1)$ des $\pi_m(\mathbf{x})$. La catégorie de référence s'impose souvent par le contexte d'étude : les non-malades contre les différents types de maladie ; le produit phare du marché contre les produits outsiders, etc. Dans notre exemple, la modalité "Référence" sera considérée comme la modalité de référence.

Dans la suite, nous considérons la première modalité comme référence. On définit alors le modèle de régression multinomiale par

$$\ln\left[\frac{\pi_m(\mathbf{x})}{\pi_1(\mathbf{x})}\right] = \theta_0^{(m)} + \theta_1^{(m)}x^{(1)} + \ldots + \theta_p^{(m)}x^{(p)} = \mathbf{x}\theta^{(m)},\ \forall m \in \{2, \ldots, M\}$$

avec $\mathbf{x} = (1, x^{(1)}, \ldots, x^{(p)})$ et $\theta^{(m)} = (\theta_0^{(m)}, \theta_1^{(m)}, \ldots, \theta_p^{(m)})'$ paramètres inconnus. Ceci revient à

$$\pi_m(\mathbf{x}) = \frac{\exp(\mathbf{x}\theta^{(m)})}{1 + \sum_{m'=2}^{M} \exp(\mathbf{x}\theta^{(m')})}.$$

On peut remarquer que pour $M = 2$, $u_1 = 0$ et $u_2 = 1$, on retrouve le modèle de régression logistique.

Pour estimer les paramètres $\theta^{(m)}$ pour $m \in \{2, \ldots, M\}$, on cherche à maximiser la vraisemblance du modèle

$$L(\underline{Y}|\theta) = \prod_{i=1}^{n}\prod_{m=1}^{M} \pi_m(\mathbf{x}_i)^{\mathbb{1}_{Y_i=u_m}}.$$

On retrouve la vraisemblance de n lois multinomiales de paramètres $(1, (\pi_1(\mathbf{x}_i), \ldots, \pi_M(\mathbf{x}_i))$ pour $1 \leq i \leq n$.

Comme pour le cas binaire, il n'y a pas de solutions explicites pour les estimateurs, on utilise donc des méthodes numériques pour les évaluer. A partir des estimateurs $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)}$, on en déduit un estimateur pour chaque $\pi_m(\mathbf{x})$:

$$\begin{cases} \hat{\pi}_m(\mathbf{x}) = \dfrac{\exp(\mathbf{x}\hat{\theta}^{(m)})}{1 + \sum_{m'=2}^{M} \exp(\mathbf{x}\hat{\theta}^{(m')})}, \forall m \in \{2, \ldots, M\} \\ \hat{\pi}_1(\mathbf{x}) = 1 - \displaystyle\sum_{m=2} \hat{\pi}_m(\mathbf{x}). \end{cases}$$

On peut ensuite s'intéresser à la prédiction. Sachant qu'un individu prend les valeurs $\mathbf{x}_0 = (1, x_0^{(1)}, \ldots, x_0^{(p)})$ on peut faire de la prédiction :

- sur la probabilité que $Y_0 = u_m$ pour $m \in \{1, \ldots, M\}$: $\hat{\pi}_m(\mathbf{x}_0)$.
- sur la modalité de Y la plus probable pour $\mathbf{x}_0$

$$\hat{Y}_0 = u_{\hat{m}_0} \text{ avec } \hat{m}_0 = \underset{m \in \{1,\ldots,M\}}{\operatorname{argmax}} \hat{\pi}_m(\mathbf{x}_0).$$

Sous des conditions similaires au cas binaire, on récupère les mêmes résultats sur le comportement asymptotique des estimateurs. On peut alors utiliser de la même façon les tests de Wald, du rapport de maximum de vraisemblance, etc.

Pour interpréter les paramètres, on peut utiliser les odds ratio. On définit l'odds d'une modalité m_1 contre une modalité m_2 par

$$\text{odds}(Y = u_{m_1} \text{ vs } Y = u_{m_2}; \mathbf{x}) = \frac{\mathbb{P}(Y = u_{m_1}|\mathbf{x})}{\mathbb{P}(Y = u_{m_2}|\mathbf{x})} = \frac{\pi_{m_1}(\mathbf{x})}{\pi_{m_2}(\mathbf{x})} = \exp[\mathbf{x}(\theta^{(m_1)} - \theta^{(m_2)})].$$

Pour deux individus x et $\tilde{x}$, on définit alors l'odds ratio par

$$\begin{aligned} \text{OR}(Y = u_{m_1} \text{ vs } Y = u_{m_2}; \mathbf{x}, \tilde{\mathbf{x}}) &= \frac{\text{odds}(Y = u_{m_1} \text{ vs } Y = u_{m_2}; \mathbf{x})}{\text{odds}(Y = u_{m_1} \text{ vs } Y = u_{m_2}; \tilde{\mathbf{x}})} \\ &= \exp[(\mathbf{x} - \tilde{\mathbf{x}})(\theta^{(m_1)} - \theta^{(m_2)})]. \end{aligned}$$

Ainsi si les deux individus $\mathbf{x}$ et $\tilde{\mathbf{x}}$ ne diffèrent que d'une unité pour la variable j, on a

$$\text{OR}(Y = u_{m_1} \text{ vs } Y = u_{m_2}; \mathbf{x}, \tilde{\mathbf{x}}) = \exp[\theta_j^{(m_1)} - \theta_j^{(m_2)}].$$

En pratique, on peut utiliser les fonctions `multinom` ou `vglm` des librairires `nnet` et `VGAM` pour ajuster un modèle de régression polytomique non-ordonnée sous R.

Sur notre exemple, une régression multinomiale est mise en oeuvre avec la fonction `multinom` :

```
> library(nnet)
> regMarq <- multinom(brand ~ femme + age,data=D,Hess=T)
```

On affiche les estimations obtenues pour les paramètres $\theta^{(m)}$ pour $m = 2, 3$:

```
> summary(regMarq)
Call:
multinom(formula = brand ~ femme + age, data = D, Hess = T)

Coefficients:
          (Intercept)      femme1        age
Enseigne     10.94688  0.05798805 -0.3177081
PetitPrix    22.72150 -0.46576724 -0.6859142

Std. Errors:
          (Intercept)    femme1        age
Enseigne     1.493166 0.1964261 0.04400704
PetitPrix    2.058030 0.2260886 0.06262666

Residual Deviance: 1405.941
AIC: 1417.941
```

Par exemple, on a $\ln[\pi_{\texttt{petitprix}}/\pi_{\texttt{Reference}}] = 22.72 - 0.47\texttt{Femme} - 0.69\texttt{age}$. Les femmes ont moins confiance dans la marque petit prix par rapport à la marque de référence et plus l'âge augmente, moins le client fait le choix de la marque petit prix par rapport à la marque de référence.

On peut alors calculer les estimations pour les $\pi_m(\mathbf{x}_i)$. Pour chaque individu du jeu de données, on récupère les probabilités avec `regMarq$fitted.values` :

```
> head(regMarq$fitted.values)
    Reference   Enseigne PetitPrix
1 0.001812569 0.05023105 0.9479564
2 0.006741608 0.09896542 0.8942930
3 0.006741608 0.09896542 0.8942930
4 0.018431742 0.20868509 0.7728832
5 0.018431742 0.20868509 0.7728832
6 0.012740144 0.13611798 0.8511419
```

Ainsi pour le premier client, il fera plutôt le choix de la marque petit prix.

On obtient les prédictions avec `> apply(regMarq$fitted.values,1,which.max)` ou directement avec la commande `> pr = predict(regMarq, D)`. La matrice de confusion vaut alors

```
> table(D[,1], pr)
           pr
        Reference Enseigne PetitPrix
  Reference   110      101        10
  Enseigne     51      238        18
  PetitPrix    13      136        58
```

On peut déterminer un intervalle de confiance pour chacun des paramètres avec la commande :

```
> confint(regMarq)
, , Enseigne

                2.5 %     97.5 %
(Intercept)  8.0203294 13.8734340
femme1      -0.3270001  0.4429762
age         -0.4039603 -0.2314558

, , PetitPrix

                2.5 %      97.5 %
(Intercept) 18.6878390 26.75516759
femme1      -0.9088928 -0.02264168
age         -0.8086602 -0.56316817
```

Pour tester la nullité de chaque coefficient $\theta_j^{(m)} = 0$ contre $\theta_j^{(m)} \neq 0$, on fait un test de Wald (ou Z-test). On obtient les p-valeurs de chaque test avec les commandes suivantes :

```
> z = summary(regMarq)$coeff / summary(regMarq)$standard.errors
> pvaleur = 2 * (1 - pnorm(abs(z), 0, 1))
> pvaleur
           (Intercept)     femme1          age
Enseigne  2.278178e-13 0.76782920 5.218048e-13
PetitPrix 0.000000e+00 0.03938811 0.000000e+00
```

On peut aussi utiliser la fonction `coeftest` de la librairie `AER` :

```
> library(AER)
> coeftest(regMarq)

z test of coefficients:

                      Estimate Std. Error  z value  Pr(>|z|)
Enseigne:(Intercept)  10.946882   1.493166   7.3313 2.279e-13 ***
Enseigne:femme1        0.057988   0.196426   0.2952   0.76783
Enseigne:age          -0.317708   0.044007  -7.2195 5.219e-13 ***
PetitPrix:(Intercept) 22.721503   2.058030  11.0404 < 2.2e-16 ***
PetitPrix:femme1      -0.465767   0.226089  -2.0601   0.03939 *
PetitPrix:age         -0.685914   0.062627 -10.9524 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On peut remarquer que les variances de chaque coefficient sont données sur la diagonale de l'inverse de la matrice hessienne. Par exemple, pour `"femme-enseigne"` on retrouve bien l'écart-type 0.1964 :

```
> Sigma <- solve(regMarq$Hessian)
> sqrt(Sigma[2,2])
[1] 0.1964261
```

Pour tester le modèle considéré contre le sous-modèle trivial (seulement l'intercept $\theta_0^{(m)}$), on considère le test du rapport de vraisemblance :

```
> regMarq0 = multinom(brand ~1,data=D)
> rv = regMarq0$deviance - regMarq$deviance
> ddl = regMarq$edf - regMarq0$edf
> pvaleur = 1 - pchisq(rv, ddl)
> print(c(rv,ddl,pvaleur))
[1] 185.8502   4.0000   0.0000
```

La différence des déviances vaut 185.85. La statistique de test suit une loi du chi-deux à $[n-(M-1)]-[n-(p+1)(M-1)]=p(M-1)=4$ degrés de liberté. La p-valeur étant très faible, on rejette l'hypothèse nulle.

On veut maintenant tester si la variable sexe joue un rôle dans le choix des clients. On va donc tester si $\mathcal{H}_0 : \theta_1^{(2)} = \theta_1^{(3)} = 0$ contre $\mathcal{H}_1 : \exists m; \theta_1^{(m)} \neq 0$. La statistique du test $(\hat{\theta}_1^{(2)}, \hat{\theta}_1^{(3)})\hat{\Sigma}_1^{-1}(\hat{\theta}_1^{(2)}, \hat{\theta}_1^{(3)})'$ suit une loi du chi-deux à 2 degrés de liberté, où $\hat{\Sigma}_1$ est la matrice de variance-covariance pour la variable `"femme"`. On peut aussi le voir comme un test de sous-modèle avec le test du rapport de vraisemblance.

```
# TEST 1
> regMarq1 = multinom(brand~age,data=D)
> rv1 = regMarq1$deviance - regMarq$deviance
> ddl = regMarq$edf - regMarq1$edf
> pvaleur = 1 - pchisq(rv1, ddl)
> print(c(rv1,ddl,pvaleur))
[1] 7.65119936 2.00000000 0.02180536

> # TEST 2
> thetafemme <- summary(regMarq)$coeff[,2]
> Sigmafemme <- Sigma[c(2,5),c(2,5)]
> Wfemme <- thetafemme %*% solve(Sigmafemme) %*% thetafemme
> Wfemme > qchisq(0.95,2)    # test à 5%
     [,1]
[1,] TRUE
```

Comme la p-valeur vaut 0.0218, on rejette l'absence d'effet de la variable sexe à 5%.

Pour interpréter les coefficients, on peut revenir aux odds ratios. Par exemple si on regarde l'odds de petits prix et marque vs la référence pour une femme avec même âge, on peut les calculer avec $\exp(\theta_1^{(m)})$:

```
> exp(summary(regMarq)$coeff[,2])
 Enseigne PetitPrix
1.0597023 0.6276534
```

Une femme a 0.628 fois plus de chances qu'un homme de préférer la marque petit prix à la marque référence.

On peut faire le même travail avec la fonction `vglm` de la librairie `VGAM`, voir Figure10.9. Notons que cette fonction prend la dernière modalité comme référence.

10.5.2 Régression polytomique ordonnée

Dans cette section, on suppose que Y est une variable qualitative ordinale, c'est-à-dire que Y admet M modalités ordonnées $u_1 \prec u_2 \prec \ldots \prec u_M$. Par exemple, si on s'intéresse au degré de satisfaction pour un produit (mauvais, moyen, bon, très bon) ; pour le stade d'évolution d'une maladie ; etc. On va illustrer cette section avec le jeu de données suivant : on s'intéresse à la qualité de 34 vins selon l'ensoleillement et la pluviométrie (les variables explicatives sont centrées réduites).

```
> library(VGAM)
> D1 <- D
> D1[,1] <- factor(D[,1], levels=c("PetitPrix","Enseigne","Reference"))
> regMarq2 <- vglm(brand ~ femme + age, data=D1, family=multinomial())
> summary(regMarq2)

Call:
vglm(formula = brand ~ femme + age, family = multinomial(), data = D1)

Pearson residuals:
                      Min      1Q  Median     3Q   Max
log(mu[,1]/mu[,3]) -5.563 -0.4433 -0.3237 0.5547 7.772
log(mu[,2]/mu[,3]) -4.722 -0.6800 -0.4469 0.9729 1.786

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 22.72140    2.05802  11.040  < 2e-16 ***
(Intercept):2 10.94674    1.49316   7.331 2.28e-13 ***
femme1:1      -0.46594    0.22609  -2.061   0.0393 *
femme1:2       0.05787    0.19643   0.295   0.7683
age:1         -0.68591    0.06263 -10.952  < 2e-16 ***
age:2         -0.31770    0.04401  -7.219 5.22e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  2

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Residual deviance: 1405.941 on 1464 degrees of freedom

Log-likelihood: -702.9707 on 1464 degrees of freedom

Number of iterations: 5

Reference group is level  3  of the response
```

FIGURE 10.9 – Résultats avec la fonction vglm

```
SunRain <- read.table("SunRain.csv", header=T, sep=";")
SunRain[,"Quality"] <- factor(SunRain[,"Quality"], levels=c("bad","medium","good"))
#centre-reduit les variables Sun et Rain
SunRain[,"Sun"] <- (SunRain[,"Sun"] - mean(SunRain[,"Sun"])) / sd(SunRain[,"Sun"])
SunRain[,"Rain"] <- (SunRain[,"Rain"] - mean(SunRain[,"Rain"])) / sd(SunRain[,"Rain"])
attach(SunRain)
```

Un résumé graphique des données est donné en Figure 10.10.

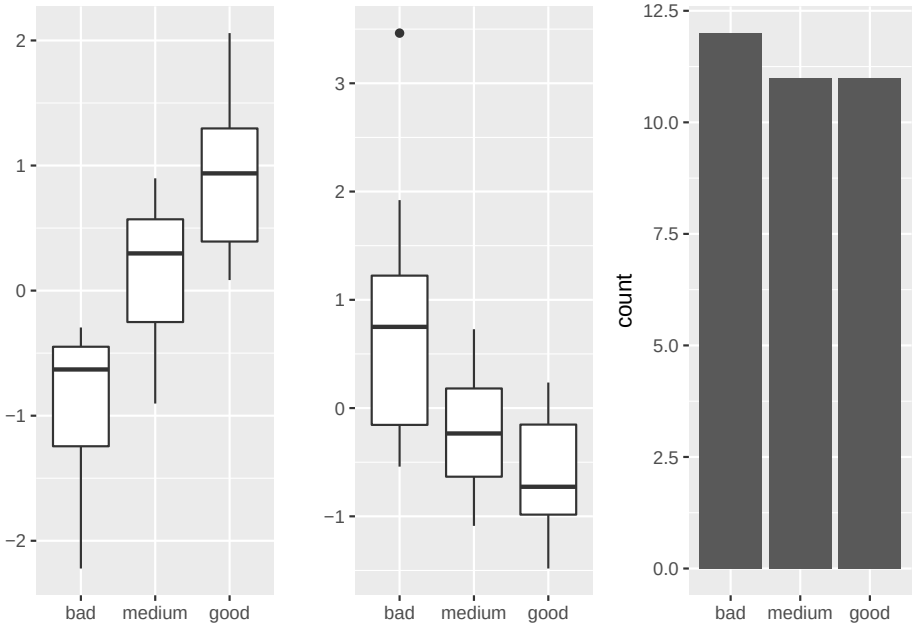


FIGURE 10.10 – Boxplot des variables centrées réduites "Sun" (à gauche) et "Rain" (au centre) en fonction de la variable "Quality". A droite, barplot pour la variable réponse "Quality".

Sous R, les fonctions `polr` et `vglm` des librairies `MASS` et `VGAM` permettent d'ajuster des modèles de **régression polytomique ordonnée**.

10.5.2.1 Modélisation par les logits cumulatifs

Pour la modélisation, on suppose qu'il existe une variable latente Z telle que

- Y s'écrit à partir de Z sous la forme suivante

$$Y = \begin{cases} u_1 & \text{si} \quad Z \in]a_0, a_1] \\ u_2 & \text{si} \quad Z \in]a_1, a_2] \\ \dots & \\ u_M & \text{si} \quad Z \in]a_{M-1}, a_M] \end{cases}$$

  avec $-\infty = a_0 < a_1 < \ldots < a_{M-1} < a_M = +\infty$ ($M-1$ coefficients inconnus),
- une liaison linéaire entre Z et les $x^{(1)}, \ldots, x^{(p)}$:

$$Z = \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)} + \gamma,$$

où $(\beta_1, \ldots, \beta_p)$ sont des paramètres inconnus et γ est une variable aléatoire symétrique de fonction de répartition F_γ.

Le modèle de régression polytomique ordonnée revient alors à modéliser pour tout $m \in \{1, \ldots, M-1\}$,

$$\mathbb{P}(Y \leq u_m|\mathbf{x}) = \mathbb{P}(Z \leq a_m|\mathbf{x}) = F_\gamma(a_m - [\beta_1 x^{(1)} + \ldots + \beta_p x^{(p)}]) = F_\gamma(\mathbf{x}\theta^{(m)}),$$

avec $\mathbf{x} = (1, x^{(1)}, \ldots, x^{(p)})$ et $\theta^{(m)} = (a_m, -\beta_1, \ldots, -\beta_p)'$. Comme dans le cadre binaire, il faut alors choisir une fonction de répartition F_γ. Si γ est supposée suivre une loi logistique, on modélise les logits cumulatifs par

$$\text{logit}\,[\mathbb{P}(Y \leq u_m|\mathbf{x})] = \mathbf{x}\theta^{(m)} = \theta_0^{(m)} + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)},$$

avec

$$\begin{aligned} \text{logit}\,[\mathbb{P}(Y \leq u_m|\mathbf{x})] &= \ln\left[\frac{\mathbb{P}(Y \leq u_m|\mathbf{x})}{\mathbb{P}(Y > u_m|\mathbf{x})}\right] \\ &= \ln\left[\frac{\pi_1(\mathbf{x}) + \ldots + \pi_m(\mathbf{x})}{\pi_{m+1}(\mathbf{x}) + \ldots + \pi_M(\mathbf{x})}\right]. \end{aligned}$$

Dans ce modèle, les coefficients des variables explicatives sont identiques et seules les constantes (intercepts) diffèrent selon les modalités de Y. Ainsi, quelque soit la modalité u_m considérée, une variable explicative donnée a la même influence sur $\mathbb{P}(Y \leq u_m|\mathbf{x})$. On dit qu'il y a égalité des pentes. Pour estimer les $M - 1 + p$ paramètres, on cherche à maximiser la vraisemblance.

```
> library(VGAM)
> levels(SunRain[,"Quality"])<-c("1","2","3")
> SunRain[,"Quality"]<-as.numeric(SunRain[,"Quality"])
> modelecumulsimpl <- vglm(Quality ~ Sun + Rain, data=SunRain,
+                          family = cumulative(parallel=T,reverse=F))
> modelecumulsimpl

Call:
vglm(formula = Quality ~ Sun + Rain, family = cumulative(parallel = T,
    reverse = F), data = SunRain)


Coefficients:
(Intercept):1 (Intercept):2           Sun          Rain
    -1.420824      2.317790     -3.265814      1.588428

Degrees of Freedom: 68 Total; 64 Residual
Residual deviance: 34.50584
Log-likelihood: -17.25292
```

On peut généraliser ce modèle en supposant que le rôle des variables dépend du niveau de la réponse, en posant

$$\text{logit}\,[\mathbb{P}(Y \leq u_m|\mathbf{x})] = \theta_0^{(m)} + \theta_1^{(m)} x^{(1)} + \ldots + \theta_p^{(m)} x^{(p)} = \mathbf{x}\theta^{(m)}$$

avec $\theta^{(m)} = (\theta_0^{(m)}, \theta_1^{(m)}, \ldots, \theta_p^{(m)})$. On se retrouve donc avec un modèle à $(M-1)(p+1)$ paramètres (estimés par maximum de vraisemblance).

```
> modelecumul <- vglm(Quality ~ Sun + Rain, data = SunRain,
+                     family = cumulative(parallel=F,reverse=F))
> modelecumul

Call:
vglm(formula = Quality ~ Sun + Rain, family = cumulative(parallel = F,
    reverse = F), data = SunRain)


Coefficients:
(Intercept):1 (Intercept):2         Sun:1         Sun:2        Rain:1        Rain:2
    -1.739086      2.005517     -4.020702     -2.814860      1.795255      1.326761

Degrees of Freedom: 68 Total; 62 Residual
Residual deviance: 34.02694
Log-likelihood: -17.01347
```

Pour la suite, on va exploiter les résultats de la modélisation simplifiée pour les illustrations. Avec `modelecumulsimpl@predictors`, on récupère les valeurs des logit $[\mathbb{P}(Y \leq u_m|\mathbf{x})]$. On peut facilement ensuite calculer les probabilités suivantes :

$$\begin{cases} \mathbb{P}(Y \leq u_m|\mathbf{x}) = \dfrac{\exp[\mathbf{x}\theta^{(m)}]}{1+\exp[\mathbf{x}\theta^{(m)}]} \\ \mathbb{P}(Y = u_m|\mathbf{x}) = \mathbb{P}(Y \leq u_m|\mathbf{x}) - \mathbb{P}(Y \leq u_{m-1}|\mathbf{x}) \\ \mathbb{P}(Y \leq u_M|\mathbf{x}) = 1. \end{cases}$$

```
> probacumul <- exp(modelecumulsimpl@predictors)/(1+exp(modelecumulsimpl@predictors))
> head(probacumul)
  logit(P[Y<=1]) logit(P[Y<=2])
1      0.9608757      0.9990324
2      0.9994916      0.9999879
3      0.9989072      0.9999740
4      0.9088956      0.9976213
5      0.9995916      0.9999903
6      0.4872894      0.9755831
> proba <- cbind(probacumul[,1],probacumul[,2]-probacumul[,1],1-probacumul[,2])
> head(proba)
       [,1]         [,2]         [,3]
1 0.9608757 0.0381566510 9.676064e-04
2 0.9994916 0.0004963458 1.210043e-05
3 0.9989072 0.0010668179 2.602321e-05
4 0.9088956 0.0887257897 2.378656e-03
5 0.9995916 0.0003986827 9.718527e-06
6 0.4872894 0.4882937635 2.441687e-02
```

Dans le cas où Y est une variable qualitative ordinale, on définit l'odds d'un individu $\mathbf{x}$ relativement à $Y \leq u_m$ par

$$\text{odds}(\mathbf{x}|u_m) = \frac{\mathbb{P}(Y \leq u_m|\mathbf{x})}{1-\mathbb{P}(Y \leq u_m|\mathbf{x})} = \exp[\mathbf{x}\theta^{(m)}].$$

L'odds ratio entre deux individus $\mathbf{x}$ et $\tilde{\mathbf{x}}$ relativement à $Y \leq u_m$ s'écrit alors

$$\text{OR}(\mathbf{x}, \tilde{\mathbf{x}}|u_m) = \frac{\text{odds}(\mathbf{x}|u_m)}{\text{odds}(\tilde{\mathbf{x}}|u_m)} = \exp\left[\sum_{j=1}^{p} \theta_j^{(m)}(x^{(j)} - \tilde{x}^{(j)})\right].$$

On peut remarquer que cet odds ratio ne dépend pas de $\theta_0^{(m)}$. Aussi dans le cas de la modélisation avec pentes parallèles, cet odds ratio ne dépend pas de la modalité u_m. En particulier, si $\mathbf{x}$ et $\tilde{\mathbf{x}}$ ne diffèrent que d'une unité pour seulement une variable j, alors $\text{OR}(\mathbf{x}, \tilde{\mathbf{x}}|u_m) = \exp[\theta_j^{(m)}]$.

Dans notre exemple, lorsque l'on observe la valeur moyenne des variables explicatives ($\mathbf{x} = (1, 0, 0)$ car les données sont centrées), on obtient dans notre exemple que $\text{odds}(\mathbf{x}|\text{"}\texttt{bad}\text{"}) = e^{-1.420} = 0.24$: on a 0.24 fois plus de chance d'avoir un vin de qualité `"bad"` que d'une qualité meilleure. De même, $\text{odds}(\mathbf{x}|\text{"}\texttt{medium}\text{"}) = e^{2.317} = 10.15$: on a 10.15 fois plus de chance d'avoir un vin de qualité inférieure à `"medium"` que `"good"`. Lorsque les jours d'ensoleillement augmentent de 1, on a $e^{-3.265814} = 0.04$ fois plus de chances d'avoir un vin moins bon qu'il ne l'est.

10.5.2.2 Modélisation par les logits adjacents

Il est également possible de définir la modélisation à partir des logits adjacents :

$$\begin{cases} L_{M-1} = \ln\left[\frac{\pi_M(\mathbf{x})}{\pi_{M-1}(\mathbf{x})}\right] = \theta_0^{(M-1)} + \theta_1^{(M-1)}x^{(1)} + \ldots + \theta_p^{(M-1)}x^{(p)} \\ \ldots \\ L_2 = \ln\left[\frac{\pi_3(\mathbf{x})}{\pi_2(\mathbf{x})}\right] = \theta_0^{(2)} + \theta_1^{(2)}x^{(1)} + \ldots + \theta_p^{(2)}x^{(p)} \\ \\ L_1 = \ln\left[\frac{\pi_2(\mathbf{x})}{\pi_1(\mathbf{x})}\right] = \theta_0^{(1)} + \theta_1^{(1)}x^{(1)} + \ldots + \theta_p^{(1)}x^{(p)}, \end{cases}$$

C'est la même idée que pour la régression multinomiale mais la catégorie de référence change à chaque étape. On peut relier les deux en remarquant que

$$\begin{cases} \ln\left[\frac{\pi_2(\mathbf{x})}{\pi_1(\mathbf{x})}\right] = L_1 \\ \\ \ln\left[\frac{\pi_3(\mathbf{x})}{\pi_1(\mathbf{x})}\right] = L_2 + L_1 \\ \ldots \\ \ln\left[\frac{\pi_M(\mathbf{x})}{\pi_1(\mathbf{x})}\right] = L_{M_1} + \ldots + L_2 + L_1 \end{cases}$$

Comme précédemment, on peut considérer une modélisation simplifiée pour limiter le nombre de paramètres :

$$\ln\left[\frac{\pi_{m+1}(\mathbf{x})}{\pi_m(\mathbf{x})}\right] = \theta_0^{(m)} + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)}.$$

Pour notre exemple, on ajuste le modèle "complet" et le modèle "simplifié", les résultats sont reportés en Figure 10.11.

On définit l'odds d'une modalité u_{m+1} par rapport à u_m relativement à un individu $\mathbf{x}$ par

$$\text{odds}(Y = u_{m+1} \text{ vs } Y = u_m; \mathbf{x}) = \frac{\mathbb{P}(Y = u_{m+1}|\mathbf{x})}{\mathbb{P}(Y = u_m|\mathbf{x})} = \frac{\pi_{m+1}(\mathbf{x})}{\pi_m(\mathbf{x})} = \exp\left[\mathbf{x}\theta^{(m)}\right].$$

```
> modeleadj <- vglm(Quality ~ Sun + Rain, data = SunRain,
+                   family = acat(parallel=F,reverse=F))
> summary(modeleadj)

Call:
vglm(formula = Quality ~ Sun + Rain, family = acat(parallel = F,
    reverse = F), data = SunRain)

Pearson residuals:
                       Min       1Q     Median     3Q   Max
loge(P[Y=2]/P[Y=1]) -1.128 -0.09384  0.0003796 0.1627 3.792
loge(P[Y=3]/P[Y=2]) -1.442 -0.40387 -0.0025986 0.2268 2.326

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1    1.649      1.055   1.564   0.1179
(Intercept):2   -1.700      0.894  -1.901   0.0572 .
Sun:1            4.110      2.010   2.045   0.0409 *
Sun:2            2.504      1.141   2.195   0.0281 *
Rain:1          -1.727      1.052  -1.641   0.1008
Rain:2          -1.185      1.036  -1.144   0.2526
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  2

Names of linear predictors: loge(P[Y=2]/P[Y=1]), loge(P[Y=3]/P[Y=2])

Residual deviance: 33.6889 on 62 degrees of freedom

Log-likelihood: -16.8445 on 62 degrees of freedom

Number of iterations: 7

> modeleadjsimpl <- vglm(Quality ~ Sun + Rain, data = SunRain,
+                        family = acat(parallel=T,reverse=F))
> summary(modeleadjsimpl)

Call:
vglm(formula = Quality ~ Sun + Rain, family = acat(parallel = T,
    reverse = F), data = SunRain)

Pearson residuals:
                       Min      1Q    Median     3Q   Max
loge(P[Y=2]/P[Y=1]) -1.104 -0.1716  0.001685 0.2444 2.734
loge(P[Y=3]/P[Y=2]) -1.602 -0.3217 -0.001404 0.1568 2.802

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1   1.2810     0.7439   1.722  0.08506 .
(Intercept):2  -2.0369     0.8481  -2.402  0.01632 *
Sun             3.0711     0.9924   3.095  0.00197 **
Rain           -1.4709     0.7037  -2.090  0.03658 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  2

Names of linear predictors: loge(P[Y=2]/P[Y=1]), loge(P[Y=3]/P[Y=2])

Residual deviance: 34.3157 on 64 degrees of freedom

Log-likelihood: -17.1578 on 64 degrees of freedom

Number of iterations: 7
```

FIGURE 10.11 – Résultats de la modélisation par les logits adjacents

Pour deux individus $\mathbf{x}$ et $\tilde{\mathbf{x}}$, on définit alors l'odds ratio par

$$\begin{aligned}\mathrm{OR}(Y = u_{m+1} \text{ vs } Y = u_m; \mathbf{x}, \tilde{\mathbf{x}}) &= \frac{\mathrm{odds}(Y = u_{m+1} \text{ vs } Y = u_m; \mathbf{x})}{\mathrm{odds}(Y = u_{m+1} \text{ vs } Y = u_m; \tilde{\mathbf{x}})} \\ &= \exp\left[\sum_{j=1}^{p}(x^{(j)} - \tilde{x}^{(j)})\theta_j^{(m)}\right].\end{aligned}$$

Ainsi si les deux individus $\mathbf{x}$ et $\tilde{\mathbf{x}}$ ne diffèrent que d'une unité pour la variable j, on a

$$\mathrm{OR}(Y = u_{m+1} \text{ vs } Y = u_m; \mathbf{x}, \tilde{\mathbf{x}}) = \exp[\theta_j^{(m)}].$$

Dans notre exemple, en prenant les résultats de la modélisation complète, on peut par exemple remarquer que si l'ensoleillement augmente d'une unité, on a $e^{4.11} = 60.9$ fois plus de change que le vin soit "medium" que "bad" ; et $e^{2.504} = 12.2$ fois plus de chance que le vin soir "good" que "medium". Si l'on prend la modélisation simplifiée, on a $e^{3.0711} = 21.5$ fois plus de chance de passer dans la catégorie supérieure pour la qualité du vin (que l'on ait un vin "bad" ou "medium" présentement). Lorsque l'on observe la valeur moyenne des variables explicatives ($x = (0, 0)$), on a $e^{1.281} = 3.6$ fois plus de chance d'avoir un vin "médium" que "bad" et $e^{-2.0369} = 0.13$ fois plus de chance d'avoir un vin "good" que un vin "medium".

Dans `modeleadj@predictors`, on récupère l'ensemble des valeurs des prédicteurs linéaires $\ln[\hat{\pi}_{m+1}(\mathbf{x}_i)/\hat{\pi}_m(\mathbf{x}_i)]$. A partir de ces valeurs, on peut retrouver les $\pi_m(\mathbf{x}_i)$ (disponibles dans `modeleadj@fitted.values`) par la formule

$$\begin{cases} \hat{\pi}_{m+1}(\mathbf{x}) = \dfrac{\prod_{v=1}^{m} e^{\mathbf{x}\hat{\theta}^{(v)}}}{1 + \sum_{m'=1}^{M-1}\prod_{v=1}^{m'} e^{\mathbf{x}\hat{\theta}^{(v)}}}, \quad \forall m \in \{1, \ldots, M-1\} \\ \hat{\pi}_1(\mathbf{x}) = \dfrac{1}{1 + \sum_{m=1}^{M-1}\prod_{v=1}^{m} e^{\mathbf{x}\hat{\theta}^{(v)}}}. \end{cases}$$

On définit alors les prédictions pour nos n individus par

$$\hat{Y}_i = u_{\hat{m}} \quad \text{où} \quad \hat{m} \in \underset{m=1,\ldots,M}{\mathrm{argmax}}\ \hat{\pi}_m(\mathbf{x}_i).$$

Dans notre exemple, on compare ainsi les prédiction avec les valeurs observées de la réponse :

```
> hatpi<-modeleadj@fitted.values
> hatY<-apply(hatpi,1,which.max)
> table(Quality,hatY)
       hatY
Quality  1  2  3
      1 11  1  0
      2  1  8  2
      3  0  3  8
```

Chapitre 11

Régression de Poisson / régression loglinéaire

Dans ce chapitre, on s'intéresse au cas où la variable réponse Y compte le nombre de fois qu'un certain évènement a lieu dans une période de temps donnée (e.g. nombre d'accidents de la route sur une année, nombre d'enfants par famille, le nombre de grèves d'une compagnie sur une période de trois ans, etc). Nous allons illustrer les différents points abordés dans ce chapitre avec l'exemple suivant :

Exemple : Nombre d'incidents maritimes
On s'intéresse à la variable `incidents` comptant le nombre d'incidents par mois de mise en service d'un bateau. On considère ici un échantillon de 40 bateaux et l'on souhaite expliquer la variable `incidents` à l'aide des 4 variables suivantes :

- `type` : il y a 5 types de bateaux, désignés par `A-B-C-D-E`. C'est une variable qualitative nominale, codée sous forme de facteur,
- `construction` : période de construction du bateau, à savoir entre 1960 et 1979 par périodes de 5 ans,
- `operation` : période de mise en service (entre 1960 et 1974 ou entre 1975 et 1979),
- `service` : nombre total de mois de mise en service du bateau.

Comme dans le chapitre précédent, on considère des variables explicatives quantitatives et qualitatives. Le comportement de ces variables est résumé sur la Figure 11.1, sauf la variable `type` pour laquelle chacune des modalités apparaît exactement 8 fois.

À première vue, il semblerait d'après l'histogramme que la variable réponse `incidents` suive une loi de Poisson de petit paramètre. En particulier, la probabilité d'observer peu d'incidents est très élevée, alors que la probabilité d'observer plusieurs incidents décroit exponentiellement. Notons toutefois que la distribution de Poisson est la loi la plus simple permettant de modéliser des données de comptage, mais ce n'est pas la seule.

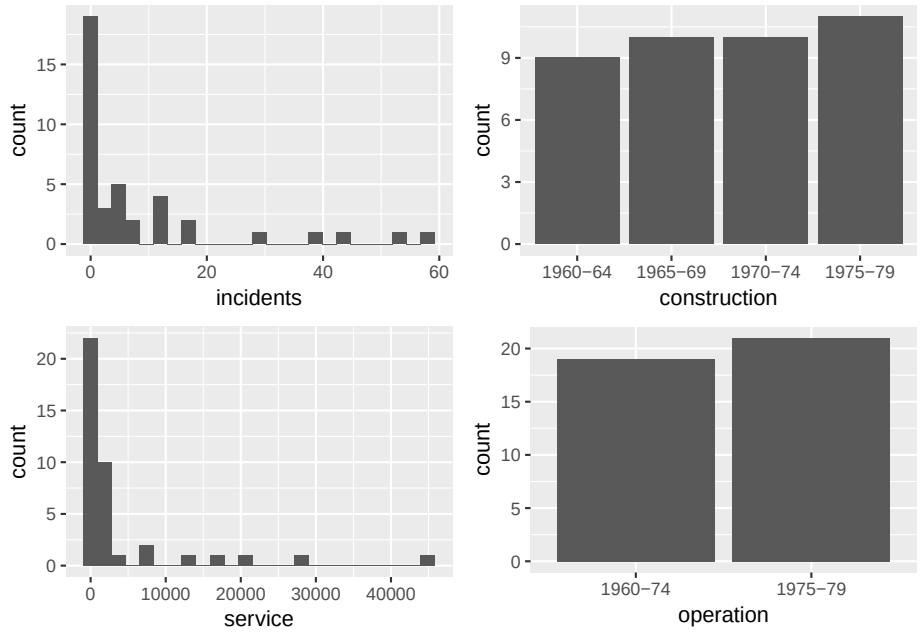


FIGURE 11.1 – Résumé des différentes variables de l'exemple d'incidents maritimes.

11.1 Modèle de régression loglinéaire

11.1.1 Pourquoi un modèle particulier ?

Dans la suite, on note $Y = (Y_1, \ldots, Y_n)'$ le vecteur des réponses, et $\mathbf{x}_i$ le vecteur ligne des variables explicatives considérées pour l'individu i pour chaque i dans $\{1 \ldots, n\}$. Les variables réponses à expliquer $Y_i|\mathbf{x}_i \sim \mathcal{P}(\lambda(\mathbf{x}_i))$ vérifient

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \lambda(\mathbf{x}_i),$$

avec $\lambda(\mathbf{x}_i) > 0$. L'objectif est de construire un modèle pour reconstituer $\lambda(\mathbf{x}_i)$ en fonction des variables explicatives.

Si on utilise le modèle de régression usuel $Y_i = \mathbf{x}_i\theta + \varepsilon_i$ pour expliquer le nombre d'incidents en fonction du nombre de mois de mise en service, on s'aperçoit d'une part que l'hypothèse de normalité des résidus n'est clairement pas réaliste (voir la figure 11.2). D'autre part, les variables ε_i étant supposées centrées,

$$\lambda(\mathbf{x}_i) = \mathbb{E}[Y_i|\mathbf{x}_i] = \mathbf{x}_i\theta.$$

Or rien n'indique que $\mathbf{x}_i\theta > 0$. Il est donc nécessaire de définir une fonction lien reliant $\lambda(\mathbf{x}_i)$ au prédicteur linéaire $\eta_i = \mathbf{x}_i\theta$. Pour garantir que l'espérance conditionnelle $\lambda(\mathbf{x}_i) = \mathbb{E}[Y_i|\mathbf{x}_i]$ est bien strictement positive, on définit le modèle par

$$\lambda(\mathbf{x}_i) = \lambda_\theta(\mathbf{x}_i) = \exp(\mathbf{x}_i\theta).$$

Cela revient à poser $\ln(\lambda_\theta(\mathbf{x}_i)) = \mathbf{x}_i\theta$. On retrouve la fonction lien logarithmique, qui est le lien canonique associé à la loi de Poisson, d'où le terme générique de **régression**

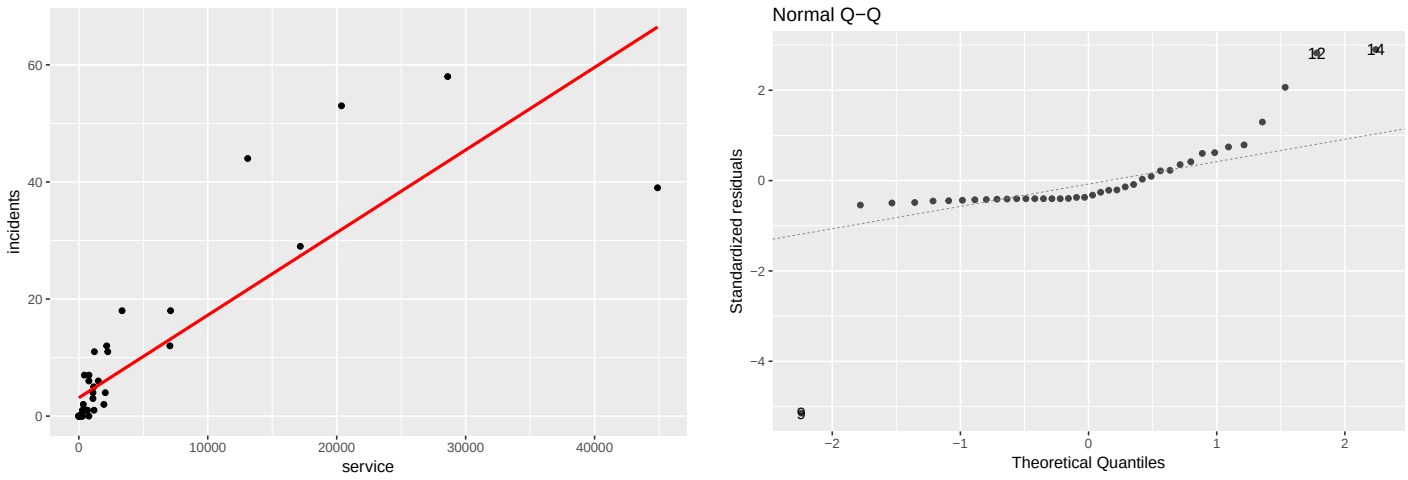


FIGURE 11.2 – Ajustement d'un modèle linéaire (à gauche) et graphique QQ-plot des résidus correspondants (à droite).

loglinéaire. Une fois le modèle posé, il reste à estimer le paramètre du modèle θ inconnu.

11.1.2 Estimation des paramètres

Comme dans le cas binaire, il faut bien faire attention à la nature des variables explicatives, et définir pour les variables qualitatives des modalités de référence. On se place ici dans un cadre très général.

Le paramètre θ est estimé par la méthode du maximum de vraisemblance. La vraisemblance des données $\underline{Y} = (Y_1, \ldots, Y_n)'$ est définie par :

$$L(\underline{Y};\theta) = \prod_{i=1}^{n} \left[\frac{\lambda_\theta(\mathbf{x}_i)^{Y_i}}{Y_i!} \exp(-\lambda_\theta(\mathbf{x}_i)) \right],$$

et la log-vraisemblance par :

$$\begin{aligned} l(\underline{Y};\theta) &= \sum_{i=1}^{n} \left[Y_i \ln(\lambda_\theta(\mathbf{x}_i)) - \lambda_\theta(\mathbf{x}_i) - \ln(Y_i!) \right] \\ &= \sum_{i=1}^{n} \left[Y_i \mathbf{x}_i \theta - e^{\mathbf{x}_i\theta} - \ln(Y_i!) \right]. \end{aligned}$$

Comme dans les chapitres précédents, on cherche alors à annuler les dérivées partielles :

$$\frac{\partial l(\underline{Y};\theta)}{\partial \theta_j} = \sum_{i=1}^{n} \left[x_i^{(j)} (Y_i - e^{\mathbf{x}_i\theta}) \right].$$

Encore une fois, le système obtenu n'admet généralement pas de solution calculable analytiquement. Un algorithme de type Newton-Raphson ou de Fisher-scoring est alors

mis en place, nécessitant l'évaluation la matrice hessienne ou la matrice d'information de Fisher. Les dérivées d'ordre secondes sont obtenues de la manière suivante :

$$\frac{\partial^2 l(\underline{Y};\theta)}{\partial\theta_j\partial\theta_k} = -\sum_{i=1}^{n} x_i^{(j)} x_i^{(k)} e^{\mathbf{x}_i\theta},$$

d'où l'écriture matricielle de l'information de Fisher

$$\mathcal{I}_n(\theta) = X'WX, \quad \text{avec} \quad W = \text{diag}[e^{\mathbf{x}_1\theta}, \ldots, e^{\mathbf{x}_n\theta}],$$

et X la matrice de design dont les lignes sont composées des vecteurs $\mathbf{x}_i$.

Remarquons encore une fois que cette matrice dépend du paramètre inconnu θ d'où la nécessité de mettre en place un algorithme itératif. Par ailleurs, notons que les dérivées secondes ne dépendant pas des variables Y_i, les algorithmes de Newton-Raphson et de Fisher-scoring sont exactement les mêmes.

11.1.3 Ajustement et prédiction

Une fois le modèle ajusté, nous obtenons une estimation pour chaque prédicteur linéaire $\eta_i = \mathbf{x}_i\theta$ par $\hat{\eta}_i = \mathbf{x}_i\hat{\theta}$ et pour chaque paramètre

$$\hat{\lambda}(\mathbf{x}_i) = \lambda_{\hat{\theta}}(\mathbf{x}_i) = \exp(\mathbf{x}_i\hat{\theta}).$$

Les valeurs ajustées $\widehat{Y}_i$ pour les Y_i sont alors définies suivant la règle

$$\widehat{Y}_i \in \underset{k\in\mathbb{N}}{\text{argmax}} \left\{ \frac{(\hat{\lambda}(\mathbf{x}_i))^k}{k!} e^{-\hat{\lambda}(\mathbf{x}_i)} \right\}.$$

$\widehat{Y}_i$ correspond donc à l'entier le plus probable pour la loi de Poisson de paramètre $\hat{\lambda}(\mathbf{x}_i)$.

Si l'on se donne maintenant un nouvel individu décrit par $\mathbf{x}_0$ alors le modèle ajusté permet de prédire son nombre moyen de "succès", donné par $\hat{\lambda}(\mathbf{x}_0) = \exp(\mathbf{x}_0\hat{\theta})$, et sa réponse prédite définie par

$$\widehat{Y}_0 \in \underset{k\in\mathbb{N}}{\text{argmax}} \left\{ \frac{[\hat{\lambda}(\mathbf{x}_0)]^k}{k!} e^{-\hat{\lambda}(\mathbf{x}_0)} \right\}.$$

11.2 Exemple de régression loglinéaire avec R

L'étude inférentielle du modèle de régression de Poisson est similaire au cas logistique, et découle directement des résultats asymptotiques étudiés dans le chapitre 9. Ils ne sont donc pas détaillés ici, mais illustrés dans cette partie sur l'exemple du nombre d'incidents maritimes.

11.2.1 Régression loglinéaire simple

Dans cette section, on modélise la variable réponse `incidents` à l'aide d'une seule variable explicative. On va distinguer selon la nature de la variable explicative.

11.2.1.1 Variable explicative quantitative

Commençons par modéliser la variable réponse `incidents` à l'aide de la variable $x =$`service` :

$$\ln[\lambda_\theta(x)] = \theta_0 + \theta_1 x.$$

```
> fit.service <- glm(incidents ~ service, data=ShipAccidents, family=poisson)
> summary(fit.service)

Call:
glm(formula = incidents ~ service, family = poisson, data = ShipAccidents)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.0040  -3.1674  -2.0055   0.9155   7.2372

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.613e+00  7.150e-02   22.55   <2e-16 ***
service     6.417e-05  2.870e-06   22.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 730.25  on 39  degrees of freedom
Residual deviance: 374.55  on 38  degrees of freedom
AIC: 476.41

Number of Fisher Scoring iterations: 6
```

L'estimation du coefficient de la variable `service` vaut 6.417×10^{-5}, et est donc très proche de 0. Cependant, la p-valeur du Z-test (basé sur l'approximation Gaussienne) étant $< 2 \times 10^{-16}$, nous rejetons la nullité de ce coefficient. Cela est probablement dû à la très grande variance de cette variable. En particulier, la variable `service` semble avoir une influence significative sur la variable `incidents`.

11.2.1.2 Variable explicative qualitative

À présent, modélisons la variable réponse `incidents` à l'aide de la seule variable qualitative `type` à 5 modalités. Comme dans le cas binaire, pour rendre le modèle identifiable, il faut choisir une modalité de référence (ici, la modalité choisie par défaut est `type=A`). Le modèle s'écrit donc

$$\ln[\lambda_\theta(x)] = \theta_0 + \theta_1 \mathbb{1}_{\texttt{type=B}} + \theta_2 \mathbb{1}_{\texttt{type=C}} + \theta_3 \mathbb{1}_{\texttt{type=D}} + \theta_4 \mathbb{1}_{\texttt{type=E}}.$$

```
> fit.type <- glm(incidents ~ type, data=ShipAccidents, family=poisson)
> summary(fit.type)

Call:
glm(formula = incidents ~ type, family = poisson, data = ShipAccidents)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.9530  -2.0616  -0.4541   1.2873   4.3425

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6582     0.1543  10.747  < 2e-16 ***
typeB         1.7957     0.1666  10.777  < 2e-16 ***
typeC        -1.2528     0.3273  -3.827  0.00013 ***
typeD        -0.9045     0.2875  -3.146  0.00165 **
typeE        -0.2719     0.2346  -1.159  0.24650
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 730.25  on 39  degrees of freedom
Residual deviance: 275.65  on 35  degrees of freedom
AIC: 383.52

Number of Fisher Scoring iterations: 6
```

L'interprétation des coefficients n'étant pas si simple, il est possible de tester la significativité de la variable `type` par un test de sous-modèle :

```
> anova(glm(incidents ~ 1, data=ShipAccidents, family=poisson), fit.type, test="Chisq")
Analysis of Deviance Table

Model 1: incidents ~ 1
Model 2: incidents ~ type
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        39     730.25
2        35     275.65  4    454.6 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ici, le test de rapport de vraisemblance (basé sur la déviance des modèles étudiés) rejette le sous-modèle nul. La variable `type` a une influence significative sur le nombre d'incidents, ce qui est en accord avec les boîtes à moustaches (boxplots) représentées en Figure 11.3

11.2.2 Régression loglinéaire multiple

Dans cette section, on modélise la variable réponse `incidents` à l'aide de toutes les variables explicatives disponibles. Comme pour le cas binaire, nous pourrions considérer toutes les interactions d'ordre 2 entre les variables explicatives. Cependant, cela mènerait à estimer 37 coefficients, ce qui semble peu raisonnable pour un échantillon de 40 bateaux. Nous considérons donc le modèle additif, les résultats sont reportés en Figure 11.4.

Les modalités de référence des deux variables qualitatives `construction` et `operation` choisies sont respectivement `1960-64` et `1960-74`. Comme dans le cas de la régression

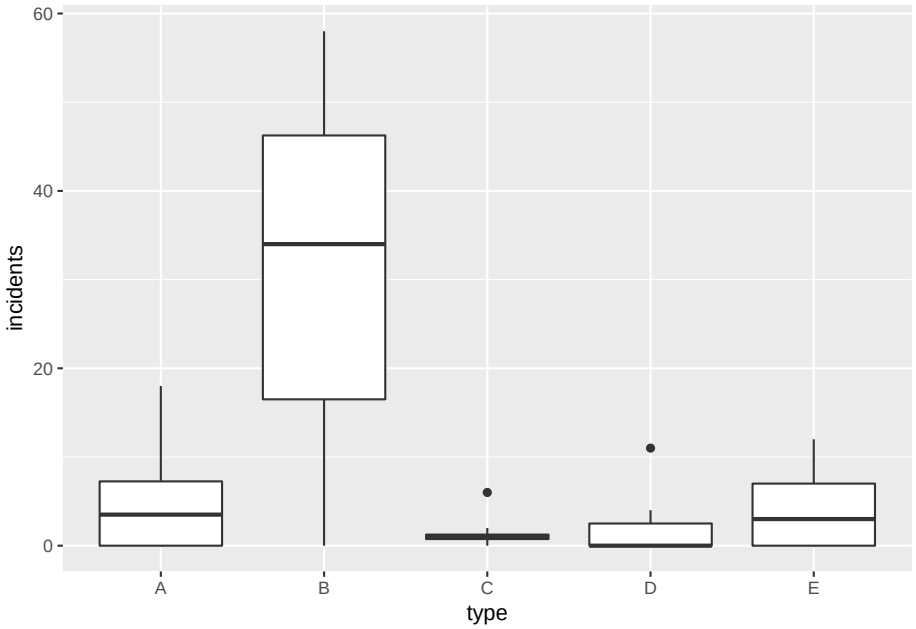


FIGURE 11.3 – Nombres d'accidents par mois de services en fonction du type de bateau.

```
> fit.add <- glm(incidents ~ . , data=ShipAccidents, family=poisson)
> summary(fit.add)

Call:
glm(formula = incidents ~ ., family = poisson, data = ShipAccidents)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5810  -1.4773  -0.8972   0.5952   3.2154

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          5.492e-04  2.787e-01   0.002 0.998427
typeB                5.933e-01  2.163e-01   2.743 0.006092 **
typeC               -1.190e+00  3.275e-01  -3.635 0.000278 ***
typeD               -8.210e-01  2.877e-01  -2.854 0.004321 **
typeE               -2.900e-01  2.351e-01  -1.233 0.217466
construction1965-69  1.148e+00  1.793e-01   6.403 1.53e-10 ***
construction1970-74  1.596e+00  2.242e-01   7.122 1.06e-12 ***
construction1975-79  5.670e-01  2.809e-01   2.018 0.043557 *
operation1975-79     8.619e-01  1.317e-01   6.546 5.92e-11 ***
service              7.270e-05  8.488e-06   8.565  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 730.253  on 39  degrees of freedom
Residual deviance:  99.793  on 30  degrees of freedom
AIC: 217.66

Number of Fisher Scoring iterations: 5
```

FIGURE 11.4 – Résultats de la régression loglinéaire multiple, modèle additif

simple, malgré un coefficient très petit (de l'ordre de 7×10^{-5}), la variable `service` semble significative (ainsi que toutes les autres variables).

11.2.2.1 Sélection de variables et sous-modèles

Il est possible de faire de la sélection de variables grâce à la commande `step(fit.add)`. Une procédure de sélection de variables descendante sur critère AIC est alors appliquée sur le modèle additif. Dans notre cas, elle renvoie exactement le même modèle.

On pourrait également tester la nullité simultanée des coefficients des variables `contructions` et `operation`, ce qui revient à faire un test de sous-modèle, en considérant seulement les variables `type` et `service`.

```
> fit.ssmod <- glm(incidents ~ type + service, data=ShipAccidents, family=poisson)
> anova(fit.ssmod, fit.add,  test="Chisq")
Analysis of Deviance Table

Model 1: incidents ~ type + service
Model 2: incidents ~ type + construction + operation + service
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        34    230.832
2        30     99.793  4   131.04 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Le test de rapport de vraisemblance rejette le sous-modèle avec une p-valeur $< 2\times10^{-16}$.

11.2.2.2 Prédiction

On souhaite prédire le nombre moyen d'incidents pour un bateau de type `A`, construit entre 1965 et 1969, et mis en service entre 1960 et 1975 au bout de 1000 mois de services.

```
> new.data = data.frame(type=factor("A"), construction=factor("1965-69"),
+                       operation=factor("1960-74"), service = 1000)
> lambda_hat = exp(predict(fit.add,new.data)) ; lambda_hat
       1
3.391016
```

Utilisant la loi de Poisson, nous pouvons alors prédire la probabilité qu'un tel bateau n'ait aucun incident, ou au plus un incident :

```
> # probabilité d'aucun incident :
> exp(-lambda_hat)
         1
0.03367446
>
> # probabilité d'au plus un incident :
> (1+lambda_hat) * exp(-lambda_hat)
        1
0.1478651
```

11.3 Sur-dispersion et modèle binomial négatif

Dans le cas du modèle de régression de Poisson, on a

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \mathrm{Var}(Y_i|\mathbf{x}_i),$$

ce qui est une hypothèse très restrictive. Si $\mathbb{E}[Y_i|\mathbf{x}_i] > \mathrm{Var}(Y_i|\mathbf{x}_i)$ (respectivement $\mathbb{E}[Y_i|\mathbf{x}_i] < \mathrm{Var}(Y_i|\mathbf{x}_i)$), nous parlons alors de *sur-dispersion* (respectivement de *sous-dispersion*). Ces deux propriétés n'étant pas autorisées par le modèle de Poisson, nous définissons une classe plus riche de modèles basée sur la loi binomiale négative.

Rappelons que la loi binomiale négative de paramètres n et p permet de modéliser le nombre d'échecs nécessaires avant l'obtention de n succès lors de la répétition de "tirage" indépendants de probabilité de succès p. Elle peut être généralisée à $n = r$ non-entier.

Le modèle binomial négatif suppose que la loi de Y_i vérifie pour tout k

$$\mathbb{P}(Y_i = k) = \frac{\Gamma(r+k)}{k!\Gamma(r)} \left(\frac{\lambda(\mathbf{x}_i)}{\lambda(\mathbf{x}_i)+r}\right)^k \left(1 - \frac{\lambda(\mathbf{x}_i)}{\lambda(\mathbf{x}_i)+r}\right)^r.$$

Nous pouvons alors démontrer que

$$\mathbb{E}[Y_i] = \lambda(\mathbf{x}_i) \quad \text{et} \quad \mathrm{Var}(Y_i) = \lambda(\mathbf{x}_i)\left(1 + \nu^2\lambda(\mathbf{x}_i)\right),$$

où $\nu = 1/r$ mesure le degré de sur-dispersion. Remarquons que le cas limite $\nu = 0$ correspond à la loi de Poisson.

Comme dans le cas Poisson, l'espérance conditionnelle est modélisée par

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \lambda_\theta(\mathbf{x}_i) = \exp(\mathbf{x}_i\theta).$$

En particulier, ce modèle appartient également à la famille des modèles de regression loglinéaire. Les paramètres inconnus θ et ν sont estimés par maximum de vraisemblance. Dans R, le modèle binomial négatif est implémenté dans la fonction `glm` pour la famille de lois `family = quasipoisson(link = "log")`.

Troisième partie

Annexes

Annexe A

Rappels de probabilités, statistiques et d'optimisation

A.1 Rappels sur les échantillons gaussiens

A.1.1 La loi normale

Définition A.1. *On dit que la variable aléatoire X suit une* ***loi normale*** *de paramètres (m, σ^2), notée $\mathcal{N}(m, \sigma^2)$, si la loi de X a pour densité*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-m)^2\right].$$

Proposition A.1.

- *Si X suit une loi $\mathcal{N}(m, \sigma^2)$ alors $\mathbb{E}[X] = m$, $Var(X) = \sigma^2$ et $(X-m)/\sigma$ suit la loi $\mathcal{N}(0,1)$. De plus, la fonction caractéristique de la loi de X est définie par*

$$\forall t \in \mathbb{R},\ \Phi_X(t) = \mathbb{E}\left[e^{itX}\right] = \exp\left(itm - \frac{\sigma^2 t^2}{2}\right).$$

- *Si $X_1, \ldots, X_n$ sont des variables aléatoires gaussiennes indépendantes, telles que, pour $i = 1, \cdots, n$, X_i suit la loi $\mathcal{N}(m_i, \sigma_i^2)$, alors pour tout $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$,*

$$\alpha_1 X_1 + \ldots + \alpha_n X_n \text{ suit la loi } \mathcal{N}(\alpha_1 m_1 + \ldots + \alpha_n m_n, \alpha_1^2\sigma_1^2 + \ldots + \alpha_n^2\sigma_n^2). \quad \text{(A.1)}$$

A.1.2 Vecteurs gaussiens

Définition A.2. *Un vecteur aléatoire X à valeurs dans $\mathbb{R}^d$ est dit* ***gaussien*** *si toute combinaison linéaire de ses composantes est une variable aléatoire gaussienne.*
Si $X = (X_1, \cdots, X_d)'$ est un vecteur gaussien, on définit son ***vecteur moyenne*** *$\mathbb{E}[X]$ par :*

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \cdots, \mathbb{E}[X_d])'$$

et sa ***matrice de variance-covariance*** *Var*(X) *par*

$$\begin{aligned} Var(X) &= \mathbb{E}\left[(X-\mathbb{E}(X))\,(X-\mathbb{E}(X))'\right] \\ &= \begin{pmatrix} Var(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_d) \\ Cov(X_2,X_1) & Var(X_2) & \cdots & Cov(X_2,X_d) \\ \vdots & \ddots & \cdots & \vdots \\ Cov(X_d,X_1) & Cov(X_d,X_2) & \cdots & Var(X_d) \end{pmatrix}. \end{aligned}$$

Remarques :

- La matrice $\mathrm{Var}(X)$ est symétrique puisque l'on a pour tout $i \neq j$:
$$\mathrm{Var}(X)_{i,j} = \mathrm{Cov}(X_i,X_j) = \mathrm{Cov}(X_j,X_i) = \mathrm{Var}(X)_{j,i}.$$
- Si $(X_1,\cdots,X_n)$ est un n-échantillon de loi gaussienne, i.e. $X_1,\cdots,X_n$ sont n variables aléatoires indépendantes et identiquement distribuées selon une loi $\mathcal{N}(\mu,\sigma^2)$, alors on a évidemment que $X=(X_1,\cdots,X_n)'$ est un vecteur gaussien de vecteur moyenne $\mathbb{E}[X]=(\mu,\cdots,\mu)'$ et de matrice de variance-covariance $\mathrm{Var}(X)=\sigma^2 I_n$ où I_n désigne la matrice identité.

On va s'intéresser à la fonction caractéristique d'un vecteur gaussien et aux conséquences importantes qui en découlent.

Théorème A.1. *Soit $X=(X_1,\cdots,X_d)'$ un vecteur gaussien. On note $m=\mathbb{E}[X]\in\mathbb{R}^d$ et $\Sigma = Var(X)\in\mathcal{M}_d(\mathbb{R})$. On a que X admet pour fonction caractéristique la fonction*

$$\forall u\in\mathbb{R}^d,\ \Phi_X(u)=\mathbb{E}\left[exp(iu'X)\right]=exp\left(iu'm-\frac{1}{2}u'\Sigma u\right).$$

La loi de X est entièrement déterminée par la donnée de m et de Σ.
On note $X\sim\mathcal{N}_d(m,\Sigma)$.

Corollaire A.1. *(***Propriété de linéarité)***
Soit $X=(X_1,\cdots,X_d)'\sim\mathcal{N}_d(m,\Sigma)$. On a pour toute matrice A de $\mathcal{M}_{pd}(\mathbb{R})$ et pour tout vecteur b de $\mathbb{R}^p$

$$AX+b\sim\mathcal{N}_p(Am+b,A\Sigma A').$$

Remarque : Soient $(X_i)_{i=1,\cdots,n}$ des variables aléatoires indépendantes de loi $\mathcal{N}(m_i,\sigma_i^2)$. Alors on a

$$X=(X_1,\cdots,X_n)'\sim\mathcal{N}_n(m,\Sigma) \text{ avec } m=(m_1,\cdots,m_n)' \text{ et } \Sigma=\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{pmatrix}.$$

En prenant $A=(\alpha_1,\cdots,\alpha_n)$ et $b=0_n$, on retrouve la Proposition A.1, équation (A.1).
En effet, on a $Am+b=\sum_{i=1}^n\alpha_i m_i$ et $A\Sigma A'=\sum_{i=1}^n\alpha_i^2\sigma_i^2$.

Corollaire A.2. (Propriété d'indépendance)
Soit $X = (X_1, \cdots, X_d)'$ un vecteur gaussien. Alors les trois propriétés suivantes sont équivalentes :

1. *Les composantes $X_1, \cdots, X_d$ sont mutuellement indépendantes.*
2. *Les composantes $X_1, \cdots, X_d$ sont deux à deux indépendantes.*
3. *La matrice de variance-covariance Σ est diagonale, i.e. $\forall i \neq j$, $Cov(X_i, X_j) = 0$.*

Remarque : Les composantes d'un vecteur gaussien sont des variables aléatoires gaussiennes mais la réciproque est fausse. En effet, on considère X et Y deux variables indépendantes telles que $X \sim \mathcal{N}(0,1)$ et $Y \sim \mathcal{B}(0.5)$. Alors $X_1 = X$ et $X_2 = (2Y-1)X$ sont des variables gaussiennes mais $(X_1, X_2)'$ n'est pas un vecteur gaussien. On note que dans cet exemple, $\mathrm{Cov}(X_1, X_2) = 0$ mais que X_1 et X_2 ne sont pas indépendantes.

A.1.3 Loi du khi-deux, loi de Student, loi de Fisher

Définition A.3. *Soient $Y_1, \ldots, Y_n$ des variables aléatoires indépendantes et de même loi $\mathcal{N}(0,1)$. La loi de $Y_1^2 + \ldots + Y_n^2$ est appelée **loi du khi-deux** à n degrés de liberté, et notée $\chi^2(n)$.*

Proposition A.2.

- *Si $V \sim \chi^2(n)$ alors $\mathbb{E}[V] = n$ et $Var(V) = 2n$.*
- *Si $V_1 \sim \chi^2(n_1)$, si $V_2 \sim \chi^2(n_2)$ et si V_1 et V_2 sont des variables aléatoires indépendantes, alors $V_1 + V_2 \sim \chi^2(n_1 + n_2)$.*

Définition A.4. *Soient U et V deux variables aléatoires telles que $U \sim \mathcal{N}(0,1)$, $V \sim \chi^2(n)$ et, U et V sont indépendantes. Alors la loi de*

$$\frac{U}{\sqrt{V/n}} = \sqrt{n}\frac{U}{\sqrt{V}}$$

*est appelée **loi de Student** à n degrés de liberté, notée $\mathcal{T}(n)$.*

Définition A.5. *Soient V_1 et V_2 deux variables aléatoires indépendantes, respectivement de loi $\chi^2(n_1)$ et $\chi^2(n_2)$. La loi de*

$$\frac{V_1/n_1}{V_2/n_2}$$

*est appelée **loi de Fisher** de paramètres (n_1, n_2). Elle est notée $\mathcal{F}(n_1, n_2)$.*

A.1.4 Estimation de la moyenne et de la variance d'un échantillon gaussien.

Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes et de même loi (i.i.d.), de loi $\mathcal{N}(m, \sigma^2)$. À partir de l'observation d'une réalisation de l'échantillon $(X_1, \ldots, X_n)$,

on souhaite estimer les paramètres inconnus m et σ^2.

• **Estimateur de m :** La moyenne empirique

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

est un estimateur de m.

▷ $\bar{X}_n$ est un estimateur sans biais de m : $\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = m$.

▷ $\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n} \underset{n\to\infty}{\longrightarrow} 0$.

▷ D'après l'inégalité de Bienaymé-Tchebytchev, $\bar{X}_n$ converge en probabilité quand n tend vers $+\infty$ vers m, i.e.

$$\bar{X}_n \underset{n\to\infty}{\xrightarrow{\mathbb{P}}} m.$$

▷ D'après la Proposition A.1, équation (A.1), $\bar{X}_n \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$.

▷ Il en résulte que la variable aléatoire

$$\sqrt{n}\frac{(\bar{X}_n - m)}{\sigma} \sim \mathcal{N}(0,1).$$

• **Estimateur de σ^2 :**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right\}$$

est un estimateur sans biais de σ^2.

De plus par la loi des grands nombres, on peut démontrer que S^2 converge en probabilité quand n tend vers $+\infty$ vers σ^2, c'est-à-dire

$$S^2 \underset{n\to\infty}{\xrightarrow{\mathbb{P}}} \sigma^2. \tag{A.2}$$

Théorème A.2 (Théorème de Cochran)**.**
Soient $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.
On note X le vecteur $(X_1, \ldots, X_n) \in \mathbb{R}^n$. Soit $E_1 \oplus E_2 \oplus \ldots \oplus E_p$ une décomposition de $\mathbb{R}^n$ en p sous-espaces orthogonaux de dimensions respectives $r_1, \ldots, r_p$. On note X_{E_i} la projection orthogonale de X sur E_i. Alors les vecteurs $X_{E_1}, X_{E_2}, \ldots, X_{E_p}$ sont indépendants, de plus, pour tout i, la variable $\|X_{E_i}\|^2$ a pour loi $\sigma^2\chi^2(r_i)$.

Proposition A.3. *Soient $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{N}(m, \sigma^2)$.*

▷ *Les variables aléatoires*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \tag{A.3}$$

sont indépendantes.

▷ $\bar{X}_n \sim \mathcal{N}(m, \sigma^2/n)$,
▷ $S^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$.
▷ *Il en résulte que la variable aléatoire*

$$\sqrt{n}\ \frac{\bar{X}_n - m}{S} \sim \mathcal{T}(n-1). \tag{A.4}$$

A.1.5 Construction d'intervalles de confiance

Notons $t_{1-\alpha/2}$ le $(1-\alpha/2)$-quantile de la loi de Student à $n-1$ degrés de liberté.

Il résulte de la Proposition A.3, équation (A.4), que

$$\mathbb{P}\left(-t_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X}_n - m)}{S} \leq t_{1-\alpha/2}\right) = 1-\alpha.$$

Ceci fournit l'intervalle de confiance pour m avec coefficient de sécurité $1-\alpha$:

$$\left[\bar{X}_n - t_{1-\alpha/2}\frac{S}{\sqrt{n}}\,;\, \bar{X}_n + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right].$$

Afin de construire un intervalle de confiance pour σ^2, nous introduisons les $\alpha/2$ et $1-\alpha/2$ quantiles de la loi du khi-deux à $n-1$ degrés de liberté, notés respectivement $u_{\alpha/2}$ et $u_{1-\alpha/2}$.

On obtient l'intervalle de confiance pour σ^2 avec coefficient de sécurité $1-\alpha$:

$$\left[\frac{(n-1)S^2}{u_{1-\alpha/2}}\,;\, \frac{(n-1)S^2}{u_{\alpha/2}}\right].$$

A.2 Estimation sans biais de variance minimale

Tous les résultats de cette section sont admis et nous renvoyons pour les détails à des ouvrages plus théoriques comme Saporta [21] ou Dacunha-Castelle et Duflo [7].

Définition A.6. *Soit U une statistique fonction de $X_1, \cdots, X_n$ de loi $g(u,\beta)$ (densité dans le cas continu ou $\mathbb{P}(U=u)$ dans le cas discret). U est dite* ***exhaustive*** *si l'on a $L(X,\beta) = g(u,\beta)h(X)$ (principe de factorisation).*

Cela signifie que la loi conditionnelle de l'échantillon est indépendante du paramètre. Par conséquent, une fois connue U, aucune valeur de l'échantillon, ni aucune autre statistique ne nous apportera de renseignements supplémentaires sur β.

Théorème A.3. *S'il existe un estimateur de β sans biais, de variance minimale, alors il est unique presque sûrement.*

Théorème A.4. ***Rao-Blackwell****. Soit T un estimateur quelconque sans biais de β et soit U une statistique exhaustive pour β. Alors $T^* = \mathbb{E}[T/\beta]$ est un estimateur sans biais de β au moins aussi bon que T.*

Théorème A.5. *S'il existe une statistique exhaustive U, alors l'estimateur T sans biais de β de variance minimale (unique d'après le théorème A.3) ne dépend que de U.*

Définition A.7. *On dit qu'une statitique U est* ***complète*** *pour une famille de lois de probabilités $f(x, \beta)$ si $\mathbb{E}[h(U)] = 0, \forall \beta \Rightarrow h = 0$ ps.*

Nous pouvons montrer que la statistique exhaustive des familles exponentielles est complète.

Théorème A.6. ***Lehmann-Scheffé****. Si T^* est un estimateur sans biais de β dépendant d'une statistique exhaustive complète U alors T^* est l'unique estimateur sans biais de variance minimale de β. En particulier si l'on dispose déjà de T, estimateur sans biais de β, alors $T^* = \mathbb{E}[T/U]$.*

En conclusion, si on dispose d'un estimateur sans biais fonction d'une statistique exhaustive complète alors c'est le meilleur estimateur possible.

A.3 La méthode de Newton-Raphson

Soit $t : \mathbb{R} \to \mathbb{R}$ une fonction $\mathcal{C}^1$ donnée. La problématique consiste à trouver $Z^\star$ tel que $t(Z^\star) = 0$. Par définition de la dérivée, on a

$$t'(Z^\star) = \lim_{h \to 0} \frac{t(Z^\star + h) - t(Z^\star)}{h}.$$

La méthode de Newton est basée sur l'heuristique suivante. Si x est suffisamment 'proche' de $Z^\star$, alors moralement

$$t'(x) \simeq \frac{t(x) - t(Z^\star)}{x - Z^\star} \;\Leftrightarrow\; x - Z^\star \simeq \frac{t(x)}{t'(x)},$$

par définition de $Z^\star$. On va utiliser cette méthode de manière itérative en initialisant un x_0 puis en posant, pour tout $n \in \mathbb{N}$,

$$x_n = x_{n-1} - \frac{t(x_{n-1})}{t'(x_{n-1})}.$$

Sous des hypothèses assez souples (fonction t deux fois différentiable au voisinage de $Z^\star$ par exemple), on peut démontrer que $x_n \to Z^\star$ quand $n \to +\infty$.

A.4 Théorème central limite : condition de Lindeberg

Le théorème suivant généralise le Théorème central limite à des suites de variables indépendantes mais non identiquement distribuées. Ce type de résultat est particulièrement intéressant pour le modèle linéaire généralisé.

Théorème A.7. *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes d'espérances et de variances respectives m_i et σ_i^2. Soient $S_n^2 = \sum_{i=1}^n \sigma_i^2$ et pour tout $i \in \{1, \ldots, n\}$, F_i la fonction de répartition des variables $X_i - m_i$. Si*

$$\forall \varepsilon > 0, \lim_{n \to +\infty} \left[\frac{1}{S_n^2} \sum_{i=1}^{n} \int_{|x| > \varepsilon S_n} x^2 dF_i(x) \right] = 0, \tag{A.5}$$

alors,

$$\frac{\sum_{i=1}^n (X_i - m_i)}{\sqrt{S_n^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ quand } n \to +\infty.$$

Annexe B

Preuves de quelques résultats du cours

B.1 Preuve pour le test de Fisher

On reprend les notations du chapitre 4. Rappelons la nature de chaque objet : $\theta \in \mathbb{R}^k$, $C \in \mathcal{M}_{qk}(\mathbb{R})$, $X_0 \in \mathcal{M}_{nk_0}(\mathbb{R})$ et $X \in \mathcal{M}_{nk}(\mathbb{R})$ avec $Im(X_0) \subset Im(X)$.

Proposition B.1. *On veut tester*

$$\mathcal{H}_0 : Y = X_0\beta + \varepsilon \ \ (M_0) \ \textit{contre} \ \mathcal{H}_1 : Y = X\theta + \varepsilon \ \ (M_1)$$

i.e

$$\mathcal{H}_0 : C\theta = 0 \ \textit{contre} \ \mathcal{H}_1 : C\theta \neq 0.$$

Test 1 *:*
La statistique de test

$$F = \frac{SCR_0 - SCR/(k-k_0)}{SCR/(n-k)} = \frac{\|X\widehat{\theta} - X_0\widehat{\beta}\|^2/(k-k_0)}{\|Y - X\widehat{\theta}\|^2/(n-k)} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(k-k_0, n-k)$$

et la zone de rejet est donnée par

$$\mathcal{R} = \{F \geq f_{1-\alpha,k-k_0,n-k}\}.$$

Test 2 *:*
La statistique

$$\tilde{F} = \frac{[C\widehat{\theta}]'[C(X'X)^{-1}C']^{-1}[C\widehat{\theta}]/q}{SCR/(n-k)} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(q, n-k)$$

et la zone de rejet est donnée par

$$\mathcal{R} = \{\tilde{F} \geq f_{1-\alpha,q,n-k}\}.$$

Ces deux tests sont identiques.

Preuve : Montrons que $\tilde{F} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(q, n-k)$

L'application $C : \mathbb{R}^k \to \mathbb{R}^q$ est surjective car $rg(C) = q$ par hypothèse.
On a $C\widehat{\theta} \sim \mathcal{N}_q(C\theta, \sigma^2\Delta)$ avec $\Delta = C(X'X)^{-1}C'$. La matrice $(X'X)^{-1}$ étant inversible, elle peut s'écrire sous la forme AA' où $A \in \mathcal{M}_k(\mathbb{R})$ inversible. De plus, $rg(\Delta) = rg(CAA'C') = rg(A'C') = q - dim(Ker(A'C'))$. Or $A'C'x = 0_k \Leftrightarrow C'x = 0_k \Rightarrow x = 0_q$ car A inversible et C' injective. Ainsi $rg(\Delta) = q$ et $\Delta = (CA)(CA)' \in \mathcal{M}_q(\mathbb{R})$. Δ étant inversible, elle se décompose en $\Delta = BB'$ avec $B \in \mathcal{M}_q(\mathbb{R})$ inversible.

Sous $\mathcal{H}_0$, $C\widehat{\theta} \sim \mathcal{N}_q(0_q, \sigma^2\Delta)$ donc $B^{-1}C\widehat{\theta} \sim \mathcal{N}_q(0_q, \sigma^2 I_q)$. On en déduit donc que $[B^{-1}C\widehat{\theta}]'[B^{-1}C\widehat{\theta}]/\sigma^2 \sim \chi^2(q)$. De plus $SCR = (n-k)\widehat{\sigma^2} \sim \sigma^2\chi^2(n-k)$ et $\widehat{\theta}$ et $\widehat{\sigma^2}$ sont indépendants. On en conclut que

$$\frac{[B^{-1}C\widehat{\theta}]'[B^{-1}C\widehat{\theta}]/q}{SCR/(n-k)} = \frac{[C\widehat{\theta}]'[C(X'X)^{-1}C']^{-1}[C\widehat{\theta}]/q}{SCR/(n-k)} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(q, n-k).$$

Montrons que $F = \tilde{F}$
Tout d'abord,

$$\begin{aligned}
\|Y - X_0\widehat{\beta}\|^2 &= \min_{\beta\in\mathbb{R}^q} \|Y - X_0\beta\|^2 \\
&= \min_{u\in Im(X_0)} \|Y - u\|^2 \\
&= \min_{u\in X(Ker(C))} \|Y - u\|^2 \\
&= \min_{\theta\in Ker(C)} \|Y - X\theta\|^2 \\
&= \|Y - X\tilde{\theta}\|^2.
\end{aligned}$$

Le vecteur $\tilde{\theta}$ minimise $\|Y - X\theta\|^2$ sous la contrainte $\theta \in Ker(C)$. Soit $\lambda \in \mathbb{R}^q$. Pour déterminer $\tilde{\theta}$ on résout,

$$\begin{aligned}
& \frac{\partial}{\partial\theta}[(Y - X\theta)'(Y - X\theta) + \lambda'C\theta] = 0_k \\
\Leftrightarrow \quad & \frac{\partial}{\partial\theta}[Y'Y - \theta'X'Y - Y'X\theta + \theta'X'X\theta + \lambda'C\theta] = 0_k \\
\Leftrightarrow \quad & -2X'Y + 2X'X\theta + C'\lambda = 0_k
\end{aligned}$$

donc $\tilde{\theta} = (X'X)^{-1}X'Y - \frac{1}{2}(X'X)^{-1}C'\lambda$. En utilisant la contrainte $C\tilde{\theta} = 0_q$, on obtient que $\frac{1}{2}\lambda = \Delta^{-1}C(X'X)^{-1}X'Y$ car Δ est inversible. Finalement, $\tilde{\theta} = (X'X)^{-1}X'Y - (X'X)^{-1}C'\Delta^{-1}C(X'X)^{-1}X'Y = \widehat{\theta} - (X'X)^{-1}C'\Delta^{-1}C\widehat{\theta}$.

Ainsi

$$\begin{aligned}
\|X\widehat{\theta} - X_0\widehat{\beta}\|^2 &= \|X\widehat{\theta} - X\tilde{\theta}\|^2 \\
&= \|X(X'X)^{-1}C'\Delta^{-1}C\widehat{\theta}\|^2 \\
&= (C\widehat{\theta})'\Delta^{-1}C(X'X)^{-1}X'X(X'X)^{-1}C'\Delta^{-1}(C\widehat{\theta}) \\
&= (C\widehat{\theta})'\Delta^{-1}(C\widehat{\theta}).
\end{aligned}$$

Montrons que $q = k - k_0$
Soit $(e_1, \ldots, e_{k-q})$ une base de $Ker(C)$ donc $(Xe_1, \ldots, Xe_{k-q})$ est une famille génératrice de $X(Ker(C))$. On montre ensuite facilement que c'est une famille libre car X est injective. Ainsi $dim(X(Ker(C))) = dim(Im(X_0)) = k - q = k_0$.

B.2 Preuve de la proposition 7.3

Preuve : On considère un modèle d'ANOVA à deux facteurs de la forme générale

$$Y = X\theta + \varepsilon = (\mathbb{1}_n, X_{(\alpha)}, X_{(\beta)}, X_{(\gamma)})\theta + \varepsilon$$

avec $\theta = (\mu, \alpha, \beta, \gamma)'$, $\alpha = (\alpha_1, \ldots, \alpha_I)$, $\beta = (\beta_1, \ldots, \beta_J)$ et $\gamma = (\gamma_{11}, \ldots, \gamma_{IJ})$.
On considère les sous-espaces vectoriels suivants de $\mathbb{R}^n$:

$$\begin{array}{l} E_\mu = Vect(\mathbb{1}_n) \\ E_\alpha = \{X_{(\alpha)}\alpha; \sum_{i=1}^I n_{i+}\alpha_i = 0\} \\ E_\beta = \{X_{(\beta)}\beta; \sum_{j=1}^J n_{+j}\beta_j = 0\} \\ E_\gamma = \{X_{(\gamma)}\gamma; \sum_{i=1}^I n_{ij}\gamma_{ij} = \sum_{j=1}^J n_{ij}\gamma_{ij} = 0\} \end{array}$$

On introduit les ensembles $A_{(\alpha)} = \{\alpha;\ \sum_{i=1}^I n_{i+}\alpha_i = 0\}$ et $A_{(\beta)} = \{\beta;\ \sum_{j=1}^J n_{+j}\beta_j = 0\}$

Commençons par caractériser que E_μ, E_α, E_β et E_γ sont orthogonaux :
soit $v^{(\mu)} \in E_\mu$, $v^{(\alpha)} \in E_\alpha$, $v^{(\beta)} \in E_\beta$ et $v^{(\gamma)} \in E_\gamma$. On a donc

$$\begin{array}{l} < v^{(\mu)}, v^{(\alpha)} >= \sum_{i,j,\ell} \mu\alpha_i = \mu \sum_{i=1}^I n_{i+}\alpha_i = 0 \\ < v^{(\mu)}, v^{(\beta)} >= \sum_{i,j,\ell} \mu\beta_j = \mu \sum_{j=1}^J n_{+j}\beta_j = 0 \\ < v^{(\mu)}, v^{(\gamma)} >= \sum_{i,j,\ell} \mu\gamma_{ij} = \mu \sum_{i=1}^I \sum_{j=1}^J n_{ij}\gamma_{ij} = 0 \\ < v^{(\alpha)}, v^{(\gamma)} >= \sum_{i,j,\ell} \alpha_i\gamma_{ij} = \sum_{i=1}^I \alpha_i(\sum_{j=1}^J n_{ij}\gamma_{ij}) = 0 \\ < v^{(\beta)}, v^{(\gamma)} >= \sum_{i,j,\ell} \beta_j\gamma_{ij} = \sum_{j=1}^J \beta_j(\sum_{i=1}^I n_{ij}\gamma_{ij}) = 0 \\ < v^{(\alpha)}, v^{(\beta)} >= \sum_{i,j,\ell} \alpha_i\beta_j = \sum_{i=1}^I \sum_{j=1}^J n_{ij}\alpha_i\beta_j \end{array}$$

On remarque que si $n_{ij} = n_{i+}n_{+j}/n$, alors

$$< v^{(\alpha)}, v^{(\beta)} >= \sum_{i=1}^I \frac{n_{i+}}{n}\alpha_i \left(\sum_{j=1}^J n_{+j}\beta_j\right) = 0.$$

Réciproquement, supposons que E_α et E_β sont orthogonaux :

$$\sum_{i=1}^I \sum_{j=1}^J n_{ij}\alpha_i\beta_j = 0,\ \forall\alpha \in A_{(\alpha)},\ \forall\beta \in A_{(\beta)}. \tag{B.1}$$

Fixons α. (B.1) est vrai pour tout $\beta \in A_{(\beta)}$ et $\sum_{j=1}^J n_{+j}\beta_j = 0$ donc

$$\sum_{j=1}^J \left(\sum_{i=1}^I n_{ij}\alpha_i\right)\beta_j = 0 = \sum_{j=1}^J n_{+j}\beta_j.$$

Ainsi, $\sum_{i=1}^{I} n_{ij}\alpha_i = c_j n_{+j}$ où c_j constante pour $j = 1, \ldots, J$. En sommant sur j,

$$\sum_{j=1}^{J}\left(\sum_{i=1}^{I} n_{ij}\alpha_i\right) = \sum_{j=1}^{J} c_j n_{+j} = \sum_{i=1}^{I} n_{i+}\alpha_i = 0$$

d'où $c_j = 0$ pour tout j donc $\sum_{i=1}^{I} n_{ij}\alpha_i = 0$.

A nouveau, pour tout $\alpha \in A_{(\alpha)}$ et pour tout j,

$$\sum_{i=1}^{I} n_{ij}\alpha_i = 0 = \sum_{i=1}^{I} n_{i+}\alpha_i$$

donc n_{ij} et n_{i+} sont proportionnels pour tout i :

$$n_{ij} = d_j n_{i+} \text{ avec } d_j \text{ constante.}$$

Ainsi $\sum_{i=1}^{I} n_{ij} = \sum_{i=1}^{I} d_j n_{i+} = c_j n = n_{+j}$ d'où $d_j = n_{+j}/n$ et $n_{ij} = (n_{+j}/n)n_{i+}$.

B.3 Preuve de la proposition 6.2

Preuve :

$$\begin{aligned}
\mathcal{R}(m, m^\star) &= \mathbb{E}\left[\|X_{(m)}\hat{\theta}_{(m)} - \mu^\star\|^2\right] \\
&= \mathbb{E}\left[\|X_{(m)}\hat{\theta}_{(m)} - \mu^\star_{(m)} + \mu^\star_{(m)} - \mu^\star\|^2\right] \qquad \text{avec } \mu^\star_{(m)} = P_{[X_{(m)}]}\mu^\star \\
&= \mathbb{E}\left[\|X_{(m)}\hat{\theta}_{(m)} - \mu^\star_{(m)}\|^2\right] + \mathbb{E}\left[\|\mu^\star_{(m)} - \mu^\star\|^2\right] \qquad \text{par Pythagore} \\
&= \mathbb{E}\left[\|X_{(m)}\hat{\theta}_{(m)} - \mu^\star_{(m)}\|^2\right] + \|\mu^\star_{(m)} - \mu^\star\|^2.
\end{aligned}$$

Or

$$X_{(m)}\hat{\theta}_{(m)} = P_{[X_{(m)}]}Y = P_{[X_{(m)}]}\left(X_{(m^\star)}\theta_{(m^\star)} + \varepsilon_{(m^\star)}\right) = \mu^\star_{(m)} + P_{[X_{(m)}]}\varepsilon_{(m^\star)},$$

donc

$$\|X_{(m)}\hat{\theta}_{(m)} - \mu^\star_{(m)}\|^2 = \|P_{[X_{(m)}]}\varepsilon_{(m^\star)}\|^2 \sim (\sigma^\star)^2\chi^2(|m| + 1)$$

d'après le théorème de Cochran. Ainsi $\mathbb{E}\left[\|X_{(m)}\hat{\theta}_{(m)} - \mu^\star_{(m)}\|^2\right] = (\sigma^\star)^2(|m| + 1)$.

B.4 Preuve de la proposition 6.3

Preuve : Sous le modèle $m^\star$, la densité de $Y = (Y_1, \ldots, Y_n)'$ vaut

$$f^\star(Y) = (2\pi\sigma^{\star\,2})^{-n/2}\exp\left(-\frac{1}{2\sigma^{\star\,2}}\|Y - \mu^\star\|^2\right).$$

Sous le modèle m, la densité de Y vaut

$$f_{(m)}(Y) = (2\pi\sigma^2_{(m)})^{-n/2} \exp\left(-\frac{1}{2\sigma^2_{(m)}}\|Y-\mu_{(m)}\|^2\right).$$

Ainsi

$$\ln\left(\frac{f^\star(Y)}{f_{(m)}(Y)}\right) = \frac{n}{2}\ln\left(\frac{\sigma^2_{(m)}}{\sigma^{\star\,2}}\right) + \frac{\|Y-\mu_{(m)}\|^2}{2\sigma^2_{(m)}} - \frac{\|Y-\mu^\star\|^2}{2\sigma^{\star\,2}}.$$

D'où

$$\begin{aligned} KL(m^\star, m) &= \mathbb{E}_{f^\star}\left[\ln\left(\frac{f^\star(Y)}{f_{(m)}(Y)}\right)\right] \\ &= \frac{n}{2}\ln\left(\frac{\sigma^2_{(m)}}{\sigma^{\star\,2}}\right) + \frac{1}{2\sigma^2_{(m)}}\mathbb{E}_{f^\star}\left[\|Y-\mu_{(m)}\|^2\right] - \frac{1}{2\sigma^{\star\,2}}\mathbb{E}_{f^\star}\left[\|Y-\mu^\star\|^2\right]. \end{aligned}$$

Or $\mathbb{E}_{f^\star}\left[\|Y-\mu^\star\|^2\right] = \mathbb{E}_{f^\star}\left[\|\varepsilon^\star\|^2\right] = n\sigma^{\star\,2}$ et

$$\begin{aligned} \mathbb{E}_{f^\star}\left[\|Y-\mu_{(m)}\|^2\right] &= \mathbb{E}_{f^\star}\left[\|Y-\mu^\star+\mu^\star-\mu_{(m)}\|^2\right] \\ &= \mathbb{E}_{f^\star}\left[\|Y-\mu^\star\|^2\right] + \|\mu^\star-\mu_{(m)}\|^2 + 2\mathbb{E}_{f^\star}\left[(\mu^\star-\mu_{(m)})'(Y-\mu^\star)\right] \\ &= n\sigma^{\star\,2} + \|\mu^\star-\mu_{(m)}\|^2 \end{aligned}$$

car $\mathbb{E}_{f^\star}[Y] = \mu^\star$. Finalement, on obtient que

$$\begin{aligned} KL(m^\star, m) &= \frac{n}{2}\ln\left(\frac{\sigma^2_{(m)}}{\sigma^{\star\,2}}\right) + \frac{n\sigma^{\star\,2}+\|\mu^\star-\mu_{(m)}\|^2}{2\sigma^2_{(m)}} - \frac{n\sigma^{\star\,2}}{2\sigma^{\star\,2}} \\ &= \frac{n}{2}\left[\ln\left(\frac{\sigma^2_{(m)}}{\sigma^{\star\,2}}\right) + \frac{\sigma^{\star\,2}}{\sigma^2_{(m)}} - 1\right] + \frac{\|\mu^\star-\mu_{(m)}\|^2}{2\sigma^2_{(m)}}. \end{aligned}$$

B.5 Critère du C_p de Mallows

Soit $m \in \mathcal{M}$ fixé. On rappelle que d'après la proposition 6.2, le risque quadratique entre m et $m^\star$ vaut :

$$\mathcal{R}(m, m^\star) = \|\mu^\star_{(m)} - \mu^\star\|^2 + (\sigma^\star)^2(|m|+1).$$

Commençons par essayer d'estimer le terme de biais. D'après le théorème de Pythagore et le théorème de Cochran, on a :

$$\begin{aligned} \mathbb{E}\left[\|Y-\widehat{Y}_{(m)}\|^2\right] &= \mathbb{E}\left[\|Y-\mu^\star_{(m)}\|^2\right] - \mathbb{E}\left[\|\widehat{Y}_{(m)}-\mu^\star_{(m)}\|^2\right], \\ &= \mathbb{E}\left[\|Y-\mu^\star+\mu^\star-\mu^\star_{(m)}\|^2\right] - \mathbb{E}\left[\|\widehat{Y}_{(m)}-\mu^\star_{(m)}\|^2\right], \\ &= \mathbb{E}\left[\|Y-\mu^\star\|^2\right] + \|\mu^\star-\mu^\star_{(m)}\|^2 - (|m|+1)\sigma^{\star\,2}, \\ &= \|\mu^\star-\mu^\star_{(m)}\|^2 + n\sigma^{\star\,2} - (|m|+1)\sigma^{\star\,2}, \end{aligned}$$

ou encore

$$\|\mu^\star - \mu^\star_{(m)}\|^2 = \mathbb{E}\left[\left\|Y - \widehat{Y}_{(m)}\right\|^2\right] + (|m|+1)\sigma^{\star\,2} - n\sigma^{\star\,2}. \tag{B.2}$$

D'après (B.2), le terme de biais $\|\mu^\star - \mu^\star_{(m)}\|^2$ peut donc être estimé par $\|Y - \widehat{Y}_{(m)}\|^2 + (|m|+1)\sigma^{\star\,2}$ (on néglige le terme en $n\sigma^{\star\,2}$ puisque ce dernier ne dépend pas de m et n'interviendra donc pas dans la minimisation).

Si la variance est connue, on obtient alors le critère :

$$C_p(m) = \|Y - \widehat{Y}_{(m)}\|^2 + 2|m|\sigma^{\star\,2}.$$

On retiendra alors le modèle $\hat{m}_{CP}$ vérifiant :

$$\hat{m}_{CP} = \arg\min_{m\in\mathcal{M}} C_p(m).$$

Dans le cas où la variance est inconnue, on utilisera l'estimateur $\widehat{\sigma^2} = \widehat{\sigma^2}_{(m_p)}$ où $m_p = \{1,\ldots,p\}$ est le modèle prenant en compte tous les régresseurs.

B.6 Preuve de la proposition 9.5

Dans le cas d'une famille exponentielle,

$$l(\underline{Y};\theta) = \sum_{i=1}^n \left\{\frac{Y_i\omega_i - b(\omega_i)}{\gamma(\phi)} + c(Y_i,\phi)\right\} = \sum_{i=1}^n \ell_i$$

avec $\mu_i = b'(\omega_i)$, $\eta_i = g(\mu_i) = \mathbf{x}_i\theta$, $Var(Y_i) = b''(\omega_i)\gamma(\phi)$.
Calculons

$$\frac{\partial \ell_i}{\partial \theta_j} = \frac{\partial \ell_i}{\partial \omega_i}\frac{\partial \omega_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \theta_j} :$$

Comme

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \omega_i} &= [Y_i - b'(\omega_i)]/\gamma(\phi) = (Y_i - \mu_i)/\gamma(\phi),\\
\frac{\partial \omega_i}{\partial \mu_i} &= 1/b''(\omega_i) = \gamma(\phi)/\mathrm{Var}(Y_i),\\
\frac{\partial \eta_i}{\partial \theta_j} &= x_i^{(j)} \quad \text{car} \quad \eta_i = \mathbf{x}_i\boldsymbol{\theta},\\
\frac{\partial \mu_i}{\partial \eta_i} &\qquad \text{dépend de la fonction lien} \quad \eta_i = g(\mu_i),
\end{aligned}$$

on obtient que

$$S_j = \frac{\partial l(\underline{Y};\theta)}{\partial \theta_j} = \sum_{i=1}^n \frac{(Y_i - \mu_i)x_i^{(j)}}{\mathrm{Var}(Y_i)}\,\frac{\partial \mu_i}{\partial \eta_i} \quad \forall j = 0,\ldots,p.$$

Bibliographie

[1] Agresti, A. *Categorical data analysis*, vol. 482. John Wiley & Sons, 2003.

[2] Akaike, H. A bayesian analysis of the minimum aic procedure. *Annals of the Institute of Statistical Mathematics 30*, 1 (1978), 9–14.

[3] Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[4] Antoniadis, A., Berruyer, J., and Carmona, R. *Régression non linéaire et applications*. Economica, 1992.

[5] Azaïs, J.-M., and Bardet, J.-M. *Le modèle linéaire par l'exemple : régression, analyse de la variance et plans d'expériences illustrés avec R, SAS et Splus*. Dunod, 2005.

[6] Birgé, L., and Massart, P. Gaussian model selection. *Journal of the European Mathematical Society 3*, 3 (2001), 203–268.

[7] Castelle, D. D., and Duflo, M. *Probabilités et statistiques : tome 1 : problèmes à temps fixe*. Masson, 1994.

[8] Cornillon, P.-A., Husson, F., Jégou, N., Matzner-Lober, E., Josse, J., Guyader, A., Rouviére, L., and Kloareg, M. *Statistiques avec R*. Rennes (Presses universitaires de), 2010.

[9] Cornillon, P.-A., and Matzner-Løber, É. Théorie et applications.

[10] Coursol, J. Technique statistique des modèles linéaires.

[11] Daudin, J.-J. Le modèle linéaire et ses extensions-modèle linéaire général, modèle linéaire généralisé, modèle mixte, plans d'expériences (niveau c), 2015.

[12] Draper, N. R., and Smith, H. *Applied regression analysis*, vol. 326. John Wiley & Sons, 1998.

[13] Droesbeke, J.-J., Lejeune, M., and Saporta, G. *Modèles statistiques pour données qualitatives*. Editions Technip, 2005.

[14] Guyon, X. Modele linéaire et économétrie. *Ellipse, Paris* (2001).

[15] Hoerl, A. E., Kannard, R. W., and Baldwin, K. F. Ridge regression : some simulations. *Communications in Statistics-Theory and Methods 4*, 2 (1975), 105–123.

[16] Hoerl, A. E., and Kennard, R. W. Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics-Theory and Methods 5*, 1 (1976), 77–88.

[17] Mallows, C. L. Some comments on cp. *Technometrics 42*, 1 (2000), 87–94.

[18] McCullagh, P. *Generalized linear models*. Routledge, 2018.

[19] McDonald, G. C., and Galarneau, D. I. A monte carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association 70*, 350 (1975), 407–416.

[20] Pinheiro, J., and Bates, D. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2006.

[21] Saporta, G. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.

[22] Schwarz, G., et al. Estimating the dimension of a model. *The annals of statistics 6*, 2 (1978), 461–464.

[23] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological) 58*, 1 (1996), 267–288.

[24] Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society : series B (statistical methodology) 67*, 2 (2005), 301–320.