Table des matières

Chapitre 1

Introduction

Ce module fait suite aux précédents modules sur les analyses descriptives en adoptant un esprit différent puisqu'ici nous allons aborder la statistique inférentielle.

Un modèle linéaire est une expression qui relie une variable quantitative (la variable à expliquer ; par exemple un poids, une expression de gènes) à des variables, quantitatives et/ou qualitatives (les variables explicatives ; par exemple l'âge, le sexe, des conditions expérimentales).

Les analyses des modèles linéaires portent des noms différents selon la nature des variables explicatives utilisées dans le modèle. Le tableau suivant contient le nom des différentes analyses par nature des variables explicatives.

Variables explicatives	Nom de l'analyse
1 quantitative	régression simple
plusieurs quantitatives	régression multiple
plusieurs qualitatives	analyse de variance
1 ou plusieurs quantitatives et plusieurs qualitatives	analyse de covariance

Il existe cependant une théorie statistique englobant ces divers types de modèles : le modèle linéaire.

Notons que si non plus une, mais plusieurs variables quantitatives sont à expliquer conjointement, on se place dans le cadre de la régression multivariée, qui est fortement liée à l'analyse canonique. D'autre part, si la variable à expliquer est qualitative plutôt que quantitative, d'autres modèles sont à mettre en place comme la régression logistique ou la régression loglinéaire qui s'intégrent dans la famille du modèle linéaire général.

Dans la suite, nous aborderons en détail le modèle de régression simple, puis nous passerons en revue les autres modèles avec leurs spécificités, en gardant en mémoire que les méthodes d'estimation des paramètres, les tests et les analyses diagnostics sont identiques.

Chapitre 2

La régression

2.1 Le modèle de régression simple

Le but d'une analyse de régression est d'étudier les relations qui existent entre des facteurs/variables mesurables à partir d'observations (données) prises sur ces facteurs. Des objectifs plus précis d'une telle analyse peuvent être :

- la prévision (ex : étant donné l'âge, fumeur/non fumeur, le poids, etc ..., combien d'années un individu devrait-il survivre ?) ;
- la sélection de variables (ex : Parmi la température, l'ensoleillement, la pluie, l'altitude, le bruit ambiant, etc ..., quels facteurs ont une influence significative sur la croissance des pins des landes ?) ;
- la spécification de modèle (ex : Comment la durée de vie de transformateurs électriques varie-t-elle en fonction de leur grosseur ?) ;
- l'estimation de paramètres (ex : la luminosité en fonction de la distance des étoiles d'une certaine galaxie est de la forme $L = K_1 + K_2 d + \sigma\epsilon$, où K_1, K_2 et σ sont des paramètres inconnus à estimer à partir des observations).

Données pH. On veut étudier sur des carpes le pH (x) du milieu ambiant et le pH (y) de leur sang. Les données ont été simulées selon le code R suivant :

*> x<- round(runif(30)*5+1,1)*
*> y<-2+3*x+rnorm(30,0,1)*

Les données consistent en 30 unités statistiques (u.s.).
Pour l'u.s. i, on a (x_i, y_i).

Au vu de la figure 2.1 (code R : *> plot(x,y)*), on pressent qu'il existe une relation linéaire entre x et y : $y \doteq \beta_1 + \beta_2 x$. On écrit donc le modèle de régression suivant, expliquant y par une combinaison linéaire de paramètres à estimer (β_1 et β_2) :

$$y_i = \beta_1 + \beta_2 x_i + e_i, \text{ pour } i = 1, ...30; \tag{2.1}$$

où e_i est un résidu que l'on espère le plus petit possible. La variable y est appelée variable endogène (variable réponse, variable dépendante) ; les variables x_i sont appelées variables exogènes (variables explicatives, facteurs, covariables, variables indépendantes).

Hypothèses : Les observations y_i sont des réalisations de 30 variables aléatoires indépendantes Y_i de moyenne $\beta_1 + \beta_2 x_i$ et de variance σ^2. De manière équivalente, les résidus e_i sont

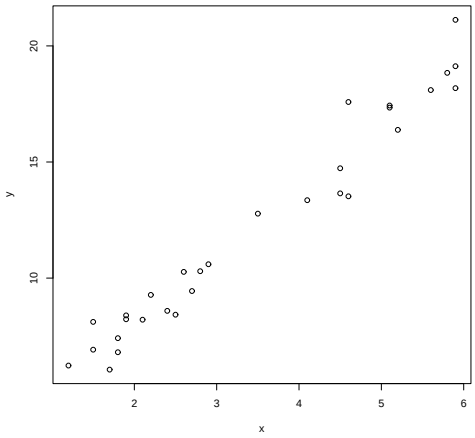


FIGURE 2.1 – Données pH : le pH sanguin de 30 carpes *vs.* le pH ambiant.

des réalisations de 30 variables aléatoires indépendantes E_i de moyenne 0 et de variance σ^2.

2.1.1 Ecriture et hypothèses du modèle

Nous observons n paires (y_1, x_1), ..., (y_n, x_n) et supposons que :

$$y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, ..., n;$$

où
$y_1, y_2, ..., y_i, ..., y_n$ sont les n observations de la variable endogène (ou variable à expliquer),
$x_1, ..., x_n$ sont les n observations de la variable exogène (ou variable explicative),
$e_1, ..., e_n$ sont les n termes d'erreur,
β_1 est le paramètre d'ordonnée à l'origine (la valeur moyenne de y lorsque x prend la valeur 0),
β_2 est le paramètre de pente (si x augmente d'une unité, alors y augmente de β_2 unités en moyenne).

Trois hypothèses sont essentielles à faire sur la distribution des termes d'erreur :

(i) les résidus sont de moyenne nulle (hypothèse de linéarité)

$$\mathrm{E}(e_i) = 0, \forall i = 1, ..., n \tag{2.2}$$

(ii) les e_i ont une variance identique (homoscédasticité)

$$\mathrm{Var}(e_i) = \sigma^2, \forall i = 1, ..., n \tag{2.3}$$

(iii) les e_i sont indépendants (donc non corrélés)

$$\mathrm{Cov}(e_i, e_j) = 0, \forall i \neq j \tag{2.4}$$

En supposant les valeurs x_i, $(i = 1, ..., n)$ comme étant non aléatoires (déterministes), les hypothèses ci-dessus impliquent que $\mathrm{E}(y_i) = \beta_1 + \beta_2 x_i$; $\mathrm{Var}(y_i) = \sigma^2$ et $\mathrm{Cov}(y_i, y_j) = 0$. On voit bien que quelle que soit la valeur de la variable explicative, seule l'espérance de y dépend de x, c'est à dire que la variance de y et la covariance entre deux observations de la variable à expliquer ne dépendent pas de la valeur de la variable explicative.

La droite de régression $(\beta_1 + \beta_2 x)$ représente la valeur attendue de y en fonction de la valeur de x. Les valeurs observées de y sont distribuées de façon aléatoire autour de cette droite. Les termes d'erreur sont les différences entre les valeurs observées de y et la droite. Comme la variance de ces termes d'erreur est constante en x, la distance moyenne des points à la droite est la même pour toute valeur de x. Finalement la non-corrélation entre les termes d'erreur signifie que la valeur d'un terme d'erreur n'est pas influencée par la valeur des autres termes d'erreur.

L'expression 2.1 se décompose de manière classique en :

$$\text{observation} = \text{modèle} + \text{résidu}.$$

2.1.2 Le modèle linéaire gaussien

Dans le modèle linéaire expliqué au paragraphe précédent, seuls les deux premiers moments des termes d'erreur (espérance et variance) sont supposés connus. Dans un modèle linéaire gaussien, on se donne une hypothèse supplémentaire : la distribution des résidus est supposée normale.

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

Cette hypothèse implique que les variables aléatoires y_i sont normalement distribuées.

2.1.3 Estimation des paramètres β_1 et β_2

Dans cette partie, la variance σ^2 est supposée connue. Les paramètres inconnus sont β_1 et β_2 et ils sont estimés à partir des données observées $(y_i, x_i, i = 1, ..., n)$. Deux grandes méthodes sont utilisées :

- la méthodes des moindres carrés, qui ne suppose connues que l'espérance et la variance de y ;
- la méthode du maximum de vraisemblance, qui suppose les résidus gaussiens.

On notera classiquement les estimations avec un "chapeau". Par exemple, $\hat{\beta}_1$ désigne l'estimation de β_1, c'est à dire une fonction de y (et de x). Rappelons ici que x est supposé connu, alors que y est une variable aléatoire (ou sa réalisation).

On appelle *ième* valeur ajustée ou prédite ou attendue la fonction de y suivante :

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

et le résidu correspondant vaut :

$$\hat{e}_i = y_i - \hat{y}_i.$$

Les moindres carrés ordinaires

L'idée est de trouver la droite qui explique la plus grande partie possible de la relation entre la variable endogène et la variable exogène, c'est à dire trouver la droite qui minimise la partie inexpliquée ou la partie due à la fluctuation aléatoire. On cherche donc la droite qui passe le plus près possible de tous les points ou, en d'autres termes, la droite qui minimise la distance des points à la droite (les termes d'erreurs).

La méthode des moindres carrés consiste à trouver les valeurs de $\hat{\beta}_1$ et $\hat{\beta}_2$ (estimateurs de β_1 et β_2) qui minimisent la somme des carrés des écarts entre les valeurs observées et les valeurs ajustées (somme des carrés des résidus) :

$$\min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} \hat{e}_i^2 = \min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \tag{2.5}$$

Puisque la fonction à minimiser a de bonnes propriétés (lisse, convexe), elle se minimise en prenant les dérivées de la somme par rapport à $\hat{\beta}_1$ et $\hat{\beta}_2$, en posant ces dérivées égales à zéro et en résolvant le système de deux équations à deux inconnues. On obtient :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

et

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

où $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ représentent les moyennes des observations y_i et x_i respectivement.

Prenons l'exemple des données pH. Avec le logiciel R, on peut ajuster le modèle (2.1) :

> reg1 <- lm(y ~ x)

```
Coefficients :
 (Intercept)           x
    5.572096   0.1951841

Degrees of freedom : 7 total ; 5 residual
Residual standard error : 0.05754663
```

> plot(x,y,xlab="pH ambiant (x)",ylab="pH sanguin (y)")
> lines(x,reg1$fitted.values)

On vérifie graphiquement que l'ajustement est cohérent en comparant valeurs observées et valeurs ajustées (figure 2.2)

La méthode du maximum de vraisemblance

Cette méthode nécessite l'ajout de l'hypothèse de normalité des résidus : $e_i \sim iidN(0, \sigma^2)$, ce qui implique que les y_i sont des variables aléatoires normales indépendantes :

$$y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2).$$

Cette méthode repose sur l'idée suivante : si les données de l'échantillon ont été observées, cela provient du fait que ces données sont les plus vraisemblables. Les estimateurs des paramètres

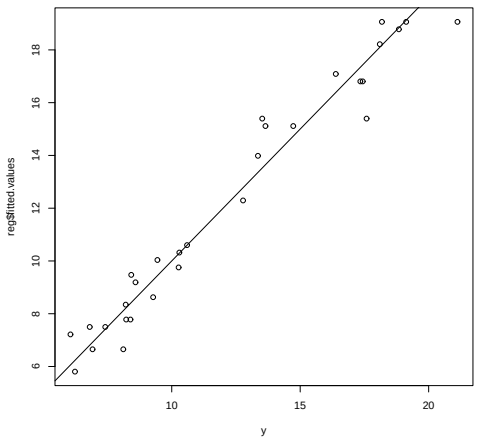


FIGURE 2.2 – Données pH : droite de régression du modèle (2.1).

inconnus du modèle sont donc calculés en maximisant une quantité (vraisemblance) qui mesure la probabilité d'observer l'échantillon. Dans le cadre de la régression linéaire simple, on cherche donc à maximiser la fonction de vraisemblance :

$$\begin{aligned} L(\beta_1, \beta_2, \sigma^2) &= \prod_{i=1}^{n} f(y_i; x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_1 - \beta_2 x_i)^2} \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2 \right\} \end{aligned}$$

Le logarithme de la vraisemblance, multiplié par (-2), s'écrit

$$l(\beta_1, \beta_2, \sigma^2) = -2 \ln L(\beta_1, \beta_2, \sigma^2) = n \ln(2\pi) + n \ln \sigma^2 + \sigma^{-2} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

Maximiser la vraisemblance équivaut donc à faire des moindres carrés.

2.1.4 Propriétés des estimateurs

Pour le modèle de régression simple, on peut montrer que :

$\mathrm{E}(\hat{\beta}_1) = \beta_1$ et $\mathrm{E}(\hat{\beta}_2) = \beta_2$

On dit alors que les estimateurs des paramètres sont sans biais. Rappelons que l'espérance est calculée par rapport à la loi de y.

D'autre part,

$$\mathrm{Var}(\hat{\beta}_2) = \mathrm{Var}(\frac{S_{xy}}{S_{xx}}) = \frac{\sigma^2}{S_{xx}}$$

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\bar{y} - \hat{\beta}_2\bar{x}) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$$

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \mathrm{Cov}(\bar{y} - \hat{\beta}_2\bar{x}, \hat{\beta}_2) = -\frac{\bar{x}\sigma^2}{S_{xx}}$$

On remarque ici que les deux estimateurs peuvent être très fortement corrélés. Pour éviter cela, on peut opérer une reparamétrisation :

$$\beta_0 + \beta_2(x_i - \bar{x})$$

avec $\beta_0 = \beta_1 + \beta_2\bar{x}$. Les estimateurs de ces deux paramètres ne sont pas corrélés : $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_2) = 0$, et la variance d'estimation de β_0 est plus faible que celle de β_1 : $\mathrm{Var}(\hat{\beta}_0) = \sigma^2/n < \mathrm{Var}(\hat{\beta}_1)$. Cette remarque souligne l'importance d'une bonne paramétrisation sur la précision des estimations. Des problèmes numériques sont aussi évités.

2.1.5 Estimation ponctuelle de σ^2

En général, on ne connait pas la valeur de σ^2, il faut alors l'estimer. Comme les résidus $\hat{e}_i = y_i - \hat{y}_i$ peuvent être vus comme des estimateurs des e_i, la variance d'échantillonnage des $\hat{e}_i$ est un estimateur raisonnable de $\sigma^2 = \mathrm{Var}(e_i)$. Un estimateur sans bais est donné par :

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

où $SSE = \sum_{i=1}^n \hat{e}_i^2$ est la somme des carrés résiduels.

2.1.6 Tests d'hypothèse et intervalles de confiance

Si les y_i sont des variables aléatoires normales, puisque les $\hat{\beta}$ sont des combinaisons linéaires des y_i, alors ces estimateurs sont donc aussi des variables aléatoires normales. Plus particulièrement :

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$$

$$\hat{\beta}_2 \sim N(\beta_2, \frac{\sigma^2}{S_{xx}}).$$

On peut donc standardiser les estimateurs pour obtenir :

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim N(0,1)$$

$$\frac{\hat{\beta}_2 - \beta_2}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$$

Comme σ^2 n'est pas connue, nous remplaçons dans les expressions ci-dessus σ^2 par son estimation s^2. Ce faisant, on doit corriger la distribution (cf. Annexe, et en admettant le résultat suivant : $s^2 \sim \chi^2_{n-2}$) afin d'obtenir :

$$\frac{\hat{\beta}_1 - \beta_1}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_2 - \beta_2}{s/\sqrt{S_{xx}}} \sim t_{n-2}$$

Les deux équations ci-dessus nous mènent aux intervalles de confiance à $(1-\alpha)100\%$ suivants pour β_1 et β_2 :

$$\left[\hat{\beta}_1 \pm t_{\alpha/2;n-2}\, s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right]$$

$$\left[\hat{\beta}_2 \pm t_{\alpha/2;n-2}\, s/\sqrt{S_{xx}}\right]$$

On peut aussi tester l'hypothèse de nullité d'un des paramètres. Par exemple, pour tester l'hypothèse nulle $H_0 : \beta_1 = 0$ *vs* l'hypothèse alternative $H_1 : \beta_1 \neq 0$, on utilise la statistique :

$$t_1 = \frac{\hat{\beta}_1}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

qui est distribuée selon une loi de Student à $n-2$ degrés de liberté. On rejettera donc l'hypothèse nulle si $t_1 > t_{\alpha/2;n-2}$ au niveau α. Il est clair que ce test correspond à la présence (H_0) ou à l'absence (H_1) de 0 dans l'intervalle de confiance.

2.1.7 Vérification des hypothèses

Tous les résultats (estimation, tests, intervalle de confiance) du modèle linéaire reposent sur des hypothèses fondamentales faites sur la distribution des termes d'erreur. Les résidus du modèle sont donc des outils privilégiés pour vérifier ces hypothèses. Voici un plan de vérification qu'il serait bon de suivre après chaque analyse (les hypothèses à vérifier sont classées par ordre d'importance décroissante) :
1) Vérifier que les résidus sont centrés : $\mathrm{E}(\boldsymbol{e}) = 0$ (hypothèse de linéarité) ;
2) Vérifier l'homoscédasticité (la variance des résidus doit être constante) ;
3) Vérifier l'indépendance des observations (hypothèse de non corrélation) ;
4) Vérifier la normalité des résidus.

Plusieurs versions de ces résidus existent :

Résidus ordinaires : le i^{eme} résidu (ordinaire) est défini comme étant $e_i = y_i - \hat{y}_i$. Si l'hypothèse 1) est vraie, alors $\mathrm{E}(e_i) = 0$. Si les hypothèses 2) et 3) sont vraies, alors

$$\mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2$$

et

$$\mathrm{Cov}(e_i, e_j) = -h_{ij}\sigma^2,$$

où $h_{ij} = 1/n + (x_i - \bar{x})(x_j - \bar{x})/S_{xx}$ et $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.
Enfin, si l'hypothèse 4)est vraie, $e_i \sim N(0, \sigma^2(1 - h_{ii}))$.

Avec le logiciel R, ces résidus sont obtenus avec le code suivant :
monmodele$residuals, où *monmodele* est le résultat de la fonction lm :
monmodele <- lm(y ~ x)

Résidus standardisés : le i^{eme} résidu standardisé est défini comme étant $r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$.

Avec le logiciel R, on obtient ces résidus avec le code : *rstandard(monmodele)*.

Résidus studentisés : il s'agit de légères modifications des résidus précédents : $t_i = \frac{e_i}{s_i\sqrt{1-h_{ii}}}$, où s_i^2 est une estimations sans biais de $Var(\hat{e}i)$. On montre que les t_i suivent une loi de Student à $n-3$ degrés de liberté.

Avec le logiciel R, on obtient ces résidus avec le code : *rstudent(monmodele)*.

Vérification de la linéarité

Graphique des résidus vs valeurs ajustées : $\hat{y}_i$ vs $\hat{e}_i$ (figure 2.3)

Ce graphique permet surtout de cerner les problèmes avec l'hypothèse 1) de linéarité. Si l'hypothèse est raisonnable, ce graphique devrait montrer un nuage de points centrés horizontalement autour de 0. Le graphique devrait avoir une allure complètement aléatoire, c'est à dire qu'il ne devrait y avoir aucune tendance discernable (e_i croissant ou décroissant avec $\hat{y}_i$, graphique à l'allure quadratique, etc...). Ce graphique peut également cerner des problèmes avec les autres hypothèses, mais les graphiques basés sur les résidus studentisés sont plus appropriés dans ces cas.

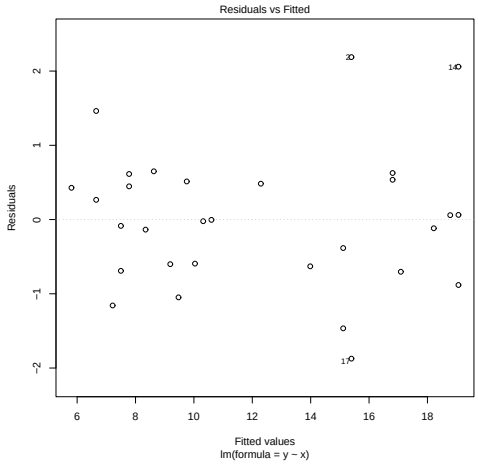


FIGURE 2.3 – Données pH : plot des résidus vs valeurs ajustées, pour le modèle (2.1).

Graphique des résidus vs variable explicative : x_i vs $\hat{e}_i$ (figure 2.4)

Encore une fois, ce type de graphique permet de détecter des problèmes avec l'hypothèse de linéarité ; il devrait avoir l'air d'un nuage de points dispersés horizontalement de façon aléatoire autour de 0.

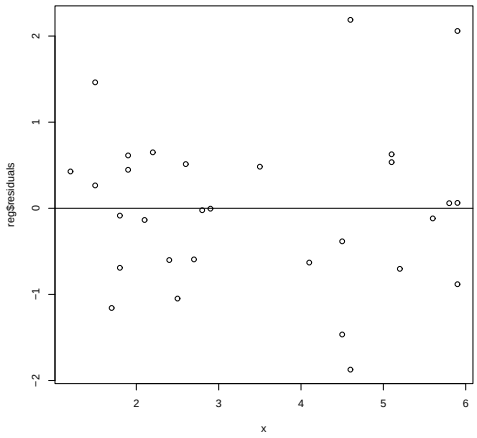


FIGURE 2.4 – Données pH : plot des résidus vs variable explicative, pour le modèle (2.1).

Vérification de l'homoscédasticité

Cette hypothèse est importante. Une forte violation de cette dernière entraîne des conséquences désastreuses sur : les erreurs standards des paramètres, les risques des tests, les intervalles de confiance. La méthode la plus couramment utilisée est la vérification graphique. Elle consiste à représenter les résidus en fonction des valeurs ajustées, des valeurs observées ou des valeurs de x. On peut également utiliser les résidus studentisés pour vérifier l'hypothèse d'homoscédasticité. Un graphique ayant une apparence d'entonnoir indique que la variance ne semble pas constante (problème d'hétéroscédasticité). Si certains résidus ont des valeurs plus grandes que 2 en valeur absolue, ceci peut indiquer un manque de normalité ou la présence de données atypiques qu'il est nécessaire de corriger ou d'enlever de l'étude.

Vérification de l'indépendance

Graphique des résidus vs numéro d'observations : $\hat{e}_i$ vs i

Ce graphique sert à vérifier l'hypothèse de non corrélation des résidus. Si les résidus de grande (faible) valeur ont tendance à suivre des résidus de grande (faible) valeur, alors il y a un problème d'autocorrélation positive. Si les résidus de grande (faible) valeur ont tendance à suivre des résidus de faible (grande) valeur, alors il y a un problème d'autocorrélation négative.

Quand la régression est réalisée sur des données qui varient au cours du temps, les observations peuvent ne pas être indépendantes. Pour vérifier l'indépendance, un test est habituellement

utilisé : le test de Durbin-Watson. Il est basé sur la statistique :

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{2.6}$$

Les e_i sont les résidus de la régression et n est le nombre d'observations. On peut montrer que $0 \leq d \leq 4$ et que $d \simeq 2 - 2\frac{\sum_{i=2}^{n}(e_i e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \simeq 2 - 2\rho_e$ où ρ_e est le coefficient d'autocorrélation d'ordre 1 des résidus. Il est obtenu en calculant la corrélation entre la série des résidus et la même série décalée de 1.

Si $\rho_e = 0$ soit $d \simeq 2$ alors les résidus sont non corrélés. Si par contre $\rho_e \neq 0$ ou encore $d \neq 2$ alors les résidus sont corrélés.

Vérification de la normalité

Cette étape n'est pas aussi importante qu'on le croit généralement. La normalité est une propriété qui permet aux estimateurs de converger rapidement. Le théorème central limite nous assure que pour des échantillons assez grands, les estimateurs que nous utilisons sont normalement distribués. La symétrie des distributions observées est un critère important qui assure une convergence rapide vers la loi normale. Les méthodes pour vérifier la normalité sont nombreuses, parmi celles-ci on peut citer les méthodes graphiques (QQplot, PPplot, histogrammes, boxplot, etc...) et les tests (Chi2, Shapiro-Wilk, Kolmogorov-Smirnov, ...).

Graphique des résidus studentisés vs quantiles de la loi normale : t_i vs u_i (figure 2.5)
Ce graphique permet de détecter les problèmes avec le postulat de normalité. Il est parfois appelé QQplot normal ou droite de Henry, tout dépend de la forme utilisée pour les u_i. Dans le QQplot, il s'agit des quantiles de la loi normale standard. Dans le cas de la droite de Henry, il s'agit de l'espérance des statistiques d'orde de la loi normale standard. Dans les deux cas, si l'hypothèse de normalité est raisonnable, le graphique devrait avoir la forme d'une ligne droite de pente positive. Des graphiques à l'allure de courbe concave ou convexe indiquent une distribution non symétrique des résidus, alors qu'un graphique en forme "d'intégrale inversée couchée" indique que les résidus proviennent d'une distribution ayant des queues plus épaisses que celles de la loi normale.

Boxplot des résidus : Le Boxplot des résidus (ordinaires ou studentisés) sert à déterminer si ces derniers proviennent d'une distribution symétrique et si certains résidus sont de valeur extrême. Une distribution non symétrique est indiquée par une moustache plus longue que l'autre, ou une ligne médiane proche d'une extremité de la boîte. Un résidu extrême est indiqué par un point à l'extérieur des moustaches.

Détection et élimination de valeurs atypiquess

Un examen critique des données est une étape importante en statistique. Il existe deux grands types de données généralement classées comme atypiques : les données qui ne sont pas habituelles, et les données qui violent une hypothèse de l'analyse statistique utilisée. Différentes attitudes devraient être adoptées suivant la nature du problème rencontré. Les données provenant d'erreurs grossières de mesures ou d'erreurs de frappe doivent être supprimées de l'analyse. Seul un jugement biologique permet de déclarer une valeur comme aberrante. Souvent, après un examen attentif des données on trouve des valeurs inhabituelles. Un expérimentateur prudent doit alors

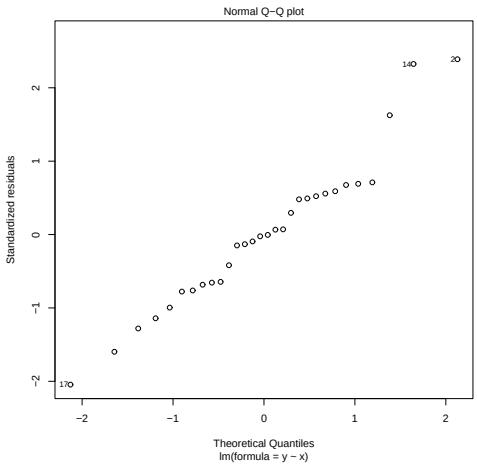


FIGURE 2.5 – Données pH : QQplot du modèle (2.1).

rechercher la (les) cause(s) de telles valeurs. Deux cas de figures se présentent alors : soit la cause est identifiée et il faut changer la donnée ou la méthode d'analyse ; soit la cause n'est pas identifiée et un test statistique peut être utilisée pour détecter une valeur atypique.

L'examen graphique des résidus est un bon outil (graphique des $\hat{e}$ en fonction de $\hat{y}$). Une autre technique consiste à calculer des indices pour chaque résidu. La plupart des indices calculés par les logiciels de statistique ont une signification inférentielle. Les trois les plus couramment usités sont : les résidus standardisés, les distances de Cook, les contributions.

Avec les résidus standardisés, il est donc possible de tester l'"aberrance" de chaque résidu en utilisant un test de Student. Attention toutefois aux tests multiples (voir plus loin pour le traitement des tests multiples).

Les contributions (leverage) et les mesures de Cook mesurent la contribution de chaque résidu à la variance résiduelle (non expliquée par le modèle). Sous les hypothèses usuelles (hypothèses du modèle), les distances de Cook suivent une loi de Fisher à p et $n-p$ degrés de liberté. Elles s'obtiennent par :

$$D_i = \frac{h_{ii}}{2(1-h_{ii})} r_i^2,$$

tenant compte ainsi de l'importance du résidu i et de l'influence h_{ii} de l'observation i sur la droite de régression (effet "levier"). Une méthode pour identifier les observations qui contribuent trop à la variance résiduelle consiste à réaliser un test de Fisher sur le résidu de Cook (i.e. de comparer sa valeur limite à un seuil donné d'une loi de Fisher à p et $n-p$ ddl).

Pour des données gaussiennes, les leverage devraient être voisines de $\frac{p}{n}$; p représente le nombre de paramètres indépendants estimés dans le modèle. Si pour un résidu, le leverage correspondant

est supérieur à $\frac{2p}{n}$, la donnée peut être considérée comme suspecte.

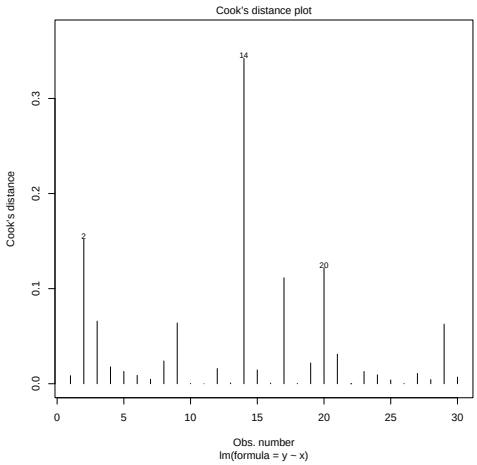


FIGURE 2.6 – Distance de Cook pour le modèle (2.1) sur les données pH.

Comment régler les problèmes ?

Manque de linéarité : Ceci est en général dû à une mauvaise spécification de la forme de la relation entre la variable endogène et la variable exogène. Le problème peut être réglé par une ou plusieurs options suivantes :
- transformer la variable exogène, ajouter au modèle des termes en x_i^2, x_i^3, ... ;
- ajouter au modèle de nouvelles variables exogènes ;
- transformer la variable endogène.

Hétéroscédasticité : La transformation de Box-Cox pourra souvent prescrire une transformation de la variable endogène qui règlera ce problème (transformation stabilisatrice de la variance). Si la transformation de Box-Cox ne fonctionne pas, alors la régression pondérée ou l'utilisation d'une autre méthode statistique peut être utile.

Méthode de Box-Cox : Cette méthode suppose un modèle de régression général de la forme : $g(y_i;\lambda) = \beta_1 + \beta_2 x_i + e_i$ où λ est un paramètre inconnu et qu'il faut estimer à partir des données ; $g(y;\lambda) = \frac{y^\lambda - 1}{\lambda}$ si $\lambda \neq 0$ et $g(y;\lambda) = \ln \lambda$ si $\lambda = 0$.

On estime λ en même temps que les autres paramètres par la méthode du maximum de vraisemblance.

Auto-corrélation des résidus : ce problème est plus difficile à régler. Il est parfois possible de le régler en ajoutant une variable exogène qui explique pourquoi il y a autocorrélation. Par

exemple, si les premières mesures sont faites sur l'individu A, les mesures suivantes sur l'individu B, etc ; alors on peut ajouter une variable exogène dénotant l'individu sur lesquelles les mesures ont été faites, ainsi on aura régler le problème. Mais en général, il faut avoir recours à un modèle plus complexe que le modèle linéaire simple (séries temporelles, modèle linéaire mixte).

Manque de normalité : Encore une fois, la transformation de Box-Cox règle souvent le problème. Parfois, le manque de normalité est tout simplement dû à quelques observations extrêmes. Dans ces cas, nous pourrons réglér le problème en traitant ces observations de façon appropriée. Une autre option consiste à utiliser des méthodes de régression robustes ou non paramétriques (non abordées dans ce cours).

2.2 Régression lineaire multiple

Quand on dispose de plusieurs variables explicatives, on peut mettre en oeuvre un modèle de régression linéaire multiple. Supposons que l'on dispose de n observations sur une variable continue y et p variables continues $x_1, ..., x_j, ...x_p$. On note y_i (resp. $x_{j,i}$) la $ième$ observation de y (resp. x_j). Le modèle est le suivant :

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_j x_{j,i} + ... + \beta_p x_{p,i} + e_i \tag{2.7}$$

est un modèle de régression multiple, cherchant à expliquer la variable y par p variables explicatives $x_1, ...x_j, ...x_p$, à l'aide d'un échantillon de taille n.

Les résidus e_i sont supposés de moyenne nulle, de variance égale σ^2, et mutuellement indépendants, éventuellement gaussiens.

Les paramètres inconnus $\beta_1, ...\beta_j, ...\beta_p$ (et éventuellement σ^2) sont estimés par moindres carrés (la somme des carrés des résidus la plus petite possible), ou par maximum de vraisemblance si les résidus sont supposés gaussiens, exactement comme dans le cas de la régression simple.

2.2.1 Multicolinéarité

Des problèmes d'estimation des paramètres et de variance peuvent arriver lorsque dans le modèle de régression, on utilise des variables explicatives corrélées entre elles. On parle alors de multicolinéarité et cela conduit à des estimations biaisées des paramètres avec des variances importantes. Pour diagnostiquer ces situations, une des méthodes est de faire la régression de chaque variable en fonction des autres variables explicatives et de mesurer les liaisons à l'aide du coefficient R_j^2 de chacune de ces régressions (où R_j est le coefficient de corrélation multiple obtenu en régressant la variable x_j sur les $(k-1)$ autres variables explicatives). On appelle tolérance, la valeur $1 - R_j^2$. Une tolérance qui est proche de 1 signifie une absence de multicolinéarité entre les variables explicatives. En revanche, si la tolérance tend vers 0, alors on détecte un problème de multicolinéarité entre les variables explicatives.

2.2.2 Critères de sélection de modèle

Pour obtenir un compromis satisfaisant entre un modèle trop simple (grands résidus) et un modèle faisant intervenir beaucoup de variables (donc très instable), on dispose de plusieurs critères qui ne donnent pas nécessairement le même résultat.

Coefficient de détermination et ses variantes

Pour mesurer la qualité d'un modèle de régression linéaire, ou pour comparer des modèles de régression linéaire entre eux, on définit le coefficient de détermination :

$$R^2 = \frac{SS_{Reg}}{SS_{Tot}} = 1 - \frac{SS_E}{SS_{Tot}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} = corr^2(y, \hat{y}) \tag{2.8}$$

On a que $0 \leq R^2 \leq 1$. Quand $R^2 = 0$, toute la variabilité est due à l'erreur aléatoire et le modèle n'explique absolument rien de la valeur de y_i. Quand $R^2 = 1$, tous les points tombent sur la droite de régression, c'est à dire que l'ajustement du modèle est parfait et que la valeur de y_i est une fonction exacte de x_i.

Le coefficient de détermination R^2 peut donc être interprété comme la proportion de la variabilité dans les y_i qui est expliquée par le modèle de régression.

Bien que facilement interprété et naturellement attrayant, le coefficient de détermination souffre de quelques problèmes qui font qu'il ne peut pas être utilisé pour comparer n'importe quels modèles de régression l'un avec l'autre. L'inconvénient principal est que dès que l'on ajoute un terme à un modèle de régression, le coefficient de détermination augmente. Afin de circonscrire à ce problème, nous pouvons utiliser le coefficient de détermination ajusté :

$$R^2_{ajust} = 1 - \frac{(n-1)}{(n-p-1)}(1 - R^2) = \frac{(n-1)R^2 - p}{n-p-1}$$

avec n le nombre d'observations et p le nombre de paramètres. Avec le R^2_{ajust}, l'ajout d'une variable explicative peut aussi résulter en une diminution de la statistique. La comparaison de modèles sur la base de ce critère revient à comparer deux modèles sur la base de leur estimé de la variance des termes d'erreur s^2. Le meilleur modèle sera celui ayant le R^2_{ajust} le plus grand.

C_p de Mallows

Une autre critère appelé, le coefficient C_p de Mallows peut être utilisé. Il est défini par :

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

où SS_E est la somme des carrés résiduels du modèle et $\hat{\sigma}^2$ est l'estimation de la variance résiduelle sous le modèle complet. On choisira la modèle pour lequel le coefficient C_p est minimum.

Test de Fisher pour modèles emboités

Il se peut qu'on veuille tester si le modèle à p variables explicatives peut être réduit à q (q petit devant p) variables ; c'est à dire que l'on veut tester si un sous-modèle plus simple explique une partie suffisamment grande de la variabilité dans les y_i pour qu'il ne soit pas nécessaire d'utiliser le modèle le plus complexe (car trop de paramètres à estimer). Cela revient à tester l'hypothèse de nullité de $k(= p - q)$ paramètres du modèle :

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \text{ avec } k \text{ petit devant } p$$

Sous l'hypothèse alternative, au moins un des paramètres $\beta_1, ..., \beta_k$ est non-nul.

Ce test peut être formulé comme la comparaison de deux modèles emboités, l'un à $p+1$ paramètres et l'autre à $q+1$ paramètres. L'hypothèse H_0 peut être testée au moyen de la statistique :

$$F_{cal} = \frac{SS_{E_0} - SS_{E_1}}{SS_{E_1}} \frac{n-p-1}{k} \sim F(k, n-p-1)$$

où SS_{E_0} est la somme des carrés résiduelles du modèle réduit sous H_0 et SS_{E_1} est la somme des carrés résiduelles du modèle de référence (modèle complet à p variables explicatives).

On compare F_{cal} à la valeur limite de la statistique d'une loi de Fisher $F_\alpha(k, n-p-1)$. Si $F_{cal} > F_\alpha(k, n-p-1)$ alors on rejette H_0.

Remarque : Dans le cas où $k = 1$, on teste la nullité d'un seul paramètre du modèle. Etant donné la propriété selon laquelle une variable aléatoire distribuée selon une loi $F(1, m)$ est le carré d'une variable aléatoire de Student à m degré de liberté ; le test de Fisher ci-dessus et le test de Student donnent les mêmes conclusions.

2.3 Annexe : Quelques rappels sur les lois

Loi du Chi-deux

Si $X_1, X_2, \ldots, X_n$ sont des variables aléatoires $N(0,1)$ et indépendantes alors

$$Q_n = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

avec $\mathrm{E}(Q_n) = n$ et $\mathrm{Var}(Q_n) = 2n$

Remarques :
La somme de 2 chi-deux indépendantes est aussi un chi-deux.
Une variable du chi-deux est toujours positive.

Loi de Student

Si $X \sim N(0,1)$ et $Q \sim \chi_n^2$ avec X et Q deux variables indépendantes alors

$$T_n = \frac{X}{\sqrt{Q/n}} \sim t_n$$

Remarque :
Si $n \to +\infty$, t_n tend vers une loi normale réduite.

Loi de Fisher

Si $Q_1 \sim \chi_{n_1}^2$ et $Q_2 \sim \chi_{n_2}^2$ avec Q_1 et Q_2 deux variables indépendantes alors

$$F_{n_1;n_2} = \frac{Q_1/n_1}{Q_2/n_2} \sim F_{n_2}^{n_1}$$

Théorème de Cochran

Soient $X \sim N(\mu, \Sigma)$; A et B deux matrices carrées, symétriques ($A' = A$ et $B' = B$) et idempotentes ($AA = A$ et $BB = B$) d'ordre n ; a un vecteur aléatoire de $\mathbb{R}^n$; $Q_1 = X'AX$ et $Q_2 = X'BX$, deux formes quadratiques,
ALORS

1) $A\Sigma B = 0 \Longrightarrow Q_1$ et Q_2 indépendantes
2) $A\Sigma a = 0 \Longrightarrow Q_1$ et $a'X$ indépendantes
3) $\parallel X - \mu \parallel^2 \sim \chi^2_n$
4) Si $rang(A) = r$, $\Sigma = I$ et $\mu = 0$ alors $X'AX \sim \chi^2_r$

Chapitre 3

Modèle linéaire : Analyse de variance

L'analyse de variance est un cas particulier de la régression. La différence essentielle est la structure que possèdent les variables explicatives. L'objectif de l'analyse de variance est la recherche de relations entre une variable quantitative et des variables qualitatives (appelées aussi facteurs de variation). Quand un seul facteur de variation est utilisé pour expliquer les variations de y, on réalise une analyse de la variance à un facteur (à une voie). Dans le cas général, plusieurs facteurs (p) sont présents pour expliquer les variations de y, on parle alors d'analyse de variation à p facteurs.

3.1 ANOVA à un facteur

3.1.1 Un exemple

Données Ampoules. Dans cet exemple, on considère 6 procédés de fabrication de lampes à ultra-violets :

F1	1602 1615 1624 1631
F2	1472 1477 1485 1493 1496 1504 1510
F3	1548 1555 1559 1563 1575
F4	1435 1438 1448 1449 1454 1458 1467 1475
F5	1493 1498 1509 1516 1521 1523
F6	1585 1592 1598 1604 1609 1612

TABLE I – Observations de durées de vie d'ampoules échantillonnées pour 6 procédés de fabrication.

On numérote les u.s. (i, j), où i est le numéro du procédé de fabrication et j est le numéro de la lampe à i fixé. On note y_{ij} la durée de vie de la j *ème* lampe fabriquée suivant le procédé i, et μ_i la durée de vie moyenne d'une lampe fabriquée suivant le procédé i.

Le modèle s'écrit :

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, ...6 \quad j = 1, ...n_i \tag{3.1}$$

où e_{ij} est un résidu tel que $e_{ij} \sim N(0, \sigma^2)$ et n_i le nombre d'observations pour le procédé i. Les résidus sont supposés être indépendants. Le modèle peut également s'écrire comme celui d'une régression linéaire multiple :

$$y_{ij} = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_6 x_6 + e_{ij}$$

avec $x_i = 1$ si y_{ij} est une observation faite dans la classe i et $x_i = 0$ sinon. On a la relation suivante entre l'écriture des deux modèles d'analyse de variance : $\beta_i = \mu_i$.

3.1.2 Diverses paramétrisations

Analysons les données Ampoules avec le logiciel R :

> options(contrasts="contr.treatment")

> reg2 <-lm(dvie ~ proc)

> summary(reg2)

```
Call : lm(formula = dvie   proc)
Residuals :
 Min   1Q     Median    3Q   Max
 -19   -9   4.996e-15   9.5    22

   Coefficients :
                 Value   Std.Error    t value   Pr(>|t|)
 (Intercept)  1618.0000      6.2022  260.8773     0.0000
      procF2  -127.0000      7.7748  -16.3348     0.0000
      procF3   -58.0000      8.3211   -6.9703     0.0000
      procF4  -165.0000      7.5961  -21.7218     0.0000
      procF5  -108.0000      8.0069  -13.4883     0.0000
      procF6   -18.0000      8.0069   -2.2480     0.0321

Residual standard error : 12.4 on 30 degrees of freedom
Multiple R-Squared : 0.9644
F-statistic : 162.7 on 5 and 30 degrees of freedom, the p-value is 0
```

> reg2$fitted.values

```
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15
 1618  1618  1618  1618  1491  1491  1491  1491  1491  1491  1491  1560  1560  1560  1560
   16    17    18    19    20    21    22    23    24    25    26    27    28    29    30
 1560  1453  1453  1453  1453  1453  1453  1453  1453  1510  1510  1510  1510  1510  1510
   31    32    33    34    35    36
 1600  1600  1600  1600  1600  1600
```

Analysons maintenant les données Ampoules avec le logiciel SAS :

proc glm data=ampoules ;
class proc ;
model dvie=proc / solution p ;
run ;

```
                                The GLM Procedure

   Dependent Variable : dvie
                              Sum of
 Source             DF       Squares   Mean Square   F Value   Pr > F
 Model               5   125144.7500    25028.9500    162.67   <.0001
 Error              30     4616.0000      153.8667
 Corrected Total    35   129760.7500
```

```
R-Square  Coeff Var  Root MSE  dvie Mean
0.964427   0.812021  12.40430   1527.583

Source  DF    Type I SS  Mean Square  F Value  Pr > F
proc     5  125144.7500   25028.9500   162.67  <.0001

Source  DF  Type III SS  Mean Square  F Value  Pr > F
proc     5  125144.7500   25028.9500   162.67  <.0001

                          Standard
Parameter       Estimate       Error  t Value  Pr > |t|
Intercept  1600.000000 B  5.06403440   315.95    <.0001
proc F1      18.000000 B  8.00694143     2.25    0.0321
proc F2    -109.000000 B  6.90111562   -15.79    <.0001
proc F3     -40.000000 B  7.51117686    -5.33    <.0001
proc F4    -147.000000 B  6.69908783   -21.94    <.0001
proc F5     -90.000000 B  7.16162613   -12.57    <.0001
proc F6       0.000000 B           .        .         .
```

```
NOTE : The X'X matrix has been found to be singular, and a generalized inverse was used
to solve the normal equations. Terms whose estimates are followed by the letter 'B' are
not uniquely estimable.
```

```
Observation      Observed     Predicted     Residual
          1  1602.0000000  1618.0000000  -16.0000000
          2  1615.0000000  1618.0000000   -3.0000000
          3  1624.0000000  1618.0000000    6.0000000
          4  1631.0000000  1618.0000000   13.0000000
          5  1472.0000000  1491.0000000  -19.0000000
          6  1477.0000000  1491.0000000  -14.0000000
          7  1485.0000000  1491.0000000   -6.0000000
          8  1493.0000000  1491.0000000    2.0000000
          9  1496.0000000  1491.0000000    5.0000000
         10  1504.0000000  1491.0000000   13.0000000
etc ...
```

On observe que les valeurs ajustées (*reg2$fitted.values* pour R et *colonne Predicted* pour SAS) sont les mêmes avec R et SAS, et pourtant les estimations des β (*Value* pour R, *Estimate* pour SAS) sont différentes. Pourquoi ?

Explications : En fait, il existe plusieurs paramétrisations possibles, que nous allons décrire.

Paramétrisation du modèle : Les paramètres sont les μ_i pour $i = 1, ...p$.

Décomposition centrée : On écrit $\mu_i = \mu + \alpha_i$ avec $\sum \alpha_i = 0$.

Décomposition SAS/R = une cellule de référence : On écrit $\mu_i = \mu_p + a_i$ avec $a_p = 0$ dans SAS : le dernier niveau du facteur sert de référence ; soit $a_i = (\mu_i - \mu_p)$ le contraste entre le niveau i et le niveau p. Ou $\mu_i = \mu_1 + a_i$ avec $a_1 = 0$ dans R : le premier niveau du facteur sert de référence, soit $a_i = (\mu_i - \mu_1)$.

Toutes ces paramétrisations sont équivalentes car on peut passer de l'une à l'autre par une bijection.

3.1.3 Vérification des hypothèses - Diagnostics

Comme dans le cadre de la régression, des vérifications sont à effectuer : normalité des résidus, homoscédasticité, valeurs atypiques, ...

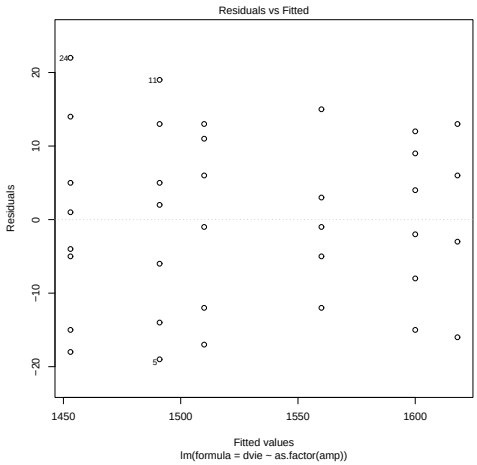


FIGURE 3.1 – Données Ampoules : visualisation des résidus.

3.1.4 Estimation des paramètres

Comme pour la régression, les paramètres du modèle d'analyse de variance peuvent être estimés par la méthode des moindre carrés ou par la méthode du maximum de vraisemblance.

μ_i est estimé par la moyenne empirique :

$$\hat{\mu}_i = y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}.$$

Cette estimation est d'autant plus précise que le nombre d'observations pour la cellule i est grand :

$$Var(\hat{\mu}_i) = \frac{\sigma^2}{n_i}.$$

La variance résiduelle est estimée par :

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

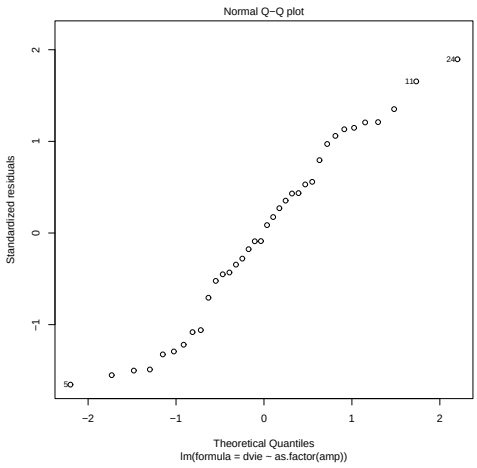


FIGURE 3.2 – Données Ampoules : QQplot des résidus.

3.1.5 Intervalle de confiance et tests d'hypothèses

Soit le modèle $y_{ij} = \mu_i + e_{ij}$ où les e_{ij} sont iid suivant une loi centrée de variance σ^2 qui sera supposée $N(0, \sigma^2)$ pour la construction des tests. Dans le cadre général du modèle gaussien, on a montré que les estimateurs des paramètres du modèle sont distribués selon une loi normale, donc :

$$\hat{\mu}_i \sim N(\mu_i, \sigma^2/n_i)$$

On peut en déduire un intervalle de confiance de μ_i de sécurité $1 - \alpha$:

$$\left[\hat{\mu}_i \pm t_{(n-I),(1-\alpha/2)}\sqrt{\frac{\hat{\sigma}^2}{n_i}}\right]$$

L'hypothèse $H_0 : \mu_1 = ... = \mu_I$ revient à dire que la moyenne est indépendante du niveau ou encore que le facteur n'a pas d'effet et l'hypothèse alternative H_1 est définie par $\exists(i, k)$ tel que $\mu_i \neq \mu_k$. Cette dernière hypothèse revient à reconnaître un effet ou une influence du facteur sur la variable Y. L'étude de cette hypothèse revient à comparer par un test de Fisher un modèle complet (les moyennes sont différentes) avec un modèle réduit supposant la nullité des paramètres et donc l'égalité des moyennes à celle de la dernière cellule ou à la moyenne générale. Les résultats nécessaires à la construction du test qui en découle sont résumés dans la table d'analyse de variance :

Source de variation	ddl	Somme des carrés	Variance	F
Modèle (inter)	p-1	SSB	MSB=SSB /(p-1)	MSB/MSW
Erreur (intra)	n-p	SSW	MSW=SSW/(n-p)	
Total	n-1	SST		

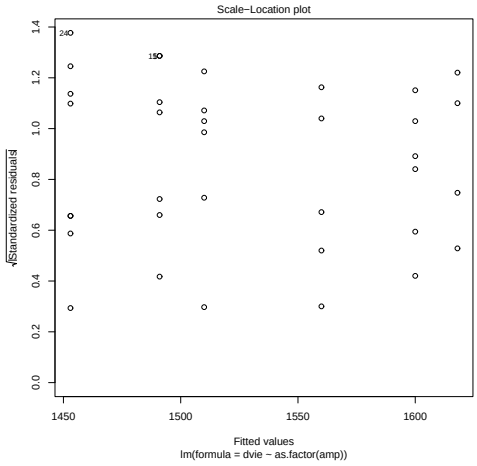


FIGURE 3.3 – Données Ampoules : recherche de liaison moyenne-variance.

Avec SSB= $\sum_{i,j}(y_{i.}-y_{..})^2$; SSW= $\sum_{i,j}(y_{ij}-y_{i.})^2$; SST= $\sum_{i,j}(y_{ij}-y_{..})^2$; un point à la place d'un indice veut dire la moyenne sur l'indice considéré ($y_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$).

La statistique F peut s'interpréter comme le rapport de la variabilité inter-groupe sur la variabilité intra-groupe. En effet, le carré moyen du modèle mesure l'écart des moyennes des groupes à la moyenne générale (c'est une mesure de variabilité inter). Le carré moyen résiduel mesure l'écart de chaque unité statistique à la moyenne du groupe (c'est une mesure de la variabilité intra). Si le facteur a un effet sur la variable à expliquer, la variation INTER sera importante par rapport à la variation INTRA.

Dans le cas d'un facteur à 2 classes ($p = 2$), on retrouve un test équivalent au test de Student de comparaison des moyennes de deux échantillons indépendants.

3.2 ANOVA à deux facteurs croisés

La considération de deux (ou plus) facteurs explicatifs dans un modèle d'analyse de variance engendre plusieurs complications dont en particulier la notion d'interaction entre variables explicatives. La présence d'interaction atteste du fait que les effets d'un des facteurs dépend des effets de l'autre facteur.

Les niveaux du premier facteur sont notés par un indice i variant de 1 à p, ceux du deuxième facteur par un indice j variant de 1 à q. Pour chaque combinaison, on observe un même nombre $n_{ij} = c > 1$ de répétition, ce qui nous place dans le cas particulier d'un plan équilibré ou équirépété. Ceci introduit des simplifications importantes dans les estimations des paramètres ainsi que dans la décomposition des variances.

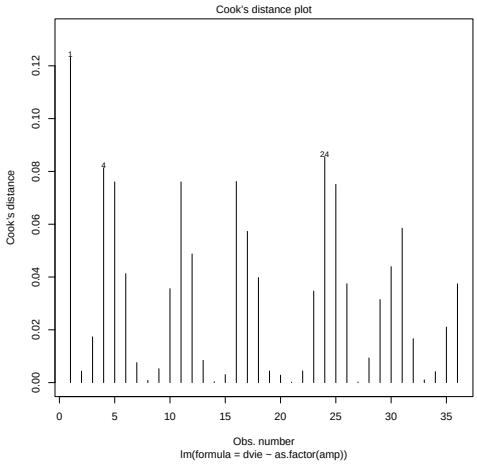


FIGURE 3.4 – Données Ampoules : distance de Cook pour détection d'observations aberrantes.

Le modèle général s'écrit :

$$y_{ijk} = \mu_{ij} + e_{ijk}$$

On suppose que les termes d'erreur e_{ijk} sont mutuellement indépendants et de même loi gaussienne. Le modèle d'analyse de variance à deux facteurs s'écrit également de la manière suivante :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

avec les contraintes : $\sum_{i=1}^{p} \alpha_i = \sum_{j=1}^{q} \beta_j = 0$, $\forall j, \sum_{i=1}^{p} \gamma_{ij} = 0$ et $\forall i, \sum_{j=1}^{q} \gamma_{ij} = 0$.

Lorsque les paramètres d'interaction γ_{ij} sont tous nuls, le modèle est dit $additif$, ce qui correspond à une situation très particulière. Ceci signifie que les écarts relatifs au premier facteur sont indépendants du niveau k du 2ème facteur et vice versa. Dans le cas équirépété, les tests des effets sont résumés dans la table d'analyse de variance suivante :

Source de variation	ddl	Somme des carrés	Variance	F
1er facteur	p-1	SS1	MS1=SS1/(p-1)	MS1/MSE
2ème facteur	q-1	SS2	MS2=SS2/(q-1)	MS2/SSE
Interaction	(p-1)(q-1)	SSI	MSI=SSI/(p-1)(q-1)	MSI/MSE
Erreur	n-pq	SSE	MSE=SSE/(n-pq)=$\hat{\sigma}^2$	
Total	n-1	SST		

avec

$SS1 = qc\sum_i (y_{i..} - y_{...})^2$;
$SS2 = pc\sum_j (y_{.j.} - y_{...})^2$;
$SSI = c\sum_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2$;

$SSE = \sum_{ijk}(y_{ijk} - y_{ij.})^2$
$SST = \sum_{ijk}(y_{ijk} - y_{...})^2$.

Ici aussi, plusieurs paramétrisations sont possibles et sont en correspondance bijective.

Paramétrisation du modèle : les paramètres sont ici μ_{ij} pour $i = 1, ...p$ et $j = 1, ...q$.

Décomposition centrée : on écrit $\mu_{ij} = \mu + \alpha_i^l + \alpha_j^c + \alpha_{ij}^x$ avec $\sum_i \alpha_i^l = \sum_j \alpha_j^l = \sum_j \alpha_{ij}^x = \sum_i \alpha_{ij}^x = 0$.

Décomposition SAS/R : on écrit : $\mu_{ij} = \mu_{pq} + a_i^l + a_j^c + a_{ij}^x$ avec $a_p^l = a_q^c = a_{pj} = a_{iq} = 0$ (dans SAS),
ou $\mu_{ij} = \mu_{11} + a_i^l + a_j^c + a_{ij}^x$ avec $a_1^l = a_1^c = a_{1j} = a_{i1} = 0$ (dans R).

L'estimation et l'inférence sur les paramètres, ainsi que les analyses post-modélisations sont les mêmes que dans le cadre de la régression.

On peut se faire une idée de la présence d'interactions en traçant le graphe des moyennes empiriques $\bar{y}_{ij}$ en fonction de i, pour chaque j (figure 3.2). Si les droites sont parallèles, l'effet du premier facteur s'additionne à l'effet du deuxième, il n'y a donc pas d'interaction. Si par contre des droites se croisent, on peut suspecter la présence d'interactions.

3.3 Analyse de covariance

L'analyse de covariance se situe encore dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par plusieurs variables à la fois quantitatives et qualitatives. Dans les cas les plus complexes, on peut avoir plusieurs facteurs (variables qualitatives) avec une structure croisée ou hiérarchique ainsi que plusieurs variables quantitatives intervenant de manière linéaire ou polynomiale. Le principe général est toujours d'estimer des modèles $intra - groupes$ et de faire apparaître (tester) des effets différentiels $inter - groupes$ des paramètres des régressions. Ainsi, dans le cas simple où seulement une variable parmi les explicatives est quantitative, nous sommes amenés à tester l'hétérogénéité des constantes et celle des pentes (interaction) entre différents modèles de régression linéaire.

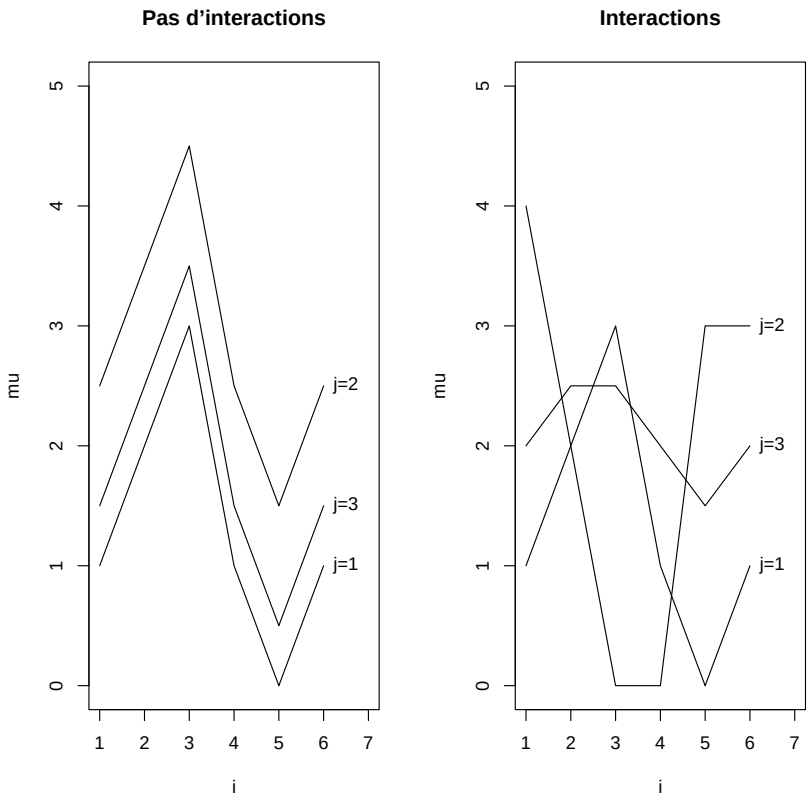
Prenons un cas simple, le modèle est explicité dans le cas élémentaire où une variable quantitative Y est expliquée par une variable qualitative T à q niveaux et une variable quantitative, appelée encore covariable, X. Pour chaque niveau j de T, on observe n_j valeurs $x_{1j}, ..., x_{n_j j}$ de X et n_j valeurs $y_{1j}, ..., y_{n_j j}$ de Y ; n est la taille de l'échantillon. Le modèle s'écrit :

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$$

où les e_{ij} sont iid suivant une loi centrée de variance σ_e^2 qui sera supposée $N(0, \sigma_e^2)$ pour la construction des tests.

Différentes hypothèses peuvent alors être testées par un test de Fisher :
(1) Test des interactions : les droites partagent la même pente ;
(2) Test de l'influence du facteur quantitatif ;
(3) Test de la significativité des différences des termes constants.

On commence par tester l'hypothèse (1), si le test n'est pas significatif, on regarde (2) qui s'il n'est pas non plus significatif, conduit à l'absence d'effet de la variable X. De même, toujours si (1) n'est pas significatif, on s'intéresse à (3) pour juger de l'effet du facteur T.


Pas d'interactions
Interactions
mu
i
j=1
j=2
j=3
0
1
2
3
4
5
1
2
3
4
5
6
7

Chapitre 4

Premières notions sur les tests multiples

4.1 Rappels sur les risques de première et seconde espèce

Risque de première espèce : noté α. Le risque de première espèce est le risque de rejeter (avec la règle décision) l'hypothèse H_0 alors qu'en réalité cette hypothèse est vraie.

Risque de seconde espèce : noté β. le risque de seconde espèce est le risque d'accepter (avec la règle de décision) l'hypothèse H_0 alors qu'en réalité cette hypothèse est fausse.

	Décision	
Réalité	H_0	H_1
H_0	$1-\alpha$	α
H_1	β	$1-\beta$

La quantité $1-\beta$ est une probabilité de bonne décision appelé puissance du test.

Remarque : Accepter H_0 ne signifie pas que cette hypothèse est vraie mais seulement que les observations disponibles ne sont pas incompatibles avec cette hypothèse et que l'on n'a pas de raison suffisante de lui préférer l'hypothèse H_1 compte tenu des résultats expérimentaux.

4.2 Tests multiples

Supposons que p moyennes $(m_1, m_2, ..., m_p)$ soient à comparer et que ces p moyennes soient respectivement estimées par : $\bar{X}_1, \bar{X}_2, ..., \bar{X}_p$ et que ces moyennes soient estimées sur des échantillons de tailles respectives $n_1, n_2, ..., n_p$. En comparant les moyennes deux à deux, il faut faire $\frac{p(p-1)}{2}$ comparaisons. Chaque comparaison de 2 moyennes est effectuée en utilisant la règle de décision suivante : si

$$\frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\hat{\sigma}^2(1/n_i + 1/n_j)}} > t_{1-\alpha/2;n_i+n_j-2}$$

alors on rejette l'hypothèse $H_0 : m_i = m_j$.

Si deux comparaisons sont réalisées avec un risque de première espèce de α, il est faux de penser que la décision globale peut être prise avec un risque α. Cela provient du fait qu'une succession de tests de risque α ne permet pas de prendre une décision globale avec ce même risque.

	2	3	4	5	6
Erreur nominale de type I	5%	5%	5%	5%	5%
Erreur globale de type I	5%	12.2 %	20.3 %	28.6%	36.6%

La méthode de Bonferroni est une méthode qui ne permet pas un strict contrôle de α mais en revanche elle en donne une majoration. L'idée de Bonferroni est de se placer dans le pire des cas (pour α). Par exemple si on a $p = 5$ moyennes à comparer, il faut effectuer 10 comparaisons. Pour avoir un risque global α, il faut que chacune des 10 comparaisons soit effectuée avec un risque $\alpha' = \alpha/10$.

En pratique, Bonferroni fournit une liste de gènes différentiellemnt exprimés dans laquelle on contrôle le nombre de faux positifs. Quand le nombre des gènes est grand, cette liste est souvent vide.

A l'opposé, le LSD (Least Square Difference), c'est à dire le test de Student sans correction, est le plus laxiste : il va détecter des gènes différentiellement exprimés qui en fait ne le sont pas.

En général, on présente ces taux d'erreurs dans le tableau suivant :

Réalité	*Décision*		
	H_0 vraie	H_1 vraie	Total
H_0 vraie	U	V	m_0
H_1 vraie	T	S	m_1
	W	R	m

où m tests sont effectués. Pour une analyse de données transcriptomiques dans laquelle on teste les effets différentiels de m gènes, m_1 est le nombre de gènes déclarés différentiellement exprimés, alors que R est le nombre réel (mais inconnu) de gènes différentiellement exprimés.

Diverses méthodes sont proposées pour contrôler ces divers taux d'erreurs.

Le FWER (Family Wise Error Rate) représente la probabilité d'effectuer au moins une erreur de première espèce sur l'ensemble des comparaisons :

$$P[V \geq 1] = m_0\alpha.$$

On prend donc un seuil nominal de $\alpha' = \alpha/m_0$.

Au même titre que Bonferroni, plus il y a de tests (soit de gènes à tester), moins on rejette H_0 (moins de gènes déclarés différentiellement exprimés). La notion suivante est très utile pour pallier à cet inconvénient.

La FDR (False Discovery Rate) contrôle l'espérance du taux de faux positifs, ou le nombre de faux positifs parmi les différences déclarées significatives. Pratiquement, on ordonne les m p-values des m tests (les gènes) et on recherche le plus haut rang k des p-values tel que $p - value(k) \geq \alpha k/m$.

Il existe d'autres approches récentes ou en cours de développement pour contrôler la FDR positive, le nombre moyen d'erreurs, etc ... . Pour cela, nous conseillons au lecteur curieux de parcourir la littérature abondante sur ce sujet au cours de ces dernières années.

Chapitre 5

Un exemple : données de nutrition chez la souris

5.1 Les données

Les données ont été fournies par Thierry Pineau et Pascal Martin de l'unité TOXALIM (INRA, Toulouse - site de Saint-Martin). Elles proviennent d'une étude de nutrition chez la souris. Pour 40 souris, nous disposons :

- des données d'expression de 120 gènes recueillies sur membrane nylon avec marquage radioactif,
- des mesures de 20 acides gras hépatiques.

Par ailleurs, les 40 souris sont réparties selon deux facteurs :

- Génotype (2 modalités) : les souris sont soit de type sauvage (wt) soit génétiquement modifiées (PPAR) ; 20 souris dans chaque cas.
- Régime (5 modalités) : les 5 régimes alimentaires sont notés `ref`, `efad`, `dha`, `lin`, `tournesol` ; 4 souris de chaque génotype sont soumises à chaque régime alimentaire.

Le modèle linéaire va nous permettre de répondre à quelques questions :

- Quels sont les gènes différentiellement exprimés ?
- Quelle est l'influence des facteurs "génotype" et "régime" sur l'expression des gènes ?

5.2 Principe des analyses de variance

L'analyse de variance (ANOVA) permet d'apprécier l'effet d'une ou plusieurs variables qualitatives (les facteurs) sur une variable quantitative (la variable réponse, ici le niveau d'expression des gènes). Dans le domaine de l'analyse transcriptomique, cette approche a été largement développée, en particulier par Kerr *et al.* (2000). Pour l'analyse de nos données, un modèle d'ANOVA à trois facteurs (génotype, régime, gène) permet de mettre en évidence des effets d'interaction d'ordre 3 très significatifs à l'aide du test de Fisher. Cela signifie qu'il existe des gènes régulés simultanément par le régime et le génotype, les effets du régime et du génotype étant non additifs. Le modèle d'ANOVA considéré s'écrit

$$y_{ijkl} = g_i + r_j + G_k + gr_{ij} + gG_{ik} + rG_{jk} + grG_{ijk} + e_{ijkl} \tag{5.1}$$

où y_{ijkl} représente le logarithme du niveau d'expression du gène k ($k = 1, ..., 120$), pour le régime j ($j = 1, ..., 5$) et le génotype i ($i = 1, 2$), mesuré chez la souris l ($l = 1, ..., 4$) ; g_i représente l'effet du génotype i, r_j celui du régime j, G_k celui du gène k, gr_{ij} représente l'effet de l'interaction du génotype i et du régime k, gG_{ik} l'effet de l'interaction du génotype i et du gène k, rG_{jk} l'effet de l'interaction du regime j et du gène k et grG_{ijk} représente l'interaction d'ordre

3 combinant le génotype i, le régime j et le gène k. On suppose que les résidus e_{ijkl} du modèle sont indépendants et identiquement distribués suivant une loi normale de moyenne nulle et de variance σ^2. L'écriture d'un tel modèle suppose que les gènes sont tous de même variabilité. Dans cet exemple, cette hypothèse est discutable. En effet, si vous tracez la figure représentant le diagramme en boîtes pour les 120 gènes, vous verrez que quelques gènes semblent fortement variables. Nous verrons par la suite comment lever cette hypothèse. À partir de ce modèle, on peut estimer les effets principaux des 120 gènes, effectuer des comparaisons de moyennes à l'aide du test de Fisher, puis opérer des corrections pour des tests multiples afin de repérer les gènes surexprimés ou sous-exprimés selon le génotype et le régime.

Dans cette séquence de tests, les variances des gènes sont supposées égales, contrairement aux tests de Student de comparaison de moyennes par régime et génotype pour un gène fixé. Ce dernier cas revient à écrire une modèle d'ANOVA par gène, sous la forme suivante

$$y_{ijl} = g_i + r_j + gr_{ij} + e_{ijl} \tag{5.2}$$

où les notations utilisées ici sont identiques à celles du modèle (5.1). Ici, il est nécessaire de faire autant d'analyses de variance que de gènes étudiés (soit 120 dans notre exemple) mais nous disposerons d'une variance estimée par gène. Toutefois une telle analyse n'est pas toujours recommandée car en règle générale le nombre d'observations par gène est très faible, ce qui conduit à des estimations de variance très peu précises. Notons cependant que ces 120 analyses conduisent à 120 estimations des 10 effets $genotype_i \times regime_j$. Un modèle équivalent, mais utilisant simultanément l'ensemble des données pour estimer les paramètres, s'écrit comme le modèle (5.1) en posant

$$\mathrm{var}(e_{ijkl}) = \sigma^2_{e,k}. \tag{5.3}$$

D'autre part, entre le modèle (5.1), supposant toutes les variances des gènes égales, et le modèle (5.3) supposant une variance différente pour chaque gène, il est possible d'ajuster un modèle intermédiaire prenant en compte les hétérogénéités de variances de l'expression des gènes, en définissant simplement des groupes de gènes de variabilité homogène (Robert-Granié *et al.*, 1999 ; Foulley *et al.*, 2000 ; San Cristobal *et al.*, 2002). Ainsi, sur les 120 gènes analysés, un histogramme des variances nous a conduit à définir trois groupes de gènes ayant des variabilités très différentes : un groupe contenant les gènes `FAS`, `G6Pase`, `PAL` et `S14`, présentant des variabilités résiduelles importantes (variances supérieures à 0.02) ; un deuxième groupe à variabilité modérée (variances comprises entre 0.009 et 0.02), comprenant les gènes `CYP2c29`, `CYP3A11`, `CYP4A10`, `CYP4A14`, `CYP8b1`, `GSTmu`, `GSTpi2`, `L-FABP`, `Lpin`, `Lpin1`, `TRa` et `cHMGCoAS` ; enfin un dernier groupe à faible variabilité (variances inférieures à 0.009), contenant l'ensemble des autres gènes. À partir de ces trois groupes de gènes, nous pouvons construire un modèle dont la variance dépend de cette nouvelle variable à trois classes. Le modèle s'écrit encore comme les modèles (5.1) et (5.3) en posant cette fois

$$\mathrm{var}(e_{ijkl}) = \sigma^2_h, \tag{5.4}$$

où $h = \{1, 2, 3\}$ représente l'indice d'hétérogénéité de variance.

Nous pouvons ainsi comparer les gènes différentiellement exprimés selon les 3 modèles :
- Modèle (5.1), modèle d'ANOVA avec une unique variance pour l'ensemble des gènes ;
- Modèle (5.3), modèle d'ANOVA avec une variance différente par gène ;
- Modèle (5.4), modèle d'ANOVA avec trois groupes de variances différentes.

Notons que le modèle (5.3) implique l'estimation de 120 variances différentes, alors que le modèle (5.4) ne nécessite l'estimation que de trois paramètres de variances ; ce dernier est donc

beaucoup plus économe en nombre de paramètres à estimer. Enfin, d'un point de vue technique et opérationnel, la mise en oeuvre de ces modèles peut être réalisée en utilisant la fonction `lme` du logiciel statistique R ou la procédure `mixed` du logiciel SAS.

5.3 Synthèse des tests multiples

L'objectif de l'analyse statistique est de déterminer quels sont les gènes différentiellement exprimés entre les 2 génotypes et les 5 régimes. Quelle que soit la méthode statistique utilisée, il existera une probabilité non nulle (risque de première espèce α) de détecter des faux positifs (gènes déclarés différentiellement exprimés alors qu'ils ne le sont pas) et une autre probabilité non nulle (risque de deuxième espèce β) de ne pas être capable de détecter des gènes réellement différentiellement exprimés (faux négatifs). Il est bien entendu souhaitable de minimiser ces deux probabilités d'erreur sachant que, toutes choses égales par ailleurs, la seconde augmente quand la première diminue et réciproquement. Le test de Student est couramment utilisé pour tester l'égalité de deux moyennes (l'hypothèse nulle étant de considérer que les moyennes des intensités des signaux d'un gène donné dans chaque condition 1 et 2 sont égales). Ainsi, quand la statistique de Student excède un certain seuil (dépendant du risque de première espèce α choisi, généralement 5%), les niveaux d'expression du gène étudié entre les deux populations testées sont considérées comme significativement différentes. Lorsque l'on souhaite tester plus de deux conditions, le test de Fisher, qui est une extension du test de Student, est utilisé. L'hypothèse nulle constitue l'absence d'expression différentielle d'un gène entre les diverses conditions et l'hypothèse alternative montre une différence d'expression.

Enfin, prendre un risque de 5% dans une expérimentation où 10 000 gènes, par exemple, sont étudiés simultanément peut conduire à obtenir 500 faux positifs, ce qui est parfaitement inacceptable. C'est pourquoi ont été proposées des modifications du test de Student adaptées à l'analyse du transcriptome (méthodes de Bonferroni, FWER, FDR...). Le lecteur souhaitant des détails sur ces approches peut se référer, par exemple, à Benjamini & Hochberg (1995), Bland & Altman (1995), Dudoit *et al.* (2002) ou Speed (2003).

La méthode de Bonferroni, rappelons-le, est une méthode qui ne permet pas un strict contrôle de α, mais qui en donne une majoration. Pour avoir un risque global α, il faut que chacune des p comparaisons soit effectuée avec un risque $\alpha' = \alpha/p$. En pratique, Bonferroni fournit une liste de gènes différentiellemnt exprimés dans laquelle on contrôle le nombre de faux positifs. Mais, lorsque le nombre des gènes est grand, cette liste est souvent vide.

Pour revenir à notre étude, à partir de chaque modèle proposé dans le paragraphe précédent, nous pouvons rechercher les gènes différentiellement exprimés entre les deux génotypes à régime fixé (120 comparaisons pour chacun des 5 régimes) ou entre régime à génotype fixé (1200 comparaisons par génotype), ce qui conduit à effectuer 3000 comparaisons. Le tableau I présente le nombre de gènes sélectionnés selon les trois modèles considérés et selon le test ou l'ajustement utilisée (Student, Bonferroni, Benjamini-Hochberg qui correspond à l'approche FDR).

On peut remarquer que le nombre de gènes sélectionnés est peu différent selon le modèle utilisé et que, globalement, les trois modèles sélectionnent le même groupe de gènes. Les petites différences sont principalement liées à l'ordre de sélection de ces gènes.

D'autre part, on peut, à partir de critères de sélection de modèle tels que le critère d'Akaike (AIC ; Akaike, 1974) ou le critère de Schwarz (BIC ; Schwarz, 1978), ou encore en effectuant un test du rapport de vraisemblance, choisir le modèle le plus adéquat.

TABLE I – *Nombre de gènes sélectionnés selon le modèle et le test utilisés.*

Tests	Modèle (5.1)	Modèle (5.3)	Modèle (5.4)
Student à 5%	85	103	97
Student à 1%	55	65	67
Benjamini-Hochberg à 5%	44	56	59
Benjamini-Hochberg à 1%	35	40	38
Bonferroni à 5%	53	62	65
Bonferroni à 1 pour mille	18	19	21

Le tableau II présente les valeurs des critères AIC et BIC pour les trois modèles mis en compétition.

TABLE II – *Valeurs des critères AIC et BIC.*

Modèles	-2AIC	-2BIC
(5.1)	-6576.9	-6570.7
(5.3)	-6946.6	-6612.1
(5.4)	-7044.5	-7036.2

Le meilleur modèle est celui pour lequel les valeurs des critères -2AIC ou -2BIC sont les plus petits. Dans les deux cas, il s'agit du modèle (5.4).

Le test du rapport de vraisemblance consiste, quant à lui, à comparer deux modèles emboîtés (par exemple, (5.1) *vs* (5.3)) ; l'hypothèse nulle considérée suppose alors que toutes les variances sont égales. La statistique du rapport de vraisemblance nécessite de calculer la différence entre les logarithmes des vraisemblances sous chacun des deux modèles. Sous l'hypothèse nulle, cette statistique suit asymptotiquement une loi de khi-deux dont le nombre de degré de liberté est égal à la différence des nombres de paramètres à estimer sous chacun des deux modèles considérés. Si nous effectuons ces différents tests du rapport de vraisemblance ((5.1) *vs* (5.3), (5.1) *vs* (5.4), (5.3) *vs* (5.4)), il en ressort que le modèle (5.4), avec trois groupes de variances, est encore le meilleur.

À partir de ce modèle (5.4), on peut estimer les différents effets du modèle, et s'intéresser aux différences d'expression des gènes entre génotypes à régime fixé ou encore aux différences d'expression des gènes entre régimes à génotype fixé.

En raison de la multiplicité des tests, la correction proposée par Benjami & Hochberg (1995) a été utilisée. Lorsque nous considérons les différences d'expression des gènes entre génotypes à régime fixé, l'hypothèse nulle représente l'absence d'expression différentielle d'un gène entre les deux génotypes. On peut visualiser l'ensemble des résulats des *p-values* de ces différents tests en effectuant une ACP centrée sur le logarithme des *p-values*, les gènes en ligne et les régimes en colonne. La figure 5.1 présente le premier plan principal des gènes différentiellement exprimés entres les deux génotypes à régime fixé. Les deux premiers axes principaux représentent 93% de la variance totale. Pour des raisons de visibilité, les résultats sont présentés sur les 59 gènes différentiellement exprimés selon le modèle (5.4) et en utilisant la correction de Benjamini & Hochberg à 5% (Tab. I).

On observe que les gènes `CYP3A11`, `CYP4A10`, `CYP4A14`, `L.FABP`, `PMDCI` et `THIOL` différencient

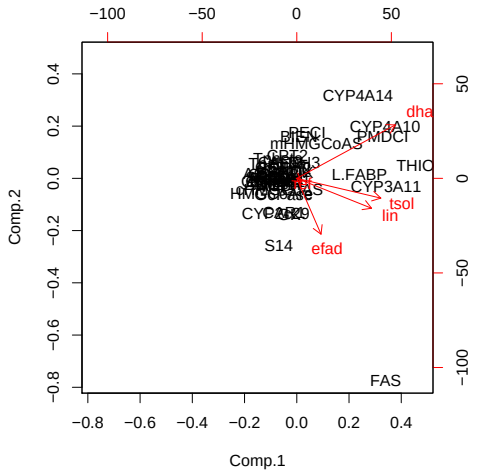


FIGURE 5.1 – *Représentation sur le premier plan principal de l'ACP du logarithme des p-value des gènes différentiellement exprimés entre les deux génotypes à régime fixé.*

les deux génotypes pour les régimes `dha`, `lin` et `tsol`. Certains de ces gènes présentent des expressions constitutives différentielles entre les souris des deux génotypes. De plus ces gènes sont régulés positivement par ces trois régimes riches en acides gras polyinsaturés d'une famille particulière (*Oméga 3* pour `dha` et `lin` et *Oméga 6* pour `tsol`) chez les souris `WT` alors que la régulation de plusieurs de ces gènes est altérée chez les souris PPAR. Les gènes `mHMGCoAS`, `PECI` et `BIEN` apparaissent dans le contraste entre génotypes pour le régime `dha`, alors que les gènes `S14` et `FAS` apparaissent pour le régime `efad`. Les souris des deux génotypes présentent là encore des régulations différentielles de ces gènes, soulignant ainsi le rôle du récepteur PPARα dans ces modulations d'expression provoquées par le s régimes alimentaires.

La même approche sur les effets différentiels entre couples de régimes, à génotype fixé, est réalisée. Les représentations de la figure 5.2 et de la figure 5.3 présentent le premier plan principal des gènes différentiellement exprimés entre régime pour le génotype `WT` et pour le génotype `PPAR`. Les deux premiers axes, pour chacune des figures, représentent respectivement 79% et 78% de la variance totale. Les gènes `Lpin` et `Lpin1` apparaissent dans des contrastes impliquant le régime `efad` pour le génotype `WT`, et le régime `tsol` pour le génotype `PPAR`. Le gène `CYP3A11` est impliqué dans le régime `dha`, quel que soit le génotype. Les gènes `FAS` et `S14` apparaissent dans les contrastes impliquant le régime `efad` pour le génotype `WT`, alors que le même gène `FAS` apparaît dans les contrastes impliquant le régime `ref` pour le génotype `PPAR`. L'ensemble de ces résultats confirme les résultats obtenus pour l'ACP.

Remerciements

Un grand merci à Pascal Martin, Gwenola Tosser-Klopp et Agnès Bonnet (INRA) pour les discussions scientifiques autour de leurs données respectives.

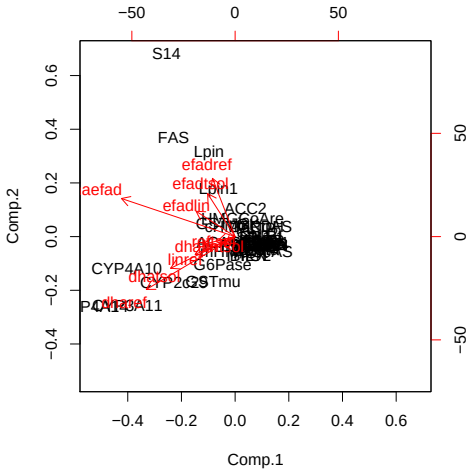


FIGURE 5.2 – *Représentation sur le premier plan principal de l'ACP du logarithme des p-value des gènes différentiellement exprimés entre les régimes pour le génotype* `WT`.

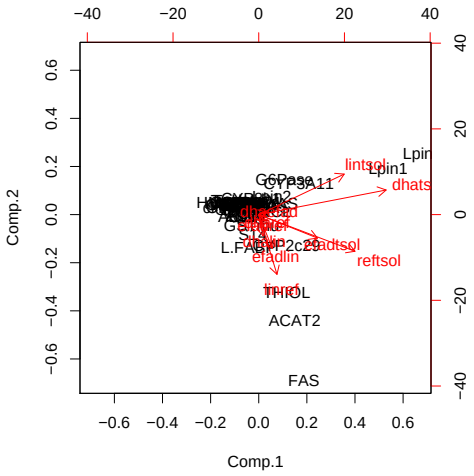


FIGURE 5.3 – *Représentation sur le premier plan principal de l'ACP du logarithme des p-value des gènes différentiellement exprimés entre les régimes pour le génotype* `PPAR`.

Chapitre 6

Quelques Références

- Benjamini Y., Hochberg Y. (1995), Controlling the false discovery rate : a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, 85, 289-300
- Bland J., Altman D. (1995), Multiple significance tests : the Bonferroni method, *British medical Journal*, 310, 170.
- Bonnet A., Benne F., Dantec C., Gobert N., Frappart P.O., SanCristobal M., Hatey F., Tosser-Klopp G. (2004), Identification of genes and gene networks involved in pig ovarian follicular development, by using c-DNA microarrays,*XIII International Workshop on the Development and Function of Reproductive organs*, 12-15 July 2004, Copenhagen, Denmark.
- Dalgaard P. (2003) *Introductory Statistics with R.* Springer. ISBN 0-387-95475-9 (accessible)
- Dudoit S.,Yang Y., Speed T., Callow M. (2002), Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, 12(1), 111-139.
- Faraway J.J. (2002) Practical Regression and Anova using R (chercher sur le site de Bioconductor) (Très complet et détaillé)
- Foulley J.-L., Jaffrezic F., Robert-Granié C. (2000), EM-REML estimation of covariances parameters in Gaussian mixed models for longitudinal data analysis, *Genetics Selection Evolution*, 32, 129-141.
- Kerr K., Martin M., Churchill G. (2000), Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, 7, 819-837.
- Lee S.S., Pineau T., Drago J., Lee E.J., Owens J.W., Kroetz D.L., Fernandez-Salguero P.M., Westphal H., Gonzalez F.J. (1995), Targeted disruption of the alpha isoform of the peroxisome proliferator-activated receptor gene in mice results in abolishment of the pleiotropic effects of peroxisome proliferators, *Molecular and Cellular Biology* 15(6), 3012-22.
- Martin P.G.P., Lasserre F., Calleja C., Van Es A., Roulet A., Concordet D., Cantiello M., Barnouin R., Gauthier B. and Pineau T. (2005), Transcriptional modulations by RXR agonists are only partially subordinated to PPARalpha signaling and attest additional, organ-specific, molecular cross-talks.s, *Gene Expression*, 12(3) :177-92.
- McLachlan G.J., Do K.-A., Ambroise C. (2004), *Analysing microarray gene expression data*, Wiley.
- Plateforme Biostatistique : http ://www.math.univ-toulouse.fr/ biostat/Accueil.html (pour toute aide pour vos analyses statistiques des données).
- Robert-Granié C., Bonaiti B., Boichard D., Barbat A. (1999), Accounting for variance heterogeneity in French dairy cattle genetic evaluation, *Livestock Production Science*, 60, 343-357.
- San Cristobal M., Robert-Granié C., Foulley JL. (2002), Hétéroscédasticité et

modèles linéaires mixtes : théorie et applications en génétique quantitative, *Journal de la Société Française de Statistique*, 143, 1-2.
- Saporta G. (1990), *Probabilités analyse des données et statistique*, Technip.
- Searle SR (1971) Linear Models. Wiley. ISBN 0-471-76950-9 (il faut maîtriser l'algèbre linéaire)
- Speed T. (2003), *Statistical Analysis of Gene Expression Microarray Data*, Interdisciplinary Statistics, Chapman & Hall/CRC.