ÉCOLE NORMALE SUPÉRIEURE DU BURUNDI

DÉPARTEMENT DES SCIENCES NATURELLES

SECTION : MATHÉMATIQUE

MODÈLES LINÉAIRES GÉNÉRALISÉS

CODE : MSM1106

UE1 : MATHÉMATIQUES APPLIQUÉES III

VOLUME : 60H (4 ECTS)

COURS MAGISTRAL (CM) : 45H

TRAVAUX DIRIGÉS (TD) : 0H

TRAVAUX PRATIQUES (TP) : 15H

SYLLABUS DE L'ÉLÉMENT CONSTITUTIF DE L'UNITÉ D'ENSEIGNEMENT (ECUE) DESTINÉ AUX ÉTUDIANTS DE MASTER 1 EN SCIENCES MATHÉMATIQUES ET ENSEIGNEMENT

Titulaire : Prof. Emmanuel BARANKANIRA

Docteur en Biostatistique de l'Université de Montpellier (France, 2016)

Master en Mathématiques-Biostatistique de SupAgro (France, 2012)

DEA en Statistique, parcours Épidémiologie et Biostatistique de l'UCL (Belgique, 2008)

Licencié en Pédagogie Appliquée, Agrégé de l'Enseignement en Maths (UB, 2004)

Bujumbura, 10 octobre 2024

Descriptif du cours

Processus	Paramètres	Description
Élaboration	Titre de l'ECUE	**Modèles linéaires généralisés**
	Objectif général	Modéliser des phénomènes de la vie à l'aide des fonctions mathématiques particulières
	Objectifs spécifiques	À la fin de l'ECUE, l'étudiant doit être capable de : – Estimer matriciellement et tester la significativité les paramètres d'un modèle linéaire multiple par la méthode des moindres carrés ordinaires (trois façons) ; – Faire la sélection des modèles à l'aide des critères statistiques adéquats ; – Appliquer des modèles linéaires généralisés (régression logistique, régression binomiale, régression Poisson) avec ou sans effets aléatoires à l'analyse des données ; – Construire un modèle logistique ordinal ; – Construire un modèle logistique multinomial ; – Faire une analyse de la variance à un, deux et trois critères de classification avec ou sans effet aléatoire ; – Restituer les résultats des analyses statistiques sous forme d'un compte-rendu ou d'un rapport.
	Prérequis	Logiciels de statistique
	Organisation de l'ECUE	CM : 45h; TD : 0h; TP : 15h; VHP : 60 h (4 crédits)
	Bref contenu de l'ECUE	Dans cet ECUE, les notions de statistique descriptive et de statistique analytique seront d'abord rappelées. Ensuite, le modèle linéaire multiple sera étudié en mettant un accent particulier sur la sélection et la validation des modèles, la vérification de

		l'adéquation du modèle, la détection des facteurs de confusion et des facteurs modificateurs d'effet, et sur l'interprétation des résultats. Cette analyse sera suivie par la construction des modèles logistiques (binaire, ordinal, multinomial), la régression binomiale et la régression de Poisson en tant que modèles linéaires généralisés. Les paramètres du modèle seront estimés et testés, puis le meilleur modèle sera sélectionné sans oublier l'étude de l'adéquation et de la validation du modèle. Les modèles logistiques faisant intervenir des effets aléatoires et les modèles logistiques nécessitant la pondération des données seront aussi abordés. Enfin, l'analyse de la variance à un critère, à deux critères et à trois critères de classification, avec ou sans interaction, hiérarchique et non hiérarchique, avec ou sans effet aléatoire sera abordée. Pour ces modèles, il s'agira de la spécification des modèles, de la construction de la table de l'analyse de la variance, de la démonstration de la formule fondamentale et l'application à l'analyse des données tirées principalement du domaine de l'éducation.
Méthodologie et supports pédagogiques	Méthodologie	Active et participative
	Supports	Syllabus de cours Logiciel R
Modes d'évaluation	Evaluation formative	Projet de cours : 40 %
	Evaluation sommative	Examen écrit : 60 %

Table des matières

Liste des tableaux

Liste des figures

Introduction

En statistique, un modèle de régression linéaire est un modèle de régression d'une variable expliquée (ou variable dépendante) sur une ou plusieurs variables explicatives en faisant l'hypothèse que la fonction qui lie les variables explicatives à la variable expliquée est linéaire dans ses paramètres. Un modèle de régression linéaire est aussi appelé tout simplement modèle linéaire. Son objectif est d'étudier l'association entre la variable réponse et chacune des variables explicatives d'une part et d'expliquer cette variable réponse par une combinaison des variables explicatives afin de faire des prévisions d'autre part. Dans le cas où la fonction qui lie la variable dépendante et les variables explicatives est linéaire, un modèle linéaire classique peut être utilisé.

Le modèle linéaire suppose certains postulats entre autres la normalité des résidus du modèle, la variance constante pour les résidus en fonction de la variable explicative et l'indépendance des observations. La variable réponse suit une loi normale $N(\mu,\sigma)$ où μ est une fonction linéaire des variables explicatives. Lorsque la variable réponse est quantitative et les variables explicatives qualitatives, l'analyse de la variance permet de comparer les moyennes (plus de deux) dans les catégories des variables explicatives. Lorsque le lien linéaire entre deux variables quantitatives X et Y peut être comparé dans les niveaux d'une variable qualitative. Dans ce cas, il s'agit de l'analyse de la covariance. Le modèle linéaire multiple, le modèle de l'analyse de la variance et le modèle de l'analyse de la covariance sont des modèles linéaires de type général. Ces modèles ont comme fonction de lien définie entre la variable réponse et les variables explicatives la fonction identité. Dans le cas d'une relation fonctionnelle différente de l'identité, le modèle linéaire devient un modèle linéaire généralisé. C'est notamment le cas du modèle de Poisson et du modèle logistique. Pour le modèle de Poisson, la variable réponse est considérée comme une composante aléatoire, la combinaison des variables explicatives comme une composante déterministe et une fonction de lien, le logarithme, est utilisée entre l'espérance de la variable réponse et les variables explicatives. Ce modèle fait partie des modèles log-linéaires.

Dans le cas où la variable réponse est de type qualitatif, le modèle utilisée est souvent la régression logistique. Dans ce cas, la variable réponse peut avoir deux ou plusieurs modalités. Dans le cas où elle prend deux modalités, la variable réponse suit une loi de Bernoulli (régression logistique binaire). Dans le cas où elle prend plus deux modalités, il s'agira de la régression logistique multinomiale ou polytomique (nominale ou ordinale). Dans la plupart des cas, les variables explicatives sont des effets fixes, ou un mélange d'effets fixes et d'effets aléatoires. Pour ce dernier cas de figure, il s'agit d'un modèle linéaire mixte.

Chapitre 1 : Analyse de la variance

L'analyse de la variance (*ANalysis Of VAriance, ANOVA en anglais*) est une méthode permettant de comparer les moyennes de plusieurs échantillons (au moins 3 même si les résultats sont les mêmes deux moyennes). La variable dépendante (y) doit être quantitative continue et la variable indépendante ou le facteur (x) doit être de type qualitatif avec au moins 3 modalités. D'une manière générale, l'analyse de la variance a pour premier objectif de comparer les ensembles de plus de deux moyennes en identifiant les sources de variation qui peuvent expliquer les différences existant entre elles. À ce titre, l'analyse de la variance est un des principaux actifs de l'inférence statistique. Il s'agit de la généralisation du test de Student qui permet de comparer deux moyennes de deux échantillons indépendants ou appariés. Lorsqu'il y a un seul facteur (ou critère), alors il s'agit de l'analyse de la variance à un critère (*ANOVA 1* pour one-way analysis of variance).

Lorsqu'il y a un seul, deux ou trois facteurs fixes, alors il s'agit de l'analyse de la variance à un, à deux et à trois critères respectivement, ce qui se note ANOVA 1, ANOVA 2, ANOVA 3). Il est possible d'utiliser en ANOVA à plusieurs facteurs un modèle hiérarchique, un modèle non hiérarchique, un modèle pour des données répétées, un modèle avec un plan équilibré ou déséquilibré, un modèle à effets fixes, un modèle à effets aléatoire ou un modèle mixte (effets fixes et effets aléatoires). Les conditions d'utilisation de l'ANOVA sont l'indépendance des observations de y (ou l'indépendance des résidus), la normalité des résidus (ou de y) dans les niveaux des facteurs et l'homogénéité de la variance des résidus (ou de y), c'est-à-dire la variance constante de y (ou des résidus). Dans le cas où une de ces conditions n'est pas vérifiée, alors l'analyse de la variance non paramétrique est utilisée.

1.1. Analyse de la variance à un critère

1.1.1. Spécification du modèle

Le modèle de l'analyse de la variance à un critère (ANOVA 1) s'écrit [1,2] :

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij} \tag{1.1}$$

avec y_{ij} la j^{ème} observation de la variable dépendante y pour le i^{ème} niveau du facteur, μ la moyenne globale, α_i l'effet du facteur et ε_{ij} un terme d'erreur qui doit suivre une loi normale de moyenne nulle et de variance constante (c'est l'homoskédasticité).

Les erreurs doivent suivre une loi normale de moyenne nulle et de variance constante :

$$\varepsilon_{ij} \sim N(0,\sigma) \tag{1.2}$$

Les erreurs doivent être indépendantes avec comme contrainte somme :

$$\sum_{i=1}^{I} \alpha_i = 0 \tag{1.3}$$

1.1.2. Hypothèses de test

Les hypothèses de test sont :

$H_0 : \mu_1 = \mu_2 = ... = \mu_I$ (Les moyennes sont égales)

$H_0 : \exists i \neq j : \mu_i \neq \mu_j$ (Il existe au moins deux moyennes différentes)

Ces hypothèses peuvent aussi s'écrire : $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_I = 0$ contre $H_0 : \exists i : \alpha_i \neq 0$. Il s'agira d'utiliser le test de Fisher.

Les hypothèses de test peuvent aussi s'écrire :

$H_0 : \alpha_i = 0$ (Le facteur n'a pas d'effet sur la variable dépendante)

$H_1 : \alpha_i \neq 0$ (Le facteur a un effet sur la variable dépendante)

1.1.3. Statistiques descriptives

Les moyennes de la variable dépendante dans les niveaux du facteur sont données par :

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \tag{1.4}$$

La moyenne globale vaut :

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{i=1}^{I} n_i \overline{y}_i \tag{1.5}$$

Il convient aussi de calculer la variance et l'écart-type de la variable dépendante globalement et dans chaque niveau du facteur.

1.1.4. Formule fondamentale

La variabilité totale se décompose en une somme de la variabilité liée au facteur et de la variabilité due aux résidus comme suit :

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}\right)^2=\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_i\right)^2+\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\overline{y}_i-\overline{y}\right)^2 \tag{1.6}$$

En effet, la somme des carrés des écarts totaux $\left(SCE_T\right)$ qui montre la variabilité totale vaut :

$$\begin{aligned}
SCE_T &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij}^2-2\overline{y}y_{ij}+\overline{y}^2\right) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}y_{ij}^2-2\overline{y}\sum_{i=1}^{I}\sum_{j=1}^{n_i}y_{ij}+\sum_{i=1}^{I}\sum_{j=1}^{n_i}\overline{y}^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}y_{ij}^2-2\overline{y}\sum_{i=1}^{I}\sum_{j=1}^{n_i}y_{ij}+\sum_{i=1}^{I}\sum_{j=1}^{n_i}\overline{y}^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}y_{ij}^2-2n\overline{y}^2+n\overline{y}^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_u}y_{ij}^2-n\overline{y}^2
\end{aligned} \tag{1.7}$$

La somme des carrés des écarts dus au facteur (ou des effets inter ou between) vaut :

$$\begin{aligned}
SCE_F &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\overline{y}_i-\overline{y}\right)^2 \\
&= \sum_{i=1}^{I}n_i\left(\overline{y}_i-\overline{y}\right)^2 \\
&= \sum_{i=1}^{I}n_i\left(\overline{y}^2-2\overline{y}\,\overline{y}_i+\overline{y}^2\right) \\
&= \overline{y}^2\sum_{i=1}^{I}n_i-2\overline{y}\sum_{i=1}^{I}n_i\overline{y}_i+\sum_{i=1}^{I}n_i\overline{y}^2 \\
&= n\overline{y}^2-2n\overline{y}^2+\sum_{i=1}^{I}n_i\overline{y}^2 \\
&= \sum_{i=1}^{I}n_i\overline{y}^2-n\overline{y}^2
\end{aligned} \tag{1.8}$$

La somme des carrés des écarts factoriels vaut aussi :

$$SCE_F = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\overline{y}_i - \overline{y}\right)^2 = \sum_{i=1}^{I} n_i\left(\overline{y}_i - \overline{y}\right)^2 \tag{1.9}$$

La somme des carrés des écarts résiduels vaut :

$$SCE_R = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{\boldsymbol{i}}\right)^2 \tag{1.10}$$

La somme des carrés des écarts résiduels (ou intra ou within) vaut [3] :

$$\begin{aligned}
SCE_R &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij}^2 - 2\overline{y}_i y_{ij} + \overline{y}_i^2\right) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}\overline{y}_i y_{ij} + \sum_{i=1}^{I}\sum_{j=1}^{n_i}\overline{y}_i^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}\overline{y}_i y_{ij} + \sum_{i=1}^{I} n_i\overline{y}_i^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - 2\sum_{i=1}^{I}\overline{y}_i\left(\sum_{j=1}^{n_i} y_{ij}\right) + \sum_{i=1}^{I} n_i\overline{y}_i^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - 2\sum_{i=1}^{I} n_i\overline{y}_i^2 + \sum_{i=1}^{I} n_i\overline{y}_i^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{I} n_i\overline{y}_i^2
\end{aligned} \tag{1.11}$$

Il vient :

$$SCE_F + SCE_R = \sum_{i=1}^{I} n_i\overline{y}_i^2 - n\overline{y}^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{I} n_i\overline{y}_i^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - n\overline{y}^2 = SCE_T$$

La formule fondamentale de l'analyse de la variance est donc :

$$SCE_T = SCE_F + SCE_R \tag{1.12}$$

1.1.5. Test d'hypothèse

Les carrés moyens sont respectivement donnés par :

$$\begin{aligned} CMT &= \frac{SCE_T}{n-1} \\ CMF &= \frac{SCE_F}{I-1} \\ CMR &= \frac{SCE_R}{n-I} \end{aligned} \tag{1.13}$$

La statistique de Fisher observée vaut [2] :

$$F^{obs} = \frac{CMF}{CMR} \grave{a} \left(I-1, n-I\right) ddl \tag{1.14}$$

Le seuil de décision est :

$$\alpha = 5\% = 0,05 \tag{1.15}$$

La p-value (p-valeur), qui est la probabilité d'observer une statistique au moins aussi grande que celle qui aurait été observée sous l'hypothèse nulle, vaut :

$$p - value = P\left(F^{obs} \geq F^{tab}\right) \tag{1.16}$$

L'hypothèse nulle sera rejetée si la p-value est inférieure ou égale au seuil de décision.

Le tableau 1 est un tableau de l'analyse de la variance à un critère.

Tableau 1 : Tableau de l'analyse de la variance à un critère

Source de variation	**Somme des carrés**	**Degré de liberté**	**Carré moyen**	**F^{obs}**	**P-value**
Factorielle	SCE_F	$I-1$	$CMF = \frac{SCE_F}{I-1}$	$\frac{CMF}{CMR}$	$P\left(F^{obs} \geq F^{tab}\right)$
Résiduelle	SCE_R	$n-I$	$CMR = \frac{SCE_R}{n-I}$	//////////////	//////////////
Totale	SCE_T	$n-1$	$CMT = \frac{SCE_T}{n-1}$	//////////////	//////////////

Source : Pr BARANKANIRA Emmanuel

La statistique de Fisher observée est à comparer avec la statistique tabulée $F(I-1,n-I)$. Pour décider du rejet ou non de l'hypothèse nulle, il suffit de comparer la valeur observée de la statistique de Fisher à la statistique tabulée ou comparer la p-value (probabilité que la statistique observée soit au moins aussi grande que la statistique tabulée) au risque de première espèce (souvent de 5%). Si cette p-value est inférieure au seuil de signification choisi, H_0 est rejetée (test significatif), ce qui signifie que le facteur a bel et bien une influence sur la variable dépendante (ou que les moyennes sont significativement différente). Dans ce cas, il est nécessaire d'évaluer la cause de cette différence significative avec la comparaison multiple des moyennes en ANOVA 1 en se basant sur la méthode de Tukey ou d'autres méthodes telles que celles de Bonferonni, Dunet, Duncan, Scheffé, Gabriel, Sidak. Pour ces méthodes, les moyennes sont comparées deux à deux en se donnant la différence des moyennes avec le risque d'erreur de 5% (par défaut). En plus, elles donnent une probabilité ajustée sur chaque différence de moyennes comparées. Dans le cas où la p-value est supérieure ou égale au seuil de signification, H_0 ne sera pas rejetée (test non significatif), ce qui veut dire que le facteur n'a pas d'effet sur la variable dépendante (ou que les moyennes ne sont pas significativement différente).

Les conditions d'applications de l'ANOVA sont :

- ✓ Les observations (ou les résidus) doivent être indépendantes (s) (à vérifier graphiquement et par le test de dépendance des observations de Box-Ljung par exemple)
- ✓ Les observations (ou les résidus) doivent suivre une loi normale dans les niveaux du facteur (à vérifier graphiquement et par le test de normalité des erreurs de Shapiro-Wilk)
- ✓ La variance des résidus doit être constante. C'est l'homoskédasticité (à vérifier graphiquement et par le test de Bartlett).

Si ces conditions ne sont pas vérifiées, une analyse de la variance non paramétrique à un facteur (ANOVA non paramétrique de Kruskal-Wallis) est utilisée.

1.1.6. Exemple d'application pour l'ANOVA 1

Considérons les données fictives suivantes où la variable dépendante (y) est le dosage de l'aluminium par spectrométrie d'absorption selon la méthode de dosage (facteur). Y a-t-il une différence globale entre les 3 méthodes de dosage ?

```
Méthode 1 : 12.4 13.7 11.1 10.9 13.1 11.7
Méthode 3 : 14.2 13.9 11.7 14.8 12.5 13.8
Méthode 3 : 15.5 16.0 15.3 15.6 14.9 12.7
```

- Calculez les statistiques descriptives (effectif, minimum, moyenne, écart-type, maximum) de la variable dépendante globalement et dans les niveaux du facteur
- Construisez des boîtes à moustaches par catégories.
- Comment « tester » s'il existe une **différence globale sous les logiciels Excel et R** ?

```
> ## Saisie des données
> y1 <- c(12.4, 13.7, 11.1, 10.9, 13.1, 11.7)
> y2 <- c(14.2, 13.9, 11.7, 14.8, 12.5, 13.8)
> y3 <- c(15.5, 16.0, 15.3, 15.6, 14.9, 12.7)
> y  <- c(y1, y2, y3)
> m <- rep(c("M1","M2","M3"), each=6)
> mydata <- data.frame(y=y, Methode=m)
> print(mydata)
```

```
> str(mydata)
'data.frame':   18 obs. of  2 variables:
 $ y      : num  12.4 13.7 11.1 10.9 13.1 11.7 14.2 13.9 11.7 14.8 ...
 $ Methode: Factor w/ 3 levels "M1","M2","M3": 1 1 1 1 1 1 2 2 2 2 ...
```

La base de données contient 18 observations et 2 variables. Visualisons les données par des boîtes à moustaches afin de détecter d'éventuelles observations aberrantes et d'évaluer la symétrie de la distribution (**Figure 1**). La ligne rouge sur les boîtes à moustaches représente la moyenne globale. Les moyennes semblent différentes.

```
par(mar=c(4.5, 4.5, .5, .5))
boxplot(y ~ Methode, xlab="Méthode", col="white")
abline(h=13.54444, col="red")
```

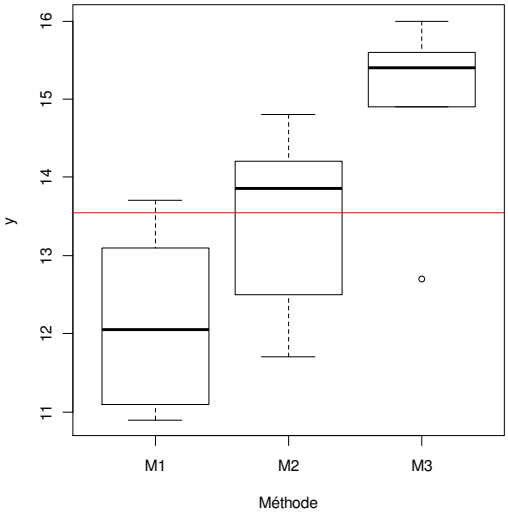


Figure 1 : Boîte à moustaches de la variable y pour chaque niveau du facteur
Source : Pr BARANKANIRA Emmanuel

Les observations de y doivent être indépendantes, de variance constante et suivre une loi normale dans les niveaux du facteur. L'hypothèse d'indépendance est vérifiée à l'aide d'un graphique d'autocorrélation simple ou partielle (**Figure 2**). Toutes les barres sont à l'intérieur de la région de confiance, ce qui montre que l'hypothèse semble être vérifiée.

```
> par(mfrow=c(1, 2))
> par(mar=c(4.5, 4.5, .5, .5))
> acf(y, main="", ylab="Autocorrélation simple")
> pacf(y, main="", ylab="Autocorrélation partielle")
```

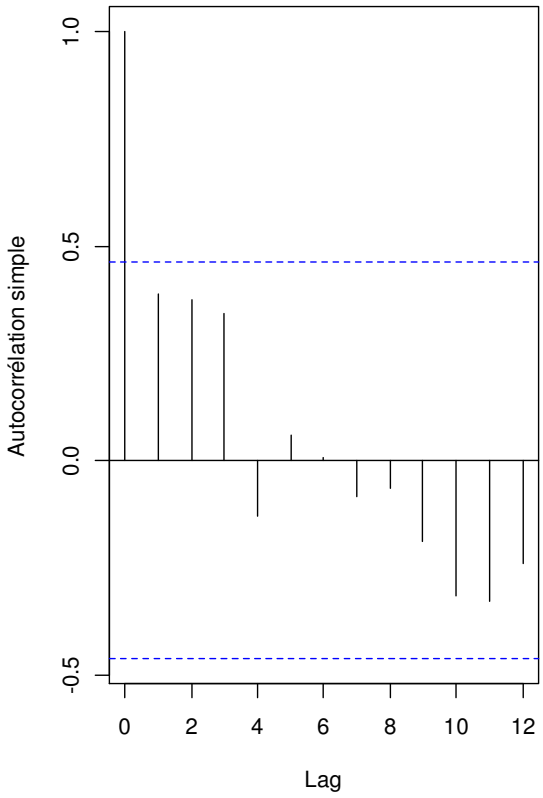


Figure 2 : Graphique d'autocorrélation simple
Source : Pr BARANKANIRA Emmanuel

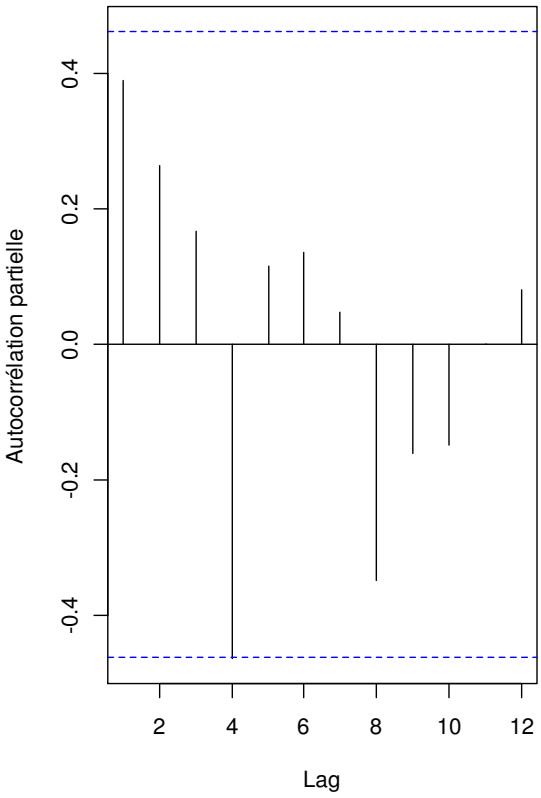


Figure 3 : Graphique d'autocorrélation partielle

Le test de Box et Pierce ne rejette pas l'hypothèse nulle d'indépendance des observations (χ^2=2,72 ; ddl=1 ; p-value=0,099). Les valeurs de y sont donc indépendantes.

Les moyennes calculées manuellement sont :

$$\overline{y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1j} = \frac{72,9}{6} = 12,15$$

$$\overline{y}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} y_{2j} = \frac{80,9}{6} = 13,48333 \simeq 13,48$$

$$\overline{y}_3 = \frac{1}{n_3}\sum_{j=1}^{n_3} y_{3j} = \frac{90}{6} = 15$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{I} n_i \overline{y}_i = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} = \frac{243,8}{18} = 13,54444 \simeq 13,54$$

Le tableau ci-après montre les statistiques descriptives (effectif ou le nombre d'observations, minimum, moyenne, écart-type ou déviation standard, médiane, maximum) de y. Il y a le même nombre d'observations dans chaque niveau du facteur et la moyenne est supérieure à la déviation standard (écart-type échantillonnal). Il est possible de créer une fonction qui permet de calculer ces statistiques descriptives globalement et pour chaque méthode de dosage et de stocker les résultats dans un objet Tableau sous forme d'une matrice.

```
> Tableau
   Effectif Minimum Moyenne Écart-type Médiane Maximum
M1        6    10.9   12.15       1.12   12.05    13.7
M2        6    11.7   13.48       1.15   13.85    14.8
M3        6    12.7   15.00       1.18   15.40    16.0
```

Le calcul statistique sur ordinateur des moyennes de la variable y globalement et dans chaque niveau du facteur se fait à l'aide des commandes R suivantes :

```
> tapply(y, Methode, mean)
      M1       M2       M3
12.15000 13.48333 15.00000
> mean(y)
[1] 13.54444
```

La variance d'une variable y est donnée par :

$$
\begin{aligned}
S_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i^2-2\bar{y}\,y_i+\bar{y}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}y_i^2-2\bar{y}\sum_{i=1}^{n}y_i+\bar{y}^2\sum_{i=1}^{n}1\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}y_i^2-2n\bar{y}^2+n\bar{y}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right)
\end{aligned}
$$

La déviation standard (écart-type) de y est :

$$
S_y=\sqrt{S_y^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}=\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right)}
$$

Le calcul statistique sur ordinateur des variances et des déviations standards de la variable y globalement et dans chaque niveau du facteur se fait à l'aide des commandes R suivantes :

```
> tapply(y, Methode, var)
      M1       M2       M3
1.247000 1.333667 1.400000
> tapply(y, Methode, sd)
      M1       M2       M3
1.116692 1.154845 1.183216
> var(y); sd(y)
[1] 2.606144
[1] 1.614356
```

Sachant que les carrés des observations de y pour chaque niveau du facteur sont :

```
$M1
[1] 153.76 187.69 123.21 118.81 171.61 136.89

$M2
[1] 201.64 193.21 136.89 219.04 156.25 190.44

$M3
[1] 240.25 256.00 234.09 243.36 222.01 161.29
```

et que leurs sommes sont :

```
> sum(tapply(y, Methode, carre)$M1)
[1] 891.97
> sum(tapply(y, Methode, carre)$M2)
[1] 1097.47
> sum(tapply(y, Methode, carre)$M3)
[1] 1357
```

alors, les variances de chaque groupe et la variance globale valent respectivement

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}\left(y_{1j} - n_1\overline{y}_1\right)^2 = \frac{1}{n_1 - 1}\left(\sum_{j=1}^{n_1} y_{1j}^2 - n_1\overline{y}_1^2\right) = \frac{891,97 - 6(12,15)^2}{5} = 1,247 \simeq 1,25$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(y_{2j} - n_2\overline{y}_2\right)^2 = \frac{1}{n_2 - 1}\left(\sum_{j=1}^{n_2} y_{2j}^2 - n_2\overline{y}_2^2\right) = \frac{1097,47 - 6(13,48)^2}{5} = 1,333 \simeq 1,33$$

$$S_3^2 = \frac{1}{n_3 - 1}\sum_{j=1}^{n_3}\left(y_{3j} - n_3\overline{y}_3\right)^2 = \frac{1}{n_3 - 1}\left(\sum_{j=1}^{n_3} y_{3j}^2 - n_3\overline{y}_3^2\right) = \frac{1357 - 6(15)^2}{5} = 1,40$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij} - n\,\overline{y}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - n\,\overline{y}^2\right) = \frac{3346,44 - 18(13,54)^2}{17} = 2,61$$

Les écart-types sont :

$$S_1 = \sqrt{S_1^2} = \sqrt{1,247} = 1,12$$
$$S_2 = \sqrt{S_2^2} = \sqrt{1,333667} = 1,15$$

$$S_3 = \sqrt{S_3^2} = \sqrt{1,40} = 1,18$$
$$S = \sqrt{S^2} = \sqrt{2,61} = 1,62$$

Selon la formule fondamentale de l'analyse de la variance, la somme des carrés des écarts totaux doit être égale à la somme de la somme des carrés des écarts factoriels et de la somme des carrés des écarts résiduels.

```
> round(c(SCEF, SCET, SCER), 2)
[1] 24.4 44.3 19.9
```

La somme des carrés des écarts totaux vaut :

$$\begin{aligned} SCE_T &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij}^2 - 2\bar{y}\, y_{ij} + \bar{y}^2\right) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - 2\bar{y}\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} + \bar{y}^2\sum_{i=1}^{I}\sum_{j=1}^{n_i} 1 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - 2n\bar{y}^2 + n\bar{y}^2 \end{aligned}$$

$$\begin{aligned} SCE_T &= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}^2 - n\bar{y}^2 \\ &= 3346,44 - 18\left(\frac{243,8}{18}\right)^2 \\ &= 3346,44 - \frac{(243,8)^2}{18} \\ &= 3346,44 - 3302,1355556 \\ &= 44,30444 \\ &\simeq 44,30 \end{aligned}$$

La somme des carrés des écarts dus au facteur vaut :

$$
\begin{aligned}
SCE_F &= \sum_{i=1}^{I} n_i \left(\overline{y}_i - \overline{y}\right)^2 = \sum_{i=1}^{I} n_i \overline{y}_i^2 - n\,\overline{y}^2 \\
&= 6\left[\left(12{,}15 - \frac{243{,}8}{18}\right)^2 + \left(\frac{80{,}9}{18} - \frac{243{,}8}{18}\right)^2 + \left(15 - \frac{243{,}8}{18}\right)^2\right] \\
&= 6\left[\left(\frac{-25{,}1}{18}\right)^2 + \left(\frac{-1{,}1}{18}\right)^2 + \left(\frac{26{,}2}{18}\right)^2\right] \\
&= \frac{1}{54}\left(630{,}01 + 1{,}21 + 686{,}33\right) \\
&= \frac{1317{,}66}{54} \\
&= 24{,}4011 \\
&\simeq 24{,}40
\end{aligned}
$$

La somme des carrés des écarts dus aux résidus vaut :

$$
\begin{aligned}
SCE_R &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2 \\
&= 0{,}0625 + 2{,}4025 + 1{,}1025 + 1{,}5625 + 0{,}9025 + 0{,}2025 \\
&\quad + \left(14{,}2 - \frac{80{,}9}{6}\right)^2 + \left(13{,}9 - \frac{80{,}9}{6}\right)^2 + \left(11{,}7 - \frac{80{,}9}{6}\right)^2 + \left(14{,}8 - \frac{80{,}9}{6}\right)^2 + \left(12{,}5 - \frac{80{,}9}{6}\right)^2 + \left(13{,}8 - \frac{80{,}9}{6}\right)^2 \\
&\quad + 0{,}25 + 1 + 0{,}09 + 0{,}36 + 0{,}01 + 5{,}29 \\
&= 19{,}90
\end{aligned}
$$

La formule fondamentale est vérifiée :

$$SCE_F + SCE_R = 24{,}40 + 19{,}90 = 44{,}30 = SCE_T$$

Les carrés moyens sont :

$$
\begin{aligned}
CME_T &= \frac{SCE_T}{n-1} = \frac{44{,}30}{17} = 2{,}61 \\
CME_F &= \frac{SCE_F}{I-1} = \frac{24{,}40}{2} = 12{,}20 \\
CME_R &= \frac{SCE_R}{n-I} = \frac{19{,}90}{15} = 1{,}33
\end{aligned}
$$

Les hypothèses de test sont :

$H_0 : \mu_1 = \mu_2 = \mu_3$ (Les moyennes sont égales)

$H_1 : \exists i \neq j : \mu_i \neq \mu_j$ (Il existe au moins deux moyennes différentes)

La statistique de test est :

$$F^{obs} = \frac{CME_F}{CME_R} = \frac{12,20}{1,33} = 9,172932 \simeq 9,17 \, à \, (2;15) \, ddl$$

Le tableau de l'analyse de la variance est :

Source de variation	*SCE*	ddl	*CME*	F^{obs}	F^{tab}	P-value
Méthode	24,40	2	12,20	9,17	3,68	<0,05
Résiduelle	18,90	15	1,33	/////////	/////////	/////////
Totale	44,30	17	2,61	/////////	/////////	/////////

Source : Pr BARANKANIRA Emmanuel

Ce résultat calculé à la main ressemble fortement à celui trouvé de manière informatique :

```
> summary(res)
            Df Sum Sq Mean Sq F value  Pr(>F)
Methode      2   24.4  12.201   9.195 0.00248
Residuals   15   19.9   1.327
---
```

Le rejet global de l'hypothèse nulle a été dû au fait qu'il y a une différence significative entre les moyennes obtenues avec les méthodes M1 et M2 comme cela peut s'observer sur les boîtes à moustaches et le graphique des intervalles de confiance à 95 % de la différence des moyennes obtenus lors des comparaisons multiples basées sur la méthode de Tukey (**Figure 4**).

```
$Methode
          diff        lwr      upr     p adj
M2-M1 1.333333 -0.3941236 3.060790 0.1452297
M3-M1 2.850000  1.1225431 4.577457 0.0017706
M3-M2 1.516667 -0.2107903 3.244124 0.0897668
```

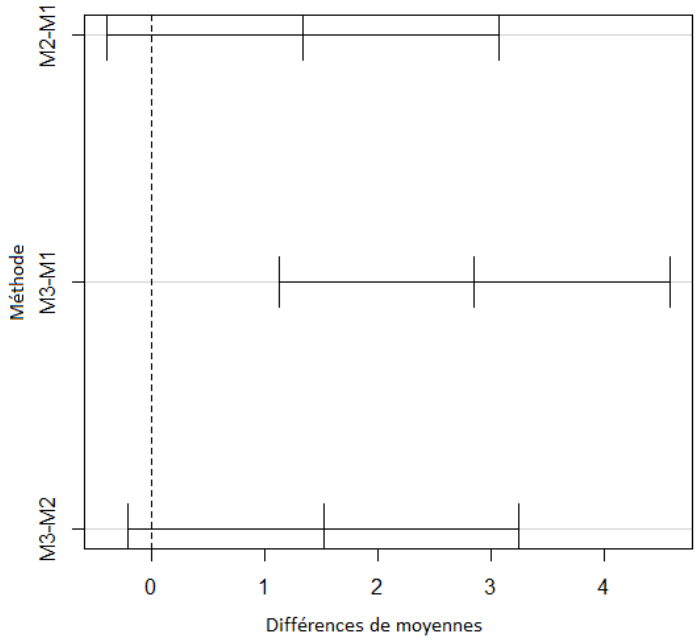
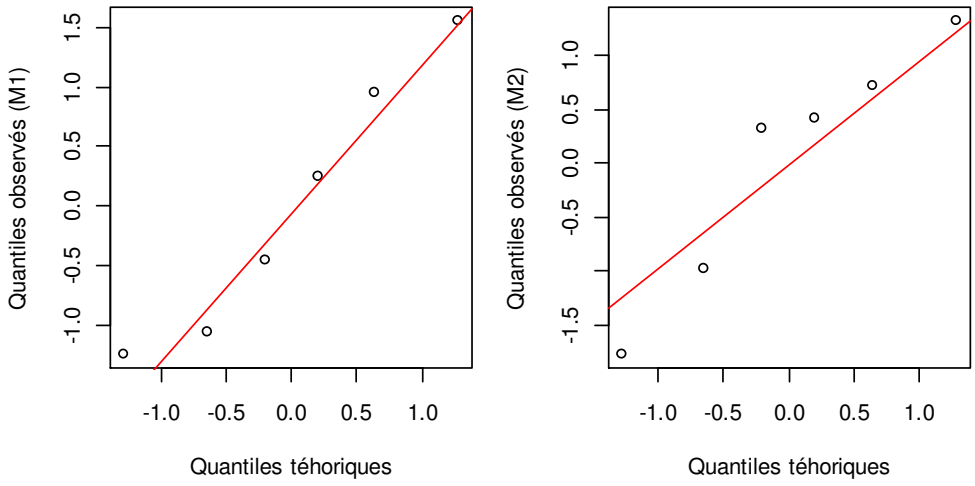
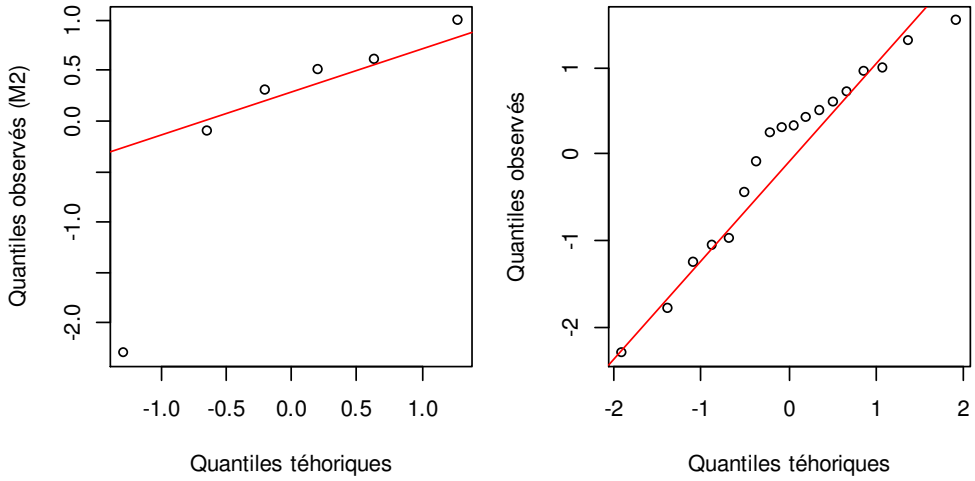


Figure 4 : Comparaisons multiples
Source : Pr BARANKANIRA Emmanuel

Les diagrammes quantile-quantile des résidus (pour la normalité) montrent que tous les points semblent alignés autour de la droite de Henri globalement et dans tous les niveaux du facteur.

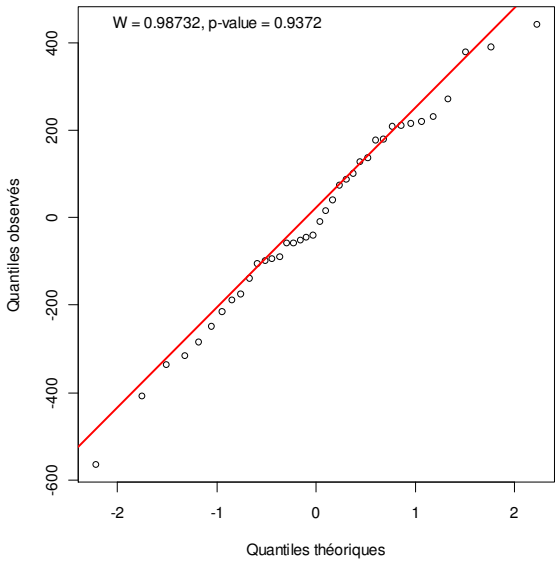


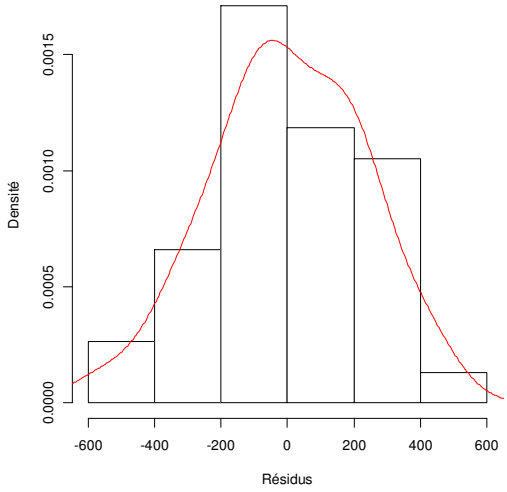


Source : Pr BARANKANIRA Emmanuel

```
> shapiro.test(residus[Methode=="M1"])

        Shapiro-Wilk normality test

data:  residus[Methode == "M1"]
W = 0.94201, p-value = 0.6755

> shapiro.test(residus[Methode=="M2"])

        Shapiro-Wilk normality test

data:  residus[Methode == "M2"]
W = 0.92177, p-value = 0.5182

> shapiro.test(residus[Methode=="M3"])

        Shapiro-Wilk normality test

data:  residus[Methode == "M3"]
W = 0.77953, p-value = 0.03816

> shapiro.test(residus)

        Shapiro-Wilk normality test

data:  residus
W = 0.93915, p-value = 0.2802
```

Le test de normalité de Shapiro et Wilk normality ne rejette pas l'hypothèse nulle de normalité des résidus pour le méthode M1 (W = 0.94201, p-value = 0,6755), pour la méthode M2 (W = 0.92177, p-value = 0,5182), mais cette hypothèse est rejetée pour la méthode M3 (W = 0.77953, p-value = 0,03816). Les résidus ne suivent donc pas une loi normale.



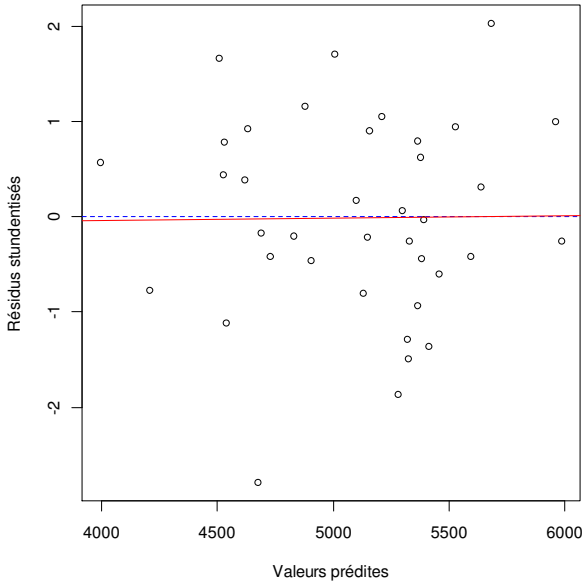


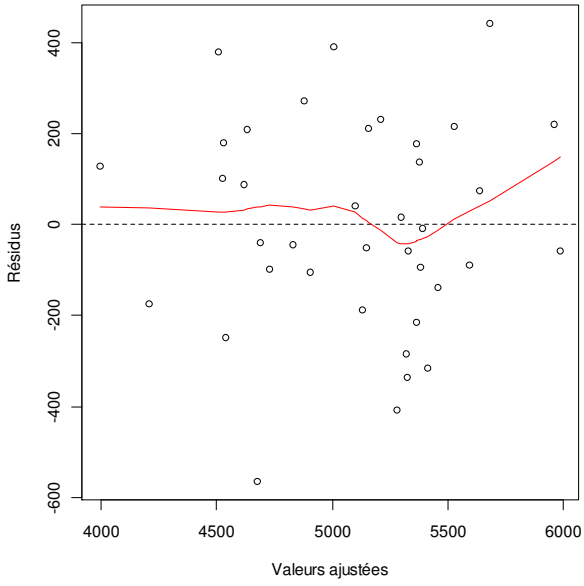
Source : Pr BARANKANIRA Emmanuel

Le test de Bertlett ne rejette pas l'hypothèse nulle qui dit que la variance est constante.

```
> bartlett.test(y ~ Methode)

        Bartlett test of homogeneity of variances

data:  y by Methode
Bartlett's K-squared = 0.015457, df = 2, p-value = 0.9923
```

Le test de Bartlett ne rejette pas l'hypothèse nulle qui qui que la variance des résidus est constant (K^2 = 0,015457 ; ddl = 2 ; p-value = 0,9923). Les résidus sont donc homoskédastiques.





Source : Pr BARANKANIRA Emmanuel

Le test de Kruskal et Wallis (analyse de la variance non paramétrique) basé sur la somme des rangs rejette l'hypothèse nulle qui dit que les rangs médians sont égaux.

Décision : On rejette H_0 (test significatif)

Conclusion : Au seuil de 5 %, les moyennes (médianes) sont significativement différentes. Il y a donc un effet de la méthode sur le résultat du dosage. Les méthodes de dosage aboutissent à des résultats significativement différents.

```
> kruskal.test(y ~ Methode)

        Kruskal-Wallis rank sum test

data:  y by Methode
Kruskal-Wallis chi-squared = 10.037, df = 2, p-value = 0.006616
```

1.2. Analyse de la variance à deux critères

1.2.1. Spécification du modèle

Le modèle de l'ANOVA à deux facteurs fixes de classification avec interaction pour le plan équilibré se matérialise par la formule suivante [1] :

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{1.17}$$

y_{ijk} étant la $k^{ème}$ de la variable dépendante pour le $i^{ème}$ niveau du premier facteur et $j^{ème}$ le niveau du deuxième facteur, μ la moyenne globale, α_i l'effet du premier facteur, β_j l'effet du deuxième facteur, $(\alpha\beta)_{ij}$ l'effet d'interaction et ε_{ijk} les erreurs du modèle avec $\varepsilon_{ijk} \sim N(0, \sigma^2)$ indépendants et comme contraintes :

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0 \tag{1.18}$$

Le nombre de modalités du premier facteur (ou le nombre de moyennes à comparer) vaut *I* et *J* pour le deuxième facteur. Le nombre d'observations dans chaque case vaut *K* et, de ce fait, le nombre d'observations vaut $n=IJK$. Ici, le plan est équilibré.

Il est à préciser que les conditions d'utilisation de l'ANOVA restent la normalité des résidus, l'homogénéité de la variance des résidus et l'indépendance des résidus.

Dans ce cas, il vient :

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \tag{1.19}$$

Si ces conditions ne sont pas vérifiées, une analyse de la variance non paramétrique à deux facteurs est utilisée.

1.2.2. Hypothèses de test

Les hypothèses de test pour le terme d'interaction sont :

$H_0 : (\alpha\beta)_{ij} = 0$ (Absence d'interaction)

$H_1 : (\alpha\beta)_{ij} \neq 0$ (Présence de l'interaction)

Les hypothèses de test pour le premier facteur sont :

$H_0 : \alpha_i = 0$ (Le premier facteur n'a pas d'effet sur la réponse)

$H_1 : \alpha_i \neq 0$ (Le premier facteur a un effet sur la réponse)

Les hypothèses de test pour le deuxième facteur sont :

$H_0 : \beta_j = 0$ (Le deuxième facteur n'a pas d'effet sur la réponse)

$H_1 : \beta_j \neq 0$ (Le deuxième facteur a un effet sur la réponse)

1.2.3. Formule fondamentale

Les moyennes se calculent comme suit :

$$\overline{y}_{\bullet\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk} \tag{1.20}$$

$$\overline{y}_{i\bullet\bullet} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk} \tag{1.21}$$

$$\overline{y}_{\bullet j \bullet} = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} y_{ijk} \tag{1.22}$$

La somme des carrés dus au premier facteur, la somme des carrés dus au deuxième facteur, la somme des carrés dus à l'interaction, la somme des carrés résiduels et la somme des carrés totaux sont données respectivement par :

$$SCE_A = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{i\bullet\bullet} - \overline{y}\right)^2 \tag{1.23}$$

$$SCE_B = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{\bullet j \bullet} - \overline{y}\right)^2 \tag{1.24}$$

$$SCE_{AB} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{ij\bullet} - y_{i\bullet\bullet} - y_{\bullet j \bullet} + \overline{y}\right)^2 \tag{1.25}$$

$$SCE_R = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{ijk} - y_{ij\bullet}\right)^2 \tag{1.26}$$

$$SCE_T = SCE_A + SCE_B + SCE_{AB} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{ijk} - \overline{y}\right)^2 \tag{1.27}$$

La formule fondamentale s'écrit alors :

$$SCE_T = SCE_A + SCE_B + SCE_{AB} + SCE_R \tag{1.28}$$

1.2.4. Test d'hypothèses

Le test de comparaison des moyennes est le test de Fisher (ou de Fisher-Snedecor) dont les statistiques de test sont, sous l'hypothèse nulle :

$$F_A = \frac{CM_A}{CM_R} = \frac{\dfrac{SCE_A}{I-1}}{\dfrac{SCE_R}{IJ(K-1)}} \sim F\left[I-1, IJ(K-1)\right] \tag{1.29}$$

$$F_B = \frac{CM_B}{CM_R} = \frac{\dfrac{SCE_B}{J-1}}{\dfrac{SCE_R}{IJ(K-1)}} \sim F\left[J-1, IJ(K-1)\right] \tag{1.30}$$

$$F_{AB} = \frac{CM_{AB}}{CM_R} = \frac{\frac{SCE_{AB}}{(I-1)(J-1)}}{\frac{SCE_R}{IJ(K-1)}} \sim F\left[(I-1)(J-1), IJ(K-1)\right] \quad (1.31)$$

où CM représente le carré moyen.

Il est d'usage de résumer ces informations dans un tableau, appelé tableau de l'analyse de la variance donné par le tableau 2.

Tableau 2 : Tableau de l'analyse de la variance à deux facteurs

Source de variation	Somme des carrés	ddl	Carrés moyens	F^{obs}	p-value
Facteur A	SCE_A	$I-1$	$CM_A = \frac{SCE_A}{I-1}$	$\frac{CM_A}{CM_R}$	?
Facteur B	SCE_B	$J-1$	$CM_B = \frac{SCE_B}{J-1}$	$\frac{CM_B}{CM_R}$	?
Interaction	SCE_{AB}	$(I-1)(J-1)$	$CM_{AB} = \frac{SCE_{AB}}{(I-1)(J-1)}$	$\frac{CM_{AB}}{CM_R}$	?
Erreur	SCE_R	$IJ(K-1)$	$CM_R = \frac{SCE_R}{IJ(K-1)}$	//////////	//////////
Totale	SCT	$n-1$	//////////	//////////	//////////

Pour décider du rejet ou non de l'hypothèse nulle, il suffit de comparer la valeur de la statistique de Fisher à la statistique tabulée ou la p-value au risque de première espèce (souvent de 5%). Si cette p-value est inférieure au seuil de signification choisi, H_0 est rejetée, ce qui signifie que le facteur a bel et bien une influence sur la variable dépendante (ou que les moyennes sont significativement différente). Le principe est de d'abord construire le graphique d'interaction entre ces facteurs et de tester la significativité de ce terme ensuite.

1.2.5. Exemple d'ANOVA 2 sans répétitions

On a dosé l'aluminium dans un emballage, par absorption atomique à la sortie de machines dans une usine de production d'emballages. Il y a trois machines distinctes, une par unité de production et

chaque emballage a été dosé par cinq laboratoires. Le but est de tester si les laboratoires sont cohérents dans leurs mesures :

	Laboratoire				
Méthode	1	2	3	4	5
1	120	120	130	150	110
2	60	70	60	70	75
3	60	50	50	60	54

- Calculer les statistiques descriptives (effectif, minimum, moyenne, écart-type, maximum) de la variable dépendante globalement et dans les niveaux de chaque facteur
- Construire des boîtes à moustaches par catégories des facteurs.
- Comment « tester » s'il existe une **différence globale sous les logiciels Excel et R** ?

1.2.6. Exemple d'ANOVA 2 avec répétitions

On a étudié la durée du développement d'un parasite (en jours) à l'intérieur de trois organismes hôtes (facteur A), en fonction de quatre températures d'élevage (facteur B). Calculez manuellement la somme des carrés des écarts totaux.

	t1	t2	t3	t4
Hôte 1	15 14 17	18 17	12 13 12	14 15
Hôte 2	16 19	23 24	15 14	12 11
Hôte 3	18 17	20 21 21	17 19	12 13

a) Donnez la spécification de l'analyse de la variance à deux critères (ANOVA 2) avec interaction
b) Construisez l'ANOVA 2 avec interaction
c) Testez l'interaction entre les deux facteurs

1.3. Analyse de la variance à trois critères

L'analyse de la variance à trois critères de classification (ANOVA 3) est un ensemble de modèles d'analyse d'expériences où interviennent simultanément trois facteurs de variation contrôlés ; chacun d'eux est présenté dans l'expérience par différents niveaux de facteurs ou variantes. Les facteurs sont dits croisés si le choix des niveaux se fait indépendamment pour chacun d'eux. Toutes les combinaisons possibles obtenues en prenant chaque niveau des trois différents facteurs sont considérées. Le but poursuivi est d'étudier l'effet de chacun d'eux pris globalement mais également la façon dont ils interagissent sur la caractéristique mesurée.

Un facteur A se présente sous I modalités, chacune d'entre elles étant noté α_i. Un facteur B se présente sous J modalités, chacune d'entre elles étant noté β_j. Un facteur C se présente sous K modalités, chacune d'entre elles étant noté ω_k. Pour chacun des couples de modalités (α_i, β_j, ω_k), effectuons une mesure d'une réponse Y qui est une variable continue.

1.3.1. Spécification du modèle

Le modèle de l'ANOVA 3 avec interaction est :

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \omega_k + (\alpha\beta)_{ij} + (\alpha\omega)_{ik} + (\beta\omega)_{jk} + (\alpha\beta\omega)_{ijk} + \varepsilon_{ijkl} \tag{1.32}$$

où y_{ijkl} est la valeur prise par la réponse Y(la variable quantitative à expliquer) dans les conditions $(\alpha_i, \beta_j, \omega_k)$, les y_{ijkl} étant indépendantes et suivent une loi normale, μ étant une moyenne globale, α_i est l'effet du i^{ème} niveau du premier facteur(facteur A), β_j est l'effet du j^{ème} niveau du second facteur(facteur B), ω_k est l'effet du k^{ème} niveau du troisième facteur(facteur C), $(\alpha\beta)_{ij}$, $(\alpha\omega)_{ik}$ et $(\beta\omega)_{jk}$ sont les interactions des facteurs deux à deux, $(\alpha\beta\omega)_{ijk}$ est l'interaction des trois facteurs et ε_{ijk} est le terme d'erreur.

Pour faire la décomposition de la variance, la première étape de l'analyse de la variance consiste à expliquer la variance totale sur l'ensemble des échantillons en fonction de la variance due aux facteurs (la variance expliquée par le modèle), de la variance due à l'interaction entre les facteurs et de la variance résiduelle aléatoire (la variance non expliquée par le modèle). S_n^2 étant un estimateur biaisé de la variance, la somme des carrés des écarts est utilisé pour les calculs et l'estimateur non biaisé de la variance S^2_{n-1} (également appelé carré moyen).

1.3.2. Statistiques descriptives

La moyenne des valeurs de y_{ijk} par rapport aux indices j et k est :

$$\overline{y}_{i..} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijkl} \tag{1.33}$$

La moyenne des valeurs de Y_{ijk} par rapport aux indices i,j et k est la moyenne $\overline{y}$, parfois aussi appelée grande moyenne du tableau :

$$\overline{y}_{...} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijkl} \tag{1.34}$$

La moyenne de ces valeurs de y_{ijk} par rapport aux indices i et k est :

$$\overline{y}_{.j.} \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} y_{ijkl} \tag{1.35}$$

La moyenne de ces valeurs de y_{ijk} par rapport aux indices i et j est :

$$\overline{y}_{..k} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ijkl} \tag{1.36}$$

1.3.3. Formule fondamentale

La variation théorique due au facteur A (somme des carrés dus au facteur A) est définie par :

$$SCF_{\alpha} = JK \sum_{i=1}^{I} \left(\overline{y}_{i..} - \overline{y}_{...} \right)^2 \tag{1.37}$$

La variation théorique due au facteur B (somme des carrés dus au facteur B) est définie par :

$$SCF_{\beta} = IK \sum_{j=1}^{J} \left(y_{.j.} - \overline{y}_{...} \right)^2 \tag{1.38}$$

La variation théorique due au facteur C (somme des carrés dus au facteur C) est définie par :

$$SCF_{\omega} = IJ \sum_{k=1}^{K} \left(Y_{..k} - Y_{...} \right)^2 \tag{1.39}$$

La variation théorique due à l'interaction de deux facteurs A et B (somme des carrés dus à l'interaction de deux facteurs) est définie par :

$$SCF_{\alpha\beta} = K \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{..} \right)^2 \tag{1.40}$$

La variation théorique due à l'interaction de deux facteurs B et C (somme des carrés dus à l'interaction de deux facteurs) est définie par :

$$SCF_{\beta\omega} = I \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\overline{y}_{.jk} - \overline{y}_{.j.} - \overline{y}_{..k} + \overline{y}_{...} \right)^2 \tag{1.41}$$

La variation théorique due à l'interaction de deux facteurs A et C (somme des carrés dus à l'interaction de deux facteurs) est définie par :

$$SCF_{\alpha\omega} = J\sum_{i=1}^{I}\sum_{k=1}^{K}\left(y_{i.k} - y_{i..} - y_{..k} + y_{...}\right)^2 \tag{1.42}$$

La variation résiduelle théorique (somme des carrés résiduels) est, quant à elle, définie par :

$$SCF_R = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{ijk} - \overline{y}_{ij.} - \overline{y}_{i.k} - \overline{y}_{.jk} + \overline{y}_{i..} + \overline{y}_{.j.} + \overline{y}_{..k} - \overline{y}_{...}\right)^2 \tag{1.43}$$

La variation totale théorique (somme des carrés totaux) est égale à :

$$SCF_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{ijk} - \overline{y}_{...}\right)^2 \tag{1.44}$$

La relation fondamentale de l'ANOVA3 est alors :

$$SCF_T = SCF_\alpha + SCF_\beta + SCF_\omega + SCF_T + SCF_{\alpha\beta} + SCF_{\alpha\omega} + SCF_{\beta\omega} + SCF_{\alpha\beta\omega} + SCR \tag{1.45}$$

Le tableau 3 est le tableau synthétique de l'analyse de la variance à trois critères de classification.

Tableau 3 : Synthèse de l'analyse de la variance à trois critères de classification

Source de variation	**Somme des carrés**	**Degrés de liberté**	**Carré moyen**	**F**
Facteur A	SCF_α	$n_\alpha = I-1$	$CMF_\alpha = SCF_\alpha / n_\alpha$	CMF_α / CMR
Facteur B	SCF_β	$n_\beta = J-1$	$CMF_\beta = SCF_\beta / n_\beta$	CMF_β / CMR
Facteur C	SCF_ω	$n_\omega = K-1$	$CMF_\omega = SCF_\omega / n_\omega$	CMF_ω / CMR
Interaction AB	$SCF_{\alpha\beta}$	$n_{\alpha\beta} = (I-1)(J-1)$	$CMF_{\alpha\beta} = SCF_{\alpha\beta} / n_{\alpha\beta}$	$CMF_{\alpha\beta} / CMR$
Interaction AC	$SCF_{\alpha\omega}$	$n_{\alpha\omega} = (I-1)(K-1)$	$CMF_{\alpha\omega} = SCF_{\alpha\omega} / n_{\alpha\omega}$	$CMF_{\alpha\omega} / CMR$
Interaction BC	$SCF_{\beta\omega}$	$n_{\beta\omega} = (J-1)(K-1)$	$CMF_{\beta\omega} = SCF_{\beta\omega} / n_{\beta\omega}$	$CMF_{\beta\omega} / CMR$
Interaction ABC	$SCF_{\alpha\beta\omega}$	$n_{\alpha\beta\omega} = (I-1)(J-1)(K-1)$	$CMF_{\alpha\beta\omega} = SCF_{\alpha\beta\omega} / n_{\alpha\beta\omega}$	$CMF_{\alpha\beta\omega} / CMR$
Résidus	SCR	$n_R = IJK(L-1)$	$CMF_R = SCR / n_R$	
Totale	SCT	$n_T = IJK - 1$	CMT	

Le test de l'ANOVA teste l'hypothèse nulle qui correspond au cas où toutes les moyennes sont égales. Et l'hypothèse alternative est qu'il existe au moins une distribution dont la moyenne s'écarte des autres moyennes.

Pour décider du non rejet ou du rejet de l'hypothèse nulle, il reste à comparer la valeur de la statistique de Fisher à la statistique calculée ou la p-value. Donc, si la p-value est inférieure au seuil de significativité choisi (souvent 5%), l'hypothèse nulle H0 est rejetée, ce qui signifie que le facteur α ou le facteur β ou le facteur ω ont bien une influence sur la variable dépendante ou encore qu'il existe une interaction entre ces trois facteurs deux à deux ou tous en même temps. Dans le cas contraire, l' *H0* n'est pas rejetée, ce qui traduit l'absence d'influence des facteurs pris séparément ou en interaction sur la variable dépendante.

Lorsque toutes les interactions du premier et du second ordre ne sont pas significatives, le modèle s'écrit :

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \omega_k + \varepsilon_{ijkl} \tag{1.46}$$

Lorsque toutes les interactions du premier ordre uniquement ne sont pas significatives, le modèle s'écrit :

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \omega_k + (\alpha\beta\omega)_{ijk} + \varepsilon_{ijkl} \tag{1.47}$$

Lorsque l'interaction du second ordre uniquement n'est pas significative, le modèle s'écrit :

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \omega_k + (\alpha\beta)_{ij} + (\alpha\omega)_{ik} + (\beta\omega)_{jk} + \varepsilon_{ijkl} \tag{1.48}$$

Le modèle de l'analyse de la variance à trois critères fixes de classifications sans répétition se matérialise de la forme suivante :

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \omega_k + (\alpha\beta)_{ij} + (\alpha\omega)_{ik} + (\beta\omega)_{jk} + \varepsilon_{ijkl} \tag{1.49}$$

Le tableau 4 est le tableau de l'analyse de la variance à trois critères de classifications.

Tableau 4 : Tableau de l'analyse de la variance à trois facteurs

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F
Facteur A	SCF_α	$n_\alpha = I-1$	CMF_α	CMF_α / CMR
Facteur B	SCF_β	$n_\beta = J-1$	CMF_β	CMF_β / CMR
Facteur C	SCF_ω	$n_\omega = K-1$	CMF_ω	CMF_ω / CMR
Interaction AB	$SCF_{\alpha\beta}$	$n_{\alpha\beta} = (I-1)(J-1)$	$CMF_{\alpha\beta}$	$CMF_{\alpha\beta}$/ CMR
Interaction AC	$SCF_{\alpha\omega}$	$n_{\alpha\omega} = (I-1)(K-1)$	$CMF_{\alpha\omega}$	$CMF_{\alpha\omega}$/ CMR
Interaction BC	$SCF_{\beta\omega}$	$n_{\beta\omega} = (J-1)(K-1)$	$CMF_{\beta\omega}$	$CMF_{\beta\omega}$/ CMR
Résidus	$SCF_{\alpha\beta\omega}$	$n_R = IJK(L-1)$	CMR	
Totale	SCR	$n_T = IJK-1$	CMT	

Nous rappelons la relation fondamentale de notre ANOVA :

$$SCT = SCF_\alpha + SCF_\beta + SCF_\omega + SCF_{\alpha\beta} + SCF_{\alpha\omega} + SCF_{\beta\omega} + SCR$$

Pour décider du rejet ou non de l'hypothèse nulle, il reste à comparer la valeur de la statistique de Fisher à la statistique tabulée ou la p-value au risque de première espèce (souvent de 5%). Donc, si la p-value est inférieure au seuil de signification choisi, H_0 est rejetée, ce qui signifie que le facteur α, le facteur β, le facteur ω ont une influence sur la variable dépendante ou encore qu'il existe une interaction entre ces trois facteurs deux à deux ou tous en même temps. Dans le cas contraire, l'H0 n'est pas rejetée, ce qui traduit l'absence d'influence des facteurs pris séparément ou en interaction sur la variable dépendante.

1.4. Analyse de la variance à trois facteurs avec répétition pour le modèle à effets mixtes

Il existe deux possibilités : soit deux facteurs sont fixes et un est aléatoire, soit un facteur est fixe et deux sont aléatoires. Pour notre cas, nous avons : un facteur fixe α se présente sous I modalités, chacune d'entre elles étant notée α_i. Un autre facteur fixe β se présente sous J modalités, chacune d'elles étant notée β_j. Un dernier facteur aléatoire ω se présente sous K modalités, chacune d'elles étant notée ω_K et ces derniers représentent un échantillon prélevé dans une population importante. Pour chacun des couples de modalités (α_i, β_j, ω_K) nous effectuons une mesure d'une réponse Y qui est une variable continue.

Nous introduisons le modèle :

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \omega_k + (\alpha\beta)_{ij} + (\alpha\omega)_{ik} + (\beta\omega)_{jk} + (\alpha\beta\omega)_{ijk} + \varepsilon_{ijkl} \tag{1.50}$$

Le tableau 5 est le tableau de l'analyse de la variance à trois critères de classifications à effets mixtes.

Tableau 5 : Tableau de l'analyse de la variance modèle mixte

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F
Facteur A	SCF_α	$n_\alpha = I-1$	$CMF_\alpha = SCF_\alpha / n_\alpha$	$CMF_\alpha / CMF_{\alpha\omega}$
Facteur B	SCF_β	$n_\beta = J-1$	$CMF_\beta = SCF_\beta / n_\beta$	$CMF_\beta / CMF_{\beta\omega}$
Facteur C	SCF_ω	$n_\omega = K-1$	$CMF_\omega = SCF_\omega / n_\omega$	CMF_ω / CMR
Interaction AB	$SCF_{\alpha\beta}$	$n_{\alpha\beta} = (I-1)(J-1)$	$CMF_{\alpha\beta} = SCF_{\alpha\beta} / n_{\alpha\beta}$	$CMF_{\alpha\beta} / CMF_{\alpha\beta\omega}$
Interaction AC	$SCF_{\alpha\omega}$	$n_{\alpha\omega} = (I-1)(K-1)$	$CMF_{\alpha\omega} = SCF_{\alpha\omega} / n_{\alpha\omega}$	$CMF_{\alpha\omega} / CMR$
Interaction BC	$SCF_{\beta\omega}$	$n_{\beta\omega} = (J-1)(K-1)$	$CMF_{\beta\omega} = SCF_{\beta\omega} / n_{\beta\omega}$	$CMF_{\beta\omega} / CMR$
Interaction ABC	$SCF_{\alpha\beta\omega}$	$n_{\alpha\beta\omega} = (I-1)(J-1)(K-1)$	$CMF_{\alpha\beta\omega} = SCF_{\alpha\beta\omega} / n_{\alpha\beta\omega}$	$CMF_{\alpha\beta\omega} / CMR$
Résidus	SCR	$n_R = IJK(L-1)$	$CMF_R = SCR / n_R$	
Totale	SCT	$n_T = IJK-1$	CMT	

Nous supposons que les conditions d'utilisation de ce modèle sont bien remplies. Pour la décision du rejet ou non rejet de l'hypothèse nulle, il reste à comparer la statistique de Fisher à la statistique calculée ou la p-value au seuil choisi.

Le tableau 6 compile les résultats de l'analyse de la variance à trois facteurs fixes et croisés à savoir la variété de la pomme de terre, le site et le traitement sans effet interaction du second ordre et un du premier ordre (Variété : Traitement).

Tableau 6 : Résultats de l'analyse de la variance à 3 facteurs

Source de variation	Somme des carrés	Degré de liberté	Carrés moyens	F	P-value
Variété	7	20980	2997,1	147,75	<0,001
Site	2	2088	1043,9	51,46	<0,001
Traitement	1	82	82,4	4,062	0,046
Variété : Site	14	4655	332,5	16,39	<0,001
Site : Traitement	2	89	44,5	2,19	0,116
Résidus	117	2373	20,3	//////	/////////
Totale	143	30267	4520,7	//////	/////////

Pour ce modèle, l'interaction entre le site et le traitement n'a pas d'effet significatif sur le rendement car leur p-value est supérieure au seuil de 5 % mais l'interaction du premier ordre entre la variété et le site présente un effet sur le rendement de la pomme de terre hautement significatif (p-value <0,001). Tous les effets principaux présentent des effets significatifs sur le rendement. L'AIC de ce modèle vaut 868,18.

Le tableau 7 compile les résultats de l'analyse de la variance à trois facteurs fixes et croisés à savoir la variété, le site et le traitement sans effet interaction du second ordre et deux du premier ordre (Variété : Traitement ; Site : Traitement).

Tableau 7 : Résultats de l'analyse de la variance à trois facteurs

Source de variation	**Somme des carrés**	**Degré de liberté**	**Carres moyens**	**F**	**P-value**
Variété	20980	7	2997,1	144,84	<0,001
Site	2088	2	1043,9	50,45	<0,001
Traitement	82	1	82,4	3,98	0,048
Variété : Site	4655	14	332,5	16,07	<0,001
Résidus	2462	119	20,7	////////	/////////
Totale	30267	143	211,6	**/////// **	**////////**

Les facteurs Variété et Site ont un effet hautement significatif sur le rendement de la pomme de terre car les p-values sont beaucoup inférieures au seuil (5 %) ; il en est de même pour leur interaction. Pour le facteur Traitement, son influence sur le rendement de la pomme de terre est faiblement significatif car la p-value est inférieure au seuil mais proche du seuil de significativité de 5 %. L'AIC de ce modèle est égal à 869,48. Ce modèle est choisit car c'est celui qui minimise l'AIC ; il a l'AIC le plus petit.

Chapitre 2 : Modèle linéaire

2.1. Régression linéaire simple

2.1.1. Introduction

Le but de la régression linéaire est d'établir un lien entre une variable dépendante y et une variable indépendante x (ou une combinaison linéaire des variables indépendantes) pour pouvoir ensuite faire des prévisions sur y lorsque x est mesurée. La variable dépendante y doit être de type quantitatif et les variables indépendantes x doivent être des variables toutes quantitatives, des variables toutes qualitatives ou une combinaison des deux.

Il existe plusieurs exemples d'application du modèle linéaire dans la vie courante :

- Plus la pauvreté augmente, plus la mortalité augmente ;
- La taille d'un fils est influencée par la taille du père ;
- Plus l'âge du nourrisson augmente, plus sa taille augmente ;
- Plus l'offre augmente, plus la demande diminue ;
- Les revenus influencent les dépenses ;
- Plus les prix des boissons diminuent, plus le nombre de consommateurs augmente.

Le modèle linéaire simple s'écrit :

$$y = ax + b \tag{2.1}$$

Le modèle linéaire sans constante s'écrit :

$$y = ax \tag{2.2}$$

Le modèle log-linéaire s'écrit :

$$y = bx^a \Rightarrow \ln(y) = a\ln(x) + \ln(b) \tag{2.3}$$

Le modèle logarithmique s'écrit :

$$y = a\ln(x) + b \tag{2.4}$$

Le modèle logistique s'écrit :

$$\ln\left(\frac{p}{1-p}\right) = ax + b \tag{2.5}$$

avec p la probabilité que la variable dépendante qualitative prenne la valeur 1**.**

En statistique, un modèle de régression linéaire est un modèle de régression d'une variable expliquée (ou variable dépendante) sur une ou plusieurs variables explicatives en faisant l'hypothèse que la fonction qui lie les variables explicatives à la variable expliquée est linéaire dans ses paramètres. Un modèle de régression linéaire est aussi appelé tout simplement modèle linéaire.

Les objectifs pédagogiques sont les suivants :

1) Utiliser dans la pratique les concepts de corrélation, de régression, de coefficient de régression, de coefficient de détermination ;
2) Calculer les coefficients de régression avec un logiciel approprié et les interpréter ;
3) Faire de l'inférence statistique des coefficients de régression avec un logiciel approprié ;
4) Identifier les différentes méthodes de sélection de variables à utiliser en régression ;
5) Utiliser les techniques adéquates pour valider les modèles de régression ;

Le but de la régression linéaire est double :

- Étudier l'association entre les variables quantitatives ;
- Faire la prédiction de l'une par l'autre.

2.1.2. Spécification du modèle

Le modèle linéaire simple s'écrit [1,4] :

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.6}$$

où y est la variable dépendante (variable à expliquer, variable expliquée, variable à prédire, variable réponse, variable endogène en économie), β_0 l'intercept (ordonnée à l'origine), β_1 la pente, x la variable indépendante (variable explicative, variable prédictrice, prédicteur, facteur, régresseur, variable exogène en économie) et ε l'erreur du modèle (erreur aléatoire) ou la perturbation. Ce dernier terme intervient pour résumer toute information non prise en compte dans la relation linéaire entre y et x.

Pour chaque observation, ce modèle s'écrit [5] :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2.7}$$

Si, par exemple, la variable dépendante est le revenu annuel (*variable y*) et la variable indépendante est l'âge (*variable x*), les questions qui se posent sont les suivantes :

- Existe-t-il un lien statistique entre l'âge et le revenu annuel ?
- Si ce lien existe, quelle est sa nature ? Est-il possible de l'exprimer à l'aide d'une fonction linéaire (modèle mathématique) ?
- Cette fonction qui lie la variable indépendante et la variable dépendante permet-elle de faire des prévisions à des valeurs proches de celles obtenues lors de l'échantillonnage ?

2.1.3. Écriture matricielle

Matriciellement, ce modèle s'écrit :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{1i} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{2.8}$$

En posant :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} ; \; X = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{1i} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix} ; \; \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} ; \; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{2.9}$$

ce modèle s'écrit :

$$Y = X\beta + \varepsilon \tag{2.10}$$

Dans cette relation, la variable y (ou Y) est considérée comme variable aléatoire et son caractère aléatoire provient du fait que la variable indépendante x ne parvient pas (insuffisances) à expliquer les valeurs de y.

2.1.4. Hypothèses du modèle

Les conditions d'utilisation sont constituées par des hypothèses stochastiques qui portent sur les résidus du modèle et des hypothèses structurelles du modèle lui-même. Les hypothèses stochastiques sont [6] :

1) Les erreurs doivent être de moyenne nulle (modèle bien spécifié) : $E(\varepsilon) = 0$
2) Les erreurs doivent être de variance constante (homoskédasticité) : $Var(\varepsilon) = \sigma^2 I_n = c^{te}$
3) Les erreurs doivent être de loi normale : $\varepsilon \sim N(0, \sigma^2 I_n)$
4) Les erreurs doivent être indépendantes (non-autocorrélées) : $cov(\varepsilon_i, \varepsilon_j) = 0 \; \forall i \neq j$
5) Les erreurs et les variables explicatives doivent être indépendantes : $cov(\varepsilon_i, x^j) = 0$

Les hypothèses structurelles ou déterministes sont :

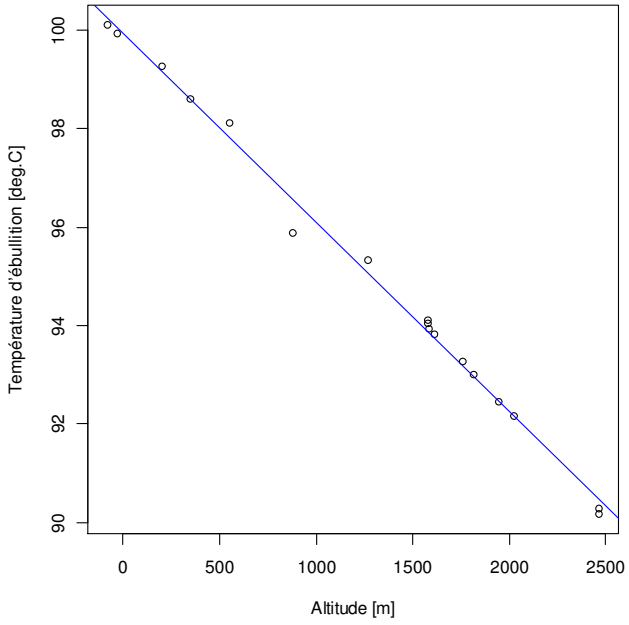
6) La matrice X doit être de rang complet : $rang(X) = p$
7) Les variables explicatives (de même que la variable dépendante) sont fixes (mesurées sans erreurs) : $E(x_j) = x_j$
8) La relation entre la variable dépendante et la variable indépendante doit être linéaire : $Y = X\beta$
9) Les observations de y doivent être i.i.d. (indépendantes et identiquement distribuées).

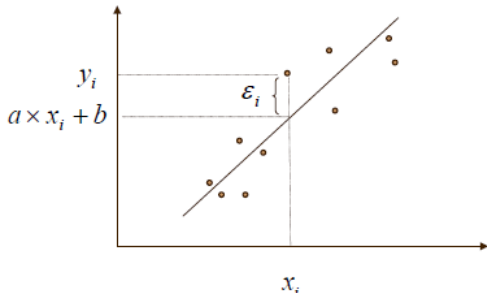
2.1.5. Estimation des paramètres

L'estimation des paramètres $\beta = (\beta_0, \beta_1)$ ainsi que celle de la variance résiduelle (σ^2) se fait par la méthode des moindres carrés ordinaires qui consiste à minimiser la somme des carrés des erreurs commises en considérant que tous les points appartiennent à la droite de régression alors que non. Il convient aussi de vérifier que le modèle est adéquat.

La fonction de perte est une somme des carrés des écarts entre les valeurs observées de la variable dépendante et les valeurs prédites. La différence entre ces écarts constitue ce qui est appelé erreur du modèle ou résidu.

Considérons le nuage de points (diagramme de dispersion) de la température en ébullition (en °C) en fonction de l'altitude (en mètres) ci-après, l'équation cartésienne de la droite étant y=ax+b où b est l'intercept (ordonnée à l'origine) et b la pente de la droite.





Ce graphique est allongé et montre qu'un ajustement linéaire est envisageable. Considérons un point n'appartenant pas à la droite de régression (en bleu) et de coordonnées (x_i, y_i). Un point appartenant à la droite de régression et qui a la même abscisse que le point précédent aura pour coordonnées $(x_i, \hat{y}_i)$. L'écart entre la valeur observée y_i et la valeur prédite $\hat{y}_i$ est appelé erreur ou résidu du modèle et est noté ε_i :

$$\varepsilon_i = y_i - \hat{y}_i \tag{2.11}$$

Le principe des moindres carrés ordinaires (MCO) consiste à minimiser la somme des carrés des écarts entre les valeurs observées et les valeurs prédites de la variable réponse, ce qui conduit à minimiser la fonction de perte :

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.12}$$

Ainsi, la dérivation de la fonction de perte par rapport à β_0 donne :

$$\begin{aligned}\frac{\partial S}{\partial \beta_0} &= \frac{\partial}{\partial \beta_0}\left[\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2\right] \\ &= \frac{\partial}{\partial \beta_0}\left[\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2\right] \\ &= -2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)\end{aligned} \tag{2.13}$$

En égalant cette dérivée à zéro, il vient :

$$\begin{aligned}\frac{\partial S}{\partial \beta_0} = 0 &\Leftrightarrow -2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)=0 \\ &\Leftrightarrow \sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)=0 \\ &\Leftrightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\beta_0 - \beta_1\sum_{i=1}^{n} x_i = 0\end{aligned}$$

Or

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \sum_{i=1}^{n} x_i = n\bar{x}$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \Rightarrow \sum_{i=1}^{n} y_i = n\bar{y}$$

Donc

$$\begin{aligned}\frac{\partial S}{\partial \beta_0} = 0 &\Leftrightarrow n\bar{y} - n\beta_0 - n\beta_1\bar{x} = 0 \\ &\Leftrightarrow n\left(\bar{y}-\beta_0-\beta_1\bar{x}\right)=0 \\ &\Leftrightarrow \bar{y}-\beta_0-\beta_1\bar{x}=0 \\ &\Leftrightarrow \hat{\beta}_0 = \bar{y}-\hat{\beta}_1\bar{x}\end{aligned}$$

Pour la suite, retenons l'équation :

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\beta_0 - \beta_1\sum_{i=1}^{n} x_i = 0 \Leftrightarrow \sum_{i=1}^{n}\beta_0 + \beta_1\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \Leftrightarrow n\beta_0 + \beta_1\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

La dérivation de la fonction de perte par rapport à β_1 donne :

$$\begin{aligned}\frac{\partial S}{\partial \beta_1} &= \frac{\partial}{\partial \beta_1}\left[\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2\right] \\ &= \frac{\partial}{\partial \beta_1}\left[\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2\right] \\ &= -2\sum_{i=1}^{n} x_i\left(y_i-\beta_0-\beta_1 x_i\right) \\ &= -2\left(\sum_{i=1}^{n} x_i y_i-\beta_0\sum_{i=1}^{n} x_i-\beta_1\sum_{i=1}^{n} x_i^2\right)\end{aligned}$$

En égalant cette dérivée à zéro, il vient :

$$\begin{aligned}\frac{\partial S}{\partial \beta_1}=0 &\Leftrightarrow -2\left(\sum_{i=1}^{n} x_i y_i-\beta_0\sum_{i=1}^{n} x_i-\beta_1\sum_{i=1}^{n} x_i^2\right)=0 \\ &\Leftrightarrow \sum_{i=1}^{n} x_i y_i-\beta_0\sum_{i=1}^{n} x_i-\beta_1\sum_{i=1}^{n} x_i^2=0 \\ &\Leftrightarrow \sum_{i=1}^{n} x_i y_i-\left(\bar{y}-\beta_1\bar{x}\right)\sum_{i=1}^{n} x_i-\beta_1\sum_{i=1}^{n} x_i^2=0 \\ &\Leftrightarrow \sum_{i=1}^{n} x_i y_i-\bar{y}\sum_{i=1}^{n} x_i-\beta_1\bar{x}\sum_{i=1}^{n} x_i-\beta_1\sum_{i=1}^{n} x_i^2=0 \\ &\Leftrightarrow \sum_{i=1}^{n} x_i y_i-n\bar{x}\bar{y}-\beta_1 n\bar{x}^2-\beta_1\sum_{i=1}^{n} x_i^2=0 \\ &\Leftrightarrow \sum_{i=1}^{n} x_i y_i-n\bar{x}\bar{y}-\beta_1\left(\sum_{i=1}^{n} x_i^2-n\bar{x}^2\right)=0\end{aligned}$$

Donc

$$\begin{aligned}\frac{\partial S}{\partial \beta_1}=0 &\Leftrightarrow \sum_{i=1}^{n} x_i y_i-n\bar{x}\bar{y}-\beta_1\left(\sum_{i=1}^{n} x_i^2-n\bar{x}^2\right)=0 \\ &\Leftrightarrow \hat{\beta}_1=\frac{\sum_{i=1}^{n} x_i y_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2-n\bar{x}^2}\end{aligned}$$

Sachant que la covariance entre deux variables x et y s'écrit :

$$\begin{aligned} cov(x,y) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \overline{xy}) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i y_i - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + \sum_{i=1}^{n}\overline{xy}\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i y_i - n\overline{xy} - n\overline{xy} + n\overline{xy}\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i y_i - n\overline{xy}\right) \end{aligned}$$

$$Cov(x,y) = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i y_i - n\overline{xy}\right) \Rightarrow \sum_{i=1}^{n}x_i y_i - n\overline{xy} = (n-1)Cov(x,y)$$

et la variance de x :

$$\begin{aligned} S_x^2 = Var(x) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 + n\bar{x}^2\right) \end{aligned}$$

$$Var(x) = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right) \Rightarrow \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 = (n-1)Var(x)$$

Il vient finalement :

$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{S_{xy}}{S_x^2}$$

Pour la suite, retenons l'équation :

$$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \Leftrightarrow \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Cela conduit au système d'équations normales qui se résout numériquement :

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Cela s'écrit matriciellement :

$$\underbrace{\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\beta} = \underbrace{\begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}}_{B}$$

En pré-multipliant membre à membre par la matrice inverse, il vient :

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

soit :

$$\hat{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} \tag{2.14}$$

Cette relation est équivalente à [6] :

$$\begin{aligned} Y = X\beta + \varepsilon &\Leftrightarrow \left(X^t X\right)^{-1} X^t Y = \left(X^t X\right)^{-1} X^t X \beta \\ &\Leftrightarrow \hat{\beta}_{MCO} = \left(X^t X\right)^{-1} X^t Y \end{aligned} \tag{2.15}$$

Avant d'estimer les paramètres, il convient de jeter un coup d'œil sur le nuage de points afin de se rendre compte qu'il n'est pas nécessaire de faire subir aux variables des transformations mathématiques. De plus, il convient de calculer la notion de coefficient de corrélation linéaire simple à l'aide de la notion de covarience. L'étude de la corrélation s'intéresse à l'étude de la force de la relation linéaire. La corrélation répond à l'existence d'une association linéaire entre deux variables et la force de la relation est mesurée par le coefficient de corrélation.

Les dérivées secondes de la fonction de perte sont :

$$\frac{\partial^2 S}{\partial \beta_0^2} = -2\frac{\partial}{\partial \beta_0}\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)\right] = -2\sum_{i=1}^{n}(-1) = 2n$$

$$\frac{\partial^2 S}{\partial \beta_1^2} = -2\frac{\partial}{\partial \beta_1}\left[\sum_{i=1}^{n}\left(x_i y_i - \beta_0 x_i - \beta_1 x_i^2\right)\right] = -2\sum_{i=1}^{n}\left(-x_i^2\right) = 2\sum_{i=1}^{n} x_i^2$$

$$\frac{\partial^2 S}{\partial \beta_0 \partial \beta_1} = -2\frac{\partial}{\partial \beta_1}\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)\right] = -2\sum_{i=1}^{n}\left(-x_i\right) = 2\sum_{i=1}^{n} x_i = 2n\overline{x}$$

$$\frac{\partial^2 S}{\partial \beta_1 \partial \beta_0} = -2\frac{\partial}{\partial \beta_0}\left[\sum_{i=1}^{n}\left(x_i y_i - \beta_0 x_i - \beta_1 x_i^2\right)\right] = -2\sum_{i=1}^{n}\left(-x_i\right) = 2\sum_{i=1}^{n} x_i = 2n\overline{x}$$

La matrice hessienne, qui est constituée des dérivées secondes, est :

$$H = 2\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} = 2X^t X \tag{2.16}$$

Pour tout vecteur $u \in \mathbb{R}^2$, les formes quadratiques $u'Hu$ peuvent s'écrire $v^t v$ en posant $v = Xu$. Comme $v^t v$ est toujours positif, alors la matrice H est définie positive.

La moyenne des résidus est nulle. En effet :

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{y}_i\right) = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \overline{y} - \overline{y} = 0$$

De plus, $\sum_{i=1}^{n} x_i\varepsilon_i = 0$. En effet :

$$\sum_{i=1}^{n} x_i\varepsilon_i = \sum_{i=1}^{n} x_i\left(y_i - \hat{y}_i\right) = \sum_{i=1}^{n} x_i\left(y_i - \beta_0 - \beta_1 x_i\right) = \sum_{i=1}^{n} x_i\left[y_i - \overline{y} - \beta_1\left(x_i - \overline{x}\right)\right]$$

Sachant que :

$$\begin{aligned}\sum_{i=1}^{n} x_i\left[y_i - \overline{y} - \beta_1\left(x_i - \overline{x}\right)\right] &= \sum_{i=1}^{n} x_i\left[y_i - \overline{y} - \beta_1\left(x_i - \overline{x}\right)\right] \\ &= \sum_{i=1}^{n} x_i\left(y_i - \overline{y}\right) - \beta_1\sum_{i=1}^{n} x_i\left(x_i - \overline{x}\right) \\ &= \sum_{i=1}^{n} x_i\left(y_i - \overline{y}\right) - \beta_1\sum_{i=1}^{n} x_i\left(x_i - \overline{x}\right)\end{aligned}$$

$$\sum_{i=1}^{n} x_i\left(y_i - \overline{y}\right) = \sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} = (n-1)S_{xy}$$

$$\sum_{i=1}^{n} x_i\left(x_i - \overline{x}\right) = \sum_{i=1}^{n} x_i^2 - \overline{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = (n-1)S_x^2$$

D'où :

$$\sum_{i=1}^{n} x_i\varepsilon_i = \sum_{i=1}^{n} x_i\left[y_i - \overline{y} - \beta_1\left(x_i - \overline{x}\right)\right] = (n-1)S_{xy} - (n-1)\frac{S_{xy}}{S_x^2}S_x^2 = 0$$

Également :

$$\sum_{i=1}^{n}\hat{y}_i\varepsilon_i = \sum_{i=1}^{n}\left(\beta_0 + \beta_1 x_i\right)\varepsilon_i = \beta_0\sum_{i=1}^{n}\varepsilon_i + \beta_1\sum_{i=1}^{n} x_i\varepsilon_i = 0$$

2.1.6. Estimateur du maximum de vraisemblance

La fonction de vraisemblance est une fonction qui dépend de trois paramètres à estimer et est associée aux observations $y_1, y_2, ..., y_n$ de y [2,5] :

$$L\left(\beta_0, \beta_1, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}\left(y_i - \beta_0 - \beta_1 x_i\right)^2 \right]$$
$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \right]$$

La log-vraisemblance vaut alors :

$$l\left(\beta_0, \beta_1, \sigma^2\right) = ln\, L\left(\beta_0, \beta_1, \sigma^2\right)$$
$$= ln\left\{ \left(2\pi\sigma^2\right)^{-\frac{n}{2}} exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \right] \right\}$$
$$= -n\left(ln\sqrt{2\pi} + \frac{1}{2} ln\,\sigma^2 \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

La dérivée partielle de la log-vraisemblance par rapport à vaut β_0 :

$$\frac{\partial l}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)$$

La dérivée partielle de la log-vraisemblance par rapport à vaut β_1 :

$$\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i \left(y_i - \beta_0 - \beta_1 x_i\right) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(x_i y_i - \beta_0 x_i - \beta_1 x_i^2\right)$$

La dérivée partielle de la log-vraisemblance par rapport à vaut β_1 :

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

L'annulation de ces trois dérivées partielles donne le système d'équations de vraisemblance :

$$\begin{cases} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right) = 0 \\ \sum_{i=1}^{n} \left(x_i y_i - \beta_0 x_i - \beta_1 x_i^2\right) = 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 = 0 \end{cases}$$

2.1.7. Coefficient de corrélation

Le coefficient linéaire de Bravais-Pearson entre deux variables quantitatives s'écrit :

$$r = \frac{\mathrm{Cov}(x, y)}{\sqrt{Var(x)}\sqrt{Var(y)}} = \frac{S_{xy}}{S_x S_y} \tag{2.17.a}$$

Ce coefficient est un nombre sans unité. De plus, son signe donne une information sur le sens de la droite (directe ou inverse). Également, il varie entre 1 et -1 comme le cosinus et le sinus. S'il est négatif, cela signifie qu'aux plus grandes valeurs de la variable indépendante x correspondent les plus grandes valeurs de la variable réponse y, sans pour autant parler de notion de causalité. La qualification de l'association entre les deux variables se fait comme suit :

- Si r est compris entre 0 et 0,1, alors l'association est nulle ;
- Si r est compris entre 0,1 et 0,3, alors l'association est faible ;
- Si r est compris entre 0,3 et 0,5, alors l'association est moyenne ;
- Si r est compris entre 0,5 et 0,75, alors l'association est bonne ;
- Si r est compris entre 0,75 et 1, alors l'association est excellente.

Un coefficient de corrélation linéaire positif montre un lien linéaire positif, ce qui veut dire qu'aux plus grandes valeurs de la variable indépendante correspondent les plus grandes valeurs de la variable dépendante et vice-versa. De même, un coefficient de corrélation linéaire négatif traduit un lien linéaire négatif, ce qui veut dire qu'aux plus grandes valeurs de la variable indépendante correspondent les plus petites valeurs de la variable dépendante et vice-versa. Il faut, toutefois, éviter de parler d'un lien causal. Le lien est fort lorsque ce coefficient d'approche de -1 ou de 1.

Le rapport entre la pente de la droite de régression et ce coefficient de corrélation est :

$$\frac{\hat{\beta}_1}{r} = \frac{S_y}{S_x} \Rightarrow \hat{\beta}_1 = r\frac{S_y}{S_x} \tag{2.17.b}$$

Cela signifie que la pente de la droite de régression et le coefficient de corrélation ont même signe. Il convient aussi de tester la significativité du coefficient de corrélation linéaire simple.

Les hypothèses du test de corrélation linéaire sont :

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

L'hypothèse nulle stipule que la corrélation est nulle et l'hypothèse alternative (ou contraire) que la corrélation n'est pas nulle.

La statistique de test est celle de Student :

$$t^{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \text{ à } (n-2) \text{ ddl} \tag{2.18}$$

L'hypothèse nulle H_0 sera rejetée lorsque la p-value (probabilité que la statistique de test observée soit au moins aussi grande que la statistique tabulée sous l'hypothèse nulle H_0) est inférieure au seuil de décision (coefficient de risque) qui est généralement choisi à 5 %. Dans ce cas, la conclusion sera que la corrélation n'est pas nulle et il sera possible d'estimer les paramètres du modèle. A contrario, lorsque la p-value est supérieure ou égale au seuil, alors l'hypothèse nulle H_0 ne sera pas rejetée. Il est à noter que l'utilisation du coefficient de corrélation de Bravais-Pearson nécessite que la variable dépendante suive une loi normale et que les observations de la variable réponse soient indépendantes. Si cela n'est pas vérifié, le coefficient de corrélation de Spearman sera calculé.

2.1.8. Propriétés de l'estimateur MCO

Les moments de Y sont :

$$E(Y) = E(X\beta + \varepsilon) = E(X\beta) + E(\varepsilon) = E(X\beta) = X\beta$$

$$Var(Y) = Var(X\beta + \varepsilon) = Var(\varepsilon) = \sigma_\varepsilon^2$$

Cela montre que :

$$Y \sim \mathrm{N}(X\beta, \sigma_\varepsilon^2)$$

Les moments de l'estimateur des paramètres sont :

$$E\left(\hat{\beta}\right) = E\left(\left(X^t X\right)^{-1} X^t Y\right) = \left(X^t X\right)^{-1} X^t E\left(Y\right) = \left(X^t X\right)^{-1} X^t X \beta = \beta$$

$$\begin{aligned} Var\left(\hat{\beta}\right) &= Var\left[\left(X^t X\right)^{-1} X^t Y\right] \\ &= \left(X^t X\right)^{-1} X^t \left[\left(X^t X\right)^{-1} X^t\right]^t Var\left(Y\right) \\ &= \left(X^t X\right)^{-1} X^t X \left(X^t X\right)^{-1} \sigma_\varepsilon^2 \\ &= \sigma_\varepsilon^2 \left(X^t X\right)^{-1} \end{aligned}$$

Cela montre que :

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \sigma_\varepsilon^2 \left(X^t X\right)^{-1}\right) \tag{2.19}$$

L'estimateur de la pente peut s'écrire :

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \omega_i \varepsilon_i$$

Dans ce cas, son espérance mathématique est :

$$E\left(\hat{\beta}_1\right) = \beta_1 + E\left(\sum_{i=1}^{n} \omega_i \varepsilon_i\right) = \beta_1 + \sum_{i=1}^{n} \omega_i E\left(\varepsilon_i\right) = \beta_1$$

L'estimateur $\hat{\beta}_1$ est donc un estimateur sans biais de β_1.

De même,

$$E\left(\hat{\beta}_0\right) = E\left[\beta_0 + \overline{\varepsilon} - \left(\hat{\beta}_1 - \beta_1\right)\overline{x}\right] = E\left(\beta_0\right) = \beta_0$$

L'espérance mathématique montre que l'estimateur moindres carrés ordinaires $\hat{\beta}$ est un estimateur sans biais de β. Cela n'est possible que lorsque la variable indépendante n'est pas stochastique (non aléatoire) et que le modèle est bien spécifié ($E\left(\varepsilon\right) = 0$).

La variance de l'estimateur de la pente vaut :

$$
\begin{aligned}
Var\left(\hat{\beta}_1\right) &= E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} \omega_i \varepsilon_i\right)^2\right] \\
&= E\left(\sum_{i=1}^{n} \omega_i^2 \varepsilon_i^2 + 2\sum_{i=1}^{n}\sum_{i'=1}^{n} \omega_i \omega_{i'} \varepsilon_i \varepsilon_{i'}\right) \\
&= \sum_{i=1}^{n} \omega_i^2 E\left(\varepsilon_i^2\right) + 2\sum_{i=1}^{n}\sum_{i'=1}^{n} \omega_i \omega_{i'} E\left(\varepsilon_i \varepsilon_{i'}\right) \\
&= \sum_{i=1}^{n} \omega_i^2 Var\left(\varepsilon_i\right) + 0 \\
&= \sum_{i=1}^{n} \omega_i^2 \sigma_\varepsilon^2 \; avec \; \omega_i = \frac{x_i - \overline{x}}{\sum_j \left(x_j - \overline{x}\right)^2} \\
&= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}
\end{aligned}
$$

Si $n \to +\infty$, alors $\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 \to +\infty$ et de ce fait $Var\left(\hat{\beta}_1\right) \to 0$. L'estimateur $\hat{\beta}_1$ est donc un estimateur convergent de β_1.

De même, la variance de l'estimateur de l'intercept est :

$$
Var\left(\hat{\beta}_0\right) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} \left(x_j - \overline{x}\right)^2}\right] \xrightarrow[n \to +\infty]{} 0 \qquad (2.20)
$$

Les estimateurs MCO sont donc des estimateurs **sans biais** et **convergents**. Parmi les estimateurs sans biais, les estimateurs MCO sont à variance minimale. Il est impossible de trouver d'autres estimateurs sans biais et à plus petite variance. Les estimateurs MCO sont donc des estimateurs BLUE (Best Linear Unbiased Estimators : Meilleurs Estimateurs Linéaires Sans Biais). Ce sont des estimateurs efficaces.

Il est à noter que la droite de régression passe par le point moyen $g(\overline{x}, \overline{y})$ qui est le centre de gravité du nuage car :

$$\hat{y}(\overline{x}) = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x} = \overline{y}$$

De plus, les estimateurs MCO (Moindres Carrés Ordinaires) ou OLS (Ordinary Least Squares) sont d'autant plus précis que :

- La variance de l'erreur est faible (la droite de régression passe bien au milieu des points du nuage) ;
- La dispersion des x est forte (les x couvrent bien l'espace de représentation).

2.1.9. Décomposition fondamentale

Comme en analyse de la variance, la somme des carrés des écarts totaux est égale à la somme de la somme des carrés des écarts expliqués (de la régression) et de la somme des carrés des écarts résiduels, sachant que les formules de calcul changent :

$$SCE_F + SCE_R = (n-1)\frac{S_{xy}^2}{S_x^2} + (n-1)S_y^2 - (n-1)\frac{S_{xy}^2}{S_x^2} = (n-1)S_y^2 = SCE_T$$

Telle est la formule fondamentale de l'analyse de régression : la somme des carrés des écarts totaux est égale à la somme de la somme des carrés des écarts factoriels et de la somme des carrés des écarts résiduels.

Cela peut aussi se démontrer comme suit (développement télescopique) :

$$\begin{aligned}
&y_i - \overline{y} = y_i - \hat{y}_i + \hat{y}_i - \overline{y} \\
&\Leftrightarrow (y_i - \overline{y})^2 = (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 \\
&\Leftrightarrow \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 \\
&\Leftrightarrow \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} \left[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) \right] \\
&\Leftrightarrow \underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{SCE_T} = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SCE_R} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \hat{y}_i)^2}_{SCE_F} + 2\underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \overline{y})}_{0}
\end{aligned}$$

En effet :

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y}) &= \sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)\left(\hat{\beta}_0+\hat{\beta}_1x_i-\overline{y}\right)\\
&= \sum_{i=1}^{n}\left(y_i-\overline{y}-\hat{\beta}_1\overline{x}-\hat{\beta}_1x_i\right)\left(\overline{y}-\hat{\beta}_1\overline{x}+\hat{\beta}_1x_i-\overline{y}\right)\\
&= \sum_{i=1}^{n}\left(y_i-\overline{y}-\hat{\beta}_1\overline{x}-\hat{\beta}_1x_i\right)\left(\hat{\beta}_1\overline{x}+\hat{\beta}_1x_i\right)\\
&= \hat{\beta}_1\sum_{i=1}^{n}\left[(y_i-\overline{y})-\hat{\beta}_1(x_i-\overline{x})\right](x_i-\overline{x})\\
&= \hat{\beta}_1\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})-\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\overline{x})^2\\
&= \frac{S_{xy}}{S_x^2}(n-1)S_{xy}-\left(\frac{S_{xy}}{S_x^2}\right)^2(n-1)S_x^2\\
&= (n-1)\frac{S_{xy}^2}{S_x^2}-(n-1)\frac{S_{xy}^2}{S_x^2}\\
&= 0
\end{aligned}$$

La somme des carrés des écarts totaux s'écrit [3] :

$$\begin{aligned}
SCE_T &= \sum_{i=1}^{n}(y_i-\overline{y})^2\\
&= \sum_{i=1}^{n}\left(y_i^2-2y_i\overline{y}+\overline{y}^2\right)\\
&= \sum_{i=1}^{n}y_i^2-2\overline{y}\sum_{i=1}^{n}y_i+\overline{y}^2\sum_{i=1}^{n}1\\
&= \sum_{i=1}^{n}y_i^2-2n\overline{y}^2+n\overline{y}^2\\
&= \sum_{i=1}^{n}y_i^2-n\overline{y}^2
\end{aligned}$$

La somme des carrés des écarts factoriels (expliqués) vaut :

$$\begin{aligned}
SCE_F &= \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y} \right)^2 \\
&= \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y} \right)^2 \\
&= \sum_{i=1}^{n} \left(\overline{y} + \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i - \overline{y} \right)^2 \\
&= \hat{\beta}_1^2 \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2 \\
&= \left(\frac{S_{xy}}{S_x^2} \right)^2 (n-1) S_x^2 \\
&= (n-1) \frac{S_{xy}^2}{S_x^2}
\end{aligned}$$

La somme des carrés des écarts résiduels vaut :

$$\begin{aligned}
SCE_R &= \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \\
&= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \\
&= \sum_{i=1}^{n} \left(y_i - \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i \right)^2 \\
&= \sum_{i=1}^{n} \left[\left(y_i - \overline{y} \right)^2 + \hat{\beta}_1^2 \left(x_i - \overline{x} \right)^2 - 2 \left(x_i - \overline{x} \right) \left(y_i - \overline{y} \right) \right] \\
&= \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2 - 2 \hat{\beta}_1 \sum_{i=1}^{n} \left(x_i - \overline{x} \right) \left(y_i - \overline{y} \right) \\
&= (n-1) S_y^2 + \left(\frac{S_{xy}}{S_x^2} \right)^2 (n-1) S_x^2 - 2(n-1) \frac{S_{xy}}{S_x^2} S_{xy} \\
&= (n-1) S_y^2 + (n-1) \frac{S_{xy}^2}{S_x^2} - 2(n-1) \frac{S_{xy}^2}{S_x^2} \\
&= (n-1) S_y^2 - (n-1) \frac{S_{xy}^2}{S_x^2}
\end{aligned}$$

2.1.10. Coefficient de détermination non ajusté

Le coefficient de détermination non ajusté peut se calculer comme suit :

$$R^2 = \left(r\right)^2 \tag{2.21.a}$$

ou

$$R^2 = \frac{SCE_F}{SCE_T} = 1 - \frac{SCE_R}{SCE_T} \tag{2.21.b}$$

Il exprime le pourcentage de la variabilité de la variable dépendante expliqué par les variations de la variable indépendante. Il peut donc s'exprimer en pourcentage (proportion) et de ce fait, il prend ses valeurs entre 0 et 1. Si ce coefficient est proche de 1, alors le modèle est excellent et s'il est proche de 0, alors le modèle ne sert à rien. C'est un nombre sans unité et qui donne la force de la relation entre x et y.

2.1.11. Coefficient de détermination ajusté

Le coefficient de détermination ajusté, quant à lui, se calcule comme suit :

$$R_a^2 = 1 - \frac{\frac{SCE_R}{n-2}}{\frac{SCE_T}{n-1}} = 1 - \frac{SCE_R}{SCE_T}\frac{n-1}{n-2} \tag{2.22}$$

Il est possible de calculer aussi le coefficient d'indétermination φ et le coefficient d'amélioration A comme suit :

$$\varphi = 1 - R^2 \tag{2.23}$$

$$A = 1 - \sqrt{\varphi} = 1 - \sqrt{1 - R^2} \tag{2.24}$$

2.1.12. Estimation de la variance résiduelle

La variance résiduelle vaut :

$$\hat{\sigma}^2 = \frac{SCE_R}{n-2} = CMR \tag{2.25}$$

Si la variance résiduelle est petite, alors la droite de régression ajuste bien le nuage de points.

2.1.13. Inférence statistique

Pour vérifier la pertinence du modèle, les hypothèses du test global sont :

$$\begin{aligned} &H_0 : \beta_0 = \beta_1 = 0 \\ &H_1 : \beta_1 \neq 0 \end{aligned} \tag{2.26}$$

L'hypothèse nulle stipule que tous les paramètres sont nuls alors que l'hypothèse alternative (ou contraire) dit qu'il existe au moins un paramètre (c'est la pente) qui n'est pas nul.

La statistique de test est celle de Fisher :

$$F^{obs} = \frac{\frac{SCE_F}{1}}{\frac{SCE_R}{n-2}} = \frac{CMF}{CMR} \text{ à (1, n-2) degrés de liberté} \tag{2.27}$$

sachant que :

$$\begin{aligned} &SCE_F \sim \chi^2_1 \\ &SCE_R \sim \chi^2_{n-2} \end{aligned} \tag{2.28}$$

Le tableau 8 ci-après est un tableau de l'analyse de la variance (ANOVA) en régression.

Tableau 8 : Tableau de l'analyse de la variance (ANOVA) en régression

Source de	Somme des	ddl	Carré	F^{obs}	F^{tab}	p-value
Factorielle	SCE_F	1	CMF	$\frac{CMF}{CMR}$	$F(1, n-1)$	$P(F^{obs} \geq F^{tab})$
Résiduelle	SCE_R	$n-2$	CMR	/////////	//////////////////	//////////////////
Totale	SCE_T	$n-1$	CMT	/////////	//////////////////	//////////////////

Les hypothèses du test individuel sur l'intercept sont [5] :

$$H_0 : \beta_0 = 0 \text{ (hypothèse nulle)}$$
$$H_1 : \beta_0 \neq 0 \text{ (hypothèse alternative)}$$

La variance et l'erreur-type de l'intercept sont :

$$Var\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \qquad (2.29)$$
$$S\left(\hat{\beta}_0\right) = \sqrt{Var\left(\hat{\beta}_0\right)}$$

La statistique de test est :

$$t^{obs} = \frac{\hat{\beta}_0}{S\left(\hat{\beta}_0\right)} \text{ à } (n-2) \text{ ddl} \qquad (2.30)$$

Les hypothèses du test individuel sur la pente sont :

$$H_0 : \beta_1 = 0 \text{ (hypothèse nulle)}$$
$$H_1 : \beta_1 \neq 0 \text{ (hypothèse alternative)}$$

La variance et l'erreur-type de la pente sont :

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \qquad (2.31)$$
$$S\left(\hat{\beta}_1\right) = \sqrt{Var\left(\hat{\beta}_1\right)}$$

Sous l'hypothèse nulle, la statistique de test est :

$$t^{obs} = \frac{\hat{\beta}_1}{S\left(\hat{\beta}_1\right)} \text{ à } (n-2) \text{ ddl} \qquad (2.32)$$

Le calcul de la p-value se fait sur base du nombre de degrés de liberté observé dans la table de la loi de Student et du seuil de décision. Lorsque la p-value est inférieure au seuil, alors l'hypothèse nulle

est rejetée (test significatif). Par contre, si cette p-value est supérieure ou égale au seuil de décision, alors l'hypothèse nulle n'est pas rejetée (test non significatif).

Il est aussi possible des construire des intervalles de confiance de niveau $1-\alpha$ autour des paramètres [5] :

$$\begin{aligned} IC_{1-\alpha}\left(\beta_0\right) &= \hat{\beta}_0 \pm t_{n-2;1-\frac{\alpha}{2}} S\left(\hat{\beta}_0\right) \\ IC_{1-\alpha}\left(\beta_1\right) &= \hat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} S\left(\hat{\beta}_1\right) \end{aligned} \tag{2.33}$$

Si l'intervalle de confiance contient la valeur que l'on teste, on décidera de ne pas rejeter l'hypothèse nulle (test non significatif) et on conclura que le paramètre n'est pas significativement différent de 0.

2.1.14. Intervalle de prévision

La valeur prédite (prédiction ponctuelle) d'une nouvelle observation i^* s'écrit :

$$\hat{y}_{i^*} = \hat{y}\left(x_{i^*}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{i^*} \tag{2.34}$$

Cet estimateur est sans biais car :

$$E\left(\hat{y}_{i^*}\right) = E\left[\hat{y}\left(x_{i^*}\right)\right] = E\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i^*}\right) = E\left(\hat{\beta}_0\right) + x_{i^*} E\left(\hat{\beta}_1\right) = \beta_0 + \beta_1 x_{i^*} = y_{i^*} \tag{2.35}$$

L'erreur pour la nouvelle observation i^* est :

$$\begin{aligned} \hat{\varepsilon}_{i^*} &= \hat{y}_{i^*} - y_{i^*} \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_{i^*} - \left(\beta_0 + \beta_1 x_{i^*} + \varepsilon_{i^*}\right) \\ &= \left(\hat{\beta}_1 - \beta_1\right) x_{i^*} + \left(\hat{\beta}_0 - \beta_0\right) - \varepsilon_{i^*} \end{aligned} \tag{2.36}$$

Son espérance mathématique vaut :

$$\begin{aligned} E\left(\hat{\varepsilon}_{i^*}\right) &= E\left[\left(\hat{\beta}_1 - \beta_1\right) x_{i^*} + \left(\hat{\beta}_0 - \beta_0\right) - \varepsilon_{i^*}\right] \\ &= x_{i^*} \underbrace{E\left(\hat{\beta}_1 - \beta_1\right)}_{0} - \underbrace{E\left(\hat{\beta}_0 - \beta_0\right)}_{0} - \underbrace{E\left(\varepsilon_{i^*}\right)}_{0} \\ &= 0 \end{aligned} \tag{2.37}$$

La variance estimée de l'erreur de prévision est :

$$Var\left(\hat{\varepsilon}_{i^*}\right) = E\left(\hat{\varepsilon}_{i^*}^2\right) = \hat{\sigma}_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{\left(x_{i^*} - \overline{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \right) = \hat{\sigma}_\varepsilon^2 \left(1 + h_{i^*}\right) = \hat{\sigma}_{\hat{\varepsilon}_{i^*}}^2 \tag{2.38}$$

où h_{i^*} est le levier de l'observation i^*.

La loi de $\hat{\varepsilon}_{i^*}$ est :

$$\hat{\varepsilon}_{i^*} \sim N\left(0, \hat{\sigma}_\varepsilon \sqrt{1 + h_{i^*}}\right) \tag{3.39}$$

Si la quantité $\left(x_{i^*} - \overline{x}\right)^2$ est petite, alors le point est proche du centre de gravité du nuage et si la quantité $\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ est petite, alors la dispersion des points est grande.

L'intervalle de confiance de niveau $1-\alpha$ de l'espérance y, notée μ_h (ou de la valeur prédite pour une valeur observée x_h) est :

$$IC_{1-\alpha}\left(\mu_h\right) = \hat{y}_h \pm t_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}_\varepsilon^2 \left(\frac{1}{n} + \frac{\left(x_h - \overline{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \right)} \tag{2.40}$$

L'intervalle de prédiction de niveau $1-\alpha$ de la nouvelle observation x_h est :

$$IP_{1-\alpha}\left(y_h\right) = \hat{y}_h \pm t_{n-2;1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{\left(x_h - \overline{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \right)} \tag{2.41}$$

Le modèle linéaire simple peut aussi s'écrire :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{2.42}$$

soit

$$Y = X\beta + \varepsilon$$

avec

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta 1 \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Il vient :

$$X^t Y = X^t X\beta + \underbrace{X^t \varepsilon}_{0}$$
$$\Leftrightarrow X^t Y = X^t X\beta$$
$$\Leftrightarrow \left(X^t X\right)^{-1} X^t Y = \underbrace{\left(X^t X\right)^{-1} X^t X}_{I_2} \beta$$
$$\Leftrightarrow \hat{\beta} = \left(X^t X\right)^{-1} X^t Y$$

2.1.15. Exemple d'application

Un nutritionniste s'intéresse à la liaison pouvant exister entre l'âge x (en mois) et la taille des nourrissons (en cm). Il relève 12 couples de données consignés dans le tableau ci-après.

x	18	19	20	21	22	23	24	25	26	27	28	29
y	76	77	78,1	78,2	78,8	79,7	79,9	81,1	81	81,8	82,8	83,5

La première étape consiste à déterminer les variables indépendante (âge des nourrissons) et dépendante (taille des nourrissons). La deuxième étape consiste à construire le nuage de points des valeurs observées. Le nuage de points (ou diagramme de dispersion) des valeurs est une représentation à deux dimensions de la série statistique double comme le montre la figure 5.

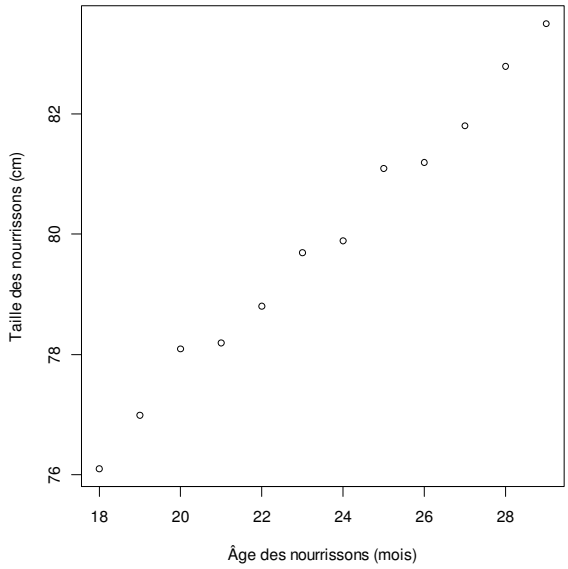


Figure 5 : Nuage de points du rendement en fonction de la quantité d'engrais
Source : Pr BARANKANIRA Emmanuel

Ce nuage de points présente une forme allongée, ce qui montre qu'un ajustement linéaire est envisageable. Avant de faire cet ajustement linéaire, il convient de calculer le coefficient de corrélation linéaire et de tester sa significativité (test de corrélation nulle ou de nullité du coefficient de corrélation linéaire de Bravais-Pearson) comme troisième étape. Ce coefficient traduit ou mesure l'intensité d'un lien linéaire entre ces deux variables quantitatives. Comme troisième étape, il s'agit d'estimer les paramètres, de tester leur significativité, de tester l'hypothèse globale, des vérifier les conditions de validité et de faire des prévisions.

Le coefficient de corrélation linéaire entre l'âge et la taille des nourrissons vaut :

$$r = cor(x, y) = \frac{S_{xy}}{S_x S_x} = \frac{8,25}{3,61(2,30)} = 0,99$$

Aux plus grandes valeurs de l'âge des nourrissons correspondent les plus grandes valeurs de la taille de ces nourrissons. L'association entre ces deux variables est excellente.

Le coefficient de détermination vaut :

$$R^2 = (r)^2 = (0,99)^2 = 0,99$$

Cela signifie que 99 % de la variabilité de la taille des nourrissons sont expliqués par les variations de leur taille.

La statistique du test de corrélation linéaire vaut :

$$t^{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,99\sqrt{10}}{\sqrt{0,99}} = 29,66 \text{ à } 10 \text{ ddl}$$

L'hypothèse nulle qui stipule que le coefficient de corrélation entre x et y est nul est rejetée au seuil de 5 % (t = 29,7 ; ddl = 10, p-value = $4,43\times10^{-11}$). Donc, la corrélation n'est pas nulle.

La matrice de design ou matrice du plan d'expériences est :

$$X = \begin{pmatrix} 1 & 18 \\ 1 & 19 \\ 1 & 20 \\ 1 & 21 \\ 1 & 22 \\ 1 & 23 \\ 1 & 24 \\ 1 & 25 \\ 1 & 26 \\ 1 & 27 \\ 1 & 28 \\ 1 & 29 \end{pmatrix}$$

Sa transposée est :

$$X^t = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 & 26 & 27 & 28 & 29 \end{pmatrix}$$

Il vient :

$$A = X^t X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 & 26 & 27 & 28 & 29 \end{pmatrix} \begin{pmatrix} 1 & 18 \\ 1 & 19 \\ 1 & 20 \\ 1 & 21 \\ 1 & 22 \\ 1 & 23 \\ 1 & 24 \\ 1 & 25 \\ 1 & 26 \\ 1 & 27 \\ 1 & 28 \\ 1 & 29 \end{pmatrix} = \begin{pmatrix} 12 & 282 \\ 282 & 6770 \end{pmatrix}$$

Le déterminant de la matrice A est :

$$\det(A) = \begin{vmatrix} 12 & 282 \\ 282 & 6770 \end{vmatrix} = 81240 - 79524 = 1716 \neq 0$$

$\Rightarrow$ *La matrice A est régulière*

$\Rightarrow A^{-1} \exists$

Les cofacteurs des éléments de A sont :

$$A_{11} = (-1)^2 |6770| = 6770$$
$$A_{12} = (-1)^3 |282| = -282$$
$$A_{21} = (-1)^3 |282| = -282$$
$$A_{22} = (-1)^4 |12| = 12$$

La comatrice (matrice des cofacteurs des éléments de A est :

$$com(A) = \begin{pmatrix} 6770 & -282 \\ -282 & 12 \end{pmatrix}$$

La matrice adjointe de A est :

$$A^{ad} = \left[com(A)\right]^t = \begin{pmatrix} 6770 & -282 \\ -282 & 12 \end{pmatrix}$$

La matrice inverse de A est :

$$A^{-1} = \frac{1}{\det(A)} A^{ad} = \frac{1}{1716}\begin{pmatrix} 6770 & -282 \\ -282 & 12 \end{pmatrix} = \frac{1}{858}\begin{pmatrix} 3385 & -141 \\ -141 & 6 \end{pmatrix}$$

$$X^t Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 & 26 & 27 & 28 & 29 \end{pmatrix} \begin{pmatrix} 76,1 \\ 77,0 \\ 78,1 \\ 78,2 \\ 78,8 \\ 79,7 \\ 79,9 \\ 81,1 \\ 81,2 \\ 81,8 \\ 82,8 \\ 83,5 \end{pmatrix} = \begin{pmatrix} 958,2 \\ 22608,5 \end{pmatrix}$$

Le vecteur des paramètres est :

$$\begin{aligned} \hat{\beta} &= \left(X^t X\right)^{-1} X^t Y \\ &= A^{-1} X^t Y \\ &= \frac{1}{858}\begin{pmatrix} 3385 & -141 \\ -141 & 6 \end{pmatrix}\begin{pmatrix} 958,2 \\ 22608,5 \end{pmatrix} \\ &= \frac{1}{858}\begin{pmatrix} 3243507 - 3187798,5 \\ -135106,2 + 135651 \end{pmatrix} \\ &= \frac{1}{858}\begin{pmatrix} 55708,5 \\ 544,8 \end{pmatrix} \\ &= \begin{pmatrix} 64,93 \\ 0,64 \end{pmatrix} \end{aligned}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 64,93 \\ 0,64 \end{pmatrix} \Rightarrow \begin{cases} \hat{\beta}_0 = 64,93 \\ \hat{\beta}_1 = 0,64 \end{cases}$$

Paramètres	Estimation	Erreur standard	t^{obs}	P-value
Intercept	64,93	0,51	127,71	$< 2\times10^{-16}$
Pente	0,64	0,02	29,66	$4,43\times10^{-11}$

Les éléments de la matrice de variances-covariances sont :

$$Var\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{0,066\left(6770\right)}{12\left(143\right)} = 0,24$$

$$S\left(\hat{\beta}_0\right) = \sqrt{Var\left(\hat{\beta}_0\right)} = \sqrt{0,24} = 0,51$$

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{0,066}{12\left(11\right)\left(143\right)} = 0,0005$$

$$S\left(\hat{\beta}_1\right) = \sqrt{Var\left(\hat{\beta}_1\right)} = \sqrt{0,0005} = 0,02$$

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{0,066\left(23,5\right)}{12\left(11\right)\left(143\right)} \simeq 0,00$$

Le modèle construit est :

$$y = 64,93 + 0,64\, x + \varepsilon$$
$$\Leftrightarrow Taille = 64,93 + 0,64\, \hat{A}ge + \varepsilon$$

Le nuage de points ajusté est représenté par la figure 6.

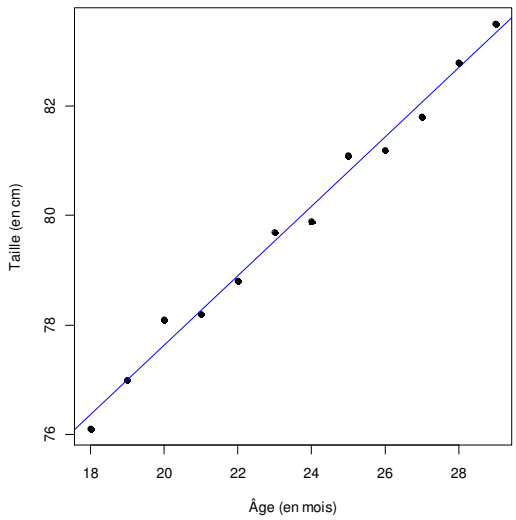


Figure 6 : Nuage de points ajusté
Source : Pr BARANKANIRA Emmanuel

La figure 7 suivante montre que les résidus ne semblent pas suivre une loi normale. Cependant, le test de normalité de Kolmogorov-Smirnov (W = 0,92 ; p-value = 0,315). Les résidus suivent donc une loi normale.

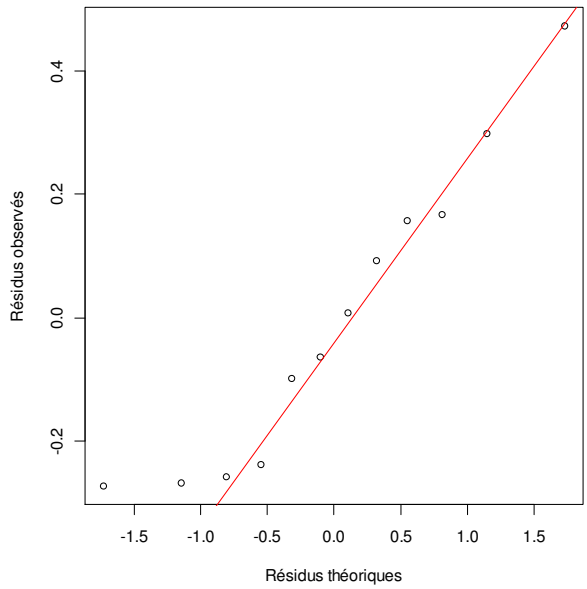


Figure 7 : Normalité des résidus
Source : Pr BARANKANIRA Emmanuel

La figure 8 ci-après montre l'intervalle de confiance (en bleu) et l'intervalle de prédiction (en rouge). Tous les points sont dans leurs intervalles de prédiction.

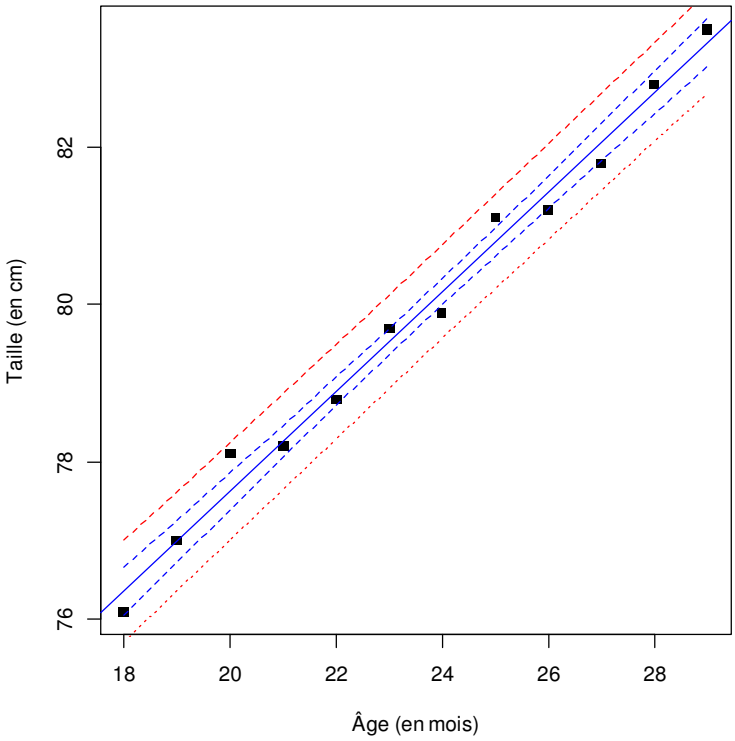


Figure 8 : Intervalle de confiance et intervalle de prédiction
Source : Pr BARANKANIRA Emmanuel

2.2. Régression linéaire multiple

2.2.1. Introduction

Comme souligné au chapitre précédent, le coefficient de corrélation linéaire entre deux variables mesure la liaison linéaire entre ces variables. Pour plus de deux variables, la matrice des corrélations est construite dont il faut extraire la partie triangulaire inférieure ou supérieure. Il s'agit d'une matrice carrée dont l'ordre est égal au nombre de variables en présence (incluant la variable dépendante) et dont les éléments diagonaux sont égaux à 1, les éléments non diagonaux représentant les coefficients de corrélation entre les variables deux à deux.

Les acquis de l'apprentissage sont :

- Face à un problème d'analyse particulier, décider si oui ou non **si la régression linéaire est appropriée** ;
- Traduire les questions de recherche approchées par un **modèle de régression linéaire multiple** en des questions spécifiques sur les coefficients du modèle ;
- Utiliser des modèles de régression linéaire multiple pour **tester des hypothèses** sur les relations entre les variables, y compris les facteurs de confusion, de modification et d'interaction ;
- **Décrire le modèle** de régression linéaire multiple, ses **conditions de validité** et leurs implications ;
- Expliquer pourquoi les estimations sont appelées **estimations des moindres carrés** ;
- Définir l'hyper-droite de régression, les **valeurs ajustées**, les **résidus** et identifier des **points d'influence trop grande** ;
- Construire les **relations entre la corrélation et les coefficients de régression** ;
- **Utiliser un logiciel statistique (logiciel R)** pour **estimer les paramètres** d'un modèle de régression et faire des **graphiques de diagnostic** pour évaluer dans quelle mesure les **conditions de validité** d'un modèle sont remplies ;
- **Interpréter les sorties** du logiciel pour un modèle de régression multiple, y compris les estimations des coefficients de régression, les tests d'hypothèses, les intervalles de confiance et les statistiques qui quantifient l'ajustement du modèle ;
- **Interpréter les coefficients de régression** lorsque la variable à prédire, les variables de prédiction ou les deux sont transformées en log ;
- **Construire** différents modèles de régression multiple, les **comparer** à l'aide de différents critères (R2a, AIC, AICc, BIC, Cp de Mallows) [1].

En régression linéaire multiple, les problèmes qu'il faut résoudre sont les suivants :

- Estimation des coefficients de régression
- Estimation de l'écart-type du terme résiduel
- Analyse des résidus
- Mesure de la force de la liaison entre y et les variables $x_1,\ldots,x_k$
- Significativité de la liaison globale entre y et $x_1,\ldots, x_k$
- Significativité de l'apport marginal de chaque variable à l'explication de y

- Sélection manuelles de bonnes variables explicatives
- Comparaison des modèles
- Construction d'un intervalle de confiance de l'espérance conditionnelle de y
- Construction d'un intervalle de prévision de y

2.2.2. Spécification du modèle

En régression linéaire multiple, la relation entre une variable dépendante Y et une combinaison des facteurs explicatifs $x_1, x_2, \ldots, x_k$ est recherchée. Autrement dit, l'objectif est d'étudier l'association entre la variable réponse et chacune des variables explicatives d'une part et d'expliquer cette variable réponse par une combinaison des variables explicatives afin de faire des prévisions d'autre part. La variable dépendante est aussi appelée variable réponse, variable à expliquer, variable expliquée, ou variable endogène en termes économiques. Les variables explicatives, quant à elles, sont appelées variables indépendantes, prédicteurs, facteurs, régresseurs ou variables exogènes en termes économiques.

Le modèle linéaire multiple à k variables explicatives s'écrit [1] :

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \tag{2.43}$$

où y est la variable réponse, $\beta_1, \beta_2, \ldots, \beta_k$ les paramètres du modèles, $x_1, x_2, \ldots, x_k$ les variables explicatives et ε le terme d'erreur ou la perturbation. Son objectif est d'étudier l'association entre les variables en présence d'une part et d'expliquer la variable réponse par une combinaison des variables explicatives afin de faire des prévisions d'autre part.

Pour chaque observation, le modèle linéaire multiple s'écrit [4] :

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{2.44.a}$$

La forme matricielle de ce modèle est :

$$Y = \beta X + \varepsilon \tag{2.44.b}$$

avec $y=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $X=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$, $\beta=\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$ et $\varepsilon=\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$.

Les paramètres de ce modèle sont estimés par la méthode des Moindres Carrés Ordinaires (MCO). Cette méthode consiste à minimiser la somme des carrés des écarts entre une valeur observé et sa valeur prédite. La matrice X est appelée matrice de design ou matrice du plan d'expériences alors que le vecteur β est appelé hyper-paramètre (vecteur des paramètres).

Le modèle linéaire suppose certains postulats entre autres la normalité des résidus du modèle, la variance constante pour les résidus en fonction de la variable explicative et l'indépendance des observations. La variable réponse suit une loi normale $N(\mu,\sigma)$ où μ est une fonction linéaire des variables explicatives.

L'hypothèse de linéarité qui est souvent faite, s'écrit :

$$E(Y \mid X=x)=\beta_0+\beta_1 X_1+\cdots+\beta_k X_k$$

où $E(Y \mid X=x)$ est la moyenne de la variable Y pour une valeur $X=x$ fixée, β_0 l'intercept (ordonnée à l'origine), $\beta^t=(\beta_1,\beta_2,...,\beta_k)$ le vecteur des coefficients (ou des paramètres).

2.2.3. Estimation et test des paramètres

Les valeurs prédites de y sont :

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_k x_{ik} \tag{2.45}$$

Pour estimer le vecteur des paramètres β, il est plausible de minimiser la fonction de perte. Cette fonction de perte qui est minimisée représente la distance des observations de la fonction de régression ajustée en prenant comme distance la somme des écarts au carré :

$$\begin{aligned} S(\beta)&=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_{i1}-\cdots-\beta_k x_{ik})^2=(Y-X\beta)^t(Y-X\beta) \\ &=(Y-X\beta)^t(Y-X\beta) \end{aligned} \tag{2.46}$$

Pour trouver l'estimateur $\hat{\beta}$ de β, il faut trouver le minimum de la fonction de perte. Ce minimum est trouvé en dérivant la fonction de perte par rapport à chaque paramètre et en annulant chaque dérivée. Autrement dit, il suffit de trouver la solution d'un système d'équations.

Ce système est :

$$\frac{dS(\beta)}{d\beta}=0 \Leftrightarrow \begin{cases} \frac{\partial S}{\partial \beta_0}=0 \\ \vdots \\ \frac{\partial S}{\partial \beta_k}=0 \end{cases} \tag{2.47}$$

Le calcul des dérivées partielles donne le résultat suivant :

$$\frac{dS(\beta)}{d\beta_j}=-2\sum_{i=1}^{n} X_{ij}\left(Y_i-\beta_0-\beta_1 X_{i1}-\beta_2 X_{i2}-\cdots-\beta_k X_{ik}\right) \tag{2.48}$$

En forme matricielle, nous obtenons donc pour la dérivée :

$$\frac{dS(\beta)}{d\beta}=-2X^tY+2X^tX\beta \tag{2.49}$$

L'annulation de cette dérivée au point $\beta=\hat{\beta}$ donne :

$$\begin{aligned}
&\left.\frac{dS(\beta)}{d\beta}\right|_{\beta=\hat{\beta}}=0 \\
&\Leftrightarrow -2X^tY+2X^tX\beta=0 \\
&\Leftrightarrow -2\left(X^tY-X^tX\beta\right)=0 \\
&\Leftrightarrow X^tY-X^tX\beta=0 \\
&\Leftrightarrow X^tX\beta=X^tY \\
&\Leftrightarrow \underbrace{\left(X^tX\right)^{-1}X^tX}_{I_{k+1}}\beta=\left(X^tX\right)^{-1}X^tY \\
&\Leftrightarrow \hat{\beta}=\left(X^tX\right)^{-1}X^tY
\end{aligned}$$

2.2.4. Application du modèle linéaire multiple

Considérons la matrice des données suivantes (Tableau 9). Les variables indépendantes sont le marché total (MT), les remises aux grossistes (RG), les prix (PR), le budget de recherche (BR), l'investissement (INV), la publicité (PUB), les frais de ventes (FV), le total du budget publicité de la branche (TPB) et la variable dépendante représente les ventes semestrielles (VE). Créons une variable TRIM qui représente le numéro du trimestre avant l'importation des données dans R.

Tableau 9 : Matrice des données

ID	MT	RG	PR	BR	INV	PUB	FV	TPB	VE
1	398	138	56	12	50	77	229	98	5540
2	369	118	59	9	17	89	177	225	5439
3	268	129	57	29	89	51	166	263	4290
4	484	111	58	13	107	40	258	321	5502
5	394	146	59	13	143	52	209	407	4872
6	332	140	60	11	61	21	180	247	4708
7	336	136	60	25	-30	40	213	328	4627
8	383	104	60	21	-45	32	201	298	4110
9	285	105	63	8	-28	12	176	218	4123
10	277	135	62	11	76	68	175	410	4842
11	456	128	65	22	144	52	253	93	5741
12	355	131	65	24	113	77	208	307	5094
13	364	120	64	14	128	96	195	107	5383
14	320	147	66	15	10	48	154	305	4888
15	311	143	67	22	-25	27	181	60	4033
16	362	145	67	23	117	73	220	239	4942
17	408	131	66	13	120	62	235	141	5313
18	433	124	68	8	122	25	258	291	5140
19	359	106	69	27	71	74	196	414	5397
20	476	138	71	18	4	63	279	206	5149
21	415	148	69	8	47	29	207	80	5151
22	420	136	70	10	8	91	213	429	4989
23	536	111	73	27	128	74	296	273	5927
24	432	152	73	16	-50	16	245	309	4704
25	436	123	73	32	100	43	276	280	5366
26	415	119	75	20	-40	41	211	315	4630
27	462	112	73	15	68	93	283	212	5712
28	429	125	74	11	88	83	218	118	5095
29	517	142	74	27	27	75	307	345	6124
30	328	123	77	20	59	88	211	141	4787
31	418	135	79	35	142	74	270	83	5036
32	515	120	77	23	126	21	328	398	5288
33	412	149	78	36	30	26	258	124	4647
34	455	126	78	22	18	95	233	118	5316
35	554	138	81	20	42	93	324	161	6180
36	441	120	80	16	-22	50	267	405	4801
37	417	120	81	35	148	83	257	111	5512
38	461	132	82	27	-18	91	267	170	5272

Les données Excel sont d'abord enregistrées sous format texte puis importées dans R avec la commande read.table() en mettant dans les parenthèses le nom de la base de données, après avoir changé le répertoire de travail (avec la commande **setwd()**). L'option **header=TRUE** signifie que la première ligne sera considérée comme nom des colonnes et l'option sep= '\t' signifie que le séparateur des enregistrements est la tabulation. La commande **head()** permet d'afficher les 6 premières lignes de la base de données et la commande **tail()**les 6 dernières lignes.

```
> mydata <- read.table("MODLIN2024.txt", header=TRUE, sep="\t")
> head(mydata)  # 6 premières lignes
  TRIM  MT  RG PR BR INV PUB  FV TPB   VE
1    1 398 138 56 12  50  77 229  98 5540
2    2 369 118 59  9  17  89 177 225 5439
3    3 268 129 57 29  89  51 166 263 4290
4    4 484 111 58 13 107  40 258 321 5502
5    1 394 146 59 13 143  52 209 407 4872
6    2 332 140 60 11  61  21 180 247 4708
```

```
> tail(mydata)  # 6 dernières lignes
   TRIM  MT  RG PR BR INV PUB  FV TPB   VE
33    1 412 149 78 36  30  26 258 124 4647
34    2 455 126 78 22  18  95 233 118 5316
35    3 554 138 81 20  42  93 324 161 6180
36    4 441 120 80 16 -22  50 267 405 4801
37    1 417 120 81 35 148  83 257 111 5512
38    2 461 132 82 27 -18  91 267 170 5272
```

Il est possible de créer une fonction qui calcule automatiquement les statistiques descriptives voulues pour toutes les variables quantitatives et qui stocke ces résultats dans une matrice. Il est possible de créer une fonction qui calcule automatiquement les statistiques descriptives voulues pour toutes les variables quantitatives et qui stocke ces résultats dans une matrice.

```
> Tableau
    Effectif Minimum Moyenne Déviation standard Maximum
MT        38     268  406.13              70.44     554
RG        38     104  129.11              13.16     152
PR        38      56   69.18               7.65      82
BR        38       8   19.42               8.08      36
INV       38     -50   56.45              62.48     148
PUB       38      12   59.08              26.15      96
FV        38     154  232.47              44.65     328
TPB       38      60  238.16             111.97     429
VE        38    4033 5096.58             514.87    6180
```

Il en est de même pour une fonction qui calcule les coefficients de corrélations de Bravais-Pearson entre les variables deux à deux sous forme d'une matrice triangulaire inférieure ou supérieure, qui

calcule la matrice des p-values et la matrice des nuages de points ajustés par la méthode des moindres carrés ordinaires.

```
> m
       MT     RG     PR     BR    INV   PUB     FV   TPB VE
MT      1
RG  -0.07      1
PR   0.55   0.02      1
BR   0.16   0.01   0.46      1
INV  0.14  -0.09  -0.06   0.16      1
PUB   0.2  -0.12   0.26    0.1   0.24     1
FV    0.9  -0.05   0.63   0.36   0.22  0.13      1
TPB -0.02  -0.15  -0.18  -0.13  -0.12  -0.2 -0.02     1
VE   0.72  -0.08   0.29   0.08   0.45  0.57   0.64 -0.1   1
```

Une corrélation très forte est observée entre les frais de ventes (FV) et le marché total (MT).

Variables	MT	RG	PR	BR	INV	PUB	FV	TPB	VE
MT	1,00								
RG	-0,07	1,00							
PR	0,55	0,02	1,00						
BR	0,16	0,01	0,46	1,00					
INV	0,14	-0,09	-0,06	0,16	1,00				
PUB	0,20	-0,12	0,26	0,10	0,24	1,00			
FV	**0,90**	-0,05	0,63	0,36	0,22	0,13	1,00		
TPB	-0,02	-0,15	-0,18	-0,13	-0,12	-0,20	-0,02	1,00	
VE	0,72	-0,08	0,29	0,08	0,45	0,57	0,64	-0,10	1,00

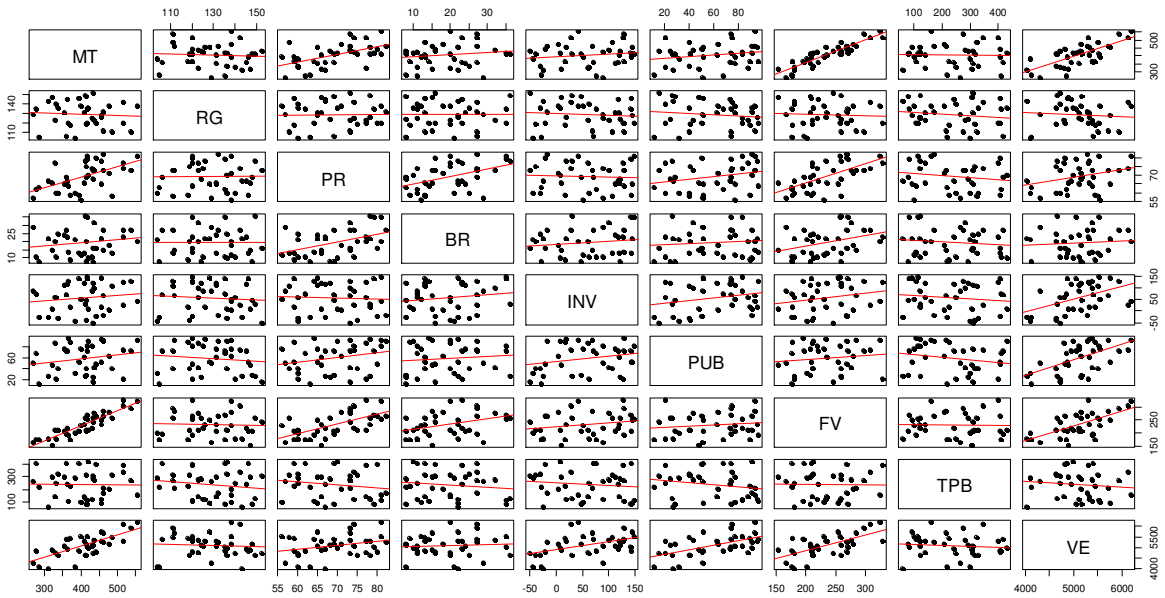
Matrice des corrélations :

```
> m
     MT     RG    PR    BR    INV    PUB     FV    TPB     VE
MT    1 -0.07 0.55 0.16  0.14   0.2    0.9 -0.02  0.72
RG        1 0.02 0.01 -0.09 -0.12 -0.05 -0.15 -0.08
PR          1 0.46 -0.06  0.26  0.63 -0.18  0.29
BR            1  0.16   0.1  0.36 -0.13  0.08
INV                1  0.24  0.22 -0.12  0.45
PUB                   1  0.13  -0.2  0.57
FV                          1 -0.02  0.64
TPB                           1  -0.1
VE                                  1
```

Matrice des p-values :

```
> corr(r)
     MT     RG     PR     BR    INV    PUB     FV    TPB     VE
MT    0 0.679      0 0.327 0.388 0.229      0 0.906      0
RG          0 0.895 0.953 0.577 0.475 0.767 0.383 0.618
PR                0 0.004 0.731 0.122      0 0.276  0.08
BR                      0 0.346 0.531 0.025 0.445 0.614
INV                           0 0.146 0.193 0.462 0.004
PUB                                 0 0.424  0.24      0
FV                                        0 0.897      0
TPB                                             0 0.567
VE                                                    0
```

Variables	MT	RG	PR	BR	INV	PUB	FV	TPB	VE
MT									
RG	0,68								
PR	<0,001	0,90							
BR	0,33	0,95	<0,001						
INV	0,39	0,58	0,73	0,35					
PUB	0,23	0,47	0,12	0,53	0,15				
FV	<0,001	0,77	<0,001	0,02	0,19	0,42			
TPB	0,91	0,38	0,28	0,44	0,46	0,24	0,90		
VE	<0,001	0,62	0,08	0,61	<0,001	<0,001	<0,001	0,57	



Ces coefficients de corrélation et les p-values sont cartographiés à travers les figures 9 et 10.

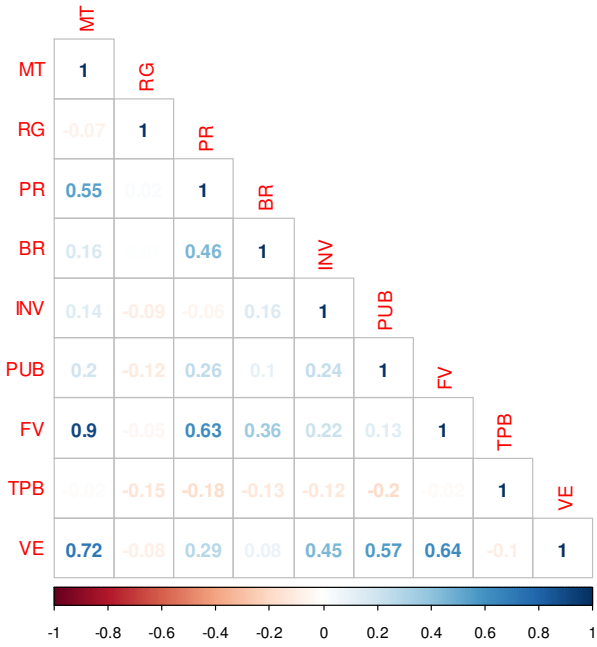


Figure 9 : Cartographie des coefficients de corrélation

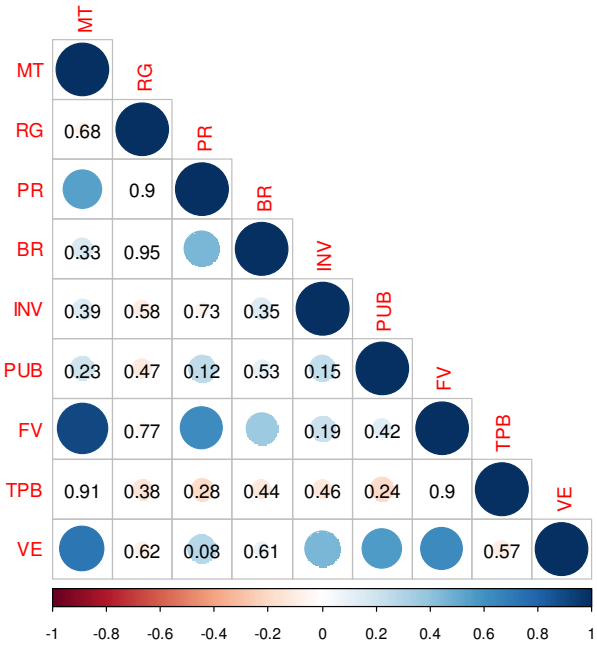


Figure 10 : Cartographie des p-values

```
> mod.vide <- lm(VE ~ 1)
> summary(mod.vide)

Call:
lm(formula = VE ~ 1)

Residuals:
     Min       1Q   Median       3Q      Max
-1063.58  -306.08    20.92   296.92  1083.42

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5096.58      83.52   61.02   <2e-16 ***
---
```

```
> mod.MT <- lm(VE ~ MT)
> summary(mod.MT)

Call:
lm(formula = VE ~ MT)

Residuals:
    Min      1Q  Median      3Q     Max
-864.71 -175.65   -1.42  294.99  548.73

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2956.8907   347.9074   8.499 3.96e-10 ***
MT             5.2685     0.8444   6.240 3.33e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 361.8 on 36 degrees of freedom
Multiple R-squared:  0.5196,    Adjusted R-squared:  0.5062
F-statistic: 38.93 on 1 and 36 DF,  p-value: 3.329e-07

> mod.RG <- lm(VE ~ RG)
> summary(mod.RG)

Call:
lm(formula = VE ~ RG)

Residuals:
     Min       1Q   Median       3Q      Max
-1068.61  -323.44    15.18   254.87  1112.48

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5518.431    843.022   6.546  1.3e-07 ***
RG            -3.268      6.497  -0.503    0.618
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 520.1 on 36 degrees of freedom
Multiple R-squared:  0.006977,  Adjusted R-squared:  -0.02061
F-statistic: 0.2529 on 1 and 36 DF,  p-value: 0.6181
```

```
> mod.PR <- lm(VE ~ PR)
> summary(mod.PR)

Call:
lm(formula = VE ~ PR)

Residuals:
     Min       1Q   Median       3Q      Max
-1021.38  -281.65    -5.23   297.47   934.37

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3759.85     747.35   5.031 1.37e-05 ***
PR             19.32      10.74   1.799   0.0804 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 500 on 36 degrees of freedom
Multiple R-squared:  0.08251,   Adjusted R-squared:  0.05702
F-statistic: 3.237 on 1 and 36 DF,  p-value: 0.08037

> mod.BR <- lm(VE ~ BR)
> summary(mod.BR)

Call:
lm(formula = VE ~ BR)

Residuals:
     Min       1Q   Median       3Q      Max
-1077.45  -303.81    51.89   301.60  1080.31

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4992.105    222.176  22.469   <2e-16 ***
BR             5.379     10.583   0.508    0.614
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 520.1 on 36 degrees of freedom
Multiple R-squared:  0.007126,  Adjusted R-squared:  -0.02045
F-statistic: 0.2584 on 1 and 36 DF,  p-value: 0.6143
```

```
> mod.INV <- lm(VE ~ INV)
> summary(mod.INV)

Call:
lm(formula = VE ~ INV)

Residuals:
    Min      1Q  Median      3Q     Max
-928.11 -325.46  -11.78  247.69 1137.36

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4885.848    102.343  47.740  < 2e-16 ***
INV            3.733      1.224   3.049  0.00429 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 465.3 on 36 degrees of freedom
Multiple R-squared:  0.2053,    Adjusted R-squared:  0.1832
F-statistic: 9.298 on 1 and 36 DF,  p-value: 0.004285

> mod.PUB <- lm(VE ~ PUB)
> summary(mod.PUB)

Call:
lm(formula = VE ~ PUB)

Residuals:
    Min      1Q  Median      3Q     Max
-716.31 -262.53  -82.38  241.48  849.52

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4436.449    174.151  25.475  < 2e-16 ***
PUB           11.174      2.701   4.136 0.000202 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 429.7 on 36 degrees of freedom
Multiple R-squared:  0.3222,    Adjusted R-squared:  0.3033
F-statistic: 17.11 on 1 and 36 DF,  p-value: 0.0002021
```

```
> mod.FV <- lm(VE ~ FV)
> summary(mod.FV)

Call:
lm(formula = VE ~ FV)

Residuals:
    Min      1Q  Median      3Q     Max
-755.29 -315.75   16.28  243.49  750.09

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3388.178    350.404   9.669 1.51e-11 ***
FV             7.349      1.481   4.962 1.69e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 402.2 on 36 degrees of freedom
Multiple R-squared:  0.4062,    Adjusted R-squared:  0.3897
F-statistic: 24.62 on 1 and 36 DF,  p-value: 1.686e-05

> mod.TPB <- lm(VE ~ TPB)
> summary(mod.TPB)

Call:
lm(formula = VE ~ TPB)

Residuals:
     Min       1Q   Median       3Q      Max
-1142.18  -319.83     6.22   324.43  1074.56

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5201.6547   200.2750  25.973   <2e-16 ***
TPB           -0.4412     0.7628  -0.578    0.567
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 519.6 on 36 degrees of freedom
Multiple R-squared:  0.009206,  Adjusted R-squared:  -0.01832
F-statistic: 0.3345 on 1 and 36 DF,  p-value: 0.5666

> mod.TRIM <- aov(VE ~ TRIM)
> summary(mod.TRIM)
             Df  Sum Sq Mean Sq F value Pr(>F)
TRIM          3  488561  162854   0.594  0.623
Residuals    34 9319643  274107
```

Les variables significatives au seuil de 20 % sont mises dans le modèle complet. Il est aussi possible de mettre toutes les variables dans le modèle et procéder à l'élimination des variables mais cette pratique est peu courante et déconseillée, à moins qu'il s'agisse d'une analyse de la variance.

```
> mod.complet <- lm(VE ~ MT+RG+PR+BR+INV+PUB+FV+TPB)
> summary(mod.complet)

Call:
lm(formula = VE ~ MT + RG + PR + BR + INV + PUB + FV + TPB)

Residuals:
    Min      1Q  Median      3Q     Max
-485.95 -115.83   -6.82  161.86  473.84

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3129.23102  641.35537   4.879 3.55e-05 ***
MT             4.42327    1.58822   2.785  0.00933 **
RG             1.67640    3.29135   0.509  0.61437
PR           -13.52623    8.30482  -1.629  0.11419
BR            -3.40966    6.56941  -0.519  0.60769
INV            1.92432    0.77775   2.474  0.01945 *
PUB            8.54684    1.82649   4.679 6.18e-05 ***
FV             1.49724    2.77061   0.540  0.59305
TPB           -0.02152    0.40059  -0.054  0.95753
---
```

Le modèle complet est meilleur que le modèle vide :

```
> BIC(mod.vide, mod.complet)
             df      BIC
mod.vide      2 588.6380
mod.complet  11 552.2961
```

La sélection des modèles conduit au modèle saturé suivant :

```
> mod.sans.FV <- lm(VE ~ MT + PR + INV + PUB)
> summary(mod.sans.FV)

Call:
lm(formula = VE ~ MT + PR + INV + PUB)
```

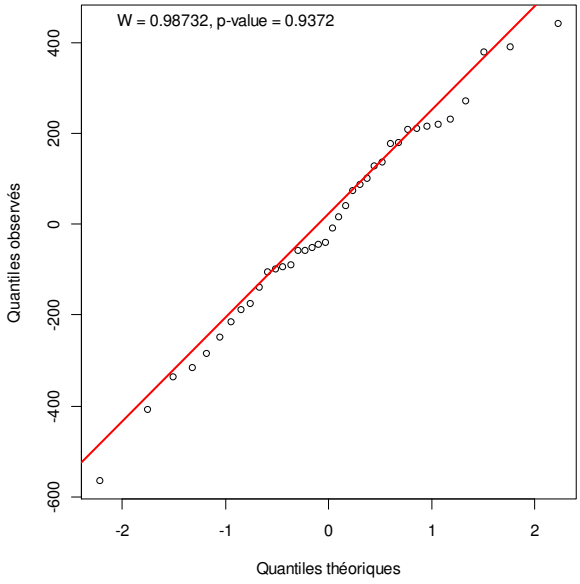
```
Residuals:
    Min      1Q  Median      3Q     Max
-565.01 -129.87  -23.71  179.10  442.16

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3302.0355   369.2121   8.943 2.45e-10 ***
MT             5.1919     0.6946   7.475 1.36e-08 ***
PR           -13.1718     6.4968  -2.027  0.05076 .
INV            1.9676     0.6778   2.903  0.00654 **
PUB            8.2294     1.6389   5.021 1.73e-05 ***
---
```

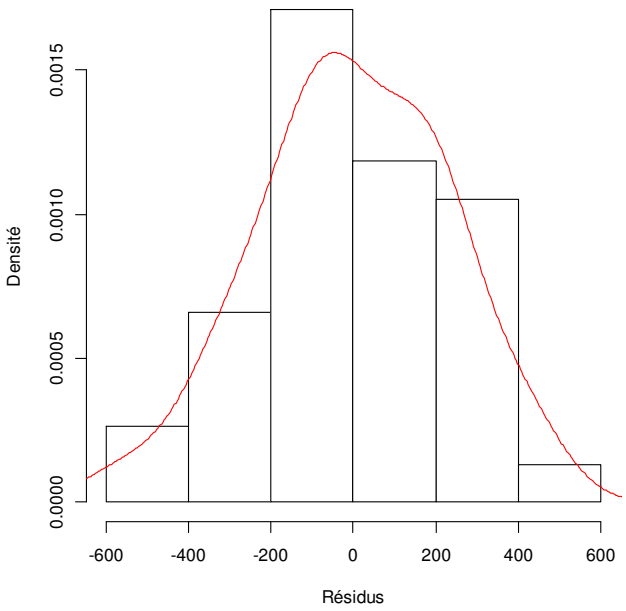
Ce modèle est meilleur que les modèles vide et complet :

```
> BIC(mod.sans.FV, mod.vide)
             df      BIC
mod.sans.FV   6 541.7969
mod.vide      2 588.6380
> BIC(mod.sans.FV, mod.complet)
             df      BIC
mod.sans.FV   6 541.7969
mod.complet  11 552.2961
```

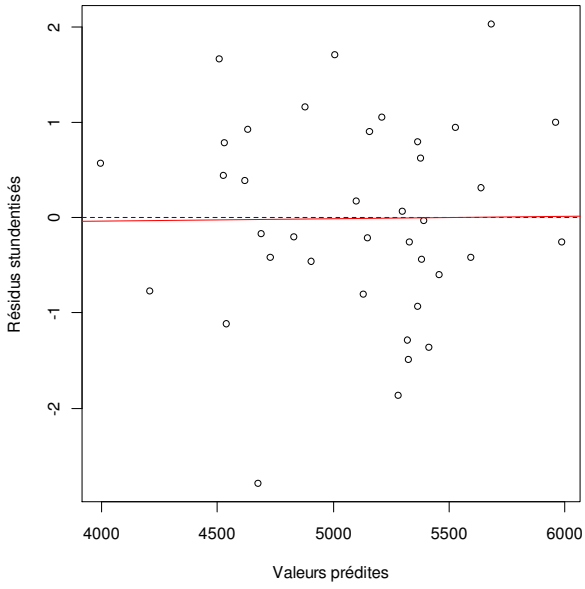
Les résidus suivent une loi normale :

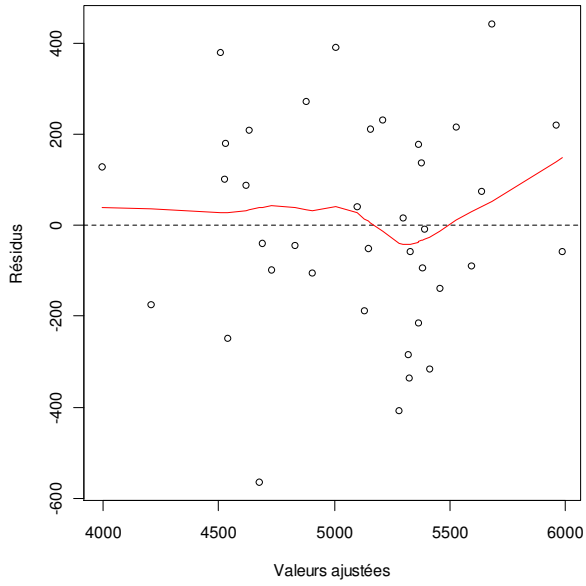
```
> ad.test(mod.sans.FV$residuals)

        Anderson-Darling normality test

data:  mod.sans.FV$residuals
A = 0.17546, p-value = 0.9176

> cvm.test(mod.sans.FV$residuals)

        Cramer-von Mises normality test

data:  mod.sans.FV$residuals
W = 0.027335, p-value = 0.8766

> lillie.test(mod.sans.FV$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  mod.sans.FV$residuals
D = 0.06938, p-value = 0.9158

> shapiro.test(mod.sans.FV$residuals)

        Shapiro-Wilk normality test

data:  mod.sans.FV$residuals
W = 0.98732, p-value = 0.9372
```





Source : Pr BARANKANIRA Emmanuel





Source : Pr BARANKANIRA Emmanuel

Chapitre 3 : Modèle logistique

3.1. Test du chi-deux

Le test de Student permet de comparer deux moyennes ou deux proportions. La comparaison de deux proportions peut aussi se faire à l'aide du test du chi-deux (chi-carré) d'homogénéité de Pearson. Il peut s'agir du test du chi-deux d'indépendance de Pearson ou du test du chi-deux d'homogénéité de Pearson. Le test du chi-deux d'indépendance de Pearson, quant à lui, est utilisé non pas pour comparer deux proportions mais plutôt pour étudier la liaison (relation) entre deux variables qualitatives afin de s'assurer que ces variables sont indépendantes ou pas.

3.1.1. Hypothèses de test

Les hypothèses du test du chi-deux d'homogénéité de Pearson sont :

$H_0 : \pi_1 = \pi_2$ (Les proportions sont égales)

$H_1 : \pi_1 \neq \pi_2$ (Les proportions sont inégales)

Les hypothèses du test du chi-deux d'indépendance de Pearson sont :

$H_0 : x \text{ et } y \text{ sont indépendantes}$
$H_1 : x \text{ et } y \text{ sont liées}$

3.1.2. Loi du chi-deux

Soit X une variable aléatoire qui suit une loi du chi-deux à k degrés de liberté. La densité de probabilité de la variable X est donnée par [4] :

$$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \text{ avec k>0} \tag{3.1}$$

avec

$$\Gamma(k) = \int_0^{+\infty} e^{-t} t^{k-1} dt$$

L'espérance mathématique de X vaut :

$$E(X) = \int_0^{+\infty} x f(x) dx$$
$$= \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \int_0^{+\infty} x^{\frac{k}{2}} e^{-\frac{x}{2}} dx$$

Par changement de variable, posons :

$$y = \frac{x}{2} \Rightarrow dy = \frac{dx}{2} \Rightarrow dx = 2dy$$

Il vient :

$$\begin{aligned} E(X) &= \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \int_0^{+\infty} x^{\frac{k}{2}} e^{-\frac{x}{2}} dx \\ &= \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \int_0^{+\infty} 2^{\frac{k}{2}+1} y^{\frac{k}{2}} e^{-y} dy \\ &= \frac{2}{\Gamma\left(\frac{k}{2}\right)} \int_0^{+\infty} y^{\frac{k}{2}} e^{-y} dy \\ &= \frac{2\Gamma\left(\frac{k}{2}+1\right)}{\Gamma\left(\frac{k}{2}\right)} \\ &= \frac{2\frac{k}{2}!}{\left(\frac{k}{2}-1\right)!} \\ &= \frac{2\frac{k}{2}\left(\frac{k}{2}-1\right)!}{\left(\frac{k}{2}-1\right)!} \\ &= k \end{aligned} \tag{3.2}$$

Sachant que :

$$Var(X) = E(X^2) - E^2(X)$$

Il vient donc que :

$$E(X^2) = \int_0^{+\infty} x^2 f(x) dx$$
$$= \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \int_0^{+\infty} x^{\frac{k}{2}+1} e^{-\frac{x}{2}} dx$$

et

$$E(X^2) = \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \int_0^{+\infty} x^{\frac{k}{2}+1} e^{-\frac{x}{2}} dx$$
$$= \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \int_0^{+\infty} 2^{\frac{k}{2}+2} y^{\frac{k}{2}+1} e^{-y} dy$$
$$= \frac{4}{\Gamma\left(\frac{k}{2}\right)} \int_0^{+\infty} y^{\frac{k}{2}+1} e^{-y} dy$$
$$= \frac{4\Gamma\left(\frac{k}{2}+2\right)}{\Gamma\left(\frac{k}{2}\right)}$$
$$= \frac{4\left(\frac{k}{2}+1\right)\left(\frac{k}{2}\right)\left(\frac{k}{2}-1\right)!}{\left(\frac{k}{2}-1\right)!}$$
$$= 4\left(\frac{k}{2}+1\right)\left(\frac{k}{2}\right)$$
$$= (2k+4)\left(\frac{k}{2}\right)$$
$$= k^2 + 2k$$

La variance de X vaut donc :

$$Var\left(X\right) = k^2 + 2k - k^2 = 2k \tag{3.3}$$

3.1.3. Statistique de test

La distance du chi-deux est notée par :

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}} \tag{3.4}$$

où $n_{i\bullet}$ sont les effectifs marginaux des lignes, $n_{\bullet j}$ les effectifs marginaux des colonnes et $n_{\bullet\bullet} = n$ l'effectif total.

En posant O_{ij} les effectifs observés et E_{ij} les effectifs attendus, il vient que :

$$\begin{cases} O_{ij} = n_{ij} \\ E_{ij} = \dfrac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}} \end{cases} \tag{3.5}$$

Pour le test du chi-deux d'homogénéité de Pearson (qui va nous intéresser ici), la distance entre les effectifs observés O_{ij} et les effectifs attendus E_{ij} se calcule à l'aide de la statistique :

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \text{ à } \textit{(I-1)}\times\textit{(J-1)} \text{ degrés de liberté} \tag{3.6}$$

où $\underline{I}$ est le nombre de modalités de la première variable qualitative et J le nombre de modalités de la deuxième variable qualitative.

Les effectifs attendus E_{ij} se calculent aussi comme suit :

$$E_{ij} = \frac{L_i C_j}{n} \tag{3.7}$$

avec L_i les effectifs marginaux des lignes et C_j les effectifs marginaux des colonnes.

Une fois que la valeur du chi-deux calculée est connue, il reste à la comparer à la valeur du chi-deux issue de la distribution du chi-carré (valeur tabulée). L'interprétation se fera à l'aide de la p-value qui représente la probabilité qu'observer une statistique du chi-deux au moins aussi grande que celle que l'on aurait observée sous l'hypothèse nulle.

Une p-value supérieure à 5 % sera interprétée comme une absence de relation (ou une indépendance) entre ces deux variables. Autrement dit, on décidera de rejeter l'hypothèse nulle (test significatif) si la p-value est inférieure au seuil de décision et de ne pas rejeter l'hypothèse nulle (test non significatif) si la p-value est supérieure ou égale au seuil de décision.

La statistique du chi-deux peut aussi s'écrire :

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}}$$

$$= n_{\bullet\bullet}\sum_{i=1}^{I}\sum_{i=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{n_{i\bullet}n_{\bullet j}}$$

$$= n_{\bullet\bullet}\sum_{i=1}^{I}\sum_{i=1}^{J} \frac{\left[n_{ij}^2 - 2\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}n_{ij} + \left(\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}\right)^2\right]}{n_{i\bullet}n_{\bullet j}}$$

$$= n_{\bullet\bullet}\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} + \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}$$

$$= n_{\bullet\bullet}\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - 2\sum_{i=1}^{I}\sum_{j=1}^{J} O_{ij} + \sum_{i=1}^{I}\sum_{j=1}^{J} E_{ij}$$

$$= n\sum_{i=1}^{I}\sum_{i=1}^{J} \frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - 2n_{\bullet\bullet} + n_{\bullet\bullet}$$

$$= n\sum_{i=1}^{I}\sum_{i=1}^{J} \frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - n$$

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

Une troisième façon de calculer la statistique du chi-deux est :

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}}$$

$$= \sum_{i=1}^{I}\sum_{i=1}^{J} \frac{\left(\frac{n_{ij}}{n_{\bullet\bullet}} - \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}^2}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}^2}}$$

$$= \sum_{i=1}^{I}\sum_{i=1}^{J} \frac{\left(\frac{n_{ij}}{n_{\bullet\bullet}} - \frac{n_{i\bullet}}{n_{\bullet\bullet}}\frac{n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\frac{n_{i\bullet}}{n_{\bullet\bullet}}\frac{n_{\bullet j}}{n_{\bullet\bullet}}}$$

$$= \sum_{i=1}^{I}\sum_{i=1}^{J} \frac{\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)^2}{p_{i\bullet}p_{\bullet j}}$$

3.2. Coefficient V de Cramer

Le test d'indépendance du chi-deux permettra de déterminer la liaison entre deux variables mais sans préciser le degré de liaison. Pour déterminer l'intensité de l'association entre la variable dépendante et chaque variable explicative, le coefficient V de Cramer est calculé.

Le V de Cramer est la racine carrée de la statistique du χ^2 divisé par le χ^2 maximal. Il mesure l'intensité d'une association entre deux variables. Sa forme mathématique est :

$$V = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} = \sqrt{\frac{\chi^2}{n \times \min\left(I-1, J-1\right)}} \tag{3.8}$$

où I et J sont respectivement le nombre de lignes et de colonnes de la table de contingence.

En principe, il prend une valeur comprise entre 0 et 1. Plus la valeur de V est proche de 0, plus il y a indépendance entre les deux variables. Lorsqu'il vaut 1, on parle de complète dépendance. Le tableau 10 montre la qualification de la relation entre deux variables qualitatives selon la valeur du V de Cramer.

Tableau 10 : Qualification selon la valeur du V de Cramer

Valeur du V de Cramer	Intensité de la relation entre les variables
Inférieure à 0,10	Relation nulle ou très faible
[0,10 ; 0,20[	Relation faible
[0,20 ; 0,30[	Relation moyenne
Supérieure ou égale à 0,30	Relation forte

3.3. Exemple d'application du test du chi-deux et du coefficient V de Cramer

Considérons les données d'une enquête qui a été menée dans la ville de Gitega sur le tabagisme.

```
> Tab <- table(Sexe, Tabac)
> dimnames(Tab) <- list(c("Masculin","Féminin"),
+ c("Non fumeur","Fumeur"))
> Tab
          Non fumeur Fumeur
Masculin         114     33
Féminin          103     20
```

Les effectifs attendus sont :

```
> chisq.test(Tab)$expected
          Non fumeur   Fumeur
Masculin  118.14444 28.85556
Féminin    98.85556 24.14444
```

```
        Pearson's Chi-squared test

data:  Tab
X-squared = 1.6258, df = 1, p-value = 0.2023
```

La statistique du chi-deux observée vaut :

$\chi^2_{obs} = 1{,}63$ à 1 ddl

La p-value vaut 0,202 < 0,05. L'hypothèse nulle d'indépendance est rejetée, ce qui montre que le tabagisme et le sexe sont indépendants.

Le coefficient V de Cramer vaut :

$V = -0{,}078$

La relation entre ces deux variables est nulle ou très faible.

3.4. Régression logistique

3.4.1. Introduction

Un modèle de régression linéaire est un modèle de régression d'une variable expliquée (ou variable dépendante) sur une ou plusieurs variables explicatives en faisant l'hypothèse que la fonction qui lie les variables explicatives à la variable expliquée est linéaire dans ses paramètres.

Un modèle de régression linéaire est aussi appelé tout simplement modèle linéaire. Ce modèle a comme fonction de lien définie entre la variable réponse et les variables explicatives la fonction identité. Dans le cas d'une relation fonctionnelle différente de l'identité, le modèle linéaire devient un modèle linéaire généralisé. C'est notamment le cas du modèle logistique.

Dans le cas où la variable réponse est de type qualitatif, le modèle utilisée est souvent la régression logistique. Dans ce cas, la variable réponse peut avoir deux ou plusieurs modalités. Dans le cas où elle prend deux modalités, la variable réponse suit une loi de Bernoulli et le modèle utilisé est la régression logistique binaire. Dans le cas où elle prend plus deux modalités, il s'agira de la régression logistique multinomiale ou polytomique. Le modèle logistique polytomique se scinde en deux : le modèle logistique ordinal pour lequel les modalités de la variable réponse sont ordonnées et modèle logistique nominal pour lequel les modalités de la variables réponse ne sont pas ordonnées.

Les variables explicatives peuvent par contre être des variables qualitatives, des variables quantitatives ou une combinaison des deux. La régression logistique est largement utilisée dans de nombreux domaines de la vie du pays. À titre d'exemple, elle est utilisée en médecine afin de permettre de trouver les facteurs qui caractérisent un groupe de sujets malades par rapport aux sujets sains (non malades). Elle est donc couramment utilisée en épidémiologie pour identifier les facteurs associés à un problème de santé ou à une pathologie.

La régression logistique est utilisée dans beaucoup de domaines :

- En médecine, elle permet par exemple de trouver les facteurs qui caractérisent un groupe de sujets malades par rapport aux sujets sains.
- Dans le domaine des assurances, elle permet de cibler une fraction de la clientèle qui sera sensible à une police des assurances sur tel ou tel risque particulier.

- Dans le domaine bancaire, elle permet de détecter les groupes à risque lors de la souscription d'un crédit.
- En économétrie, elle permet d'expliquer une variable discrète : par exemple les intentions de vote aux élections.

La régression logistique polytomique est donc une généralisation de la régression logistique binaire dans la mesure où la variable réponse possède plus de deux modalités. Il est construit une équation qui cherche à expliquer, simultanément, la probabilité de choisir chaque choix. Cela équivaut, en fait, à estimer plusieurs modèles de régression logistique binaire, une pour chaque combinaison de deux choix. Une constante, des coefficients de régression, des statistiques Wald, et un coefficient de détermination sont produits.

Comme pour la régression logistique binaire, il est possible d'utiliser la régression logistique simple ou multiple. Dans la régression logistique simple, la variable expliquée est modélisé à l'aide d'une seule variable explicative contrairement à la régression logistique multiple qui établit une relation entre une variable réponse qualitative et plusieurs variables explicatives. En régression linéaire, les valeurs prédites peuvent dépasser 1 et être plus petites que 0 selon que les valeurs de la variable explicative augmentent, ce qui n'est pas le cas en régression logistique binaire. Dans ce modèle, les résidus ne doivent pas être de loi normale puisque la variable réponse y est binaire et prend les valeurs 0 et 1. De plus, la constance de la variance des résidus n'est pas vérifiée dans la mesure où la variance de la variable réponse vaut $p(1-p)$ où $p = P(y = 1)$ est la probabilité de succès.

Dans le modèle linéaire, la relation existant entre la variable réponse et les variables explicatives est supposée linéaire et la fonction qui les lie est une fonction identité, ce qui n'est pas le cas en régression logistique. Le modèle de régression logistique nous permet donc de résoudre le problème d'absence de linéarité dans la relation entre la variable réponse et les variables explicatives. En régression linéaire, le coefficient de détermination ajusté R_a^2 montre le pourcentage de la variabilité de la variable réponse expliqué par les variations des variables explicatives. Il s'agit de la mesure de la proportion de variation chez la variable dépendante qui est expliquée par le modèle d'explication. En régression logistique, par contre, ce pourcentage est mesuré par exemple par le pseudo-R^2 de McFadden donné par :

$$pseudo - R^2 = 1 - \left[\frac{-2L(C^{te}, X)}{-2L(C^{te})} \right]$$

Remarquons que la régression logistique a introduite pour la première fois en Biostatistique par Bekarson en 1944, et puis en Économétrie par McFadden.

Beaucoup plus utilisée en épidémiologie, la régression logistique binaire est la méthode la plus utilisée pour modéliser les variables binaires. Elle permet de modéliser la guérison ou non d'un patient (dans le domaine de la santé) et l'achat ou non d'un produit (en marketing quantitatif) ainsi que dans d'autres domaines mais à condition que la variable réponse soit dichotomique (variable ayant deux modalités). Autrement dit, la régression logistique binaire permet de vérifier si les variables indépendantes peuvent prédire une variable dépendante dichotomique binaire. C'est le cas d'une variable d'intérêt binaire (1=Oui, 0=Non ; 1=Vrai, 0=Faux ; 1=Malade, 0=Sain ; 0=Vivant, 1=Décédé ; etc).

3.4.2. Spécification du modèle logistique

La régression logistique binaire est une technique permettant de modéliser une variable dépendante qualitative binaire à l'aide d'une combinaison linéaire des variables explicatives avec la fonction de lien *logit* [4] :

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \tag{3.9}$$

avec p la probabilité que $Y = 1$ et $x_1,...,x_k$ les variables explicatives, $\beta_0,\beta_1,...,\beta_k$ les paramètres du modèle et un terme d'erreur. La régression logistique dite binaire est la régression logistique ordinaire.

D'après la relation (3.9), si nous considérons que le modèle contient une seule variable explicative x, alors le modèle logistique s'écrit :

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \tag{3.10}$$

En appliquant la fonction exponentielle aux deux derniers membres de cette relation, puis en développant et en mettant en évidence p , il vient la fonction logistique [5] :

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Cette fonction est strictement croissante et prend ses valeurs dans l'intervalle [0, 1]·

En générant une séquence de -20 à 20 avec le logiciel R, puis en utilisant cette relation, nous obtenons une courbe appelée fonction logistique ou sigmoïdes de la figure 11.

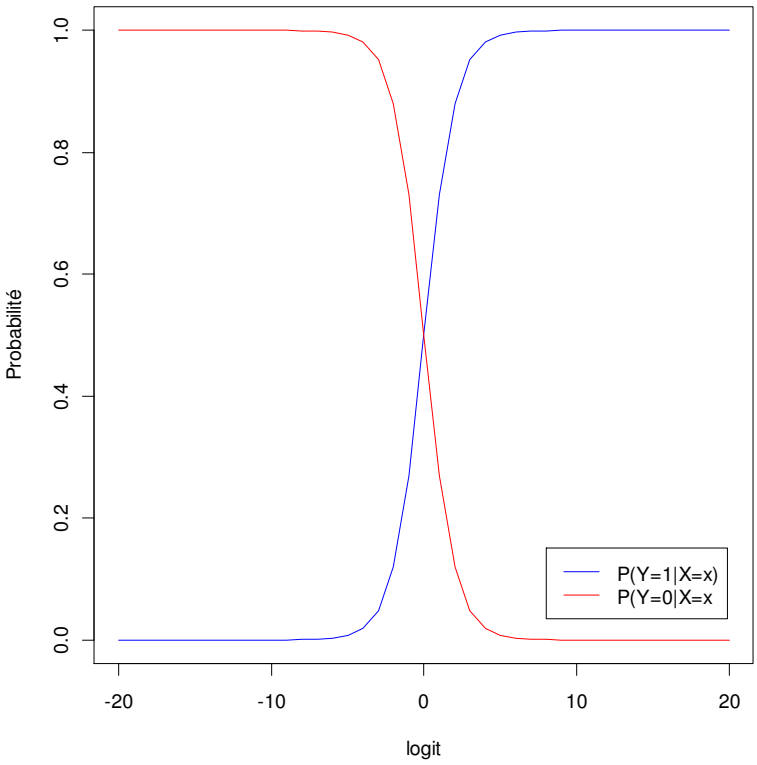


Figure 11 : Sigmoïdes
Source : Pr BARANKANIRA Emmanuel

3.4.3. Notion de cote

La cote, appelée également odds, est le rapport entre la probabilité de survenue d'un caractère considéré sachant que la variable exogène dont nous cherchons l'association est prise en compte.

Dans le cas où la variable à expliquer est binaire, de même que la variable explicative, la cote est donnée par :

$$cote = odds = \frac{P\left(y_i = 1 / X = 1\right)}{1 - P\left(y_i = 1 / X = 1\right)} \tag{3.11}$$

La quantité $\frac{P(y_i = 1/X = 1)}{1 - P(y_i = 1/X = 1)}$ exprime un odds qui signifie un rapport de chances. Si nous prenons un exemple d'un individu qui présente un odds de 3, cela montre que cet individu a trois fois plus de chances d'avoir la survenue d'un caractère de la variable d'intérêt que de ne pas en avoir. Le cote ainsi déterminée va permettre de déterminer le rapport de cotes (ou odds ratio).

3.4.4. Rapport de cotes

Le rapport de cotes (RC), appelé aussi odds ratio (OR) ou rapport de chances, est le rapport des probabilités d'avoir la survenue d'un caractère considéré pour la variable d'intérêt Y sachant que la variable explicative X_i est positif et d'avoir la survenue d'un caractère considéré pour la variable d'intérêt Y sachant que la variable explicative X_i est négatif.

Si nous considérons que cette variable explicative X qui représente le tabagisme (1= Fume, 0=Ne fume pas) par exemple et si la variable réponse Y représente le statut d'un patient (malade/non malade), alors la cote (odds) de la maladie chez les fumeurs vaut :

$$cote_F = \frac{P(Y=1|X=1)}{1-P(Y=1|X=1)} = \frac{P(Y=1|X=1)}{P(Y=0|X=1)} \tag{3.12}$$

La cote (odds) de l'absence de la maladie chez les non-fumeurs vaut :

$$cote_{NF} = \frac{P(Y=1|X=0)}{1-P(Y=1|X=0)} = \frac{P(Y=1|X=0)}{P(Y=0|X=0)} \tag{3.13}$$

L'odds ratio est le rapport entre la cote de la maladie chez les fumeurs et la cote de la maladie chez les non-fumeurs. Il est utilisé pour mesurer l'association entre deux variables cibles qualitatives (tabagisme et maladie par exemple).

Ce rapport de cotes est donné par :

$$OR = \frac{cote_F}{cote_{NF}} = \frac{\dfrac{P(Y=1|X=1)}{1-P(Y=1|X=1)}}{\dfrac{P(Y=1|X=0)}{1-P(Y=1|X=0)}} \tag{3.14}$$

Le rapport de cotes, compte tenu de sa valeur, permet de préciser les différentes variables prédictrices qui vont faire partie du modèle. Signalons qu'il est toujours supérieur ou égal à zéro.

Ainsi, nous distinguons trois cas de figures pour l'OR :

- Si $OR=1$, alors il y a égalité des risques.
- Si $OR>1$ **(IC ne contenant pas 1)**, alors le facteur étudié est un facteur de risque.
- Si $OR<1$ **(IC ne contenant pas 1)**, alors le facteur étudié est un effet protecteur.

Autrement dit, si l'OR est proche de 1, la maladie est indépendante du groupe. S'il est supérieur à 1, la maladie est fréquente dans le premier groupe que dans le second. S'il est inférieur à 1, la maladie est moins fréquente dans le second groupe que dans le premier.

Nous signalons, en passant, que les différentes valeurs de l'OR permettent d'identifier exactement les variables à inclure dans le modèle complet de la régression logistique.

Considérons le cas particulier de la relation (3.9) d'un modèle avec une seule variable explicative :

$$logit\left[p(Y=1|X=x)\right] = \beta_0 + \beta_1 x + \varepsilon$$

Dans le cas où cette variable est binaire (0/1), nous aurons, en remplaçant x respectivement par 1 et 0 :

$$logit\left[p(Y=1|X=x)\right] = \beta_0 + \beta_1$$

$$logit\left[p(Y=1|X=x)\right] = \beta_0$$

La différence membre à membre donne :

$$log\left[\frac{p\left(Y=1\middle|X=1\right)}{1-p\left(Y=1\middle|X=1\right)}\right]-log\left[\frac{p\left(Y=1\middle|X=0\right)}{1-p\left(Y=1\middle|X=0\right)}\right]=\beta_1 \tag{3.15}$$

soit :

$$log\left[\frac{\frac{p\left(Y=1\middle|X=1\right)}{1-p\left(Y=1\middle|X=1\right)}}{\frac{p\left(Y=1\middle|X=0\right)}{1-p\left(Y=1\middle|X=0\right)}}\right]=log\left(OR\right)=\beta_1 \tag{3.16}$$

et finalement :

$$OR=e^{\beta_1} \tag{3.17}$$

Le paramètre β_1 est estimé par l'expression suivante :

$$P\left(y_i=1/X=0\right)=\frac{exp\left(\beta_0\right)}{1+exp\left(\beta_0\right)}$$

De même, lorsque la variable est quantitative, nous aurons, en remplaçant x respectivement par 2 et 3 par exemple :

$$logit\left[p\left(Y=1\middle|X=2\right)\right]=\beta_0+2\beta_1$$

$$logit\left[p\left(Y=1\middle|X=3\right)\right]=\beta_0+3\beta_1$$

La différence membre à membre donne la même relation (3.16) et conduit à la même relation liant l'OR et la pente que celle donnée par (3.17).

Partant de l'expression de la fonction logistique notée :

$$log\,it(\,p\,) = ln\left(\frac{p}{1-p}\right)$$

nous trouvons l'expression du modèle logistique multiple suivante :

$$\pi(\,x\,) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

les β_i étant des paramètres et les x_i sont des variables explicatives. Il convient alors d'estimer ces paramètres.

3.4.5. Estimation de paramètres

La probabilité qu'une variable Y_i de Bernoulli prenne une valeur donnée y_i vaut :

$$P(Y_i = y_i) = p_i^{y_i}\left(1-p_i\right)^{1-y_i}$$

En régression logistique, la probabilité que la variable réponse prenne une valeur 1vaut :

$$p_i = P(Y_i = 1; x_i, \beta) \tag{3.18}$$

et son complément est donné par :

$$1-p_i = P(Y_i = 0; x_i, \beta) \tag{3.19}$$

La différence membre à membre donne la même relation (3.16) et conduit à la même relation liant l'OR et la pente

La vraisemblance ou densité jointe est alors donnée par :

$$L(\beta; y_1,...,y_n) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} p_i^{y_i}\left(1-p_i\right)^{1-y_i} \tag{3.20}$$

La log-vraisemblance vaut :

$$l = ln\,L = ln\left(\prod_{i=1}^{n} p_i^{y_i}\left(1-p_i\right)^{1-y_i}\right)$$

soit :

$$l = \sum_{i=1}^{n} \left[y_i \left(ln\, p_i \right) + \left(1 - y_i \right) ln \left(1 - p_i \right) \right] \tag{3.21}$$

En considérant que :

$$p_i = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}} = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

la relation (3.21) peut s'écrire :

$$l = \sum_{i=1}^{n} \left[y_i \left(X\beta \right) - y_i \, ln \left(1 + e^{X\beta} \right) - \left(1 - y_i \right) ln \left(1 + e^{X\beta} \right) \right]$$

ou encore :

$$l = \sum_{i=1}^{n} \left[y_i \left(X\beta \right) - ln \left(1 + e^{X\beta} \right) \right]$$

Les paramètres du modèle sont obtenus en maximisant cette log-vraisemblance, c'est-à-dire en la dérivant l par rapport à chaque composante de l'hyper-paramètre β*:*

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \left[y_i X - X \frac{e^{X\beta}}{1 + e^{X\beta}} \right]$$

Pour trouver les valeurs maximisent cette dérivée, nous résolvons le système :

$$\frac{\partial l}{\partial \beta} = 0$$

L'annulation de cette dérivée s'écrit alors :

$$\sum_{i=1}^{n} \left[y_i X - X \frac{e^{X\beta}}{1 + e^{X\beta}} \right] = 0 \tag{3.22}$$

soit :

$$\sum_{i=1}^{n} \left[X \left(y_i - p_i \right) \right] = 0 \tag{3.23}$$

avec $p_i = \frac{e^{X\beta}}{1+e^{X\beta}}$. Le système (3.22) ne peut se résoudre que numériquement, par exemple à l'aide de l'algorithme de Newton-Raphson (une des méthodes d'optimisation non linéaire).

Une autre façon de voir les choses est la suivante :

En supposant que les variables aléatoires $Y_1, Y_2, ..., Y_n$ soient indépendantes et observées aux points $x_1, x_2, ..., x_n$ de la variable *x*, et que :

$$Y_i \sim Ber\left[p(x_i)\right] \tag{3.24}$$

Dans ce cas, la fonction de probabilité de Y_i est :

$$p(y) = \left[p(x_i)\right]^y \left[1 - p(x_i)\right]^{1-y} \text{ avec } y = 1 ou\ y = 0 \tag{3.25}$$

La fonction du maximum de vraisemblance de β associée à une réalisation $(x_1, x_2, ..., x_n)$ de $(Y_1, Y_2, ..., Y_n)$ est :

$$L(\beta) = \prod_{i=1}^{n} \left[p(x_i)\right]^{y_i} \left[1 - p(x_i)\right]^{1-y_i} \tag{3.26}$$

La log-vraisemblance vaut alors :

$$l(\beta) = ln\, L(\beta) = \sum_{i=1}^{n} \left\{ y_i\, ln\, p(x_i) + (1 - y_i)\left[1 - p(x_i)\right] \right\} \tag{3.27}$$

En considérant la fonction logistique :

$$p(x_i) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)} \tag{3.28}$$

En posant :

$$u = \beta_0 + \beta_1 x_i \tag{3.29}$$

et aussi [2] :

$$g(u)=\frac{e^{u}}{1+e^{u}} \tag{3.30}$$

il vient

$$p(x_i)=g(u) \tag{3.31}$$

La dérivée devient :

$$g'(u)=\frac{e^{u}}{\left(1+e^{u}\right)^{2}}=\frac{e^{u}}{1+e^{u}}\frac{1}{1+e^{u}}=g(u)\left[1-g(u)\right] \tag{3.32}$$

La dérivation de la relation (3.32) par rapport à β_0 et β_1 respectivement donne :

$$\begin{cases}\dfrac{\partial p(x_i)}{\partial \beta_0}=p(x_i)\left[1-p(x_i)\right]\\ \dfrac{\partial p(x_i)}{\partial \beta_1}=x_i p(x_i)\left[1-p(x_i)\right]\end{cases} \tag{3.33}$$

La dérivation de la log-vraisemblance par rapport à β_0 donne :

$$\frac{\partial l(\beta)}{\partial \beta_0}=y_i\frac{p(x_i)\left[1-p(x_i)\right]}{p(x_i)}-(1-y_i)\frac{p(x_i)\left[1-p(x_i)\right]}{1-p(x_i)}=y_i-p(x_i) \tag{3.34}$$

De façon analogique, la dérivation de la log-vraisemblance par rapport à β_1 donne :

$$\frac{\partial l(\beta)}{\partial \beta_1}=x_i\left[y_i-p(x_i)\right] \tag{3.35}$$

Les équations de vraisemblance sont donc :

$$\begin{cases}\dfrac{\partial \ln L(\beta)}{\partial \beta_0}=\displaystyle\sum_{i=1}^{n}\left[y_i-p(x_i)\right]=0\\ \dfrac{\partial \ln L(\beta)}{\partial \beta_1}=\displaystyle\sum_{i=1}^{n}x_i\left[y_i-p(x_i)\right]=0\end{cases} \tag{3.36}$$

D'où, le système d'équations non résolubles analytiquement car complexe :

$$\begin{cases} \sum_{i=1}^{n} \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)} = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)} = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{3.37}$$

Rappelons que l'équation du modèle logistique multiple est :

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \tag{3.38}$$

Le signe d'un coefficient montre la direction de la relation. De plus, la valeur d'un coefficient indique l'effet spécifique produit par un mouvement d'une unité sur la variable indépendante sur le score logistique de la variable dépendante. Par score logistique, il faut comprendre l'intercept, c'est-à-dire la valeur de la variable dépendante lorsque toutes les variables indépendantes possèdent la valeur de 0.

L'estimation des paramètres par la méthode du maximum de vraisemblance (MMV) fournit des valeurs des paramètres qui rendent maximum la probabilité d'obtenir l'ensemble des données observées formant l'échantillon $\{(y_i, x_i); i = 1, 2, ..., n\}$.

Pour estimer $\beta = (\beta_0, \beta_1, ..., \beta_n)$ qui est le vecteurs des paramètres (hyper-paramètre) de la régression logistique par la méthode du maximum de vraisemblance, nous devons d'abord déterminer la loi de la distribution de $P(Y/X)$. La variable Y est une variable binaire qualitative définie dans $\{0,1\}$. Pour un individu i, nous pouvons faire la modélisation de la probabilité à l'aide de la loi binomiale $B(1,\pi)$ avec l'expression suivante :

$$P(Y_i / X_i) = (\pi_i)^{y_i} \times (1-\pi_i)^{1-y_i} \tag{3.39}$$

Dans ce cas, nous distinguons deux possibilités différentes :

- Si $Y_i = 1$ alors $P(Y_i = 1 / X_i) = \pi$
- Si $Y_i = 0$ alors $P(Y_i = 0 / X_i) = 1 - \pi$

La vraisemblance d'un échantillon considéré est donc donné par :

$$L(\beta, y_i) = \prod_{i=1}^{n} P(Y = y_i / X = x_i) \tag{3.40}$$

Pour l'ensemble de l'observation, nous avons :

$$L(\beta, y_i) = \prod_{i=1}^{n} \pi(X_i)^{y_i} \left(1 - \pi(X_i)\right)^{1-y_i} \tag{3.41}$$

Partant de cette expression donnée par l'équation ci-haut, nous déterminons le logarithme de la vraisemblance par l'expression suivante :

$$l(\beta_i, y_i) = \log L(\beta_i, y_i) \tag{3.42}$$

Il en découle que :

$$log\, L(\beta_i, y_i) = \sum_{i=1}^{n} y_i \, log\, \pi(x_i) + (1 - y_i) \, log\left(1 - \pi(x_i)\right) \tag{3.43}$$

Pour la modèle de régression logistique multiple, $\pi(x_i)$ est une fonction qui donne la probabilité $P(Y = 1 / x)$ et est donnée par l'expression suivante :

$$\pi(x_i) = \frac{e^{\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}}} \tag{3.44}$$

Dans le cas où la régression est simple, $\pi(x_i)$ est égale à [3] :

$$\pi(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_1}} \tag{3.45}$$

Pour trouver les estimateurs par la méthode du maximum de vraisemblance, nous dérivons partiellement la log-vraisemblance par rapport aux différents paramètres et nous obtenons $\hat{\beta}$ en annulant les dérivées partielles suivantes :

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} x_{ik} \left(y_i - \pi(x_i)\right) \tag{3.46}$$

En vertu des propriétés du développement, il en résulte que :

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \frac{x_i e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} \tag{3.47}$$

Nous déterminons, en définitive, les estimateurs $\hat{\beta}_j$ des paramètres β en maximisant la log-vraisemblance par rapport aux paramètres $\hat{\beta}_j$.

De là, nous résolvons le système suivant :

$$\begin{cases} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 0 \\ \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{x_i e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 0 \end{cases} \tag{3.48}$$

D'après Gilbert Saporta, ce système n'a pas de solution analytique mais se résout par des procédures de calculs numériques.

La méthode du maximum de vraisemblance nous permet d'estimer la matrice de $V\left(\hat{\beta}\right)$ de variances-covariances des estimateurs des coefficients. Nous signalons que $\hat{\pi}_i = \hat{\pi}(x_i)$.

La matrice des estimateurs est donnée par :

$$\hat{V}\left(\hat{\beta}\right) = \left[\frac{-\partial^2 l(\beta)}{\partial \beta^2}\right]^{-1}_{\beta=\hat{\beta}} \tag{3.49}$$

Cette matrice peut aussi s'écrire :

$$\hat{V}\left(\hat{\beta}\right) = \begin{pmatrix} Var\left(\hat{\beta}_0\right) & Cov\left(\hat{\beta}_0,\hat{\beta}_1\right) & \cdots & Cov\left(\hat{\beta}_0,\hat{\beta}_j\right) & \cdots & Cov\left(\hat{\beta}_0,\hat{\beta}_{p-1}\right) \\ Cov\left(\hat{\beta}_0,\hat{\beta}_1\right) & Var\left(\hat{\beta}_1\right) & \cdots & Cov\left(\hat{\beta}_1,\hat{\beta}_j\right) & \cdots & Cov\left(\hat{\beta}_1,\hat{\beta}_{p-1}\right) \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ Cov\left(\hat{\beta}_0,\hat{\beta}_j\right) & Cov\left(\hat{\beta}_1,\hat{\beta}_j\right) & \cdots & Var\left(\hat{\beta}_j\right) & \cdots & Cov\left(\hat{\beta}_j,\hat{\beta}_{p-1}\right) \\ \vdots & \vdots & & \vdots & \ddots & \\ Cov\left(\hat{\beta}_0,\hat{\beta}_{p-1}\right) & Cov\left(\hat{\beta}_1,\hat{\beta}_{p-1}\right) & \cdots & Cov\left(\hat{\beta}_j,\hat{\beta}_{p-1}\right) & \cdots & Var\left(\hat{\beta}_{p-1}\right) \end{pmatrix} \tag{3.50}$$

Cela donne :

$$\hat{V}\left(\hat{\beta}\right)=\begin{pmatrix}\sum_{i=1}^{n}\hat{\pi}_i\left(1-\hat{\pi}_i\right) & \cdots & \sum_{i=1}^{n}x_i^p\hat{\pi}_i\left(1-\hat{\pi}_i\right)\\ \vdots & \ddots & \vdots\\ \sum_{i=1}^{n}x_i^p\hat{\pi}_i\left(1-\hat{\pi}_i\right) & \cdots & \sum_{i=1}^{n}\left(x_i^p\right)^2\hat{\pi}_i\left(1-\hat{\pi}_i\right)\end{pmatrix}^{-1} \tag{3.51}$$

Finalement, nous avons :

$$\hat{V}\left(\hat{\beta}\right)=\left(\begin{pmatrix}1 & \cdots & x_1^p\\ \vdots & & \vdots\\ 1 & \cdots & x_n^p\end{pmatrix}'\begin{pmatrix}\hat{\pi}_1\left(1-\hat{\pi}_1\right) & & 0\\ & \ddots & \\ 0 & & \hat{\pi}_n\left(1-\hat{\pi}_n\right)\end{pmatrix}\begin{pmatrix}1 & \cdots & x_1^p\\ \vdots & & \vdots\\ 1 & \cdots & x_n^p\end{pmatrix}\right)^{-1} \tag{3.52}$$

De façon synthétique, nous avons :

$$\hat{V}\left(\hat{\beta}\right)=\left(X'VX\right)^{-1} \tag{3.53}$$

3.4.6. Test sur les paramètres

Les tests effectués sur les paramètres sont des méthodes qui permettent de tester l'apport d'une variable explicative x au modèle.

L'hypothèse nulle du test global s'écrit $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ et signifie que toutes les variables explicatives n'ont pas d'influence sur la probabilité d'apparition de l'événement d'intérêt alors que l'hypothèse alternative $H_0 : \exists \beta_i \neq 0$ pour dire qu'il existe au moins une variable explicative significative. Pour cela, le modèle contenant uniquement l'intercept (modèle sans variable ou modèle M_0) est comparé avec le modèle contenant toutes les variables explicatives (modèle M_1).

Ces deux modèles s'écrivent respectivement :

$$M_0 : logit\left(p\right)=log\left(\frac{p}{1-p}\right)=\beta_0+\varepsilon \tag{3.54}$$

$$M_1 : logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \tag{3.55}$$

La comparaison de ces deux modèles se fait au moyen du rapport de vraisemblance (RV) calculé en faisant la différence des déviances des deux modèles, la déviance étant égale à « moins deux fois la log-vraisemblance du modèle ». Cette différence suit une loi du khi-deux à k degrés de liberté. Si le RV est supérieur à cette statistique du chi-deux tabulée, alors la p-value est inférieure au seuil de significativité (5 %) et l'hypothèse nulle H_0 est rejetée. La conclusion est que le modèle M_1 est meilleur que le modèle M_0 et que les variables explicatives ont donc simultanément une influence sur la probabilité d'apparition de l'événement d'intérêt.

Pour le test de significativité sur un paramètre, le modèle sans une variable explicative donnée (modèle M_1) est comparé avec le modèle contenant cette variable (modèle M_2).

Ces modèles s'écrivent respectivement :

$$M_1 : logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_k x_k + \varepsilon \tag{3.56}$$

$$M_2 : logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + \cdots + \beta_k x_k + \varepsilon \tag{3.57}$$

Nous distinguons trois types de tests à savoir le test de Wald, le rapport de vraisemblance et le test du Score pour tester la significativité d'un paramètre.

Le test de Wald permet de tester l'hypothèse suivante :

$$H_0 : \beta_j = 0 \leftrightarrow H_1 : \beta_j \neq 0 \tag{3.58}$$

Le test de Wald est donc similaire avec un test de Student en régression usuelle. Pour vérifier l'hypothèse H_0, nous déterminons le statistique de Wald w donnée par l'expression suivante [4] :

$$w = \left[\frac{\hat{\beta}_j}{s\left(\hat{\beta}_j\right)}\right]^2 \sim \chi_1^2 \tag{3.59}$$

où $s\left(\hat{\beta}_j\right)$ représente l'estimation de l'écart-type de l'estimateur de $\hat{\beta}_j$. Sous l'hypothèse H_0, w suit approximativement une loi du khi-deux à un degré de liberté au seuil de signification α. Dans le cas contraire, l'hypothèse nulle est rejetée, c'est-à-dire que l'hypothèse nulle est rejetée si une fois $w \geq \chi^2_{1-\alpha}$.

Il est aussi possible d'utiliser la statistique :

$$U = \frac{\hat{\beta}_j}{s\left(\hat{\beta}_j\right)} \sim N(0,1) \tag{3.60}$$

Si $w > \chi^2_1$ ou $U > U_\alpha$ où est la quantile de la loi normale centrée-réduite de niveau égal à 1,96 (si $\alpha = 0,05$), alors l'hypothèse nulle est rejetée et on conclut que β_j est significatif et que donc M_2 est meilleur que le modèle M_1. Dans ce cas, la $j^{ème}$ variable a une influence sur la probabilité d'apparition de l'évènement, conditionnellement aux autres variables du modèle.

La statistique de Wald mesure donc la signification statistique de chaque coefficient de régression logistique. Pour que ce coefficient soit statistiquement significatif au seuil de 5 % par exemple, il faut que la valeur de la statistique de Wald dépasse 3,84, une valeur tabulée de la statistique du khi-deux à 1 degré de liberté.

S'agissant du test du rapport de vraisemblance, la méthode du rapport de vraisemblance permet, comme te test de Wald, de vérifier l'apport de la variable explicative X qui est mesurée par la statistique G. Cette statistique suit approximativement une loi du khi-deux à un degré de liberté sous l'hypothèse H_0.

Cette statistique est donnée par l'expression suivante [3] :

$$G = -2 \log \frac{vraisemblance\ sans\ la\ variable}{vraisemblance\ avec\ la\ variable} \tag{3.61}$$

Cette statistique équivaut à :

$$G = \left[-2L(C^{te})\right] - \left[-2L(C^{te}, X)\right] \tag{3.62}$$

Le test du score met à l'épreuve $H_0 : \beta = 0 \leftrightarrow \ H_1 : \beta \neq 0$. Ce test permet de vérifier la même hypothèse que le test du rapport de vraisemblance. La validation de l'hypothèse H_0 se base sur la détermination du score χ^2_{Score} qui doit nécessairement suivre une loi du khi-deux à 1 degré de liberté.

Cette statistique du score est donnée par :

$$\chi^2_{Score} = \left[\frac{\partial L}{\partial \beta}\right]'_{\beta=\hat{\beta}_{H_0}} \left[-\frac{\partial^2 L}{\partial \beta^2}\right]^{-1}_{\beta=\hat{\beta}_{H_0}} \left[\frac{\partial L}{\partial \beta}\right]_{\beta=\hat{\beta}_{H_0}} \tag{3.63}$$

3.4.7. Modification d'effet et facteurs de confusion

Considérons le modèle logistique avec deux variables explicatives avec interaction :

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \tag{3.64}$$

Les hypothèses du test d'interaction sont :

$$\begin{aligned} &H_0 : \beta_3 = 0 \\ &H_1 : \beta_3 \neq 0 \end{aligned} \tag{3.65}$$

Si l'interaction est significative, cela signifie que l'effet d'une variable modifie celui de l'autre variable explicative. En effet, si par exemple $x_2 = 0$, alors l'effet de x_1 est β_1. Si, par contre $x_2 = 1$, alors l'effet de x_1 est $\beta_1 + \beta_3$. Si l'interaction est significative, cela signifie aussi qu'il y a modification d'effet et, dans ce cas, on garde le modèle avec interaction. Dans le cas contraire, l'interaction est retirée du modèle.

La variable explicative peut aussi être considérée comme une variable confondante (facteur de confusion). Considérons, par exemple, deux modèles avec x_2 comme facteur d'ajustement :

$$M_1 : logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \varepsilon \tag{3.66}$$

$$M_2 : logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{3.67}$$

Les OR de ces deux modèles sont :

$$OR(M_1) = e^{\beta_1} \tag{3.68}$$

$$OR(M_2) = e^{\beta_1} \tag{3.69}$$

Remarquons que $OR(M_1)$ représente l'effet brut de x_1 tandis que $OR(M_2)$ représente l'effet de x_1 ajusté sur x_2. Il y aura confusion lorsque ces deux effets sont différents. Il est aussi possible de considérer qu'il y a un effet de confusion lorsque la variation relative donnée par la relation :

$$VR = \frac{OR(M_2) - OR(M_1)}{OR(M_2)} \tag{3.70}$$

dépasse un certain seuil, par exemple une valeur compris entre 10 et 20 %. Dans le cas contraire, on teste $H_0 : \beta_2 = 0$ pour décider si oui ou non x_2 sera retirée.

3.4.8. Intervalle de confiance sur les paramètres

L'intervalle de confiance (IC) est une méthode qui permet de savoir s'il ya une relation entre la variable x_j et Y.

Cet intervalle de niveau $1-\alpha$ est donné par l'expression suivante pour l'OR [5] :

$$IC_{1-\alpha}(\beta_j) = \exp\left[\hat{\beta}_j \pm u_\alpha . s(\hat{\beta}_j)\right] \tag{3.71}$$

Bien évidement, si $1 \notin IC$, alors x_j est un effet protecteur ou de risque. Par contre, si $1 \in IC$, alors il y a égalité des risques.

Sachant que:

$$\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) = \mathrm{Var}(\hat{\beta}_0) + x^2 \mathrm{Var}(\hat{\beta}_1) + 2x\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \tag{3.72}$$

il vient que l'intervalle de confiance de $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ s'écrit :

$$IC_{1-\alpha}(\pi) = \left[\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x - 1{,}96\sqrt{Var(\hat{\beta}_0+\hat{\beta}_1 x)}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x - 1{,}96\sqrt{Var(\hat{\beta}_0+\hat{\beta}_1 x)}}} ; \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x + 1.96\sqrt{Var(\hat{\beta}_0+\hat{\beta}_1 x)}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x + 1.96\sqrt{Var(\hat{\beta}_0+\hat{\beta}_1 x)}}} \right] \quad (3.73)$$

avec

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (3.74)$$

Une fois que le modèle final est sélectionné, il convient de déterminer la qualité d'ajustement du modèle aux données observées. Cela revient à comparer les valeurs prédites et les valeurs observées de la variable réponse (Test de Hosmer-Lemeshow). Malheureusement, ce test n'est pas puissant.

3.4.9. Application en santé

Les variables qualitatives (sexe, niveau d'instruction, santé, alcool, religion, entourage) sont résumées en termes de fréquences des modalités dans le tableau 11 ci-après.

Tableau 11 : Fréquence en % de différentes modalités

	Tabagisme		
Caractéristiques	**Fumeur**	**Non fumeur**	**Total**
Sexe			
Masculin	12,22	42,22	54,44
Féminin	7,41	38,15	45,56
Niveau d'instruction			
Instruit	7,04	58,15	65,19
Non instruit	12,59	22,22	34,81
Consommation d'alcool			
Buveur	16,67	50,37	67,04
Non buveur	2,96	30,00	32,96
Religion			
Chrétien	16,66	67,41	84,07
Non chrétien	2,97	12,96	15,93
Impact du tabac sur la santé			
Est conscient	8,89	70,00	78,89
Ne sait pas	10,74	10,37	21,11
Impact du tabac sur l'entourage			
Est conscient	9,26	71,85	81,11
Ne sait pas	10,37	8,52	18,89
Ensemble	19,63	80,37	100,00

Source : Pr BARANKANIRA Emmanuel

Ces résultats mettent en évidence que les hommes, les non instruits, les buveurs, les chrétiens, les personnes qui ne sont pas conscientes que le tabagisme a un impact sur la santé ou sur leur entourage fument plus que leurs homologues.

Ce même **Tableau 7** montre que, dans l'échantillon, il y avait plus d'hommes (54,44 %) que de femmes (45,56 %) et la majorité des hommes (42,22 %) étaient non fumeurs. Quant au niveau d'instruction, il y avait plus d'instruits que de non instruits (65,19 % contre 34,81 %). Dans l'échantillon, il y avait plus de non fumeurs que de fumeurs (80,37 % contre 19,63 %). Pour l'alcool, les buveurs étaient plus nombreux que les non buveurs (67,04 % contre 32,96 %). Parmi les buveurs, les non fumeurs étaient encore majoritaires (50,37 % contre 16,67 %). Pour ce qui concerne la religion, les chrétiens étaient plus nombreux que les non chrétiens (84,07 % contre 15,93 %). La majorité (78,89 %) des individus étaient conscients des méfaits du tabagisme sur la santé ou sur leur entourage. Concernant le revenu, il y a un cas extrême qui est de l'ordre de 3.000.000 FBU par mois alors que la majorité gagnait moins de deux cent mille francs burundais. Le pourcentage du nombre de fumeurs dans notre échantillon est estimé à 19,63 %.

Cette prévalence permet de calculer la cote (ou odds) d'être fumeur à travers la relation :

$$\text{odds} = \frac{p}{1-p} \tag{3.61}$$

où p désigne la probabilité d'être fumeur. Dans notre échantillon, la cote d'être fumeur est de 24,4 %.

Le **Tableau 8** illustre les statistiques descriptives (moyenne et écart-type) des variables continues.

Tableau 8 : Statistiques descriptives pour les variables continues

Caractéristiques	Moyenne ± écart-type
Âge (années)	33,27 ± 23,22
Fumeurs	39,72 ± 14,90
Non fumeurs	31,71 ± 15,35
Revenu (FBU)	119.000 ± 331 833,30
Fumeurs	146.154,70 ± 307 362,90
Non fumeurs	113.485,90 ± 419 914,90

Ce tableau montre que l'âge moyen des fumeurs était plus élevé que celui des non fumeurs (39,72 ans contre 31,71 ans). De même, le salaire moyen des fumeurs était plus élevé que celui des non fumeurs (146.154,70 FBU contre 113.485,90 FBU).

Les prévalences du tabagisme selon les modalités des variables qualitatives sont consignées dans le **tableau 9**.

Tableau 9 : Prévalence du tabagisme selon les facteurs socio-économiques

Variable	Modalité	Prévalence (%)
Sexe	Masculin	22,50
Niveau d'instruction	Instruit	10,80
Alcool	Buveur	24,86
Religion	Chrétien	19,82
Santé	Est conscient	11,27
Entourage	Est conscient	11,42

Nous remarquons qu'il y avait plus de fumeurs chez les hommes, chez les non instruits, chez les buveurs et chez les chrétiens comparativement à leurs homologues. Ainsi, les hommes, les non instruits, les buveurs, les chrétiens et les personnes qui ignorent les méfaits du tabac étaient plus exposés au tabagisme par rapport aux autres. Les prévalences du tabagisme sont : pour les hommes (22,50 %), pour les femmes (16,30 %), pour les instruits (10,80 %), pour les non instruits (36,17 %), pour les buveurs (24,86 %), pour les non buveurs (8,99 %), pour les chrétiens (19,82 %) et pour les non chrétiens (18,60 %). Pour mieux visualiser les résultats groupés dans les tableaux précédents, nous avons construits des graphiques appropriés.

La **figure 12** compare l'âge moyen des fumeurs et des non fumeurs tandis que la **figure 13** compare le revenu moyen de ces deux catégories précédentes selon le statut tabagique. De ces graphiques, nous déduisons que l'âge moyen des individus enquêtés est plus élevé chez les fumeurs que chez les non fumeurs pour lesquels nous observons des valeurs aberrantes pour l'âge. Par contre, la **figure 13** montre que la variable « Revenu » ne se comporte pas comme la variable « Âge ». Autrement dit, il ne semble pas y avoir de différences de revenus chez les fumeurs et les non fumeurs. Cela est dû au fait que, dans notre échantillon, deux individus avaient un revenu de 3.000.000 FBU. Le revenu moyen est de 119.899 FBU.

En faisant l'analyse de la variance, nous constatons que les fumeurs n'ont pas un salaire moyen significativement plus élevé que les non fumeurs (F=0.41, p-value=0,52), ce qui converge avec les statistiques descriptives (**Figure 13**).

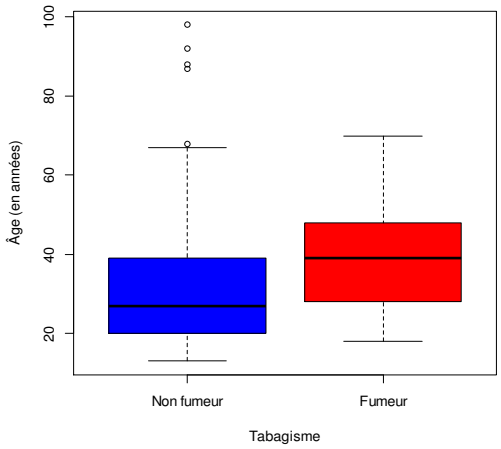


Figure 12 : **Boxplot de l'âge selon le statut tabagique**

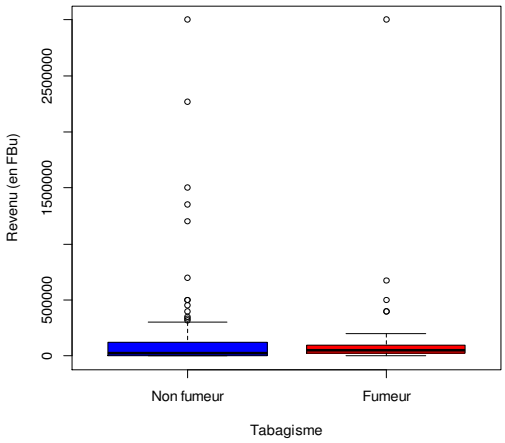


Figure 13 : **Boxplot du revenu selon le statut tabagique**

Les graphiques ci-dessous (*Figures 14 à 19*) indiquent, pour une variable exogène dichotomique, le nombre de sujets enquêtés selon les modalités de cette variable et le statut tabagique. Ces derniers montrent également que les hommes, les non instruits, les buveurs, les chrétiens et ceux qui ne savent pas que le fait de fumer a un impact sur leur santé et sur celle de l'entourage sont plus exposés au tabagisme.

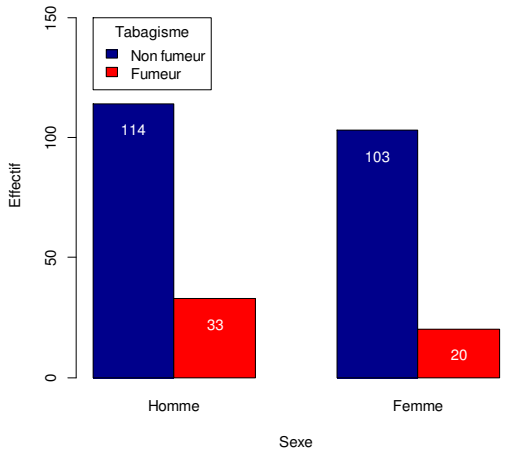


Figure 14 *:* Tabagisme et sexe

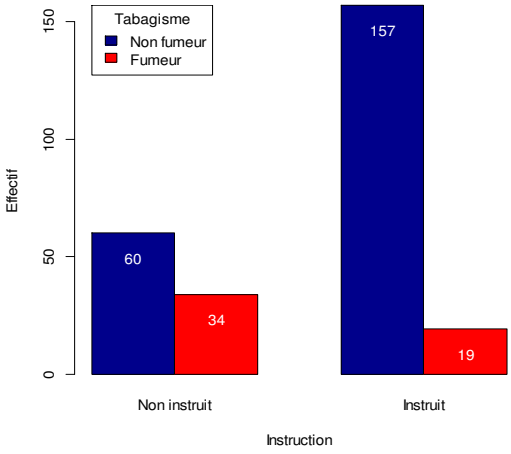


Figure 15 *:* Tabagisme et instruction

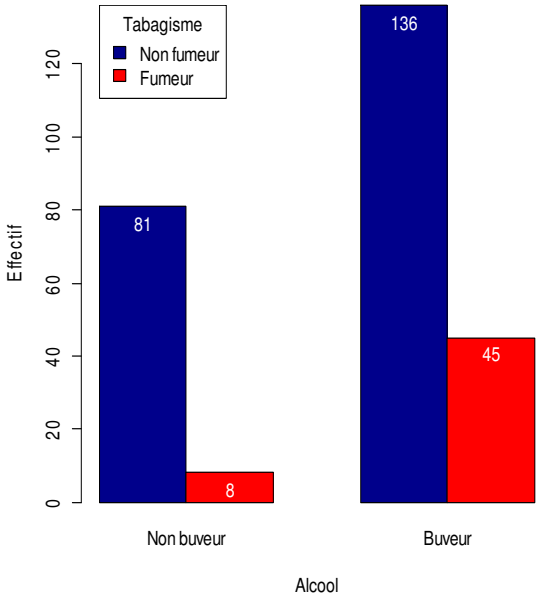


Figure 16 *:* Tabagisme et alcool

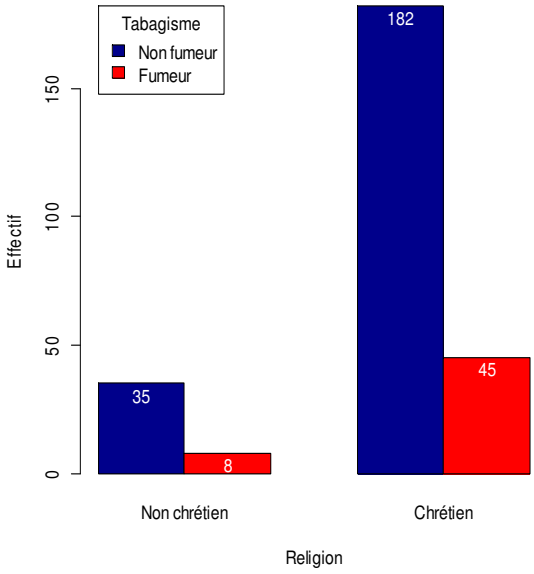
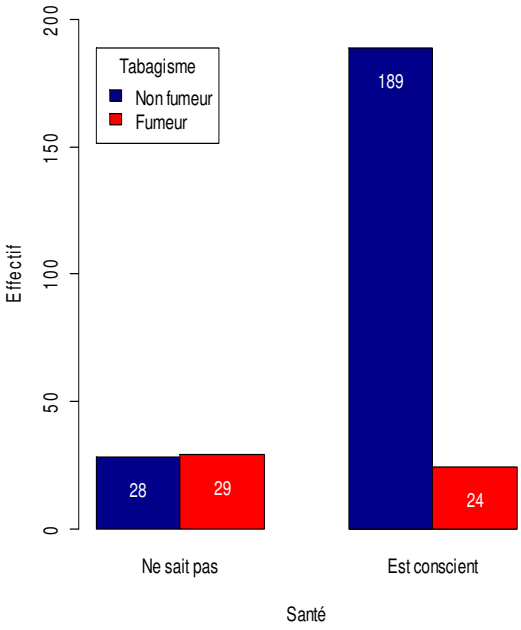


Figure 17 *:* Tabagisme et religion

Source : Pr BARANKANIRA Emmanuel



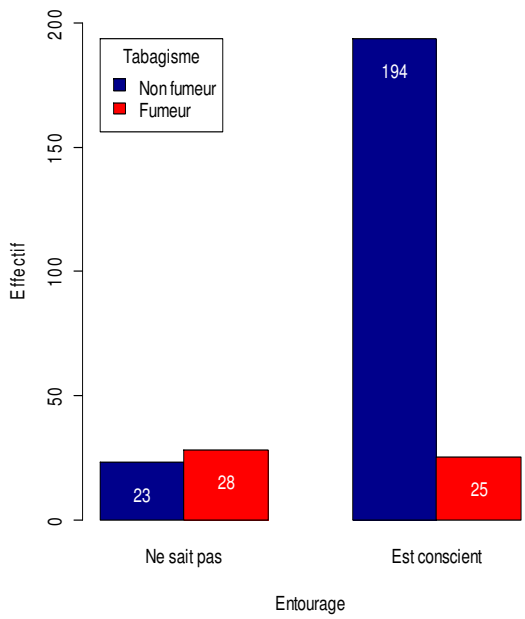


Figure 18 : Tabagisme et santé **Figure 19 : Tabagisme et entourage**
Source : **BARANKANIRA Emmanuel**

Dans le but de connaître s'il y a une relation entre la variable endogène « tabagisme » et les autres variables prises séparément deux à deux, nous avons fait le test du chi-deux d'indépendance. Les résultats obtenus sont confinés dans le **Tableau 12**.

Tableau 12 : Résultats du test d'indépendance entre le tabagisme et les autres facteurs

Variables	**P-value**
Sexe	0,26
Instruction	<0,05
Alcool	<0,05
Religion	1,00
Santé	<0,05
Entourage	<0,05

Source : Pr BARANKANIRA Emmanuel

Au seuil de 5 %, nous constatons qu'il n'y a pas d'indépendance entre le tabagisme et les variables : instruction, alcool, santé et entourage car les probabilités (P-values) relatives à ces variables sont plus petites que 5 %. Par contre, les résultats du **tableau 8** montrent qu'il n'y a pas de relation entre le tabagisme et les variables sexe et religion. En effet, leurs p-valeurs dépassent largement 5 %. L'interaction entre les variables explicatives n'était pas significative.

En statistique inférentielle, il est intéressant d'estimer les paramètres associés aux variables explicatives dans un modèle de régression logistique. Pour notre étude, en utilisant la méthode du maximum de vraisemblance, nous avons obtenu les estimations des paramètres, leurs intervalles de confiance au niveau de confiance de 95 % et les p-values correspondantes groupés dans le **tableau 13**.

Tableau 13 : Estimation des paramètres de la régression logistique univariée

Variable	Paramètre	Estimation	IC à 95 %	P-value
Sexe	Ordonnée à l'origine	-1,24	[-1,63; -0,85]	<0,001
Instruction	Ordonnée à l'origine	-0,57	[-0,99; -0,15]	<0,05
Alcool	Ordonnée à l'origine	-2,31	[-3,04; -1,59]	<0,05
Religion	Ordonnée à l'origine	-1,48	[-2,24, -0,71]	<0,05
Santé	Ordonnée à l'origine	-0,04	[-0,48; 0,55]	0,89
Entourage	Ordonnée à l'origine	0,20	[-0,35; 0,75]	0,49
Âge	Ordonnée à l'origine	-2,48	[-3,23; -1,74]	<0,05
Revenu	Ordonnée à l'origine	-1,44	[-3,46; 1,69]	<0,05

Source : Pr BARANKANIRA Emmanuel

Du tableau précédent et au seuil de 5 %, nous déduisons les équations des modèles logistiques univariés, à savoir :

$$log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = -0,\ 57\ -\ 1,54 \times Instruction + \varepsilon\ ;\ log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = -2,31 + 1,21 \times Alcool + \varepsilon$$

$$log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = -0,04\ -\ 2,10 \times Santé + \varepsilon\ ;\ log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = -0,20\ -\ 2,25 \times Entourage + \varepsilon$$

$$log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = -2,\ 48\ -\ 0,03 \times \hat{A}ge + \varepsilon$$

où x désigne l'une des valeurs des variables explicatives du tabagisme (t l'instruction, l'alcool, la santé, l'entourage ou l'âge selon le cas traité) et $\hat{p}(x)$ la probabilité correspondante. Les variables sexe, religion et revenu ont été exclues du modèle général car leurs p-values dépassaient 5 %. Ainsi, la sélection des variables pour le modèle final concernaient les variables dont leurs p-values ne dépassent pas 5 %.

Pour mieux identifier les facteurs socio-économiques du tabagisme, nous avons calculé également les rapports de cotes ou odds ratios (OR) de chacune des variables endogènes. Ces rapports ainsi que

leurs intervalles de confiance au niveau de confiance de 95% sont consignés dans le **Tableau 14**. Dans ce tableau, nous n'avons pas mentionné les rapports des côtes des modalités de référence car ils correspondent à un.

Tableau 14 : OR d'être fumeur et intervalles de confiance à 95 %

Variable	Modalité	OR	IC à 95%
Sexe	Femme	0,67	[0,36; 1,23]
Instruction	Instruit	0,21	[0,11; 0,40]
Alcool	Buveur	3,35	[1,58; 7,99]
Religion	Chrétien	1,08	[0,49; 1,65]
Santé	Conscient	0,12	[0,06; 0,24]
Entourage	Conscient	0,11	[0,05; 0,21]
Âge	Âge	1,03	[1,01; 1,05]
Revenu	Revenu	1,00	[0,99; 1,00]

Source : Pr BARANKANIRA Emmanuel

Des **tableaux 6 et 7**, nous constatons que le fait de boire de l'alcool et la variable « âge » augmentent le risque de fumer car leurs rapports de cotes dépassent 1 et leurs intervalles de confiance à 95 % ne contiennent pas 1. Par contre, l'instruction, la connaissance des méfaits du tabagisme et le fait d'être de sexe féminin diminuent ce risque de fumer car leurs rapports de cotes ne dépassent pas un et leurs intervalles de confiance à 95 % ne contiennent pas 1. Du tableau précédent, nous un groupe d'individu interrogé a 3 fois plus de chance de fumer par rapport à un groupe d'individus qui ne fume pas.

Après avoir étudié la contribution individuelle des facteurs dans l'explication du « tabagisme », nous avons voulu savoir l'influence globale de ces facteurs explicatifs. Ainsi, nous avons réalisé une analyse logistique multivariable. Cette dernière a ressorti les facteurs : âge, santé et entourage comme facteurs les plus explicatifs du tabagisme, voir le **tableau 8**.

Tableau 8 : Modèle de régression logistique multivariable final

Variable	Paramètre	Estimation	IC à 95 %	P-value
	Ordonnée à l'origine	-0,78	[-1,68; 0,13]	9 %
Âge	Âge	0,03	[0,01; 0,05]	< 5 %
Santé	Conscient	-1,23	[-2,24; -0,21]	< 5 %
Entourage	Conscient	-1,34	[-2,36; -0,31]	<5 %

Source : Pr BARANKANIRA Emmanuel

De ce tableau, nous déduisons le modèle final suivant :

$$log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = -0,\ 78 + 0,\ 03 \times \hat{A}ge - 1,23 \times Santé - 1,34 \times Entourage + \varepsilon$$

où x est une valeur de la variable « Tabagisme » conditionnée au vecteur de l'espace ayant pour composante âge, santé et entourage. Ce modèle permet de faire des prévisions. En effet, à titre d'exemple, la probabilité d'être fumeur sachant que le sujet est âgé de 30 ans (Âge=30), qu'il est conscient du fait que le tabagisme a un impact sur sa santé (Santé=1) et sur la santé de son entourage (Entourage=1) vaut 0,08.

En conclusion, cette étude a montré une prévalence du tabagisme très élevée dans la ville de Gitega. En outre, elle a montré qu'il existe un lien entre le tabagisme et les variables âge, santé du sujet et celle de l'entourage à travers le modèle logistique. Les variables qualitatives représentant le niveau d'instruction, l'alcoolisme, la santé du sujet et la santé de l'entourage n'étaient pas indépendantes du tabagisme au seuil de 5 %. Notre étude a permis d'estimer à 0,08 la probabilité d'être fumeur sachant que le sujet est âgé de 30 ans, qu'il est conscient du fait que le tabagisme a un impact sur sa santé et sur la santé de son entourage. Notre étude pourrait servir de référence aux décideurs de santé publique lors de la sensibilisation contre le tabagisme.

3.5. Courbe ROC et aire sous la courbe

3.5.1. Construction de la courbe

La courbe ROC (Receiving Operating Characteristics) est un outil statistique très riche. Ainsi, son champ d'application est très vaste. Elle est très utilisée en épidémiologie. Elle présente surtout des caractéristiques très intéressantes pour l'évaluation et la comparaison des performances des modèles. Elle propose un outil graphique qui permet d'évaluer et de comparer globalement le comportement des modèles. Un indicateur synthétique peut également lui être associé, le critère AUC (aire sous la courbe, en anglais *Area Under Curve*).

La courbe ROC met en relation le taux de vrais positifs TVP (la sensibilité) et le taux de faux positifs TFP (TFP = 1 - Spécificité) sur un graphique. Habituellement, les $\hat{\pi}(x)$ sont comparées à un seuil S=0,5 pour effectuer une prédiction $\hat{y}(x)$. Il est également possible de construire la matrice de confusion et en extraire les 2 indicateurs précités. La courbe ROC généralise cette idée en faisant

varier [illegible] sur tout le continuum des valeurs possibles entre 0 et 1. Pour chaque configuration, une matrice de confusion est construite et les TVP et TFP sont calculés. C'est l'idée directrice. Elle est un peu lourde à mettre en place mais en pratique, il n'est pas nécessaire de construire explicitement la matrice de confusion, le processus est celui-ci :

- Calculer le score $\hat{\pi}(x)$ de chaque individu à l'aide du modèle de prédiction.
- Trier le fichier selon un score décroissant.
- Considérons qu'il n'y a pas d'ex-aequo. Chaque valeur du score peut être potentiellement un seuil s. Pour toutes les observations dont le score est supérieur ou égal à s, les individus sont dans la partie haute du tableau.
- La courbe ROC correspond au graphique nuage de points qui relie les couples (TVP, TFP). Le premier point est forcément (0,0), le dernier est (1,1).

Deux situations extrêmes peuvent survenir. La discrimination est parfaite. Tous les positifs sont situés devant les négatifs, la courbe ROC est collée aux extrémités Ouest et Nord du repère. Les scores sont totalement inopérants, le modèle attribue des valeurs au hasard, dans ce cas les positifs et les négatifs sont mélangés. La courbe ROC se confond avec la première bissectrice.

3.5.2. Aire sous la courbe

Il est possible de caractériser numériquement la courbe ROC en calculant la surface située sous la courbe. C'est le critère AUC. Elle exprime la probabilité de placer un individu positif devant un négatif. Ainsi, dans le cas d'une discrimination parfaite, les positifs sont sûrs d'être placés devant les négatifs, nous avons AUC = 1. Au contraire, si le modèle attribue des scores au hasard, il y a autant de chances de placer un positif devant un négatif que l'inverse, la courbe ROC se confond avec la première bissectrice, nous avons AUC = 0,5. C'est la situation de référence. Généralement différents paliers sont proposés pour donner un ordre d'idées sur la qualité de la discrimination. L'aire sous la courbe ROC (AUC) est une mesure globale de la performance du test parmi les plus utilisées. Elle varie entre 0.5 dans le cas d'un test non informatif à 1 dans le cas d'une performance parfaite. Ainsi, une AUC de 0,50 signifie que le test est mauvais et qu'il ne fait pas mieux que la chance pour classer les individus. Plus l'aire sous la courbe est élevée, plus le test est performant. Lorsque l'aire sous la courbe vaut 0,5, alors il n'y a pas de discrimination.

Si elle est comprise entre 0,7 et 0,8, alors la discrimination est acceptable. Si, par contre, elle comprise entre 0,8 et 0,9, alors la discrimination est excellente et si elle supérieure à 0,9, la discrimination est exceptionnelle.

La **figure 12** montre la courbe ROC obtenue à partir des résultats d'un modèle logistique saturé. Ce modèle a un pouvoir prédictif, avec une discrimination acceptable donnée par l'aire sous la courbe (AUC) égale à 0,81 (**Figure 12**). Cela montre que la discrimination entre les patients qui font un bon contrôle glycémique et ceux qui font un mauvais contrôle glycémique est excellente, une étude réalisée chez les patients admis au service de Médecine Interne de l'Hôpital Militaire de Kamenge. Comme conséquence, la prédiction est possible. La méthode bootstrap a permis de ré-échantillonner 100 fois l'échantillon pour construire des intervalles de confiance bootstrap. L'indice de Youden est de l'ordre de 0,53.

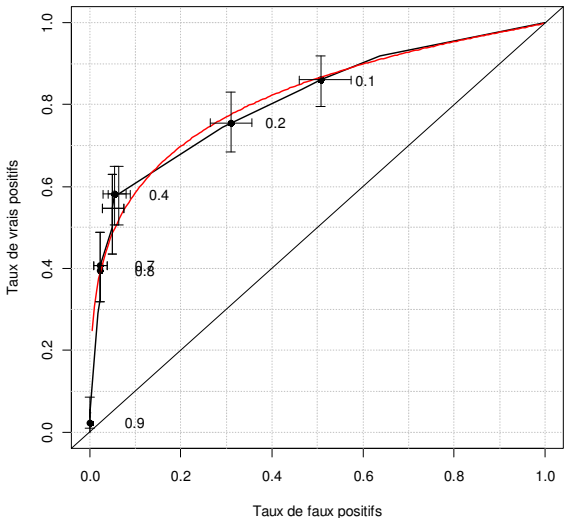


Figure 12 : Courbe ROC et intervalles de confiance bootstrap

Source : Pr BARANKANIRA Emmanuel

Chapitre 4. Modèle linéaire mixte

4.1. Introduction

Le modèle linéaire général à effets fixes et sous la forme matricielle s'écrit :

$$Y = X\beta + \varepsilon \tag{4.1}$$

avec $Y : n \times 1$ la variable dépendante, $X : n \times p$ la matrice de design, $\beta : p \times 1$ le vecteur des paramètres et $\varepsilon : n \times 1$ le vecteur des erreurs qui suit une loi normale de moyenne nulle et de variance constante :

$$\varepsilon \sim N_n\left(0, \sigma^2 I_n\right) \tag{4.2}$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} \tag{4.3}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & x_{i2} & & x_{ij} & \cdots & x_{i,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{n,p-1} \end{pmatrix} \tag{4.4}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_{p-1} \end{pmatrix} \tag{4.5}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim N_n\left(0, \Gamma_\theta\right) \tag{4.6}$$

et $\theta \in \mathbb{R}^k$ un paramètre inconnu.

4.2. Estimateur des moindres carrés ordinaires

L'estimateur des moindres carrés ordinaires des paramètres est [6] :

$$\hat{\beta}_{MV} = \left(X^t X\right)^{-1} X^t Y \tag{4.7}$$

Dans le cas où $\Gamma_\theta = \sigma^2 I_n$, alors l'estimateur de la variance des paramètres est :

$$\sigma^2 = \frac{1}{n-p}\left\|Y - X\hat{\beta}\right\|^2 \tag{4.8}$$

Le facteur $\left\|Y - X\hat{\beta}\right\|^2$ se développe comme suit [6] :

$$\left\|Y - X\hat{\beta}\right\|^2 = \left\|Y - X\left(X^t X\right)^{-1} X^t Y\right\|^2 = \left\|\left[Id - X\left(X^t X\right)^{-1} X^t\right]Y\right\|^2 = \|MY\|^2 = Y^t M^t M Y \tag{4.9}$$

avec $M = Id - X\left(X^t X\right)^{-1} X^t$ un projecteur sur l'espace orthogonal, c'est-à-dire une matrice symétrique ($M = M^t$), idempotente ($MM = M$) et :

$$MX = 0 \tag{4.10}$$

Il vient alors :

$$\left\|Y - X\hat{\beta}\right\|^2 = Y^t M Y = \left(X\beta + \varepsilon\right)^t M \left(X\beta + \varepsilon\right) = \varepsilon^t M \varepsilon \tag{4.11}$$

et alors :

$$E\left\{\left\|Y - X\hat{\beta}\right\|^2\right\} = E\left(\varepsilon^t M \varepsilon\right) = \sigma^2 tr\left(M\right) = \left(n - p\right)\sigma^2 \tag{4.12}$$

Le modèle linéaire classique suppose la normalité des erreurs ou de la réponse. Autrement dit, le modèle suppose que les erreurs suivent une loi normale de moyenne nulle et de matrice de variances-covariances $\sigma_i^2 \Gamma$ avec Γ connue, par exemple une matrice identité I_n d'ordre n. Si $\Gamma = I_n$, l'hypothèse d'homoskédasticité des erreurs est vérifiée. Cependant, il est possible d'ajouter des effets aléatoires au modèle ou de considérer que les variances ne sont pas homogènes. Dans ce genre de situations, le modèle linéaire mixte devient adapté pour modéliser le phénomène d'intérêt.

4.3. Spécification du modèle linéaire mixte

Le modèle linéaire général de l'équation (4.1) s'écrit [2] :

$$Y = X\beta + ZU + \varepsilon \tag{4.13}$$

avec $Y : n \times 1$ le vecteurs des observations, $X : n \times p$ la matrice des effets fixes, $\beta : p \times 1$ le vecteur des paramètres des effets fixes, $Z : n \times k$ la matrice des effets aléatoires, $U : k \times 1$ le vecteur des paramètres des effets aléatoires et $\varepsilon : n \times 1$ le vecteur des erreurs.

Les vecteurs U et ε sont supposés suivre une loi normale multivariée tandis que l'espérance mathématique de ces vecteurs doit être nulle :

$$E\left[\begin{pmatrix} U \\ \varepsilon \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{4.14}$$

La matrice de variances-covariances doit être une matrice diagonale dont les éléments diagonaux sont des matrices symétriques :

$$Var\left[\begin{pmatrix} U \\ \varepsilon \end{pmatrix}\right] = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \tag{4.15}$$

4.4. Estimateur du maximum de vraisemblance

Selon Christian Lavergne et Catherine Trottier, la vraisemblance ou densité jointe du modèle à variance paramétrée et séparable ($\beta \perp \theta$) s'écrit :

$$f_{MV}(\beta,\theta;y) = \frac{1}{(2\pi)^{\frac{n}{2}}} \left|\Gamma_\theta\right|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(y-X\beta)^t \Gamma_\theta^{-1}(y-X\beta)\right\} \tag{4.16}$$

La déviance (-2×log-vraisemblance) du modèle vaut :

$$l_{MV}(\beta,\theta) = -2\,log\left[f_{MV}(\beta,\theta;y)\right] \tag{4.17}$$

Les estimateurs du maximum de vraisemblance des paramètres sont :

$$\hat{\beta}_{MV} = \underset{\beta}{arg\,min}\left\{(y-X\beta)^t \Gamma_\theta^{-1}(y-X\beta)\right\} \tag{4.18}$$

$$\hat{\theta}_{MV} = \underset{\theta}{arg\,min}\left\{(y-X\beta)^t \Gamma_\theta^{-1}(y-X\beta)\right\} + log\left(\left|\Gamma_\theta\right|\right) \tag{4.19}$$

La fonction score est :

$$U_\beta\left(\hat{\theta}_{MV}\right) = \frac{\partial l(\beta,\theta)}{\partial\beta}\bigg|_{\theta=\hat{\theta}_{MV}} = -2X^t\Gamma_\theta^{-1}(y-X\beta)\bigg|_{\theta=\hat{\theta}_{MV}} \tag{4.20}$$

Il vient alors [6] :

$$\hat{\beta}_{MV} = \left(X^t\Gamma_{\hat{\theta}_{MV}}^{-1}X\right)^{-1} X^t\Gamma_{\hat{\theta}_{MV}}^{-1}Y \tag{4.21}$$

De plus

$$\begin{aligned} U_\theta^{MV}\left(\hat{\beta}_{MV}\right) &= \frac{\partial l(\beta,\theta)}{\partial\theta}\bigg|_{\beta=\hat{\beta}_{MV}} \\ &= -(y-X\beta)^t\Gamma_\theta^{-1}\frac{\partial\Gamma_\theta}{\partial\theta}(y-X\beta) + tr\left(\Gamma_\theta^{-1}\frac{\partial\Gamma_\theta}{\partial\theta}\right)\bigg|_{\beta=\hat{\beta}_{MV}} \\ &= -y^t P_\theta\frac{\partial\Gamma_\theta}{\partial\theta}P_\theta y + tr\left(\Gamma_\theta^{-1}\frac{\partial\Gamma_\theta}{\partial\theta}\right) \end{aligned} \tag{4.22}$$

avec

$$P_\theta = \Gamma_\theta^{-1}\left[Id - X\left(X^t \Gamma_\theta^{-1} X\right)^{-1} X^t \Gamma_\theta^{-1}\right] \tag{4.23}$$

Dans le cas où $Var(\varepsilon) = \sigma^2 I_n$, avec Γ une matrice symétrique, définie positive et donc diagonalisable à valeurs propres positives :

$$\Gamma = U^t \Lambda U \; avec \; U^t U = Id \tag{4.24}$$

Posons $\sqrt{\Lambda}$ la matrice diagonale des racines carrées des valeurs propres de Γ et faisons un changement de variable $Z = HY = HX\beta + \varepsilon_z$ avec $H = \sqrt{\Lambda^{-1}} U$, alors :

$$Var(Z) = \sigma^2 Id \tag{4.25}$$

En posant :

$$\gamma = (\beta, \theta) \tag{4.26}$$

la matrice d'information s'écrit :

$$\begin{aligned} I_{\beta,\theta} &= E_{\beta,\theta}\left[-\frac{\partial^2 log\, f(\beta,\theta; y)}{\partial\gamma \partial\gamma^t} \right] \\ &= E_{\beta,\theta}\left[\frac{1}{2} \frac{\partial^2 l}{\partial\gamma \partial\gamma^t} \right] \\ &= E_{\beta,\theta}\left[\frac{1}{2} \begin{pmatrix} \frac{\partial U_\beta}{\partial \beta^t} & \frac{\partial U_\beta}{\partial \theta^t} \\ \frac{\partial U_\theta^{MV}}{\partial \beta^t} & \frac{\partial U_\theta^{MV}}{\partial \theta^t} \end{pmatrix} \right] \end{aligned} \tag{4.27}$$

Il est possible de vérifier que :

$$\begin{cases} \dfrac{\partial U_\beta}{\partial \beta^t} = 2 X^t \Gamma_\theta^{-1} X \\ E_{\beta,\theta}\left(\dfrac{\partial U_\beta}{\partial \theta^t} \right) = 0 \end{cases} \tag{4.28}$$

et

$$\begin{aligned}\left[\frac{\partial U_\theta^{MV}}{\partial \theta^t}\right]_{ij} &= tr\left[\Gamma_\theta^{-1}\left(\frac{\partial^2 \Gamma_\theta}{\partial\theta_i\partial\theta_j}+\frac{\partial\Gamma_\theta}{\partial\theta_i}\Gamma_\theta^{-1}\frac{\partial\Gamma_\theta}{\partial\theta_j}\right)\right] \\ &\quad -(y-X\beta)^t\Gamma_\theta^{-1}\left(\frac{\partial^2\Gamma_\theta}{\partial\theta_i\partial\theta_j}-2\frac{\partial\Gamma_\theta}{\partial\theta_i}\Gamma_\theta^{-1}\frac{\partial\Gamma_\theta}{\partial\theta_j}\right)\Gamma_\theta^{-1}(y-X\beta)^t\end{aligned} \tag{4.29}$$

D'où :

$$I_{\beta,\theta}=\begin{pmatrix} I_\beta = X^t\Gamma_\theta^{-1}X & 0 \\ 0 & I_\theta^{MV}=\left[\frac{1}{2}tr\left(\Gamma_\theta^{-1}\frac{\partial\Gamma_\theta}{\partial\theta_i}\Gamma_\theta^{-1}\frac{\partial\Gamma_\theta}{\partial\theta_j}\right)\right]_{ij=1,2,\ldots,K}\end{pmatrix} \tag{4.30}$$

4.5. Estimateur du maximum de vraisemblance restreinte

La vraisemblance restreinte ou vraisemblance marginale après intégration sur le paramètre β vaut :

$$f_{RE}(\theta;y)=\frac{1}{(2\pi)^{\frac{n-p}{2}}}\left|\Gamma_\theta\right|^{-\frac{1}{2}}\left|X^t\Gamma_\theta^{-1}X\right|^{-\frac{1}{2}}exp\left\{-\frac{1}{2}y^tP_\theta y\right\} \tag{4.31}$$

La déviance (-2×log-vraisemblance) du modèle vaut :

$$l_{RE}(\beta,\theta)=-2log\left[f_{RE}(\theta;y)\right] \tag{4.32}$$

L'estimateur du maximum de vraisemblance restreinte des paramètres est :

$$\hat{\beta}_{MVR}=\arg\min_\theta\left\{y^tP_\theta y+log\left(\left|\Gamma_\theta\right|\right)+log\left(\left|X^t\Gamma_\theta^{-1}X\right|\right)\right\} \tag{4.33}$$

La fonction score est :

$$\begin{aligned}U_\theta^{MVR} &= \left.\frac{\partial l_{MVR}(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_{MV}} \\ &= -y^tP_\theta\frac{\partial\Gamma_\theta}{\partial\theta}P_\theta y+tr\left(\Gamma_\theta^{-1}\frac{\partial\Gamma_\theta}{\partial\theta}\right)-tr\left[\left(X^t\frac{\partial\Gamma_\theta}{\partial\theta}X\right)^{-1}X^t\Gamma_\theta^{-1}\frac{\partial\Gamma_\theta}{\partial\theta}X\right]\end{aligned} \tag{4.34}$$

Il vient alors que $\hat{\beta}_{MVR}$ est la solution de $U_\beta\left(\hat{\theta}_{MVR}\right)=0$.

La matrice d'information s'écrit :

$$\begin{aligned} I_\theta^{MVR} &= E_\theta\left[-\frac{\partial^2 log\ f_{RE}(\theta; y)}{\partial\theta\partial\theta^t}\right] \\ &= E_\theta\left[\frac{1}{2}\frac{\partial^2 l_{RE}(\theta)}{\partial\theta\partial\theta^t}\right] \\ &= E_\theta\left[\frac{1}{2}\frac{\partial U_\theta^{RE}}{\partial\theta^t}\right] \\ &= \left[\frac{1}{2}tr\left(P_\theta\frac{\partial\Gamma_\theta}{\partial\theta_i}P_\theta\frac{\partial\Gamma_\theta}{\partial\theta_j}\right)\right]_{ij=1,2,...,K} \end{aligned} \tag{4.35}$$

Il est à noter que $\hat{\theta}_{MV}$ est la solution de la minimisation du critère :

$$Critère_{MV} = y^t P_\theta y + log\left(\left|\Gamma_\theta\right|\right) \tag{4.36.a}$$

et $\hat{\theta}_{MVR}$ est la solution de la minimisation du critère :

$$Critère_{MVR} = Critère_{MV} + log\left(\left|X^t\Gamma_\theta^{-1}X\right|\right) \tag{4.36.b}$$

et

$$\hat{\beta}_{MVR} = \left(X^t\Gamma_{\hat{\theta}}^{-1}X\right)^{-1}X^t\Gamma_{\hat{\theta}}^{-1}Y \tag{4.37}$$

De plus

$$E_\theta\left(y^t P_\theta\frac{\partial\Gamma_\theta}{\partial\theta}P_\theta y\right) = tr\left(P_\theta\frac{\partial\Gamma_\theta}{\partial\theta}\right) \tag{4.38}$$

Puisque

$$E_\theta\left(P_\theta y\right) = 0 \tag{4.39}$$

et

$$P_\theta\Gamma_\theta P_\theta = P_\theta \tag{4.40}$$

La fonction score U_θ^{MVR} est une statistique centrée alors que la fonction score U_θ^{MV} est une statistique non centrée.

4.6. Choix des modèles

Les critères à minimiser sont principalement le Critère de l'Information d'Akaike (AIC pour Akaike Information Criterion) et le Critère de l'Information Bayesienne (BIC pour Bayesian Information Criterion) donnés respectivement par :

$$AIC(M) = -2\log L + 2q(M) \tag{4.41}$$

$$BIC(M) = -2\log L + \log(n) \times q(M) \tag{4.42}$$

où M est le modèle, L la vraisemblance (*likelihood* en anglais), $q(M)$ le nombre de paramètres du modèle, n le nombre d'observations avec $q(M) = p + K$. Tous ces critères sont construits en minimisant la log-vraisemblance en $\hat{\beta}$ et $\hat{\theta}$. Le lecteur pourra compléter cet ECUE par des recherches approfondies sur des aspects non visités ici tels que la régression biaisée (régression de ridge, régression lasso, régression elastic net, régression sur composantes principales, régression aux moindres carrés partiels) et la régression bayesienne.

Références bibliographiques

1. Cornillon P-A, Matzner-Løber É. Régression: théorie et applications. Paris Berlin Heidelberg [etc.]: Springer; 2006. 314 p. (Statistiques et probabilités appliquées).

2. Faraway JJ. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition. 2nd ed. Boca Raton: CRC Press; 2016. 441 p. (Chapman & Hall / CRC Texts in Statistical Science).

3. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Third edition. Hoboken, New Jersey: Wiley; 2013. 508 p. (Wiley series in probability and statistics).

4. Wasserman L. All of statistics: a concise course in statistical inference. New York: Springer; 2004. 442 p. (Springer texts in statistics).

5. Lejeune M. Statistique: la théorie et ses applications. Deuxième éd. avec exercices corrigés. Paris Berlin Heidelberg [etc.]: Springer; 2010. 446 p. (Statistique et probabilités appliquées).

6. Magnus JR, Neudecker H. Matrix differential calculus with applications in statistics and econometrics. Rev. ed., reprinted. Chichester: John Wiley; 2002. 395 p. (Wiley series in probability and statistics).