Université Hassan 1er
Faculté des Sciences et Techniques
SETTAT

Licence Sciences et Techniques

MATHEMATIQUES APPLIQUEES

Projet de Fin d'Etudes

Régression dans l'analyse des données massives

Présenté par : NAKHLA Khaoula & TINA Djara Olivier

Soutenu le : 14 juillet 2021
Devant les membres de jury :

M. R Eljid : Professeur à la FST de Settat — Examinateur

M. M HANINI : Professeur à la FST de Settat — Examinateur

M. A BENTAHAR : Professeur à la FST de Settat

Encadrant

Année Universitaire : 2020-2021

Dédicace

À nos chers parents,
A nos amis et à toutes nos familles.

Remerciements

Au nom de DIEU le plus clément et le plus Miséricordieux. Tout d'abord, je remercie DIEU le tout puissant de nous avoir donné le courage, la volonté, la patiente, et la chance de suivre le chemin de la science.

Nons souhaitons avant tout remercier notre encadrant Pr. BEN TAHAR pour ce sujet qui a suscité vraiment notre intérêt. Nous le remercie pour le temps qu'il a consacré pour notre encadrement, pour sa patience et surtout ses judicieux conseils qui ont contribué à alimenter notre réflexion sur ce sujet et qui sans eux ce mémoire n'aurait jamais vu le jour.

Nous remercions également toute l'équipe pédagogique de la FST de Settat et les intervenants professionnels responsables de notre formation, spécialement ceux de la filière LST-MA.

Nos plus vifs remerciements vont aussi aux membres de jury pour avoir accepté de juger notre présent travail.

Notre profonde gratitude va également à nos chers parents ; surtout pour leur patience et leurs encouragements constants. Vous étiez notre soutien quotidien.

Merci enfin à nos amis pour leurs sincère amitié et confiance.

Table des matières

Introduction générale

Notre présent est le témoin candide du proverbe "Un danger prévu est à moitié évité",les statistiques et ses éléments d'études jouent le rôle principal dans cette prévention en utilisant des méthodes théoriques et d'autres pratiques qui se ramènent toutes au même objectif, **La régression dans l'analyse des données massives** est l'un de ces outils.

En statistiques, un modèle de régression linéaire est un modèle de régression d'une variable expliquée sur une variable explicative(cas de la régression linéaire simple) ou plusieurs variables explicatives(cas de la régression linéaire multiple) dans lequel l'hypothèse que la fonction qui relie les variables explicatives à la variable expliquée est linéaire dans ces paramètres. Certaines méthodes, comme la régression logistique, sont à la fois des méthodes de régression au sens où il s'agit de prédire la probabilité d'appartenir à chacune des classes et des méthodes de classification.

L'analyse en composantes principales est utilisée pour réduire p variables corrélées en un nombre q de variables non corrélées de telles manières que les q variables soient des combinaisons linéaires des p variables initiales, que leur variance soit maximale et que les nouvelles variables soient orthogonales entre elles suivant une distance particulière. Elle est une méthode d'analyse des données. L'analyse des données permet de traiter un nombre très important de données et de dégager les aspects les plus intéressants de la structure de celles-ci.

Ainsi la régression et l'analyse en composante principale sont deux moyens puissants à la disposition du scientifique des données, il reste à savoir lequel utilisé quand et où ?

L'objectif de notre projet de fin d'étude est montrer que la modélisation d'une masse colossale d'information faisant intervenir plusieurs variables peu se raméner à une simple étude de quelques variables intéressantes d'écrivant le phénomène étudié. Ainsi on montre que l'ACP et la régression font bon ménage dans l'analyse des données massives.

Notre projet de fin d'étude est composé de six chapitres :

Chapitre 1 : Rappel

Ce chapitre est consacré aux rappels des probabilités, variables aléatoires, leurs caractéristiques et quelques propriétés.

Chapitre 2 : Analyse en Composantes Principales

Dans ce chapitre, on va décrire de manière succincte les liaisons entres les variables et les ressemblances entres les individus. On parlera également des étapes essentiels pour mener à bien l'ACP comme la transformation des données, la recherches des compo-

santes principales.

Chapitre 3 : Régression linéaire simple

Ce chapitre rappelle brièvement la modélisation de la régression linéaire tout en expliquant les hypothèses du modèle ainsi que les notions d'estimation des paramètres par les moindres carrées, les test de signification globale et paramétrique.

Chapitre 4 : Régression linéaire multiple

Ce chapitre généralise naturellement la régression simple. On cherche à expliquer la variable dépendante Y à l'aide de p variables indépendantes $X_1, \ldots, X_p$. Les changements majeurs se voient essentiellement dans l'écriture matricielle des estimateurs et de leur variances. Dans la pratique, on a besoin d'outils informatiques pour inverser ces matrices de grandes tailles.

Chapitre 5 : Régression logistique

Dans ce chapitre, nous présentons le modèle de régression logistique qui est un modèle de régression binomiale servant à modéliser la relation entre une variable qualitative endogène et des variables quantitatives exogènes X. On estime les paramètres de ce modèle par la méthodes du maximum de vraisemblance, puis on teste les coefficients par l'odd ratio(mesure de l'association entre la variable expliquée et les variables explicatives.

Chapitre 6 : Application sur le logiciel R

Comme toute étude théorique a la chance d'être applicable aux contextes fonctionnels, ce chapitre est dédié à l'application de la régression multiple, précédé de l'analyse univaiée et l'ACP sur la nocivité du tabac. L'objectif est d'établir une rélation entre la teneur en monoxyde carbone (CO) et la quantité de goudron(TAR), de NICOTINE ainsi que la masse(POIDS) en utilisant le logiciel R.

1 Rappel

1.1 Loi normale

C'est une loi très importante pour plusieurs raisons, tout d'abord elle apparaît dans de nombreux problèmes courants (pour les modéliser) et bien souvent, on peut approcher une loi par une loi normale. De plus, on dispose de la table de ses valeurs à laquelle on se réfère pour des calculs approchés.
Synonymes pour cette loi : loi gaussienne, loi de Gauss.

Définition 1.1.1
La loi normale standard $N(0,1)$ *est celle de densité*

$$f_{0,1}(t) = \frac{1}{\sqrt{2\pi}} \exp \frac{-t^2}{2}$$

Son espérance est $\mathbb{E}[X] = 0$ *. Sa variance est* $Var(X) = 1$.

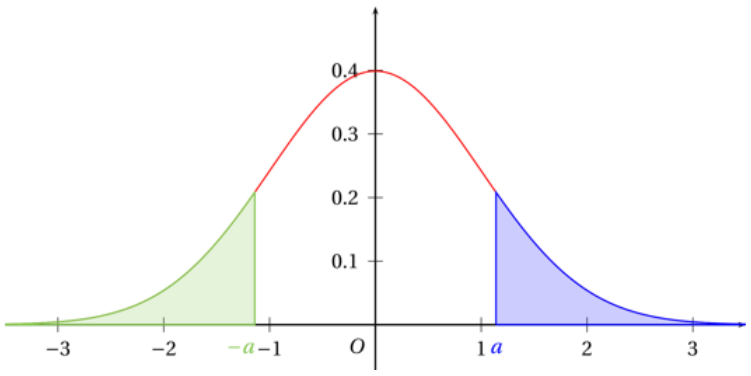


FIGURE 1.1 – Densité d'une variable normale centrée réduite

Définition 1.1.2
On dit que la v.a. X suit une loi normale $N(\mu, \sigma^2)$ si elle a pour densité la fonction :

$$f_{\mu,\sigma}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(-t-\mu)^2}{2\sigma^2} \tag{1.1}$$

Son espérance est $\mathbb{E}[X] = \mu$. Sa variance est $Var(X) = \sigma^2$.

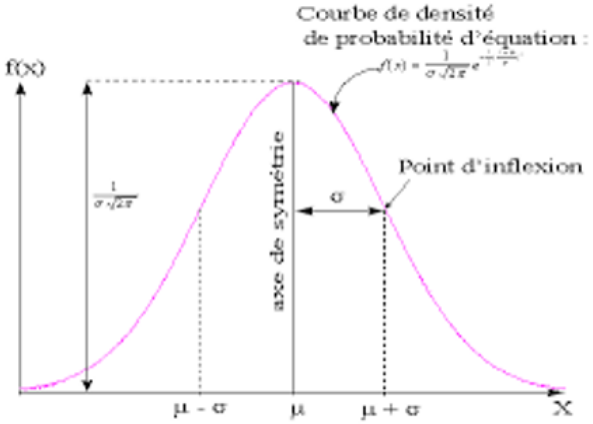


FIGURE 1.2 – Densité d'une variable normale

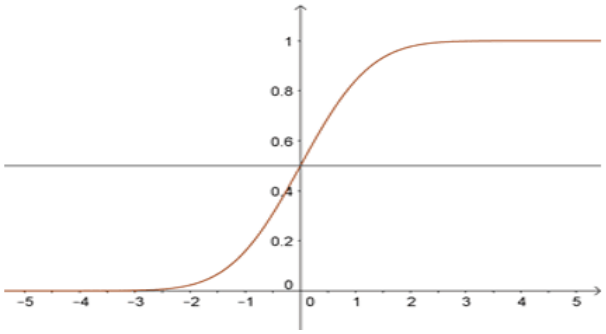


FIGURE 1.3 – Fonction de répartition d'une variable normale

Remarque
Cette loi est fondamentale en théorie des probabilités et en statistique : c'est la loi limite de la moyenne dans une suite infinie d'épreuves répétées indépendantes. En pratique elle sert à modéliser les effets additifs de petits phénomènes aléatoires indépendants répétés souvent.
On dit que la v.a. X suit une loi normale $N(\mu, \sigma^2)$ si elle a pour densité la fonction

1.1.1 Régles pour les lois normales

Si $X \sim N(\mu, \sigma^2)$ et $a \in \mathbb{R}$ alors $aX \sim N(a\mu, a\sigma^2)$, Et Quand on somme des v.a. gaussiennes indépendantes de loi $N(\mu_1, \sigma_1^2)$ et $N(\mu_2, \sigma_2^2)$,on obtient une v.a. gaussienne avec pour paramétres la somme des paramétres $N(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$,Et $X_1 \sim N(\mu_1, \sigma_1^2) \perp X_2 \sim N(\mu_2, \sigma_2^2) \implies X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
Plus généralement quand $X_1, ..., X_n$ n sont n v.a. indépendante de lois $N(m, \sigma_2^2)$, alors

$$\frac{X_1 + ... + X_n}{n} \sim N(\mu, \frac{\sigma^2}{n}).$$

Notez encore qu'on peut facilement passer d'une loi normale à la loi standard.

Propriété 1.1.1
Si la v.a. X suit une loi $N(\mu, \sigma^2)$, alors $Y : z = \frac{X-\mu}{\sigma}$ suit la loi $N(0, 1)$.
La v.a. Y s'appelle la v.a. centrée réduite associée X. En fait, pour faire des calculs effectifs de probabilité, grâce à ce résultat, on commencera systématiquement par se ramener d'une loi normale quelconque $N(\mu, \sigma^2)$ à la loi normale standard $N(0, 1)$. On pourra alors utiliser la table des valeurs pour cette loi.

1.1.2 Théorème de la limite centrale

Le théorème de la limite centrale donne la vitesse à laquelle la convergence a lieu dans loi forte des grands nombres, sous l'hypothèse supplémentaire d'intégrabilité des X_j^2 :

Théorème 1.1.1
Soit $(X_j)_{j\geq 1}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées telles que $\mathbb{E}(X_1^2) < \infty$ et $\sigma = Var(X1) > 0$.
On note $\bar{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$ la moyenne empirique.Alors pour $n \to \infty$.

$$\forall x \in \mathbb{R} \quad \lim_{n\to\infty} P\left[\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \leq x\right] = \frac{1}{\sqrt{2\pi}}\int_\infty^x \exp\frac{-z^2}{2}dz$$

Cette convergence en loi est illustrée par la figure 5.3.

Démonstration
On utilise la fonction caractéristique. Soit $u \in \mathbb{R}$.

$$\Phi_{\frac{\sqrt{n}}{\sigma}(\bar{X_n} - \mathbb{E}(X_1))}(u) = \mathbb{E}\left(\exp iu\frac{1}{\sigma\sqrt{n}}\sum = 1^n(X_j - \mathbb{E}(X_j))\right) \tag{1.2}$$

$$= \prod_{j=1}^n \mathbb{E}\left(\exp iu\frac{1}{\sigma\sqrt{n}}(X_j - \mathbb{E}(X_j))\right) \tag{1.3}$$

$$= \left(\mathbb{E}\left(\exp iu\frac{1}{\sigma\sqrt{n}}(X_j - \mathbb{E}(X_j)\right)\right)^n \tag{1.4}$$

$$= \left(\Phi_{X_i - \mathbb{E}(X_i)}\left(\frac{u}{\sigma\sqrt{n}}\right)\right)^n \tag{1.5}$$

Le passage de (1.3) à (1.4) se fait grâce à l'indépendance des X_j.
Le passage de (1.4) à (1.5) se fait car les X_j ont la même loi.
Le dernier passage se fait car les X_j ont même loi.
Comme $\mathbb{E}(X_1 - \mathbb{E}(X_1)) = 0$ et $\mathbb{E}((X_1 - \mathbb{E}(X_1))^2) = \sigma^2$, pour v au voisinage de 0,

$$\Phi_{X_i - \mathbb{E}(X_i)}(v) = 1 - \frac{\sigma^2}{2}v^2 + o(v^2).$$

Donc pour n grand,

$$\Phi_{X_i - \mathbb{E}(X_i)}\left(\frac{u}{\sigma\sqrt{n}}\right) = 1 - \frac{u^2}{2n} + o(\frac{1}{n}).$$

Ainsi

$$\Phi_{\frac{\sqrt{n}}{\sigma}(\bar{X_n} - \mathbb{E}(X_1))}(u) = \left(1 - \frac{u^2}{2n} + o(\frac{1}{n})\right)^2 \longrightarrow n \to \infty \exp\frac{-u^2}{2} = \Phi_Y(u).$$

1.2 Vecteurs aléatoires

1.2.1 Généralités

Un vecteur aléatoire de variables aléatoires continues $X = (X_1, \ldots, X_p)$ est un vecteur colonne de variables aléatoires possédant une fonction de densité jointe notée $f(x_1, ..., x_p)$ telle que

$$\int_{\mathbb{R}^p} f(x_1, ..., x_p)dx_1...dx_p = 1.$$

L'espérance mathématique d'un vecteur aléatoire est le vecteur des espérances de ces composantes :

$$\mathbb{E}(X) = \mu = \mathbb{E}\begin{pmatrix} X_1 \\ \vdots \\ X_j \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_j) \\ \vdots \\ \mathbb{E}(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_j \\ \vdots \\ \mu_p \end{pmatrix}$$

La variance(ou matrice variance-covariance) d'un vecteur aléatoire de $\mathbb{R}^p$ est définie par :

$$var(X) = \mathbb{E}[(XX')] - \mu\mu' = \begin{pmatrix} var(X_1) & \cdots & cov(X_1, X_j) & \cdots & cov(X_1, X_p) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ cov(X_1, X_j) & \cdots & var(X_j) & \cdots & cov(X_j, X_p) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ cov(X_1, X_p) & \cdots & cov(X_j, X_p) & \cdots & var(X_p) \end{pmatrix}$$

De manière générale, si X est un vecteur de variables aléatoires de moyenne μ et de matrice variance covariance σ' si $A = [aij]$ est une matrice $q \times p$ de constantes, alors

$$\mathbb{E}(AX) = \mathbb{E}\begin{pmatrix} \sum_{j=1}^{p} a_{1j}X_j \\ \vdots \\ \sum_{j=1}^{p} a_{ij}X_j \\ \vdots \\ \sum_{j=1}^{p} a_{qj}X_j \end{pmatrix} = \begin{pmatrix} \mathbb{E}(\sum_{j=1}^{p} a_{1j}X_j) \\ \vdots \\ \mathbb{E}(\sum_{j=1}^{p} a_{ij}X_j) \\ \vdots \\ \mathbb{E}(\sum_{j=1}^{p} a_{qj}X_j) \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{p} a_{1j}\mathbb{E}(X_j) \\ \vdots \\ \sum_{j=1}^{p} a_{ij}\mathbb{E}(X_j) \\ \vdots \\ \sum_{j=1}^{p} a_{1j}\mathbb{E}(X_j) \end{pmatrix} = A\mathbb{E}(X) = A\mu.$$

$$\begin{aligned} var(X) &= \mathbb{E}([AX - \mathbb{E}(AX)][AX - \mathbb{E}(AX)]') \\ &= \mathbb{E}([AX - \mathbb{E}(AX)][A'X' - \mathbb{E}(A'X')]) \\ &= \mathbb{E}([AX - A\mathbb{E}(X)][X'A' - A'\mathbb{E}(X')]) \\ &= A\mathbb{E}([X - \mathbb{E}(X)] - [X' - \mathbb{E}(X')])A' \\ &= Avar(X)A' \\ &= A\sigma'A' \end{aligned}$$

1.3 Variable normale multivariée

Le vecteur de variables aléatoires $X = (X_1, ..., X_p)'$ a une distribution normale mutivariée de moyenne $\mu = (\mu_1, ..., \mu_p)'$ et de matrice variance-covariance σ' (on suppose par simplicité que σ' est de plein rang), si sa fonction de densité est donnée par

$$f_X(x) = \frac{1}{(2\pi)^{p/2}|\sigma'|^{1/2}} exp\Big[-\frac{1}{2}(X - \mu)'(\sigma')^{-1}(X - \mu)\Big], \tag{1.6}$$

pour tout $x \in \mathbb{R}^p$.

Remarque

Si $p = 1$, on retrouve l'expression (1.1), si $p = 2$, on retrouve l'expression de f pour la loi normale bivariée.

Théorème 1.3.1

Si la matrice variance-covariance d'un vecteur multinational est diagonale, alors les variables aléatoires de ce vecteur sont indépendantes.

Démonstration

Si la matrice variance covariance on peut s'écrire $\sigma' = diag(\sigma_1^2, ..., \sigma_p^2)$,toutes les com-

posantes du vecteur X sont non-corrélées. Dans ce cas,

$$\begin{aligned} f_X(x) &= \frac{1}{(2\pi)^{p/2}|\sigma'|^{1/2}} exp\Big[-\frac{1}{2}(X-\mu)'(\sigma')^{-1}(X-\mu)\Big] \\ &= \frac{1}{(2\pi)^{p/2}\Big(\prod_{j=1}^{p}\sigma_j\Big)^{1/2}} exp\Big[-\frac{1}{2}(X-\mu)'(\sigma')^{-1}(X-\mu)\Big] \\ &= \frac{1}{(2\pi)^{p/2}\prod_{j=1}^{p}\sigma_j} exp\Big[-\sum_{j=1}^{p}\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\Big] \\ &= \prod_{j=1}^{p}\frac{1}{(2\pi)^{p/2}\sigma_j} exp\Big[-\sum_{j=1}^{p}\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\Big] \\ &= \prod_{j=1}^{p} f_{X_j}(x_j), \end{aligned}$$

où

$$f_{X_j}(x_j) = \frac{1}{(2\pi)^{p/2}\sigma_j} exp\Big[-\sum_{j=1}^{p}\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\Big],$$

est la densité de la variables X_j.
Dans le cas multinormal (et seulement dans ce cas), l'absence de corrélation implique donc l'indépendance des variables aléatoires.
Un résultat particulièrement est le suivant :

Propriété 1.3.1
Toute combinaison linéaire d'un vecteur de variables aléatoires normales est normale. Dpnc si X est un vecteur multinormal de moyenne μ et de matrice de variance covariance σ' et si A est une matrice $q \times p$ de constantes, alors on écrit

$$X = N(\mu, \sigma'),$$

et on a

$$AX = N(A\mu, A\sigma'A'),$$

Comme une projection est une combinaison linéaire, on a que :

Propriétés
Toute propriétés d'un vecteur des variables aléatoires normales est normale.

1.4 Loi dérivées de la loi normale

1.4.1 Loi de khi-deux

Soit $(X_1, ..., X_n)$ des variables aléatoires i.i.d. suivant une loi normale centrée réduite. La loi de la variable $X = \sum = 1^n X_i^2$ est appelée loi du X^2 à n degrés de liberté(ddl),noté

$X \sim X^2 n$.
On a $\mathbb{E}(X) = n$ et $Var(X) = 2n$. Lorsque n est grand, on sait par le Théorème Central Limite que X suit approximativement une loi normale de moyenne n et de variance 2n : $X \approx N(n, 2n)$.
Ainsi, pour n grand, environ 95% des valeurs de X se situent dans l'intervalle $[n - 2\sqrt{2n}, n + 2\sqrt{2n}]$. Ceci est illustré figure 8 pour $n = 50$ ddl.

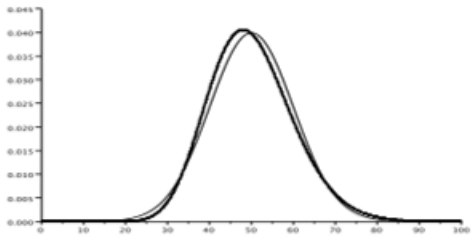


FIGURE 1.4 – Densité d'un X^2_{50}(trait gras) et densité d'une $N(50, 100)$(trait fin)

1.4.2 loi de student

Définition 1.4.1
Soit Z une variable aléatoire suivant une loi normale centrée réduite et X une variable suivant une loi du X^2 à n degrés de liberté, avec Z et X indépendantes.
La loi de la variable $T = \frac{Z}{\sqrt{\frac{X}{n}}}$ est appelée loi de Student à n degrés de liberté et on note $T \sim T_n$.

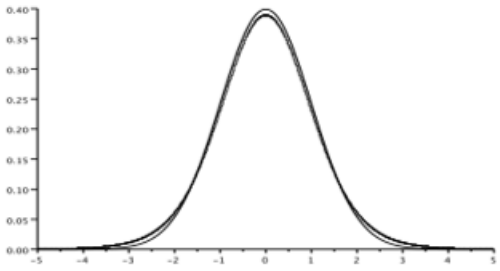


FIGURE 1.5 – Densité d'un T_{10} (trait gras) et densité d'une $N(0, 1)$ (trait fin)

Lorsque $n = 1$, T suit une loi de Cauchy et n'a donc pas d'espérance (ni, a fortiori, de variance). Pour $n = 2$, T est centrée mais de variance infinie. Pour $n \geq 3$, T est centrée et de variance $\frac{n}{n-2}$.
D'autre part, lorsque n devient grand, on sait par la Loi des Grands Nombres que le

dénominateur tend presque sûrement vers 1. De fait, on peut montrer que pour n grand, T tend en loi vers une gaussienne centrée réduite :
$T \approx N(0,1)$.
Ceci est illustré figure 1.5 pour $n = 10$ ddl. Par conséquent, lorsque n sera grand, on pourra remplacer les quantiles d'une loi de Student T_n par ceux d'une loi $N(0,1)$ (cf.tables en Annexe C.3).

1.4.3 Loi de fisher

Soit U_1 une variable aléatoire suivant une loi du X^2 à n_1 degrés de liberté et U_2 une variable aléatoire suivant une loi du X^2 à n_2 degrés de liberté, avec U_1 et U_2 indépendantes. La loi de la variable $F = \frac{\frac{U_1}{n_1}}{\frac{U_2}{n_2}}$ est appelée loi de Fisher à (n_1, n_2) degrés de liberté et on note $F \sim F_{n_2}^{n_1}$.
Pour $n_2 \succ 2$, l'espérance d'une loi de Fisher $F \sim F_{n_2}^{n_1}$ est $\frac{n_2}{n_2-2}$. Dans la suite, typiquement, n_2 sera grand, de sorte qu'à nouveau la Loi des Grands Nombres implique que $\frac{U_2}{n_2}$ tend vers 1.
Dans ce cas, F peut se voir comme un chi-deux normalisé par son degré de liberté :
$F \approx \frac{X_{n_1}^2}{n_1}$.
Ceci est illustré figure 1.6 pour $n_1 = 2$ et $n_2 = 10$.

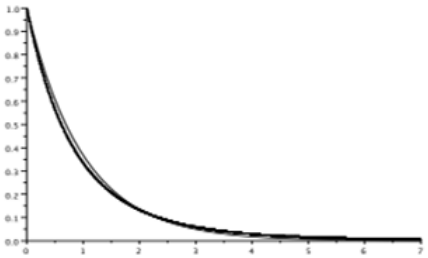
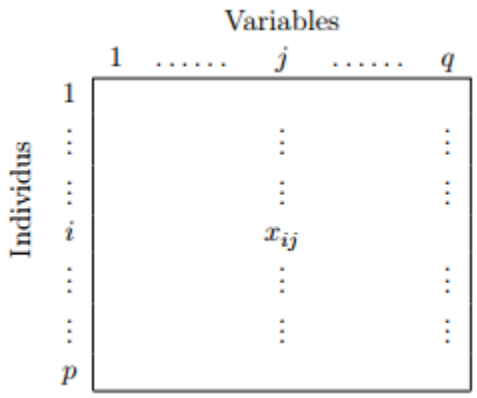


FIGURE 1.6 – Densité d'une F_{10}^2 (trait gras) et de densité d'un $\frac{X_2^2}{2}$(trait fin)

2 Analyse en Composante Principale

2.1 Données

On dispose d'un tableau de données relatives à q variables quantitatives $x_{.1}; \ldots ; x_{.q}$ portant sur p individus $x_{1.}; \ldots ; x_{p.}$. Le tableau des données X a la forme suivante :



Dans toute la suite on interprète ce tableau comme une matrice X de la forme (2.1).[1]

2.2 Principes

Définition

L'analyse en composantes principales (ACP ou PCA en anglais pour principal component analysis) est une méthode d'analyse de données et plus généralement de la statistique multi-variée, qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables qui ne sont pas corrélées les unes des autres.

Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Pour cela, le tableau de données qui contient les individus et les variables est lu sous forme de matrice.

2.2.1 Objectifs

A partir du tableau de données X, on cherche

i. Les ressemblances entre individus : les individus qui se ressemblent entre eux, et ceux qui se distinguent des autres...

ii. Les liaisons entre variables : les variables qui sont trés corrélées entre elles et celles qui ne sont pas corrélées aux autres...

Ces notions peuvent être interpréter de différentes façons, il est donc intéressant de les définir. Pour chaque $i = 1; \ldots; p$, on représente le i^{eme} individu par un point Li un point de $\mathbb{R}^q$ de composante

$$x_{i.} = (x_{i1}; ...; x_{iq})$$

L'ensemble $\{L_1, ..., L_p\}$ constitue le nuage des individus.
De même pour chaque $k = 1; ...; q$, on représente la k^{eme} variable par C_k un point de $\mathbb{R}^p$ de composante

$$x_{.k} = (x_{1k}; ...; x_{pk})$$

L'ensemble $\{C_1; ...; C_q\}$ constitue le nuage des variables

Définition 2.2.1
Deux individus se ressemblent s'ils possèdent des valeurs proches pour l'ensemble des variables c'est à dire les individus L_i et L'_i se ressemblent si $d_M(L_i; L'_i)$ est petite, où $d_M(;)$ est une M-métrique.

Définition 2.2.2
Deux variables $x_{.l}$ et $x_{.k}$ sont liées si elles ont un coéfficient de corrélation proche de 1. Le coéfficient de corrélation est données par

$$R(x_{.l}, x_{.k}) = \frac{cov(x_{.l}, x_{.k})}{\sqrt{var(x_{.k})}\sqrt{var(x_{.l})}}$$

2.2.2 Transformation des données

L'objectif de l'ACP est de rechercher pour un sous-espaces affine de dimension inférieure rapport auquel(s) le nuage NL a une inertie minimale. On sait que le meilleur sous-espace d'ajustement passe par le centre de gravité G du Nuage NL . On peut donc prendre l'origine du repère en $O = G$. Ceci revient à centrer les données, $y_{ik} = x_{ik} - \hat{x}_{ik}$. La matrice centrée Y de X est donc

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{ij} & \cdots & y_{1q} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pj} & \cdots & y_{pq} \end{pmatrix}$$

L'analyse centrée réduite ou normée pour l'ACP, qui sera présentée par la suite, est liée aussi à la transformation des données X en remplaçant les coéfficient x_{ik} par

$$z_{ik} = \frac{x_{ik} - \hat{x}_{ik}}{\sigma_{ik}}$$

On note par Z la matrice de composante z_{ik}.

2.2.3 Axes factoriels et Composantes principales

D'après le chapitre précédent pour chercher un meilleur sous espace d'ajustement du nuage NL il faut chercher une suite de vecteurs $u_1; ...; u_r$ pour fournir une représentation simplifiée de NL. Chaque vecteur us rend maximal l'inertie par rapport au centre de gravité G la projection du nuage NL sur l'axe engendré Δ_s par u_s notée ${I_{\Delta_s}}^{\perp}$. De plus ces vecteurs sont M-orthonormaux. L'inertie ${I_{\Delta_s}}^{\perp}$ est :

$${I_{\Delta_s}}^{\perp} = u_s^T MVMu_s$$

où $V = Y^T DY$ et $D = diag\{\omega_1, ..., \omega_p\}$.

Remarque

On admettra dans ce chapitre les propriétés sans les démontrer.

Propriété 2.2.1

La matrice VM est diagonalisable, ses valeurs propres sont des réels et il existe une base M-orthonormale $(u_1; \ldots; u_q)$ constituée de vecteurs propres de VM associés aux valeurs propres respectives

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

Définition 2.2.3

2.2.4 Représentations des individus

Si les projections des individus sont éloignés sur le sous espace d'ajustement, alors les points représentants ces individus sont éloignés dans l'espace. Cependant, la réciproque n'est pas forcément vraie.

Propriété 2.2.2

Si la matrice V est de rang r, alors le nuage N_L centré est inclu dans $Vect(u_1, ..., u_r)$.

Définition 2.2.4 *(Qualité de représentation)*

On note $\mathbb{U}_r = Vect(u_1, ..., u_r)$. La qualité de représentation de l'individu i sur $\mathbb{U}_r = Vect(u_1, ..., u_r)$. La qualité de représentation de l'individu i sur $\mathbb{U}_r$ est définie par

$$QLT_{\mathbb{U}_r}(i) = cos^2\theta_{y_{i\cdot},\mathbb{U}_r}$$

où $\theta_{y_{i\cdot},\mathbb{U}_r}$ est l'angle entre $y_{i\cdot}$ et $\mathbb{U}_r$.

Propriété 2.2.3
La qualité de représentation de l'individu i sur le s^{eme} axe est

$$QLT_s(i) = \frac{(F_s(i))^2}{||y_i||^2_M}$$

La qualité de représentation de l'individu i sur $\mathbb{U}_r$ est

$$QLT_{\mathbb{U}_r}(i) = \sum_{s=1}^{r} \frac{(F_s(i))^2}{||y_i||^2_M}$$

Remarque
Pour interpreter correctement la proximité des projections de deux individus sur le sous espace d'ajustement, il faut s'assurer que ces individus sont bien représentés. L'individu L_i est bien représenté sur $\mathbb{U}_r$ si la qualité $QLT_{\mathbb{U}_r}(i)$ est proche de 1.

2.2.5 Représentations des variables

Les variables $y_{.k} = (y_{1k}; \ldots; y_{pk})$ (la k^{eme} vecteur colonne de la matrice Y) de l'espace $\mathbb{R}^q$ muni de la métrique D. Pour cette métrique, la norme d'un vecteur $y.k$ est l'écart-type de la variable $y_{.k}$ et le produit scalaire entre deux vecteurs $y_{.k}$ et $y_{.l}$ est la covariance entre les deux variables $y_{.k}$ et $y_{.l}$ La composante principale F_s est un vecteur de $\mathbb{R}^p$.

Propriété 2.2.4
Supposons que le rang de la matrice V est r. Alors les composantes principales F_s, $s = 1, \ldots, r$ vérifient

- $Var(F_s) = ||F_s||^2_D = \lambda_s$
- $Cov(F_s, F_t) = \langle F_s, F_t \rangle_D = 0$ *pour* $s \neq t$
- *Soit* $v_s = \frac{F_s}{\sqrt{\lambda_s}}$ *pour* $s = 1, ..., r$. *Alors* $\{v_1, ..., v_r\}$ *est une base orthonormale pour la métrique* D_p *du* $vect(y_{.1}, ..., y_{.q}) = Img(Y)$.
- *En notant par* $G_s(k) = \langle y_{.k}, v_s \rangle_D$ *la* s^{eme} *coordonnée de* $y_{.k}$, *alors le vecteur* $G_s = (G_s(1), ..., G_s(q))$ *verifie*

$$G_s = \sqrt{\lambda_s} u_s$$

Définition 2.2.5 *(Qualité de représentation)*
La qualité de représentation de la variable $y_{.k}$ sur le sous espace $\mathbb{V}_r = Vect(v_1, ..., v_r)$ est

$$QLT_{\mathbb{V}_r(y_{.k}} = cos^2(\theta_{y_{.k}, \mathbb{V}_r})$$

où $\theta_{y_{.k}, \mathbb{V}_r}$ est l'angle entre $y_{.k}$ et $\mathbb{V}_r$.

Propriétés
La qualité de représentation de la variable $y_{.k}$ sur la s^{eme} composante principale est

$$QLT_s(y_{.k}) = R(y_{.k}, F_s)$$

La qualité de représentation de l'individu i sur $\mathbb{U}_r$ est

$$QLT_{\mathbb{U}_r}(i) = \sum_{s=1}^{r} R(y_{.k}, F_s)$$

2.2.6 Contributions aux inerties des axes

Contribution des individus à l'inertie des axes factoriels
On a vu dans le chapitre précédent que l'inertie maximale de la l'axe Δ_s sur le quel on projette le nuage des individus est

$$\lambda_s = \sum_{i=1}^{p} \omega_i F_s^2(i)$$

Définition 2.2.6
La contribution relative de l'individu i à l'inertie de l'axe s est

$$CTB_s(i) = \frac{w_i(\langle y_i, u_s\rangle)^2}{\lambda_s} = \frac{w_i(F_s(i))^2}{\lambda_s}$$

Un individu i contribue à la formation d'un axe s lorsque $CTB_s(i)$ est proche de 1 c-à-d lorsque sa projection sur cet axe sera éloignée du centre de gravité du nuage. Inversement, un individu dont la projection sur un axe sera proche du centre de gravité contribue faiblement à l'inertie portée par cette axe.

Contribution des variables à l'inertie des axes factoriels
On suppose que la matrice M est diagonale d'éléments diagonaux $m_1; ...; m_q$ strictement positifs. D'après la proposition

$$\lambda_s = \sum_{s=1}^{r} ||G_s||_M^2 = \sum_{s=1}^{r}\sum_{k=1}^{q} m_k(\langle y_{.k}, v_s\rangle_{D_p})^2$$

on peut donc énoncer la définition suivante :

Définition 2.2.7
La contribution relative de la variable $y_{.k}$ à l'inertie de l'axe r est

$$CTB_s(y_{.k}) = \frac{m_k(\langle y_{.k}, v_s\rangle_{D_p}}{\lambda_s} = m_k u_s^2(k)$$

Une variable k contribue à la formation d'un axe s lorsque $CTB_s(k)$ est proche de 1 c-à-d lorsque sa projection sur cet axe sera éloignée du centre de gravité du nuage.

2.3 Analyse en composante principale

La recherche des axes factoriels, facteurs, composantes principales d'un nuage de points dans $\mathbb{R}^p$ muni de la métrique M s'appelle Analyse en Composantes Principales (ACP).

ACP sur une matrice variance covariance

Dans ce cas on suppose que la métrique $M = I$ et les individus ont même poids $\omega_i = 1/p$, p étant le nombre d'individus. Avec ces conditions. la matrice VM est la matrice des variances covariances

$$(Cov(x_{.k}, x_{.l}))_{1\leq k,l\leq q}$$

ACP normée

On réalise l'ACP sur la matrice des corrélations des variables $x_{.1}; ...; x_{.q}$ avec $M = I$ et $\omega_i = 1/p$, c-a-d

$$Z = VM = (R(x_{.k}, x_{.l}))_{1\leq k,l\leq q}$$

où

$$R(x_{.k}, x_{.l}) = \frac{1}{n}\sum_{i=1}^{p}(\frac{x_{ik} - \bar{x}_{.k}}{\sigma_k})(\frac{x_{il} - \bar{x}_{.l}}{\sigma_l})$$

Dans les deux derniers cas on effectue souvent la représentation des variables dans le cercle de corrélations 3 c'est-à-dire au lieu de représenter les variables selon leurs covariances avec les facteurs, on les représente par leurs corrélations avec les facteurs. Si dans un plan, une variable est sur le cercle de corrélations, alors elle parfaitement représentée, donc expliquée, par les deux facteurs associés.

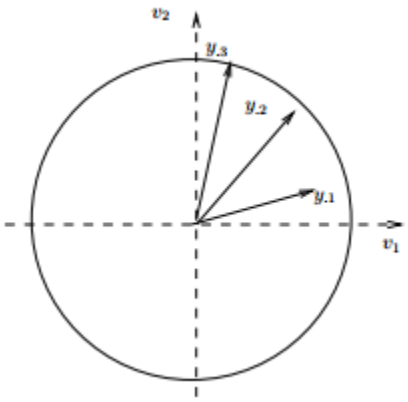


FIGURE 2.1 – Cercle de corrélation des variables

3 La régression linéaire simple

3.1 Introduction

La régression linéaire simple est l'une des méthodes les plus simples à la disposition du scientifique des données. Opérant sur un ensemble d'observations de deux dimensions (deux variables continues), la régression linéaire simple tente tout simplement d'ajuster, le mieux possible, une droite parmi les données. Ici, « le mieux possible » signifie minimiser la somme des erreurs carrées, c'est-à-dire la distance entre un point et la droite (voir la méthode des moindres carrées). Pourquoi ? Cette droite (notre modèle) devient alors un outil permettant de ressortir, si elle existe, la tendance sous-jacente dans les données en plus de servir comme modèle prédictif pour de nouveaux événements en se basant sur ceux déjà observés.

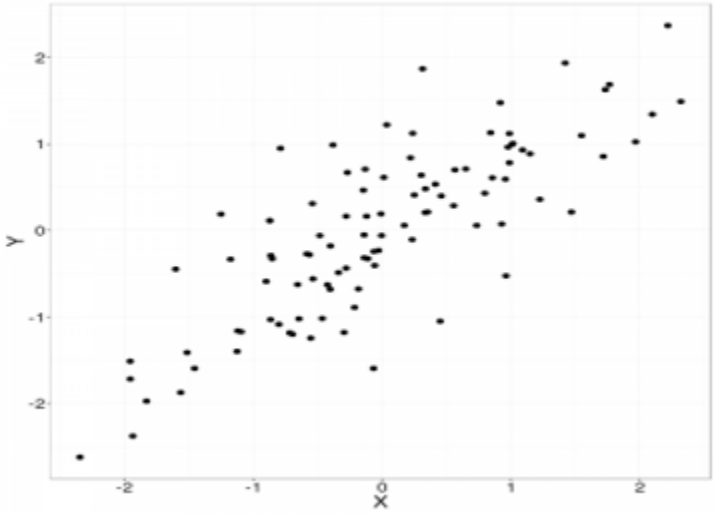


FIGURE 3.1 – Nuage de point du jeux de donnée

Bien qu'il soit tentant de suivre cette droite pour prédire des évènements à l'extérieur

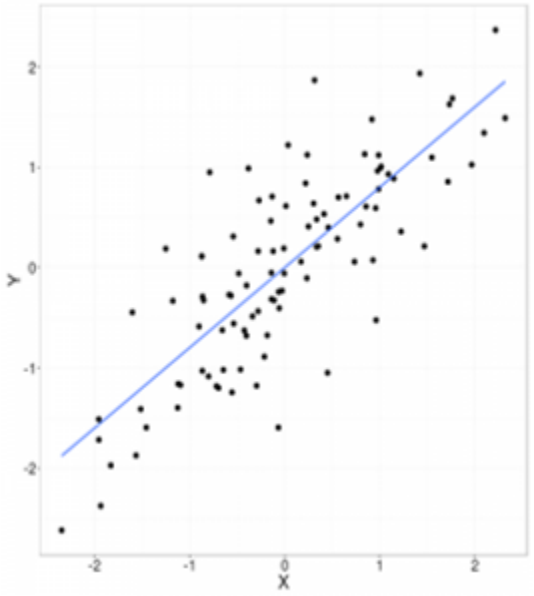


FIGURE 3.2 – Superposition d'une droite de régression (bleu)

de nos observations, il faut faire attention à de telles prédictions puisque le modèle n'a pas été bâti en prenant compte de cet espace. Aucune information n'est disponible pour supporter cette extrapolation.

3.2 Modélisation

Définition 3.2.1 *(Modèle de régression linéaire simple)*
On cherche à modéliser la relation entre deux variables quantitatives continues.[3]
Un modèle de régression linéaire simple est de la forme suivante :

$$y = \beta_1 + \beta_2 x + \varepsilon$$

où :
- y est la variable à expliquer (à valeurs dans $\mathbb{R}$) ;
- x est la variable explicative (à valeurs dans $\mathbb{R}$) ;
- ε est le terme d'erreur aléatoire du modèle ;
- β_1 et β_2 sont deux paramètres à estimer.

Explication
i. La désignation "simple" fait référence au fait qu'il n'y a qu'une seule variable explicative x pour expliquer y.
ii. La désignation "linéaire" correspond au fait que le modèle (1) est linéaire en β_1 et β_2.
Pour n observations, on peut écrire le modèle de régression linéaire simple sous la forme :

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i (2)$$

On suppose que ε est une variable aléatoire, non observée, que x_i est observée et non aléatoire et qu'enfin y_i est observée et aléatoire.
On fait les trois hypothèses additionnelles suivantes :
On peut écrire matriciellement le modèle (2) de la manière suivante [3] :

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

où

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix};$$

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- Y désigne le vecteur à expliquer de taille $n \times 1$,
- X la matrice explicative de taille $n \times 2$,
- ε le vecteur d'erreurs de taille $n \times 1$.

Les hypothèses relatives à ce modèle sont les suivantes :

$-H_1$:la distribution de l'erreur u est indépendante de X ou X est fixe.
$-H_2$:l'erreur est centrée et de variance constante (homoscédasticité) :
$\forall i = 1, ..., n\ \mathbb{E}(u_i) = 0$, $var(\varepsilon_i) = \sigma_\varepsilon^2$
$-H_3$: β_1 et β_2 sont constants, pas de rupture du modèle.
$-H_4$: Hypothese complémentaire pour les inférences : $u \sim N(0, \sigma_\varepsilon^2 I_p)$.

3.3 Moindres Carrés Ordinaires

Les points (x_i, y_i) étant donnés, le but est maintenant de trouver une fonction affine f telle que la quantité $\sum_{i=1}^{n} (L(y_i - f(x_i)))$ soit minimale. Pour pouvoir déterminer f, encore faut-il préciser la fonction de coût L. Deux fonctions sont classiquement utilisées :
- le coût absolu $L(u) = |u|$;
- le coût quadratique $L(u) = u^2$.

Les deux ont leurs vertus, mais on privilégiera dans la suite la fonction de coût quadratique. On parle alors de méthode d'estimation par moindres carrés (terminologie due à Legendre dans un article de 1805 sur la détermination des orbites des comètes).

Définition 3.3.1 *(Estimateurs des Moindres Carrés Ordinaires)*
On appelle estimateurs des Moindres Carrés Ordinaires (en abrégé MCO) $\hat{\beta}_1$ et $\hat{\beta}_2$ les valeurs minimisant la quantité :

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

Autrement dit, la droite des moindres carrés minimise la somme des carrés des distances verticales des points (x_i, y_i) du nuage à la droite ajustée

$$y = \hat{\beta}_2 + \hat{\beta}_2 x$$

.

3.3.1 Calcul des estimateurs de $\hat{\beta}_1$ et $\hat{\beta}_2$

La fonction de deux variables S est une fonction quadratique et sa minimisation ne pose aucun problème, comme nous allons le voir maintenant.

Propriété 3.3.1
(Estimateurs β_1 et β_2)
Les estimateurs des MCO ont pour expressions :

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

avec

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

Démonstration
La première méthode consiste à remarquer que la fonction S(β_1, β_2) est strictement convexe, donc qu'elle admet un minimum en un unique point (β_1, β_2), lequel est déterminé en annulant les dérivées partielles de S. On obtient les "équations normales" :

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_2} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

La première équation donne :

$$\hat{\beta_1}n + \hat{\beta}_2 \sum_{i=1}^{n}(x_i) = \sum_{i=1}^{n}(y_i)$$

d'où l'on déduit immédiatement :

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

où $\bar{x}$ et $\bar{y}$ sont comme d'habitude les moyennes empiriques des x_i et des y_i La seconde équation donne :

$$\hat{\beta}_1 \sum_{i=1}^{n}(x_i) + \hat{\beta}_2 \sum_{i=1}^{n}(x_i)^2 = \sum_{i=1}^{n}(x_i y_i)$$

et en remplaçant $\hat{\beta}_1$ par son expression (1.1), nous avons :

$$\hat{\beta}_2 = \frac{\sum x_i y_i - \sum x_i \bar{x}}{\sum x_i^2 - \sum x_i \bar{x}} = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})}$$

La seconde méthode consiste à appliquer la technique de Gauss de réduction des formes quadratiques, c'est-à-dire à décomposer S(β_1, β_2) en somme de carrés, carrés qu'il ne restera plus qu'à annuler pour obtenir les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$. Dans notre cas, après calculs, ceci s'écrit :

$$S(\beta_1, \beta_2) = n(\beta_1 - (\bar{y} - \beta_2 \bar{x}))^2 + (\sum_{i=1}^{n}(x_i - \bar{x})^2)\Big(\beta_2 - \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\Big)^2$$

$$+\Big(\sum_{i=1}^{n}(y_i - \bar{y})^2)(1 - \frac{(\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2 \sum\limits_{i=1}^{n}(y_i - \bar{y})^2}\Big)$$

où apparaissent deux carrés et un troisième terme indépendant de β_1 et β_2 :
Ce dernier est donc incompressible. Par contre, le second est nul si et seulement si β_2 = $\hat{\beta}_2$. Ceci étant fait, le premier est alors nul si et seulement si $\beta_1 = \hat{\beta}_1$.
L'expression (1.2) de $\hat{\beta}_2$ suppose que le dénominateur $\sum\limits_{i=1}^{n}(x_i - \bar{x})^2$ est non nul. Or ceci ne peut arriver que si tous les x_i sont égaux, situation sans intérêt pour notre problème et que nous excluons donc a priori dans toute la suite.

3.3.2 Quelques propriétés des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$

Sous les seules hypothèses (H_1) et (H_2) de centrages, décorrélations et homoscédasticités des erreurs ε du modèle, on peut déjà donner certaines propriétés des estimateurs $\hat{\beta}_1$et $\hat{\beta}_2$des moindres carrés.

Théorème 3.3.1 *(Estimateurs sans biais)*
$\hat{\beta}_1$ et $\hat{\beta}_2$ sont des estimateurs sans biais de $\beta 1$ et $\beta 2$.

Démonstration
Partons de l'écriture (1.3) pour $\hat{\beta}_2$:

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}$$

Dans cette expression, seuls les bruits ε sont aléatoires, et puisqu'ils sont centrés, on en déduit bien que $\mathbb{E}(\hat{\beta}_2) = \beta 2$. Pour $\hat{\beta}_1$, on part de l'expression :

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

d'où l'on tire :

$$\mathbb{E}(\hat{\beta}_1) = \mathbb{E}(\bar{y}) - \bar{x}\mathbb{E}(\hat{\beta}_2) = \beta_1 + \bar{x}\beta_2 - \bar{x}\beta_2 = \beta_1.$$

On peut également exprimer variances et covariance de nos estimateurs.

Théorème 3.3.2 *(Variances et covariance)*
Les variances des estimateurs sont :

$$var(\hat{\beta}_1) = \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2} = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2})$$

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

tandis que leur covariance vaut :

$$cov(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2 \bar{x}}{\sum(x_i - \bar{x})^2}$$

Démonstration
On part à nouveau de l'expression de $\hat{\beta}_2$ utilisée dans la preuve du non-biais :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2},$$

or les erreurs ε sont décorrélées et de même variance σ^2 donc la variance de la somme est la somme des variances :

$$var(\hat{\beta}_2) = \frac{\sum(x_i - \bar{x})^2\sigma^2}{(\sum(x_i - \bar{x})^2)^2} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Par ailleurs, la covariance entre $\bar{y}$ et $\hat{\beta}_2$ s'écrit :

$$cov(\bar{y}, \hat{\beta}_2) = cov(\frac{\sum y_i}{n}, \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}) = \frac{\sigma^2 \sum(x_i - \bar{x})}{n\sum(x_i - \bar{x})^2} = 0$$

d'où il vient pour la variance de$\hat{\beta}_1$:

$$var(\hat{\beta}_1) = var(\frac{\sum y_i}{n} - \hat{\beta}_2\bar{x}) = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum(x_i - \bar{x})^2} - 2\bar{x}cov(\bar{y}, \hat{\beta}_2)$$

c'est-à-dire :

$$var(\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}$$

Enfin, pour la covariance des deux estimateurs :

$$cov(\hat{\beta}_1, \hat{\beta}_2) = cor(\bar{y} - \hat{\beta}_2\bar{x}, \hat{\beta}_2) = cov(\bar{y}, \hat{\beta}_2) - \bar{x}var(\hat{\beta}_2) = \frac{-\sigma^2\bar{x}}{\sum(x_i - \bar{x})^2}$$

Théorème 3.3.3 *(Gauss-Markov)*
Parmi les estimateurs sans biais linéaires en y, les estimateurs $\hat{\beta}_j$ sont de variances minimales.

Démonstration
L'estimateur des MCO s'écrit $\hat{\beta}_2 = \sum_{i=1}^{n} p_i y_i$, avec $p_i = \frac{(x_i - \bar{x})}{(\sum(x_i - \bar{x})^2)}$. Considérons un autre estimateur β_2 linéaire en y_i et sans biais, c'est-à-dire :

$$\tilde{\beta}_2 = \sum_{i=1}^{n} \lambda_i y_i.$$

Montrons que $\sum \lambda_i = 0$ et $\sum \lambda_i x_i = 1$. L'égalité

$$\mathbb{E}(\tilde{\beta}_2) = \beta_1 \sum \lambda_i + \beta_2 \sum \lambda_i x_i + \sum \lambda_i \mathbb{E}(\varepsilon_i) = \beta_1 \sum \lambda_i + \beta_2 \sum \lambda_i x_i$$

est vraie pour tout β_2. L'estimateur $\tilde{\beta}_2$ est sans biais donc $\mathbb{E}(\tilde{\beta}_2) = \beta_2$ pour tout β_2, c'est-à-dire que $\sum \lambda_i = 0$ et $\sum \lambda_i x_i = 1$. Montrons que

$$var(\tilde{\beta}_2) var(\hat{\beta}_2)$$

:

$$var(\tilde{\beta}_2) = var(\tilde{\beta}_2 - \hat{\beta}_2 + \hat{\beta}_2) = var(\tilde{\beta}_2 - \hat{\beta}_2) + var(\hat{\beta}_2) + 2cov(\tilde{\beta}_2 - \hat{\beta}_2, \hat{\beta}_2).$$

Or :

$$cov(\tilde{\beta}_2 - \hat{\beta}_2, \hat{\beta}_2) = cov(\tilde{\beta}_2, \hat{\beta}_2) - var(\hat{\beta}_2) = \frac{\sigma^2 \sum \lambda_i (x_i - \bar{x})}{\sum(x_i - \bar{x})^2} - \frac{\sigma^2}{\sum(x_i - \bar{x})^2} = 0$$

la dernière égalité étant due aux deux relations $\sum \lambda_i = 0$ et $\sum \lambda_i x_i = 1$. Ainsi :

$$var(\tilde{\beta}_2) = var(\tilde{\beta}_2 - \hat{\beta}_2) + var(\hat{\beta}_2).$$

Une variance est toujours positive, donc :

$$var(\tilde{\beta}_2) \leq var(\hat{\beta}_2).$$

Le résultat est démontré. On obtiendrait la même chose pour $\hat{\beta}_1$.

3.3.3 Calcul des résidus et de la variance résiduelle

Introduction

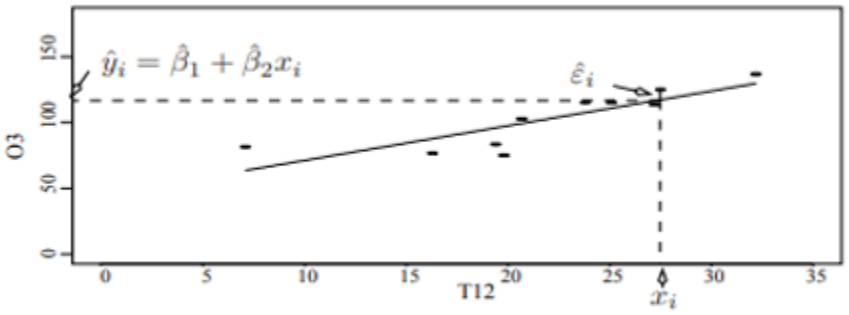


FIGURE 3.3 – Représentation des individus

Dans R2 (espace des variables x_i et y_i), $\hat{\beta}_1$ est l'ordonnée à l'origine et $\hat{\beta}_2$ la pente de la droite ajustée. Cette droite minimise la somme des carrés des distances verticales des points du nuage à la droite ajustée.[9]. Notons $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ l'ordonnée du point de la droite des moindres carrés d'abscisse x_i, ou valeur ajustée. les résidus sont définis par (figure 3.3) :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i = (y_i - \bar{y}) - \hat{\beta}_2(x_i - \bar{x}).$$

Par construction, la somme des résidus est nulle :

$$\sum_i \hat{\varepsilon}_i = \sum_i (y_i - \bar{y}) - \hat{\beta}_2 \sum_i (x_i - \bar{x}) = 0$$

Notons maintenant que les variances et covariance des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ établies en section précédente ne sont pas pratiques car elles font intervenir la variance σ^2 des erreurs, laquelle est en général inconnue. Néanmoins, on peut en donner un estimateur sans biais grâce aux résidus.

Théorème 3.3.4 *(Estimateur non biaisé de σ^2)*
La statistique $\hat{\sigma^2} = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{(n-2)}$ est un estimateur sans biais de σ^2.

Démonstration
Réécrivons les résidus en constatant que $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$ et $\beta_1 = \bar{y} - \beta_2 \bar{x} - \bar{\varepsilon}$, ce qui donne :

$$\begin{aligned} \hat{\varepsilon}_i &= \beta_1 + \beta_2 x_i + \varepsilon_i - \hat{\beta}_2 x_i \\ &= \bar{y} - \beta_2 \bar{x} - \bar{\varepsilon} + \beta_2 x_i - \bar{y} + \hat{\beta}_2 \bar{x} - \hat{\beta}_2 x_i \\ &= (\beta_2 - \hat{\beta}_2)(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \end{aligned}$$

En développant et en nous servant de l'écriture vue plus haut :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2},$$

nous avons :

$$\begin{aligned}\sum \hat{\varepsilon_i^2} &= (\beta_2 - \hat{\beta}_2)^2 \sum(x_i - \bar{x})^2 + \sum(\varepsilon_i - \bar{\varepsilon})^2 + 2(\beta_2 - \hat{\beta}_2)\sum(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) \\ &= (\beta_2 - \hat{\beta}_2)^2 \sum(x_i - \bar{x})^2 + \sum(\varepsilon_i - \bar{\varepsilon})^2 - 2(\beta_2 - \hat{\beta}_2)^2 \sum(x_i - \bar{x})^2\end{aligned}$$

Prenons-en l'espérance :

$$\mathbb{E}(\sum \hat{\varepsilon_i^2}) = \mathbb{E}(\sum(\varepsilon_i - \bar{\varepsilon})^2) - \sum(x_i - \bar{x})^2 var(\hat{\beta}_2) = (n-2)\sigma^2$$

Bien sûr, lorsque n est grand, cet estimateur diffère très peu de l'estimateur empirique de la variance des résidus, à savoir $\sum_{i=1}^{n} \frac{\hat{\varepsilon_i^2}}{n}$.

3.4 Interprétations géométriques

Introduction

Nous allons maintenant voir comment les lois précédentes interviennent dans nos estimateurs. Afin de faciliter la lecture de cette partie, fixons les notations suivantes :

$$c = \frac{-\sigma^2 \bar{x}}{\sum(x_i - \bar{x})^2}; \hat{\sigma^2} = \frac{1}{n-2}\sum \hat{\varepsilon}_i^2$$

$$\sigma_1^2 = \sigma^2(\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}); \hat{\sigma_1^2} = \hat{\sigma^2}(\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2})$$

$$\sigma_2^2 = (\frac{\sigma^2}{\sum(x_i - \bar{x})^2}); \hat{\sigma_2^2} = (\frac{\hat{\sigma^2}}{\sum(x_i - \bar{x})^2})$$

Comme nous l'avons vu, σ_1^2, σ_2^2 et c sont les variances et covariance des estimateurs des moindres carrés ordinaires. les quantités $\hat{\sigma_1^2}$ et $\hat{\sigma_2^2}$ correspondent quant à elles aux estimateurs des variances de $\hat{\beta}_1$ et $\hat{\beta}_2$.

3.4.1 Intervalle de confiance

Propriété 3.4.1 *(Lois des estimateurs avec variance connue)*
Les lois des estimateurs des MCO avec variance σ^2 connue sont :

- $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim N(\beta, \sigma^2 V) ou \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} et$

$$V = \frac{1}{\sum(x_i - \bar{x})^2}\begin{pmatrix} \frac{\sum x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} = \frac{1}{\sigma^2}\begin{pmatrix} \sigma_1^2 & c \\ c & \sigma_2^2 \end{pmatrix}$$

- $\frac{n-2}{\sigma^2}\hat{\sigma^2} \sim X_n - 2^2$, *loi du* X^2 *à* $(n-2)$ *degrés de liberté .*
- $\hat{\beta^2}$ *et* $\hat{\sigma^2}$ *sont indépendants.*

Propriété 3.4.2 *(Lois des estimateurs avec variance estimée)*

- $\frac{\hat{\beta_1}-\beta_1}{\hat{\sigma_1}} \sim T_{n-2}$ *,où* T_{n-2} *est une loi de Student à* $(n-2)$ *degrés de liberté.*
- $\frac{\hat{\beta_2}-\beta_2}{\hat{\sigma_2}} \sim T_{n-2}$
- $\frac{1}{2\hat{\sigma^2}} \sim F^2_{n-2}$ *, loi de Fisher de paramètres* $(2, n-2)$

Les lois des estimateurs des MCO avec variance σ^2 *estimée sont :*
Ces dernières propriétés nous permettent de donner des intervalles de confiance (IC) ou des régions de confiance (RC) des estimateurs. En effet, la valeur ponctuelle d'un estimateur est de peu d'intérêt en général et il est intéressant de lui associer un intervalle de confiance. Les résultats sont donnés pour un α *général, en pratique on prend typiquement* $\alpha = 0,05$.

Propriété 3.4.3 *(Intervalles et régions de confiance)*

- $IC(\beta_1) : \hat{\beta}_1 t_{n-2}\frac{1-\alpha}{2}\hat{\sigma}_1 \text{où} t_{n-2}\frac{1-\alpha}{2}\hat{\sigma}_1$, *est le quantile de niveau* $(1-\frac{\alpha}{2})$ *d'une loi de Student* T_{n-2}
- $IC(\beta_2) : \hat{\beta}_2 t_{n-2}\frac{1-\alpha}{2}\hat{\sigma}_2$.
- $RC(\beta)$:*Une région de confiance simultanée pour* β_1 *et* β_2 *au niveau* $(1-\alpha)$ *est*

$$\frac{1}{2\hat{\sigma^2}}(n(\hat{\beta}_1-\beta_1)^2 + 2n\bar{x}(\hat{\beta}_1-\beta_1)(\hat{\beta}_2-\beta_2) + \sum x_i^2(\hat{\beta}_2-\beta_2)^2) \leq f^2_{n-2}(1-\alpha)$$

où $(1-\alpha)$ *est le quantile de niveau* $(1-\alpha)$ *d'une loi*

- *Un intervalle de confiance de* σ^2 *est donné par :*

$$(\frac{(n-2)\hat{\sigma^2}}{c_{n-2}(1-\frac{\alpha}{2})}, \frac{(n-2)\hat{\sigma^2}}{c_n - 2(\frac{\alpha}{2})})$$

où $c_{n-2}(1-\frac{\alpha}{2})$ *est le quantile de niveau* $(1-\frac{\alpha}{2})$ *d'une loi* X^2_{n-2}

Remarque
Le point iii donne la région de confiance simultanée des paramètres (β_1, β_2) de la régression, appelée ellipse de confiance, tandis que i et ii donnent des intervalles de confiance pour β_1 et β_2 pris séparément. La figure 1.7 montre la différence entre ces deux notions.

3.4.2 Coefficient de détermination R^2

Définition 3.4.1 *(Coefficient de détermination R^2)*
Le coefficient de détermination R^2 est défini par :

$$R^2 = \frac{SCE}{SCT} = \frac{\|\hat{Y}-\bar{y}\|^2}{\|Y-\bar{y}\|^2} = 1 - \frac{\|\hat{\epsilon}\|^2}{\|Y-\bar{y}\|^2} = 1 - \frac{SCR}{SCT}$$

On voit sur la figure 1.3 que R^2 correspond au cos *carré de l'angle θ. De façon schématique, on peut différencier les cas suivants :*
— Si $R^2 = 1$, le modèle explique tout, l'angle β vaut zéro et Y est dans M(X) , c'est-à-dire que $y_i = \beta_1 + \beta_2 x_i$ pour tout i : les points de l'échantillon sont parfaitement alignés sur la droite des moindres carrés ;
— Si $R^2 = 0$, cela veut dire que $\sum(\hat{y}_i - \bar{y})^2 = 0$, donc $\hat{y}_i = \bar{y}$ pour tout i. Le modèle de régression linéaire est inadapté puisqu'on ne modélise rien de mieux que la moyenne ;
— Si R^2 est proche de zéro, cela veut dire que Y est quasiment dans l'orthogonal de M(X), le modèle de régression linéaire est inadapté, la variable x n'explique pas bien la variable réponse y (du moins pas de façon affine).
De façon générale, l'interprétation est la suivante : le modèle de régression linéaire permet d'expliquer $100 \times R^2\%$ *de la variance totale des données.*

3.4.3 Test de signification

3.4.3.1 Test de signification globale du modèle

Ce test permet de connaitre l'apport global de la variable X à la détermination de Y :
On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Pour tester cette l'hypothèse, on s'est basé sur la statistique de Fisher, notée par F :

$$F = \frac{SCE}{\frac{SCR}{n-2}}$$

Cette statistique indique si la variance expliquée est significativement supérieure à la variance résiduelle. Dans ce cas, on peut considérer que l'explication emmenée par la régression traduit une relation qui existe réellement dans la population.
Sous H_0,SDE est distribué selon X_i^2 et SCR selon un $X^2_{(n-2)}$, de fait pour F, on a

$$F = \frac{\frac{X_1^2}{1}}{\frac{X^2_{(n-2)}}{(n-2)}} = f_{1,n-2}^{1-\alpha}$$

Alors sous H_0,F est donc distribué selon une loi de Fisher à (1,n-2) degrés de liberté, où on rejette H_0 si :

$$F \geq f_{1,n-2}^{1-\alpha},$$

avec $f_{1,n-2}^{1-\alpha}$ est le quartile d'ordre $1-\alpha$ d'une loi de Fisher à (1,n-2)ddl.

3.4.3.2 Test de signification des paramètres

Test de signification de β_0
On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Pour tester cette l'hypothèse,on forme la statistique de test :

$$T(\hat{\beta}_0) = \frac{\hat{\beta}_0}{\hat{\sigma\beta}_0}$$

on rejette H_0 si l'observation de la statistique de test, notée ,telle que :

$$|T(\hat{\beta}_0)| \geq t_{n-2}^{1-\frac{\alpha}{2}}$$

où $t_{n-2}^{1-\frac{\alpha}{2}}$ est le quartile d'ordre $1-\frac{\alpha}{2}$ d'une loi de Student à (n-2)ddl.

Test de signification de la pente β_1
Le test de signification de la pente consiste à vérifier l'exogène X sur l'endogène Y .L'hypothèse à confronter s'écrive :

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Pour tester cette l'hypothèse,on forme la statistique de test :

$$T(\hat{\beta}_1) = \frac{\hat{\beta}_1}{\hat{\sigma\beta}_1}$$

on rejette H_0 si l'observation de la statistique de test, notée ,telle que :

$$|T(\hat{\beta}_1| \geq t_{n-2}^{1-\frac{\alpha}{2}}$$

où $t_{n-2}^{1-\frac{\alpha}{2}}$ est le quartile d'ordre $1-\frac{\alpha}{2}$ d'une loi de Student à (n-2)ddl.

3.5 Conclusion

Le modèle linéaire simple permet d'expliquer une variable Y par une fonction affine d'un seule variable.
Dans le chapitre qui suit, le variable Y se expliquer par plusieurs variables explicatives. Donc c'est la généralisation de ce chapitre, mais nous allons cette fois manipuler sytématiquement des vecteurs et des matrices à la place des scalaires.

4 La régression linéaire multiple

4.1 Introduction

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple.
Le but de la régression multiple est d'expliquer une variable Y à l'aide de plusieurs variables $(X_1, ..., X_q)$. La variable Y est appelée variable dépendante, ou variable à expliquer et les variables $X_j (j = 1, ..., q)$ sont appelées variables indépendantes, ou variables explicatives. En pratique, l'utilisation des méthodes statistiques obtenues nécessite donc, sauf dans de situations particulières, de recourir à des logiciels pour inverser des matrices(de grandes tailles). Nous allons terminer ce chapitre en introduisant la notion de régression polynomiale qui peut être considérer comme un cas particulier de la régression linéaire multiple

4.2 Modèle théorique

Étant donné un échantillon $(Y_i, X_{i1}, ..., X_{ip})$ $i \in \{1, \ldots, n\}$, on cherche à expliquer, avec le plus de précision possible, les valeurs prises par Y_i, dite variable endogène, à partir d'une série de variables explicatives $X_{i1}, ..., X_{ip}$. Le modèle théorique, formulé en termes de variables aléatoires, prend la forme :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

où ε_i est l'erreur du modèle qui exprime, ou résume, l'information manquante dans l'explication linéaire des valeurs de Y_i à partir des $X_{i1}, ..., X_{ip}$ (problème de spécifications, variables non prises en compte, etc.). Les coefficients $\beta_0, \beta_1, ..., \beta_p$ sont les paramètres à estimer.
Explication :
i. L'expression "multiple" renvoie au fait qu'il faut plusieurs variables x_i pour expliquer y.
ii. L'expression "linéaire" renvoie au fait que le modèle est linéaire en $\beta_0, \beta_1, \ldots, \beta_p$.

Le modèle se prête également à une **écriture matricielle** :

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

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Plus précisement, on suppose que x_{ij} répresente la valeur prise par la variable explicative j sur l'unité statistique i. De même, le vecteur $Y = (y_1, y_2, \ldots, y_n)$ répresente les prises par la variable dépendante sur les n unités statistiques. Dans la plupart des applications, on supposera également que la pémière variable est la constante, c'est à dire que les x_{i1} $= 1$, $i = 1, \ldots, n$.[2] Alors la matrice X sera de la forme :

$$X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{ij} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{in} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

où :
- Y est un vecteur aléatoire de dimension n.
- X est une matrice de dimension n*p connue, on l'appelle matrice du plan d'experience.
- β est le vecteur de dimension de p des paramètres inconnus du modèle.
- ε est le vecteur de dimensions n des erreurs.

On suppose que :

$-H_1$:Les ε_i sont les termes d'erreurs, d'une variable ε, non observé, indépendantes et identiquement distribués ; $E(\varepsilon_i) = 0$, $Var(\varepsilon) = \sigma_\varepsilon^2 I$
$-H_2$:Les termes x_j $j = 1, ..., p$ sont supposés déterministes (facteurs controlés) ou bien l'erreur ε est indépendante de la distribution conjointe de $X_1, ..., X_p$. On écrit dans ce dernier cas que :
$E(Y|X_1, ..., X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \qquad et \qquad Var(Y|X_1, ..., X_p) = \sigma_\varepsilon^2$
$-H_3$:Les paramètres inconnus $\beta_0, \ldots, \beta_p$ sont supposes constants.
$-H_4$:En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur $\varepsilon(N(0, \sigma_\varepsilon^2 I))$. Les ε_i sont i.i.d de loi $N(0, \sigma_\varepsilon^2 I)$.
On note que $X = [X_1, ..., X_p]$ où X_j est le vecteur de taille n correspondant à la j-eme variable.

4.3 Estimation

La problématique reste la même que pour la régression simple c'est à dire l'estimer les paramètres $\beta_0, \ldots, \beta_p$ en exploitant les observations , l'évaluer la précision de ces

estimateurs, la mesure du pouvoir explicatif du modèle, l'évaluation de l'influence des variables dans le modèle (globalement (les p variables en bloc) et, individuellement (chaque variable)) et l'évaluation de la qualité du modèle lors de la prédiction (intervalle de prédiction) ;

Remarque

- Tous les résultats du chapitre précédent se généralisent dans le cas général.
- Dans une régression multiple, il se peut que le nombre p des variables disponibles soit grand. Cette quantité d'information est parfois superfue ou redondante. Ainsi la diminution du nombre de vatiables réellement intéressantes dans la régression est envisageable. Soit on part du modèle complet et on retire des variables en utilisant un critère décrit sous Statistica (pas à pas descendant). Soit on part d'une régression simple et on ajoute des variables qui enrichissent le modèle (pas à pas ascendant). Sous Statistica, dans ces deux cas, on arrête d'enlever ou d'ajouter une variable au modèle en analysant la statistique F.

4.3.1 Estimation par la méthodes des moindres carrés

L'idée est d'estimer $\beta_0, \ldots, \beta_p$, par la méthode des moindres carrés au moyen de leurs estimateurs :

$$\hat{\beta}_0, \ldots, \hat{\beta}_p$$

qui permettent d'ajuster au mieux les y_i par les x_{ij} au sens où la somme des carées des résidus est minimisée.

Définition 4.3.1

Les valeurs ajustées(ou estimées) de Y ont pour expression :

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = P_X Y O\grave{u} P_X = X(X'X)^{-1}X$$

est une matrice idempotente sur le sous espace engendré par les colonnes de X. Le vecteur des valeurs ajustées peut être interprété comme la projection de Y sur le sous espace engendré par les colonnes de la matrice X d'autres façon, est la projection orthogonale de Y sur $Im(X)$.

Remarque $\hat{\beta}$ est appelé estimateur des moindres carrés de β.

Théorème 4.3.1

L'estimateur $\hat{\beta}$ des moindres carrés a pour expression :

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

Démonstration

Le vecteur β déterminés par la méthodes des moindres carrés qui minimise la fonction :

$$S(\beta_0, ..., \beta_n) = S(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 = ||y - X\beta||^2$$

On remarque donc que : $||\varepsilon||^2 = ||y - X\beta||^2$
Ainsi

$$||\varepsilon||^2 = ||y - X\beta||^2 = (Y - X\beta)'(Y - X\beta)$$
$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

D'autres part, la transposée d'un scalaire est égale à lui-même, on a :

$$(Y'X\beta) = \beta'X'Y$$

On a donc :

$$S(\beta_0, ..., \beta_n) = ||\varepsilon||^2 = Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

On minimise la fonction S en annulant les dérivées partielles pa rapport à β de la fonction.
On annule la dérivation matricielle pour obtient les **équations normales** :

$$\frac{\partial S(\beta)}{\partial \beta} = -2(X'Y) + 2(X'X)\beta = 0$$

Ce qui donne

$$X'Y - X'X\beta = 0$$

dont la solution correspond bien à un minimum car la matrice hessienne $2X'X$ est semi définie-positive.
Nous faisons l'hypothèse supplémentaire que la matrice $X'X$ est inversible, c'est-à-dire que la matrice X est de rang p et donc qu'il n'existe pas de colinéarité entre ses colonnes. En pratique, si cette hypothèse n'est pas vérifiée, il suffit de supprimer des colonnes de X et donc des variables du modèle. Des diagnostics de colinéarité et des aides au choix des variables sont explicités dans une présentation détaillée du modèle linéaire.
Alors, l'estimation des paramètres a_j est donnée par :

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

4.3.2 Propriétés

A l'instar de la régression simple, l'estimateur obtenu est sans biais. On trouve également une expression simpliste pour sa matrice de covariance $var(\hat{\beta})$. Rappelons que la matrice de covariance egalement appelée matrice de variance-covariance, ou encore matrice de dispersion est par definition :

$$var(\hat{\beta}) = \mathbb{E}(\hat{\beta} - \mathbb{E}(\beta))(\hat{\beta} - \mathbb{E}(\beta)') = \mathbb{E}[\hat{\beta}\hat{\beta}'] - \mathbb{E}[\hat{\beta}]\mathbb{E}[\hat{\beta}']$$

Propriété 4.3.1
L'estimateur β des moindres carrés est sans biais, c'est à dire que $\mathbb{E}[\beta] = \beta$, et sa matrice e variance-covariance est :

$$varcov(\hat{\beta} = \sigma^2(X'X)^{-1}$$

. Cette matrice des coefficients, de dimension $(p+1, p+1)$ *est donnée par :*

$$varcov(\hat{\beta} = \begin{pmatrix} var(\hat{\beta}_1) & \cdots & cov(\hat{\beta}_1, \hat{\beta}_j) & \cdots & cov(\hat{\beta}_1, \hat{\beta}_p) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_1, \hat{\beta}_j) & \cdots & var(\hat{\beta}_j) & \cdots & cov(\hat{\beta}_j, \hat{\beta}_p) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_1, \hat{\beta}_p) & \cdots & cov(\hat{\beta}_j, \hat{\beta}_p) & \cdots & var(\hat{\beta}_p) \end{pmatrix}$$

Cette matrice est symétrique, sur sa diagonale principale on observe les variances de coeficients estimés $var(\hat{\beta}_1), ..., var(\hat{\beta}_p)$

Démonstration
i. Pour le biais il suffit d'écrire :

$$\mathbb{E}((X'X)^{-1}X'Y) = (X'X)^{-1}X'\mathbb{E}(Y) = (X'X)^{-1}(X'\mathbb{E}(X\beta + \varepsilon))$$

... et comme $\mathbb{E}(\varepsilon) = 0$ on en deduit que :

$$\mathbb{E}[\beta] = (X'X)^{-1}(X'X\beta = \beta$$

ii. Pour la variance, on a : $var(\hat{\beta}) = var((X'X)^{-1}X'Y) = (X'X)^{-1}X'var(Y)X(X'X)^{-1}$.
Or

$$var(Y) = var(X\beta + \varepsilon) = var(\varepsilon) = \sigma^2 I$$

alors,

$$var(\hat{\beta}) = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

Théorème 4.3.2
L'estimateur β *des moindres carrés est de variance minimale parmi les estimateurs linéaires sans biais de* β

Démonstration
L'idée est de montrer que, pour tout estimateur $\tilde{\beta}$ de β linéaire et sans biais, $var(\tilde{\beta}) \geq var(\hat{\beta})$ Rappelons la formule générale pour la matrice de covariance de la somme de deux vecteurs U et V :

$$var(U, V) = var(U) + var(V) + cov(U, V) + cov(V, U),$$

Or $cov(U, V) = \mathbb{E}[UV'] - \mathbb{E}[U]\mathbb{E}[V]' = cov(V, U)$. Ainsi une decomposition de la variance de $\tilde{\beta}$ donne :

$$var(\tilde{\beta} - \hat{\beta}, \hat{\beta}) = var(\tilde{\beta} - \hat{\beta}) + var(\hat{\beta}) + cov(\tilde{\beta} - \hat{\beta}, \hat{\beta}) + cov(\hat{\beta}, \tilde{\beta} - \hat{\beta}).$$

Il nous suffit maintenant de montrer que $cov(\tilde{\beta} - \hat{\beta}, \hat{\beta}) = 0$ pour finir la démonstration car les variances sont semi-définies positives. Puisque ${}_b\tilde{eta}$ est linéaire, $\hat{\beta} = AY$ où A

est une matrice(n, p). On sait également qu'il est sans biais ($\mathbb{E}[\tilde{\beta}] = \beta$) donc $AX = I$. La covariance devient :

$$\begin{aligned} cov(\tilde{\beta} - \hat{\beta}, \hat{\beta}) &= cov(AX, (X'X)^{-1}X'Y) - var(\hat{\beta}) \\ &= \sigma^2 AX(X'X)^{-1}\sigma^2(X'X)^{-1} \\ &= 0 \end{aligned}$$

Une variance étant toujours positive, on a donc :

$$var(\tilde{\beta}) \geq var(\hat{\beta})$$

Le résultat est démontré. On conclut que l'estimateur dpar les moindres carrés est le meilleur car on trouve le même resultat pour les autes $\hat{\beta}$

4.3.3 Calcul des résidus et de la variance résiduelle

On appelle résidus les erreurs observées sur l'échantillon. Elles sont définis par :

$$\hat{\varepsilon} = (\hat{\varepsilon_1}, \ldots, \hat{\varepsilon_n})' = Y - \hat{Y} = (I - {P_X}^{\perp})Y = P_X\varepsilon$$

C'est la projection de Y sur le sous espace orthognal de $Vect(X)$ dans $\mathbb{R}^n$ avec

$${P_X}^{\perp} = (I - P_X)$$

${P_X}^{\perp}$ est une idempotente sur le noyau de P_X.

Proposition 4.3.1
Sous les hypotheses du modèle, on a :

- $\mathbb{E}(\hat{\varepsilon}) = 0$
- $var(\hat{\varepsilon}) = \sigma^2 {P_X}^{\perp}$
- $\mathbb{E}(\hat{Y}) = X\beta$
- $var(\hat{Y}) = \sigma^2 P_X$
- $cov(\hat{\varepsilon}, \hat{Y}) = 0$

Démonstration

- $\mathbb{E}(\hat{\varepsilon}) = \mathbb{E}({P_X}^{\perp}\varepsilon) = {P_X}^{\perp}\mathbb{E}(\varepsilon) = 0$

- $var(\hat{\varepsilon}) = P_X \perp var(\varepsilon){P'_X}^{\perp} = \sigma^2 {P_X}^{\perp}{P_X}^{\perp} = \sigma^2 {P_X}^{\perp}$

- $\mathbb{E}(\hat{Y}) = \mathbb{E}(X\hat{\beta}) = X\beta$ car $\hat{\beta}$ est un estimateur sans biais

- $var(\hat{Y}) = var(X\hat{\beta}) = Xvar(\hat{\beta})X' = \sigma^2 X(X'X)^{-1}X' = \sigma^2 P_X$

- On rappelle que la covariance entre deux vecteurs aléatoires est une application bilinéaire et que $cov(U,U) = var(U)$. Ce qui nous donne :

$$cov(\hat{\varepsilon}, Y\hat{\varepsilon}) = cov(\hat{\varepsilon}, Y) - var(\hat{\varepsilon}) = cov({P_X}^{\perp}Y, Y) - \sigma^2 P_X$$

Puisque $var(Y) = \sigma^2 I$ on a :

$$cov(\hat{\varepsilon}, Y\hat{\varepsilon}) = {P_X}^{\perp} var(Y) - \sigma^2 {P_X}^{\perp} = 0$$

Proposition 4.3.2
La statistique

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

est un estimateur sans biais de σ^2.

Démonstration

4.4 Qualité d'ajustement

4.4.1 Somme des résidus

La matrice X peut contenir une variable constante de manière explicite(une colonne contient une variable constante) ou de manière implicite(un combinaison linéaire de colonnes de X permet d'avoir une colonne. On suppose alors qu'il existe un vecteur λ de $\mathbb{R}^p$ tel que $X\lambda = I_n = (I, \ldots, I, \ldots, I)$.

Théorème 4.4.1
Si la matrice X *contient une variable constante définie de manière explicite où implicite, alors la somme des résidus est nulle.*

Démonstration
On a :

$$\sum_{i=1}^{n}\varepsilon_i = I_n'\varepsilon$$

Or, il existe un vecteur λ de $\mathbb{R}^p$ tel que $X\lambda = I_n$. On écrit donc :

$$\begin{aligned}\sum_{i=1}^{n}\varepsilon_i &= \lambda' X'\varepsilon \\ &= \lambda' X'\{I - X(X\lambda' X'X)\lambda' X^{-1}X\lambda' X'\}Y \\ &= \{\lambda' X' - \lambda' X'X(X'X)^{-1}X'\}Y\end{aligned}$$

D'autre part $'X'X(X'X)^{-1} = I$
Alors :

$$\sum_{i=1}^{n}\varepsilon_i = 0$$

4.4.2 Sommes des carrés

Théorème 4.4.2
La somme des carrées totales des écarts à la moyenne

$$SCT = (y - \bar{y})'(y - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

se décompose en une somme de deux termes :
- La somme des carrés expliquées,

$$SCT = (\hat{y} - \bar{y})'(\hat{y} - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- La somme des carrés des résidus,

$$SCT = \varepsilon'\varepsilon = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2.$$

Démonstration
En notant $\bar{y}$ le vecteur de $\mathbb{R}^n$ contenant n fois la moyenne $\bar{y}$, on a :

$$y - \bar{y} = y - \hat{y} + \hat{y} - \bar{y} = \hat{y} - \bar{y} + \varepsilon$$

donc,

$$\begin{aligned} SCT &= (y - \bar{y})'(y - \bar{y}) \\ &= (\hat{y} - \bar{y} + \varepsilon)'(\hat{y} - \bar{y} + \varepsilon) \\ &= (\hat{y} - \bar{y})'(\hat{y} - \bar{y}) + \varepsilon'\varepsilon + 2\varepsilon'(\hat{y} - \bar{y}). \end{aligned}$$

Or $\varepsilon'(\hat{y} - \bar{y}) = 0$ car ε et $(\hat{y} - \bar{y})$ sont orthogonaux. On a finalement

$$\begin{aligned} SCT &= (y - \bar{y})'(y - \bar{y}) \\ &= (\hat{y} - \bar{y})'(\hat{y} - \bar{y}) + \varepsilon'\varepsilon \\ &= SCE + SCR \end{aligned}$$

4.4.3 Tableau d'analyse de variance et coefficient de détermination

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	p	$CME = \frac{SCE}{p}$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$	$n-p-1$	$CMR = \frac{SCR}{n-p-1}$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$	$n-1$	

4.4.4 Le coefficient de Détermination R^2

R^2 est un indicateur spécifique qui permet de traduire la variance expliquée par le modèle, il s'agit du coefficient de détermination. Sa formule est la suivante :

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

Enfin, si le R^2 est certes un indicateur pertinent, il présente un défaut parfois ennuyeux, il a tendance à mécaniquement augmenter à mesure que l'on ajoute des variables dans le modèle. De ce fait, il est inopérant si l'on veut comparer des modèles comportant un nombre différent de variables. Il est conseillé dans ce cas d'utiliser **le coefficient de détermination ajusté** qui est corrigé des degrés de libertés. Le R^2 ajusté est toujours inférieur au R^2

4.5 Intervalle de confiance et Test d'hypothèses

4.5.1 Intervalle de confiance

Après l'obtention de l'estimateur, son espérance et une estimation de sa variance, il ne reste plus qu'à calculer sa loi de distribution pour construire des intervalles de confiance ou des tests d'hypothèses sur β.

Proposition 4.5.1 *(Lois des estimateurs avec variance connue)*
Les lois des estimateurs des moindres carrées avec variance connue sont :
1) $\hat{\beta}$ *est un vecteur gaussien du moyenne* β *et de variance* $\sigma(X'X)^{-1}$ *.*

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

2) $\hat{\beta}$ *et* $\hat{\sigma^2}$ *sont indépendants.*
3) $(n-p-1)\frac{\hat{\sigma^2}}{\sigma^2} \sim X^2_{n-p-1}$

Remarque
Cette proposition n'est pas satisfaisante pour avoir des intervalles de confiances car elle suppose que la variance σ^2 est connue ce qui n'est pas toujours vrai.

Proposition 4.5.2 *(Lois des estimateurs avec variance inconnue)*
Les lois des estimateurs des moindres carrées avec variance inconnue sont :
1) Pour $i = 1, \ldots, p$ *on a* t_i *suit une loi de Student à* $(n-p-1)$ *degrés de liberté, on écrit :*

$$t_i = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}(\hat{\beta}_i)} \sim T_{n-p-1}$$

2) Soit R *une matrice de taille* $q * (q+1)$ *de rang* q *avec* $(q \leq (p+1))$ *alors :*

$$\frac{1}{q\hat{\sigma}^2}(R(\hat{\beta}-\beta))'[R(X'X)^{-1}R']^{-1}R(\hat{\beta}-\beta) \sim F_{q,n-p}.$$

où $F_{q,n-p-1}^{1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à $(q, n-p-1)$ degrés de libertés.

Proposition 4.5.3 *(Intervalles de confiance)*
1)Pour tout $i \in \{1, ..., p\}$, un intervalle de confiance de niveau $(1-\alpha)$ pour β_i est :

$$\left[\hat{\beta}_i - t_{n-p-1}^{1-\frac{\alpha}{2}}\sigma(\hat{\beta}_i), \hat{\beta}_i + t_{n-p-1}^{1-\frac{\alpha}{2}}\sigma(\hat{\beta}_i)\right]$$

où $t_{n-p-1}^{1-\frac{\alpha}{2}}$ est le quantile de niveau $(1-\frac{\alpha}{2})$ d'une loi de Student $T(n-p-1)$.
2) Un intervalle de confiance de niveau $(1-\frac{\alpha}{2})$ pour σ^2 est donné par :

$$\left[\frac{(n-p-1)\hat{\sigma^2}}{c_{n-p-1}^{1-\frac{\alpha}{2}}}, \frac{(n-p-1)\hat{\sigma^2}}{c_{n-p-1}^{\frac{\alpha}{2}}}\right]$$

où $c_{n-p-1}^{\frac{\alpha}{2}}$ est le quantile de niveau $(1-\frac{\alpha}{2})$ d'une loi de Khi-deux X_{n-p-1}^2

4.5.2 Test de signification

4.5.2.1 de signification globale du modèle

L'objectif du test global de Fisher est d'étudier la liaison globale entre Y et les variables explicatives $X_j, (j = 1, ..., p)$(significative ou pas) . On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0 \\ H_1 : \exists! \beta_j \neq 0, \qquad (j = 1, ..., p) \end{cases}$$

Il est posssible d'utiliser la statistique de Fisher F pour tester cette hypothèse :

$$F = \frac{SCE/p}{\frac{SCR}{n-p-1}}$$

où F suit une loi de Fisher avec p et $(n-p-1)$ dégeé de liberté.
Si

$$F \geq f_{1,n-p-1}^{p-\alpha},$$

(avec $f_{1,n-p-1}^{p-\alpha}$, est le quantile d'ordre $1-\alpha$ d'une loi de Fisher à $(p, n-p-1)$ degrés de liberté) alors on rejette H_0
Relation mathématique entre R^2 et la statistique F

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

4.5.2.2 Test de Student de signification du paramètre du modèle

L'objectif du test de Student est d'évaluer l'influence de la variable X_j sur $Y (j = 1, ..., p)$. On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0, \qquad (j = 1, ..., p) \end{cases}$$

où β_j est le paramètre associé à la variable explicative X_j.
L'hypothèse H_0 de nullité d'un paramètre du modèle peut être testée au moyen de statistique de Student :

$$T(\hat{\beta}_j) = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$$

comparer à $t_{n-p-1}^{1-\frac{\alpha}{2}}$, où $t_{n-p-1}^{1-\frac{\alpha}{2}}$ est le quantile d'ordre $(1 - \frac{\alpha}{2})$ d'une loi de Student à $(n - p - 1)$ degrés de liberté.
- Si $|T(\hat{\beta}_j)| \geq t_{n-p-1}^{1-\frac{\alpha}{2}}$, on rejette H_0
- Si $|T(\hat{\beta}_j)| < t_{n-p-1}^{1-\frac{\alpha}{2}}$, on ne peut pas rejeter H_0

4.6 Régression polynomiale

La régression polynomiale est une analyse statistique qui décrit la variation d'une variable aléatoire expliquée à partir d'une fonction polynomiale d'une variable aléatoire explicative.[6] . C'est un cas particulier de régression linéaire multiple, où les observations sont construites à partir des puissances d'une seule variable.
Si l'on appelle (X_i, Y_i) la i-ème réalisation du couple de variables aléatoires, on recherche le polynôme

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0,$$

permettant d'écrire

$$Y_i = P_n(X_i) + \varepsilon_i$$

le résidu ε_i, ou perturbation, étant « le plus petit » dans le sens des moindres carrés.

La régression polynomiale est une régression linéaire multiple : on peut écrire la relation, pour $X_{i,p} = X_p^i$:

$$Y_i = a_n \cdot X_{i,n} + a_{n-1} \cdot X_{i,n-1} + \ldots + a_1 \cdot X_{i,1} + a_0 + \varepsilon_i.$$

Cas particuliers
La régression linéaire est une régression polynomiale de degré 1.
Resolution par la méthode des moindres carrés Considérons un jeu de données $(X_i, Y_i) 1 \leq i \leq n$. On veut effectuer une régression par un polynôme de degré trois :

$$P_3(x) = ax^3 + bx^2 + cx + d. P_3(x) = ax^3 + bx^2 + cx + d.$$

Le carré du résidu s'écrit :

$$\varepsilon(x, y)^2 = (P_3(x) - y)^2$$

soit

$$\begin{aligned}\varepsilon(x,y)^2 = \; & x^6a^2 + 2x^5ab + 2x^4ac + 2x^3ad - 2x^3ya \\ & + x^4b^2 + 2x^3bc + 2x^2bd - 2x^2yb \\ & + x^2c^2 + 2xcd - 2xyc \\ & + d^2 - 2yd \\ & + y^2.\end{aligned}$$

On note alors :

$\varepsilon_i := \varepsilon(X_i, Y_i)$ Les valeurs a, b, c, d minimisent la somme des carrés des résidus ε : $\varepsilon = \sum_i \varepsilon_i^2$ On appelle

$$\mathrm{S}_j = \sum_i \mathrm{X}_i^j$$

et

$$\mathrm{T}_j = \sum_i \mathrm{X}_i^j \mathrm{Y}_i$$

Si le paramètre a est plus élevé ou plus bas, la valeur de e augmente. La valeur de e est donc minimale pour le a recherché, c'est-à-dire que la dérivée partielle de e par rapport à a doit être nulle :

$$\frac{\partial \varepsilon}{\partial a} = 0 \Longrightarrow 2a\mathrm{S}_6 + 2b\mathrm{S}_5 + 2c\mathrm{S}_4 + 2d\mathrm{S}_3 - 2\mathrm{T}_3 = 0.$$

On peut faire de même pour chaque paramètre, ce qui donne un système d'équations linéaires :

$$\begin{pmatrix} \mathrm{S}_6 & \mathrm{S}_5 & \mathrm{S}_4 & \mathrm{S}_3 \\ \mathrm{S}_5 & \mathrm{S}_4 & \mathrm{S}_3 & \mathrm{S}_2 \\ \mathrm{S}_4 & \mathrm{S}_3 & \mathrm{S}_2 & \mathrm{S}_1 \\ \mathrm{S}_3 & \mathrm{S}_2 & \mathrm{S}_1 & \mathrm{S}_0 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} \mathrm{T}_3 \\ \mathrm{T}_2 \\ \mathrm{T}_1 \\ \mathrm{T}_0 \end{pmatrix}.$$

4.7 Conclusion

Dans une régression multiple, il se peut que le nombre p des variables disponibles soit grand. Cette quantité d'information est parfois superfue ou redondante. Ainsi la diminution du nombre de variables réellement intéressantes dans la régression est envisageable. Soit on part du modèle complet et on retire des variables en utilisant un critère décrit sous Statistica (pas à pas descendant). Soit on part d'une régression simple et on ajoute des variables qui enrichissent le modèle (pas à pas ascendant). Sous Statistica, dans ces deux cas, on arrête d'enlever ou d'ajouter une variable au modèle en analysant la statistique F.

5 La régression Logistique

5.1 Introduction

La régression logistique ou modèle logit est un modèle de régression binomiale. Comme pour tous les modèles de régression binomiale, il s'agit de modéliser au mieux un modèle mathématique simple à des observations réelles nombreuses. En d'autres termes d'associer à un vecteur de variables aléatoires $(x_1, \ldots, x_K)$ une variable aléatoire binomiale génériquement notée y. La régression logistique constitue un cas particulier de modèle linéaire généralisé.[5] Elle est largement utilisée en apprentissage automatique.

5.2 Notation

Soit Y la variable à prédire (variable expliquée) et $X = (X_1, X_2, ..., X_J)$ les variables prédictives (variables explicatives).
Dans le cadre de la régression logistique binaire, la variable Y prend deux modalités possibles$\{1, 0\}$. Les variables X_j sont exclusivement continues ou binaires.
i.Soit Ω un ensemble de n échantillons, comportant $n_1(resp.n_0)$ observations correspondant à la modalité $1(resp.0$ de Y.
ii.$P(Y = 1)(resp.P(Y = 0)$ est la probabilité a priori pour que $Y = 1(resp.Y = 0)$. Pour simplifier, cela sera par la suite noté $p(1)(resp.p(0))$.
iii.$p(X|1)(resp.p(X|0))$ est la distribution conditionnelle des X sachant la valeur prise par Y
iv.La probabilité a posteriori d'obtenir la modalité 1 de $Y(resp.0)$ sachant la valeur prise par X est notée $p(1|X)(resp.p(0|X))$.

5.3 Fondements de la régression logistique

Le modèle de la régression logistique est un membre de la famille des modèles généralisés. Les hypothèses sur lesquelles il s'appuie sont les suivantes :
1. Sachant x_i, Y_i suit une distribution provenant de la famille exponentielle, soit la loi

binomiale, $Y_i \sim Binomiale(m_i, p_i)$;dans ce chapitre, nous considérons uniquement le cas $m_i = 1 \qquad \forall i$.
2. Le prédicteur linéaire est défini par $\eta_i = x_i'\beta$. 3. La fonction de lien donnant la relation entre $\mathbb{E}[Y_i|x_i]$ et le predicteur linéaire que nous considérons est celle qui est la plus communement utilisée, cest à dire la fonction de lien logit, et cette fonction définit théoriquement entre $-\infty$ et $+\infty$.

$$\eta_i = log\Big(\frac{\mathbb{E}[Y_i|x_i]}{1 - \mathbb{E}[Y_i|x_i]}\Big) \tag{5.1}$$

La fonction de transfert logistique est non linéaire(Fig), c'est en ce sens que l'on qualifie

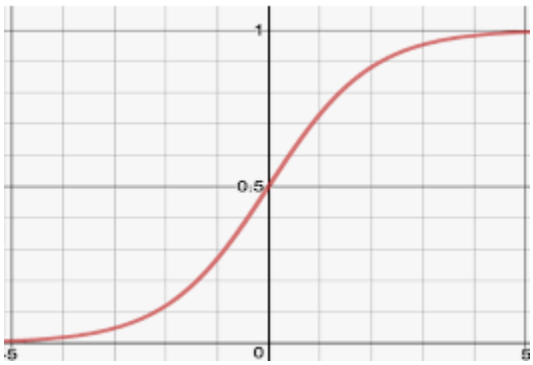


FIGURE 5.1 – La courbe de la fonction logistique

la régression logistique de régression non-linéaire dans la littérature.

5.4 Modèle de régression logistique

Le modèle de régression logistique est souvent utilisé, en pratique, afin d'évaluer l'impact de facteurs sur une variable réponse binaire. On considère préalablement que les observations sont indépendantes entre elles.

5.4.1 Ecrit du modèle pour l'individu i

$$Y_i = \begin{cases} 1 & si \quad succes \\ 0 & sinon \end{cases}$$

On définit $\pi_i = \mathbb{E}[Y_i|x_i]$, où

$$\pi_i = 0 \times P[Y_i = 0|x_i] + 1 \times P[Y_i = 1|x_i] = P[Y_i = 1|x_i]. \tag{5.2}$$

A partir de cette dernière équation, et du lien logit, tel qu'exprimé dans l'expression de η_i, on a :

$$\pi_i = \frac{exp(x_i'\beta)}{1 + exp(x_i'\beta)} = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} \tag{5.3}$$

Tel qu'on peut le déduire de 5.3, les paramètres du vecteur β s'interprètent de la façon suivante :
1. Si $\beta_j < 0$ la probabilité de succès diminue lorsque x_{ij} augmente et que la valeur de toutes les autres variables de x_i reste inchangée. Si $\beta_j = 0$, alors la variable x_{ij} n' aucun effet sur la probabilité de succès. Si $\beta_j > 0$, la probabilité d'obtenir un succès, $P[X_i = 0|x_i]$, augmente si x_{ij} augmente et que la valeur de toutes les variables de x_i demeure inchangée.
2. Si $\beta \neq 0$, la cote d'un succès, représentée par $\frac{\pi_i}{1-\pi_i}$, est multipliée par $exp(\beta_j)$ si x_{ij} croît d'une unité et que la valeur de toutes les autres variables de x_i demeure inchangée.

5.4.2 Estimation par la méthode du maximum de vraisemblance

Selon l'équation (à determiner), les paramètres à estimer sont les éléments du vecteur β. La méthode du maximum de vraisemblance est la méthode la plus commune pour estimer la valeurs des paramètres lorsque nous sommes en présence de n observations indépendantes$(Y_1, x_1), \ldots, (Y_n, x_n)$.Elle consiste prémièrement à définir la fonction de vraisemblance, soit la fonction de probabilité conjointe de $Y_1, ..., Y_n$. Celle-ci est opténue à partir de la fonction de probabilité de chaque observation individuelle en considérant l'hypothèse que les observations sont indépendantes.La fonctions de probabilité pour l'observation i est :

$$f_\beta(y_i, x_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \quad y_i \in [0, 1] \tag{5.4}$$

Donc la fonction de vraisemblance de l'échantillon est

$$L(\beta, y, x) = \prod_{i=1}^{n} \Big[\pi_i^{y_i}(1 - \pi_i)^{1-y_i}\Big]. \tag{5.5}$$

A noter que la vraisemblance correspond à la probabilité d'obtenir l'échantillon ω à partir d'un tirage dans la population, elle est donc comprise entre O et 1. La méthode du maximum de vraisemblance consiste à trouver les paramètres $\beta = (\beta_1, ..., \beta_p)$ de régression logistique qui maximisent la probabilité d'observer l'échantillon.
A cet effet, il est plus commode de travailler avec la log vraisemblance qu'est définit par :

$$\begin{aligned} l(\beta, y, x) = log\{L(\beta, y, x)\} &= \textstyle\sum_{i=1}^{n} y_i log(\pi_i) + \sum_{i=1}^{n}(1 - y_i)log(1 - \pi_i) \\ &= \textstyle\sum_{i=1}^{n} log(1 - \pi_i) + \sum_{i=1}^{n} y_i log\Big(\frac{\pi_i}{1-\pi_i}\Big) \end{aligned} \tag{5.6}$$

On sait que le logarithme est une fonction monotone, le vecteur qui β qui maximise la vraisemblance est le même qui maximise la log-vraisemblance. Cette dernière en revanche varie entre $-\infty$ et 0. On passe par la suite à la dérivation de la log-vraisemblance, que l'on appelle fonction **Score** noté $S(\beta, y, x)$:

$$S(\beta, y, x) = \frac{\partial l(\beta, y, x)}{\partial \beta} = \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \Big\{log(1 - \pi_i) + y_i log\Big(\frac{\pi_i}{1 - \pi_i}\Big)\Big\} \tag{5.7}$$

$$= \sum_{i=1}^{n} y_i x_i + \frac{\partial}{\partial \beta} \sum_{i=1}^{n}(log(1 - \pi_i)) \tag{5.8}$$

$$= \sum_{i=1}^{n} x_i(y_i - \pi_i) \tag{5.9}$$

La transition de 5.7 à 5.8 vient du fait que $log(\frac{\pi_i}{1-\pi_i}) = \eta_i = x_i'\beta$.Selon la définition de π_i, l'equation 5.9 est obtenue en déduisnat que le $log(1-\pi_i) = log[\frac{1}{1+exp(x_i\beta)}]$. L'obtention des estimateurs se fait en posant 5.9 égale à 0.

$$\sum_{i=1}^{n} x_i(y_i - \pi_i) = 0 \tag{5.10}$$

et résoudre pour les éléments de β . Sous logit, $p_i = \frac{exp(x_i'\beta)}{1-exp(x_i'\beta)'}$, l'equation 5.10 ne peut se résoudre en β avec une solution analytique. Mais on peut pallier au problème par la méthode de Newton qui est une méthode itérative.
La valeur de β qui maximise 5.5 et qui résout 5.10 est noté $\hat{\beta}$. C'est une solution unique car la log-vraisemblance est une fonction convexe.

5.4.3 Le modèle estimé

Une fois $\hat{\beta}$ estimée, on peut calculer , pour tout i, p_i : Avec le modèle Logit on écrit :

$$\begin{aligned} logit(\hat{\beta}) &= log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p \\ &= x_i'\hat{\beta} \end{aligned}$$

Notons bien que puisque $\hat{\beta}$ est un estimateur de maximum de vraisemblance, il en possède toutes les propriétés :

5.4.4 Déviance

Bien souvent, on utilise la quantité

$$D_M = -2 \times l(\beta, y, x)$$

appélée déviance(ou déviance résiduelle, en anglais residual deviance, dans le logiciel R par exemple).
Contrairement à la log-vraisemblance, elle et positive. L'objectif de l'algorithme d'optimisation est de minimise cette déviance. On peut faire le parallèle avec la somme des carrés des résidus de la régression linéaire multiple.

5.5 Test de significativité des coefficients

L'objectif des tests de significativité est d'éprouver le rôle d'une, de plusieurs, de l'ensemble, des variables explicatives. Formellement, les hypothèses nulles peuvent se décliner comme suit :

Evaluer la contibution individuelle d'une variable $H_0 : \beta_j = 0$. Ce test est systématiquement donné par les logiciels.

Normalité asymptotique des coefficients - Tests de Wald

Les estimateurs du maximum de de vraisemblance sont asymptotiquement normaux. Par conséquent lorsque les effectifs sont assez élevés, le vecteur $\hat{\beta}$ suit une loi normale multidimensionnelle.

Très facile à mettre en oeuvre puisque qu l'on dispose directement de la variance des coefficients, le test s'appuie sur la statistique de Wald W_j qui, suit une loi de X^2 à 1 degré de liberté

$$W_j = \frac{\hat{\beta}_j^2}{\hat{\sigma}_{\hat{\sigma}}^2}$$

Où $\hat{\sigma}_{\hat{\sigma}}^2$ est la variance du coefficient $\hat{\beta}_j$, lue sur la diagonale principale de la matrice de variance covariance (l'inverse du matrice hessienne).

On rejette H_0 au risque α si :

$$W_j = \frac{\hat{\beta}_j^2}{\hat{\sigma}_{\hat{\sigma}}^2} > X_\alpha^2 \Longleftrightarrow P - value < \alpha$$

5.6 Odds Ratio

D'après [7] l'odds ratio (OR), également appelé rapport des rapport des chances, rapport des cotes ou risque relatif rapproché, est une mesure statique, exprimant le degré de dépendance entre des variables aléatoires qualitatives, utilisé pour montrer les associations entre deux variables binaires.

5.6.1 Définition

Par définition, l'odds ratio correspond au rapport entre deux rapport de probabilités. Par exemple :

1er rapport : la probabilité d'être malade (ou fréquence de la maladie, notée p_1),sur la probabilité de ne pas être malade (ou fréquence des non malades, notées $1 - p_1$) chez les sujets exposés, divisé par,

2ème rapport : la probabilité d'être malade(ou fréquence de la maladie, notée p_0) sur la probabilité de ne pas être malade (ou fréquence de non malades, notée $1 - p_0$) chez les sujets non exposés.

L'odds ratio s'écrit donc :

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$$

5.6.2 Interprétation

- Si $OR > 1$, le rapport de probabilités chez les sujets exposés est plus grand que le rapport chez les sujets non exposés, on en déduit que le risque de maladie est plus

élévé chez les sujets exposés(le facteur d'exposition est un facteur de risque).
- Si $OR < 1$, le rapport de probabilités chez les sujets exposés est plus petit que le rapport chez les sujets non exposés, on en déduit que le risque de maladie est moindre chez les sujets exposés que chez les sujets no exposés(le facteur d'exposition est un facteur de protection).
- Si $OR = 1$, alors il y'a absence d'association entre maladie et exposition.[20].

5.6.3 Lien entre OR, modèle Logit et coefficient de la régression

Considerons une seule variable explicative X binaire, on a :

$$\begin{cases} logit(\pi(X)) = \beta_0 + \beta_1 X \\ logit(\pi(1)) = \beta_0 + \beta_1 \\ logit(\pi(0)) = \beta_0 \end{cases}$$

On utilise souvent en régression logistique le logarithme de l'odds ratio, plutôt sous la forme d'une différence des logits des probabilités des groupes à comparer, en remarquant que :

$$\begin{aligned} ln\left(\tfrac{p_1}{1-p_1} \times \tfrac{p_0}{1-p_0}\right) &= ln\left(\tfrac{p_1}{1-p_1}\right) - ln\left(\tfrac{p_0}{1-p_0}\right) \\ &= logit(p_1) - logit(p_0) \\ &= logit(\pi(1)) - logit(\pi(0)) \\ &= logit\left(\tfrac{\pi(1)}{\pi(0)}\right) \\ &= \beta_1 \end{aligned}$$

avec

$$OR = \left(\frac{p_1}{1-p_1} \times \frac{p_0}{1-p_0}\right)$$

On en déduit que ; L'exponentiel du coefficient peut être interprété comme un odds-ratio, et on écrit :

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

5.7 Conclusion

La régression logistique est une méthode d'analyse multivariée puissante permettant d'obtenir une quantification de l'association entre une variable expliqué qualitative binaire (une maladie par exemple) et d'autres variables explicatives quantitatives.

6 Application sur la langage R

6.1 Application de la régression linéaire multiple sur la nocivité des cigarettes

L'épidémie de tabagisme est l'une des plus graves menaces ayant jamais pesé sur la santé publique mondiale. Elle fait plus de 8 millions de morts chaque année dans le monde. Plus de 7 millions d'entre eux sont des consommateurs ou d'anciens consommateurs, et environ 1,2 million des non-fumeurs involontairement exposés à la fumée.
Ce qui est dangereux dans le tabac, ce sont les très nombreuses substances chimiques qui s'y trouvent de manière naturelle ou qui sont ajoutées par les industriels : goudrons, mercure, arsenic, acétone, acide cyanhydrique, etc. Transformées par la combustion, portées dans les poumons par la fumée, elles pénètrent dans le sang et sont à l'origine de maladies du cœur et des vaisseaux sanguins, ainsi que de nombreux cancers.
Dans ce chapitre, on va s'intéresser à étudier la relation entre CO(teneur en monoxyde de carbone)et le goudron(TAR) et le pois c'est-à-d expliquer la nocivité des cigarettes (teneur en monoxyde de carbone – CO) (y) à partir de leur composition : NICOTINE (x1) et WEIGHT (poids) (x2)

6.2 Définition des variables

- CO : Lorsqu'on brûle le tabac, il se produit une combustion. Cette combustion produit du monoxyde de carbone (CO), un gaz incolore et inodore mais très toxique. Le CO se fixe sur les globules rouges en prenant la place de l'oxygène dans le sang ce qui entraîne une moins bonne oxygénation des organes du cops(en mg).
- TAR (goudron) :Le goudron est une substance gluante et collante brun-noir contenue dans la fumée du tabac. C'est un mélange de plusieurs centaines de substances chimiques créées par la combustion du tabac (en mg).
- NICOTINE :La nicotine, contenue dans le tabac, est principalement responsable de la dépendance physique. Elle fait partie des « drogues dures ». Quand on tire sur sa cigarette, la nicotine atteint le cerveau en moins de 7 secondes. Elle va se fixer sur des

récepteurs dits « nicotiniques » dans le cerveau. Cela entraîne la libération de dopamine responsable d'effets tels que la détente, le plaisir, etc. Quand le fumeur n'a pas sa dose de nicotine, il ressent des symptômes de manque (en mg).
- WEIGHT (poids) :La masse de la cigarette (en mg).

6.3 Présentation des données

On souhaite expliquer la nocivité des cigarettes (teneur en monoxyde de carbone – CO) (y) à partir de leur composition : NICOTINE (x1) et WEIGHT (poids) (x2), soit p = 2 variables explicatives.
Nous disposons de n = 24 observations. Les observations sont sont présentés sous forme d'un tableau des données (voir figure).

	TAR	NICOTINE	POIDS	CO
Alpine	14.1	0.86	985.3	13.6
Benson&Hedges	16.0	1.06	1093.8	16.6
CamelLights	8.0	0.67	928.0	10.2
Carlton	4.1	0.40	946.2	5.4
Chesterfield	15.0	1.04	888.5	15.0
GoldenLights	8.8	0.76	1026.7	9.0
Kent	12.4	0.95	922.5	12.3
Kool	16.6	1.12	937.2	16.3
L&M	14.9	1.02	885.8	15.4
LarkLight	13.7	1.01	964.3	13.0
Marlboro	15.1	0.90	931.6	14.4
Merit	7.8	0.57	970.5	10.0
MultiFilter	11.4	0.78	1124.0	10.2
NewportLight	9.0	0.74	851.7	9.5
Now	1.0	0.13	785.1	1.5
OldGold	17.0	1.26	918.6	18.5
PallMallLight	12.8	1.08	1039.5	12.6
Raleigh	15.8	0.96	957.3	17.5
SalemUltra	4.5	0.42	910.6	4.9
Tareyton	14.5	1.01	1007.0	15.9
TrueLight	7.3	0.61	980.6	8.5
ViceroyRichLight	8.6	0.69	969.3	10.6
VirginiaSlims	15.2	1.02	949.6	13.9
WinstonLights	12.0	0.82	1118.4	14.9

6.4 Analyse univariée du Dataset

On réalise l'analyse univariée et l'ACP de notre tableau de données ci-dessus uniquement sur les variables explicatives. On dispose du tableau en fichier .txt qu'on importe dans R par la commande :

```
tab=read.table('acpprojet.txt', header = TRUE, row.names = 1)
```

```
tab
```

On visualise alors notre tableau

	TAR	NICOTINE	POIDS
Alpine	14.1	0.86	985.3
Benson&Hedges	16.0	1.06	1093.8
CamelLights	8.0	0.67	928.0
Carlton	4.1	0.40	946.2
Chesterfield	15.0	1.04	888.5
GoldenLights	8.8	0.76	1026.7
Kent	12.4	0.95	922.5
Kool	16.6	1.12	937.2
L&M	14.9	1.02	885.8
LarkLight	13.7	1.01	964.3
Marlboro	15.1	0.90	931.6
Merit	7.8	0.57	970.5
MultiFilter	11.4	0.78	1124.0
NewportLight	9.0	0.74	851.7
NOW	1.0	0.13	785.1
OldGold	17.0	1.26	918.6
PallMallLight	12.8	1.08	1039.5
Raleigh	15.8	0.96	957.3
SalemUltra	4.5	0.42	910.6
Tareyton	14.5	1.01	1007.0
TrueLight	7.3	0.61	980.6
ViceroyRichLight	8.6	0.69	969.3
VirginiaSlims	15.2	1.02	949.6
WinstonLights	12.0	0.82	1118.4

On effectue l'analyse univariée qui consiste à déterminer la moyenne et l'ecart type des variables

```
moyenne.tab = sapply(tab, mean)
```

```
moyenne.tab
```

```
      TAR    NICOTINE        POIDS
11.4833333   0.8283333 962.1708333
```

```
ecartType.tab = sapply(tab, sd)
```

```
ecartType.tab
```

```
      TAR   NICOTINE      POIDS
4.4151583  0.2655866 79.4512072
```

On calculer ensuite la matrice des corrélations entre les variables à l'aide de la commande :

```
matriceDeCorrelation.tab = round(cor(tab), 2)
```

```
matriceDeCorrelation.tab
```

	TAR	NICOTINE	POIDS
TAR	1.00	0.96	0.28
NICOTINE	0.96	1.00	0.29
POIDS	0.28	0.29	1.00

Remarque
la matrice des corrélations est à diagonale unité ce qui explique la parfaite corrélation d'une variable avec elle même.

6.5 Analyse en Composantes Principales

Pour réalisé l'ACP : on aura besoin des packages à installer tel que factomineR, factoextra, ggplot2,...
Puis on ecrit la commande :

```
analyseCompsantesPrincipales = PCA(tab)
```

```
analyseCompsantesPrincipales
```

Cette analyse réalisée sur nos 24 individus et 4 variables nous permet de voir plusieurs résultats

	name	description
1	"$eig"	"valeurs propres"
2	"$var"	"resultat pour les variables"
3	"varcoord"	"coordonnées des variables"
4	"varcor"	"correlations entres variables – dimensions"
5	"varcos2"	"qualité des variables"
6	"varcontrib"	"contributions des variables"
7	"$ind"	"resultat pour les individus"
8	"indcoord"	"coordonnées des individus"
9	"indcos2"	"cos2 for the individuals"
10	"indcontrib"	"contributions des individus"
11	"$call"	"résumés statistiques"
12	"$call$centre"	"la moyenne des variables"
13	"$call$ecart.type"	"l'erreur standard des variables"
14	"$call$row.w"	"le poids des individus"
15	"$call$col.w"	"le poids des variables"

On peut par exemple décider d'afficher les valeurs propres de notre matrice :

```
valeursPropresTableau.tab = analyseCompsantesPrincipales$eig
```

```
valeursPropresTableau.tab
```

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.10651349	70.217116	70.21712
comp 2	0.85338559	28.446186	98.66330
comp 3	0.04010092	1.336697	100.00000

$comp1, ..., comp3$ sont les composantes principales de l'ACP, les valeurs de la 1^{ere} colonne sont les valeurs propres associées aux vecteurs propres $comp1, ..., comp3$. Chaque valeur propre λ_s étant la variance de la sème composante principale.

La 2^{eme} colonne représente le pourcentage de la variance c-a-d pour chaque composante s le pourcentage de variance est

$$(\lambda_s / \sum \lambda_i) \times 100$$

Les inerties :

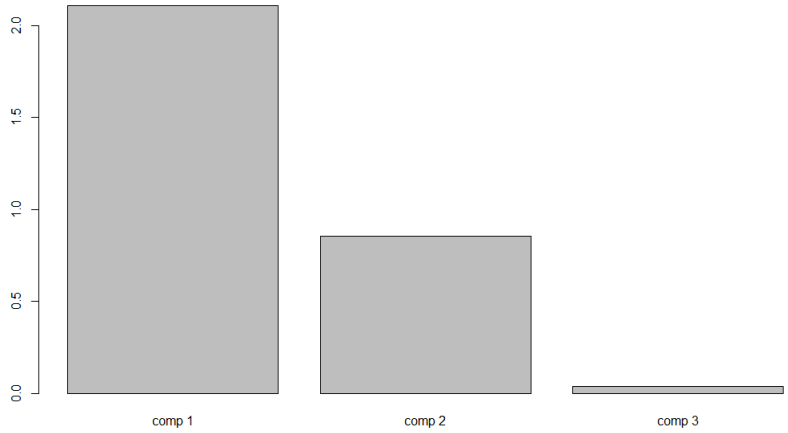
inertie = valeursPropresTableau.tab[,2]

inertie

```
   comp 1    comp 2    comp 3
70.217116 28.446186  1.336697
```

Les facteurs retenus

Eboulis des valeurs propres



Contribution et Qualités des individus

qualiteIndividus.tab = analyseCompsantesPrincipalesindcos2

qualiteIndividus.tab

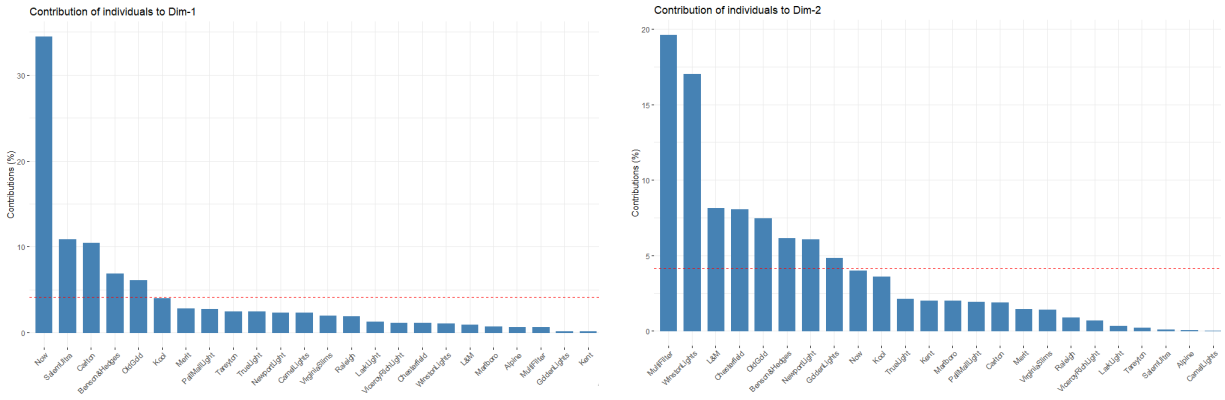
contributionIndividus.tab = analyseCompsantesPrincipalesindcontrib

contributionIndividus.tab

	Dim.1	Dim.2
Alpine	0.6764664	0.05162173
Benson&Hedges	6.8817869	6.14407262
CamelLights	2.3516634	0.02414204
Carlton	10.4629953	1.86946987
Chesterfield	1.1352043	8.04650428
GoldenLights	0.1819357	4.82285204
Kent	0.1522588	2.02106222
Kool	4.0005339	3.60639086
L&M	0.9124777	8.15339642
LarkLight	1.3120543	0.34818579
Marlboro	0.7229939	1.99512156
Merit	2.8003260	1.46014151
MultiFilter	0.6547143	19.61830236
NewportLight	2.3642737	6.05002091
Now	34.4789920	3.98453034
OldGold	6.1251690	7.46033190
PallMallLight	2.7813171	1.92215605
Raleigh	1.8941256	0.87622998
SalemUltra	10.8667925	0.10646750
Tareyton	2.5041977	0.20277227
TrueLight	2.4812521	2.12717115
ViceroyRichLight	1.1587987	0.69164082
VirginiaSlims	2.0013488	1.41503486
WinstonLights	1.0983218	17.00238091

	Dim.1	Dim.2
Alpine	0.72799816	0.022505964
Benson&Hedges	0.73244935	0.264919410
CamelLights	0.97983396	0.004075044
Carlton	0.93220683	0.067477002
Chesterfield	0.25829447	0.741701915
GoldenLights	0.08048448	0.864329151
Kent	0.14686936	0.789786173
Kool	0.73201338	0.267334443
L&M	0.21632911	0.783091167
LarkLight	0.88208542	0.094831133
Marlboro	0.39275816	0.439077322
Merit	0.82074043	0.173369718
MultiFilter	0.07584747	0.920728842
NewportLight	0.48532290	0.503119294
Now	0.95356446	0.044643033
OldGold	0.65893860	0.325136766
PallMallLight	0.69671020	0.195061411
Raleigh	0.76131367	0.142677015
SalemUltra	0.99585794	0.003952702
Tareyton	0.96823830	0.031761665
TrueLight	0.73871811	0.256561418
ViceroyRichLight	0.79543221	0.192334355
VirginiaSlims	0.77295705	0.221401605
WinstonLights	0.13710486	0.859832929

Interprétation
On relève, pour chaque axe, quels sont les individus qui ont la plus forte contribution à la formation de l'axe. Par exemple, on retient (pour l'analyse) les individus dont la contribution relative est supérieure à $\frac{100}{P}\%$. On note également si cette contribution intervient dans la partie positive ou dans la partie négative de l'axe. Soit on classe par un ordre croissant ou décroissant ces contributions ou on les représente par un histogramme. Les histogrammes des contributions par la composante 1 et la composante 2 de l'exemple précédent sont données par les figures 1.



Coordonnées, Qualités et Contributions des variables
Les tableaux suivants résument les coordonnées, qualités et contributions des variables de notre exemple.

contributionVariables.tab = analyseCompsantesPrincipalesvarcontrib

contributionVariables.tab

qualiteVariables.tab = analyseCompsantesPrincipalesvarcos2

qualiteVariables.tab

coordonneeVariables.tab = analyseCompsantesPrincipalesvarcor

coordonneeVariables.tab

	Dim.1	Dim.2
TAR	0.9640665	-0.2248178
NICOTINE	0.9646885	-0.2220544
POIDS	0.4964529	0.8680636

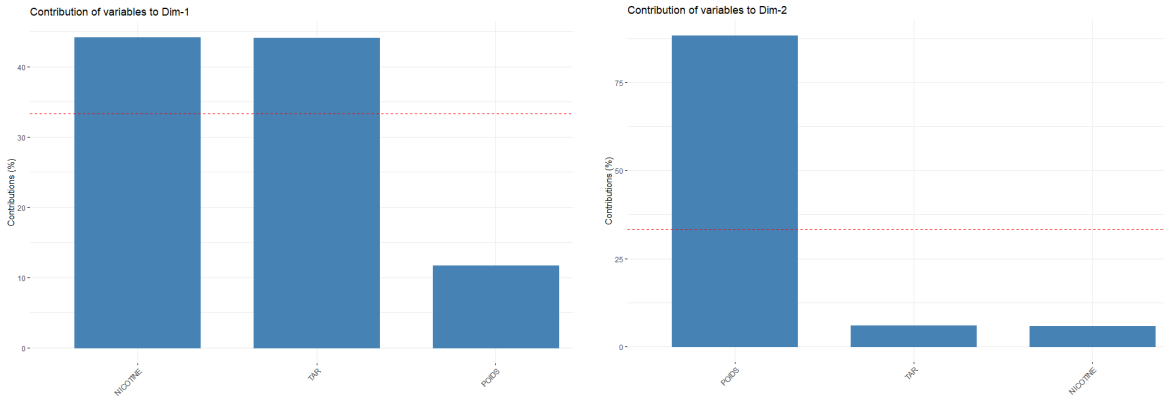
	Dim.1	Dim.2
TAR	0.9294242	0.05054306
NICOTINE	0.9306239	0.04930815
POIDS	0.2464654	0.75353438

	Dim.1	Dim.2
TAR	44.12144	5.922653
NICOTINE	44.17840	5.777945
POIDS	11.70016	88.299403

Coordonnées des variables
Les coordonnées des variables (centrées réduites initiales, $\tilde{X}_j$) dans le repère des composantes principales sont les corrélations entre ces variables et les facteurs principales

$$corr(\tilde{X}_j, F_s)$$

Plus la contribution est grande, plus la représentation est meilleure.



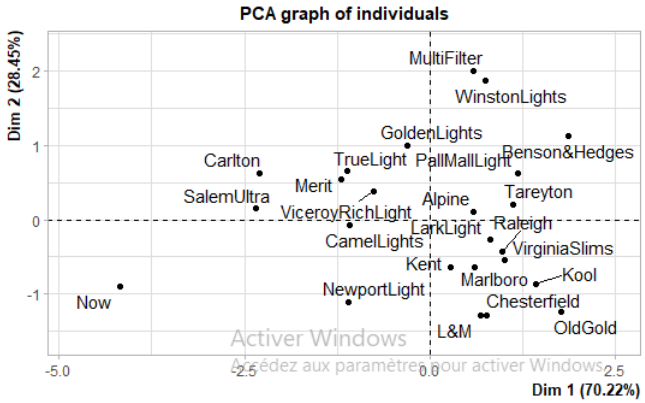
Qualité de représentation des variables

La qualité de représentation de j^{eme} variable par la s^{eme} composante principale est définie de la même façon que pour les individus

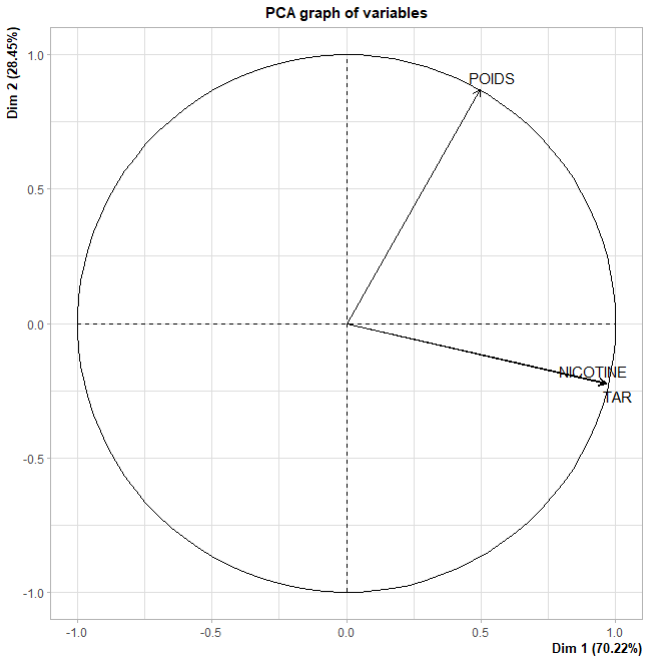
$$QLT_s(j) = \frac{(corr(\tilde{X}_j, F_s))^2}{\sum_k (corr(\tilde{X}_k, F_s))^2} = (corr(\tilde{X}_j, F_s))^2$$

Le graphe des individus :

C'est la représentation des individus sur les 2 dimensions les plus importantes



Le cercle de corrélation des variables



Synthèse

Suite aux interprétations données ci-dessus, on conclut :
La composante principale "comp 1" ou "dim 1" résume les variables TAR(goudrons) et NICOTINE. Et opposent les individus (Now, Salem, Carlon, OldGold et Benson & Hedges) qui utilisent plus de Nicotine et goudrons au reste des individus.
La composante principale "comp 2" ou "dim 2" résume le mieux la variable POIDS et oppose les individus (MultiFilter,WinstonLights, L& M, Chesterfield, Benson & Hedges, OldGold, NewportLights et GoldenLights) qui ont plus de POIDS au reste des individus.
Ce qu'on retient de cette étude est que les variables TAR(goudron) et NICOTINE sont fortement corrélées donc on peut les répresenter par la variable qui contribue le plus à la formation de l'axe 1 : c'est à dire la variable NICOTINE.

6.6 Modèle de la régression linéaire multiple

Soit le modèle de la régression linéaire multiple(RLM) avec 2 variables explicatives :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

où β_0,β_1,β_2,et β_3 sont des paramètres inconnus(les coefficients du modèle).
-x_{i1} :la i-ème valeur de la variable X1(représente la NICOTINE).
-x_{i2} :la i-ème valeur de la variable X2(correpond aux poids).
-y_i :la i-ème valeur de la variable Y(représente la teneur de monoxyde de carbone).
Le dernier terme ε_i représente la déviation entre ce que le modèle prédit et la réalité (l'erreur du modèle).
On peut récrire ce modèle sous la forme matricielle suivante :
$Y = X\beta + \varepsilon$

On modélise,sous R, le modèle de RLM par la fonction lm en faisant :

> model <- lm(CO ~ NICOTINE +POIS)

```
Call:
lm(formula = y ~ x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3532 -1.0114  0.0645  0.8522  3.5139

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.516865   4.088135  -0.616    0.545
x2          14.641152   1.321619  11.078 3.13e-10 ***
x3           0.002557   0.004418   0.579    0.569
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.613 on 21 degrees of freedom
Multiple R-squared:  0.8679,    Adjusted R-squared:  0.8554
F-statistic: 69.01 on 2 and 21 DF,  p-value: 5.859e-10
```

Comme précédemment notre but ici va etre de déterminer :
1 La valeur de la constante β_0 et des différents coefficients β_1 et β_2 qui permettent de minimiser l'erreur entre la droite de régression linéaire estimée et les valeurs réelles de Y.
2 Les variables significatives,c.à.d voir si ces différents coefficients sont différents de 0 ou non.
3 La précision de notre modèle, en utilisant entre autre. le coefficient de détermination "R-squared".

6.7 Estimation des paramètres par MC

On estime les paramètres et on obtient :

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

Il s'agit de calculer le vecteurs des estimateurs$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ définit par l'égalité suivante :

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

Les estimateurs $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ sont donnés directement,sous R, par la commande :
> model ,on obtient :

```
Call:
lm(formula = y ~ x2 + x3)

Coefficients:
(Intercept)           x2           x3
  -2.516865    14.641152     0.002557
```

Ou, on écrit l'équation , sous R, par :

$$\boxed{> (\text{hatBeta} <\text{-} (solve(t(X)\%*\%X))\%*\%(t(X)\%*\%y))}$$

où X est la matrice explicative de taille 24×4 sous R , la matrice X est écrite par la commande :
$\boxed{> \text{x<-matrix(c(rep(1,24),x1,x2),ncol=3)}}$, on obtient(voir la figure) Alors on calcule,

$$X^tX = \begin{pmatrix} 18.0896 & 19266.81 & 19.88 \\ 19266.8110 & 22363732.47 & 23092.10 \\ 19.8800 & 23092.10 & 24.00 \end{pmatrix}.$$

Donc

$$(X^tX)^{-1} = \begin{pmatrix} 0.6713505296 & -6.420687e-04 & 0.061677756 \\ -0.0006420687 & 7.501716e-06 & -0.006686085 \\ 0.0616777558 & -6.686085e-03 & 6.423732861 \end{pmatrix}.$$

Et

$$X^tY = \begin{pmatrix} 264.076 \\ 281145.080 \\ 289.700 \end{pmatrix}$$

Ainsi,on obtient :

$$\hat{\beta} = \begin{pmatrix} -2.516865419 \\ 14.641151792 \\ 0.002556661 \end{pmatrix}$$

Donc l'equation de l'hyperplan des moindres carrés est donc donnée par :

$$\hat{y_i} = -2.516865419 + 14.641151792x_{i2} + 0.002556661x_{i3}$$

6.7.1 Interprétation des résultats

le signe du coefficient nous indique le sens de la relation alors on va étudier l'impact où l'influence de chaque variables explicatives X_i sur la variable expliquée y(CO).
Pour la variable X_2 qui représente -NICOTINE : On remarque que le coefficient de régression estimée associé à la variable NICOTINE est positif cela signifie que l'augmentation de la NICOTINE implique une augmentation de la teneur de monoxyde de carbone
Pour la variable x_3 qui représente le -POIDS, le coéfficient est positif avec une valeur faible, on en deduit que l'augmentation de la masse(POIDS) du tabac entraine légèrement une augmentation de la teneur de monoxyde de carbone

6.8 Evaluation

6.8.1 Résidus

Pour tout $i \in 1.....24$, on appelle i-ème résidu la réalisation e_i de :
$e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i$
Sous R, on écrit :

```
> residuals(model)
```

On obtient,le résultat suivant(voir Figure) :

```
         1          2          3          4          5          6          7
1.00639675  0.80076867  0.53471227 -0.35870798  0.01847422 -2.23533383 -1.45074859
         8          9         10         11         12         13         14
0.02267268  0.71820024 -1.73608613  1.35804338  1.69016936 -1.57691999 -0.99509512
        15         16         17         18         19         20         21
0.10628110  0.22046533 -3.35322767  3.51386808 -1.06051388  1.05474444 -0.42129899
        22         23         24
0.53629914 -0.94491473  2.55175124
```

6.8.2 Estimation de la matrice de variance-covariance de $\hat{\beta}$

La matrice de variance-covariance, notée par $varcov(\hat{\beta})$ ou par $var(\hat{\beta})$,des coefficients est importante car elle renseigne sur la variable de chaque coefficient estimé et permet de faire les tests des hypothèses, notamment de voir si chaque coefficient est significativement différent de zéro.Elle est définie par :

$$varcov(\hat{\beta}) = \hat{\sigma^2}(X^t)(X)^{-1}$$

où $\hat{\sigma^2}$:est l'estimateur sans biais de la variance des résidus donné par :

$$\hat{\sigma^2} = \frac{\sum \hat{\varepsilon_i^2}}{n-m-1} = \frac{SCR}{n-m-1}$$

sous R, on écrit :

```
> SCR <- sum(residuals(model)^2)
> (hatSigma2_chap <- SCR/(n-m-1))
```

(où m :est le nombre des variables explicatives).
Donc, on obtient le résultat suivant :

$$\hat{\sigma^2} = \frac{SCR}{n-m-1} = \frac{54.63644}{21} = 2.601735$$

où n=24 et m=2
On peut alors calculer la matrice de variance-covariance sous R comme suit :

$> varcov <- -hatSigma2\_chap * solve(t(X)\%*\%X)$

(où varcov : est la matrice de variance covariance), ou on peut calculer la matrice de variance-covariance,sous R, directement par la commande :
> vcov(model), on obtient le résultat suivant : Les écarts-types $\hat{\sigma}(\hat{\beta}_j)$ des estimateurs $\hat{\beta}_j (j = 0, 1, 2)$ sont alors

```
             (Intercept)           x2           x3
(Intercept) 16.71285162  0.160469186 -1.739542e-02
x2           0.16046919  1.746676275 -1.670493e-03
x3          -0.01739542 -0.001670493  1.951748e-05
```

donnés par les racines carrées des éléments diagonaux de cette matrice de variance-covariance :
Sous R, on a :

> hatSigmaBetas <- sqrt(diag(vcov(model))) :

$$\hat{\sigma}(\hat{\beta}_0) = 4.088135470$$

$$\hat{\sigma}(\hat{\beta}_1) = 1.321618809$$

$$\hat{\sigma}(\hat{\beta}_2) = 0.004417859$$

6.9 Evaluation globale de la régression

6.9.1 Coefficient de détermination

Un des usages de la RLM consiste à prédire la valeur d'un Y pour un ensemble des valeurs $x_1, x_2, ..., x_p$ donné.La mesure de l'ajustement du modèle aux données est donc importante.
La proportion de variabilité expliqué par les 2 régresseurs est calculé par le coefficient de détermination R^2 qui eqt donné par la relation suivante :

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

Donc ce coefficient est un indicateur spécifique permet de traduire la variance expliquée par le modèle.
La commande R associé à R^2 est :

> R2 <- SCE / SCT

On obtient $R^2 = 86.79\%$ ce qui montre un bon ajustement.

6.10 Tests de signification

6.10.1 Test globale de Fisher

Ce test permet de répondre à la question suivante :
"Est ce que la liaison globale entre Y et les X_j est-elle significative ? "
On veut tester l'hypothèse nulle :

$$H_0 : \beta_1 = \beta_2 = \beta_3,$$

Contre l'hypothèse alternative :
$H_i : \ni j \in 1, 2, 3$ tel que $\beta_i \neq 0$
C'est-à-dire que Y dépend d'au moins une variable X_j :
On calcule la statistique de test :

$$F_{obs} = \frac{n-m-1}{m} \frac{R^2}{1-R^2} = 69.01$$

L'écriture sous R est :

> R2 <- SCE / SCT
> *F_obs* <- (R2 / m) / ((1-R2) / (n-m-1))
> *F_tab* <- qf(m = 1-0.05, df1 = m, df2 = n-m-1)

Interprétation du test
Comme le $F_{obs} = 69.01$ est de loin supérieur à la valeur critique(valeur théorique 3.403). On conclut que le test est significatif, on rejette l'hypothèse nulle H_0 au seuil de significativité $\alpha = 5\%$. Ce resultat montre qu'au moins une de nos deux variables contribue à expliquer la nocivité du tabac. Il n'explique ne nous explique pas si toutes les variables y contribuent car il est global.

6.10.2 Test de Student sur le paramètre β_j

L'objectif du test de Student est d'évaluer l'influence de la variable X_j sur Y. Donc, il permet de répondre à la question suivante :
"L'apport marginal d'une variable X_j est-il significatif ?"
Pour $j \in 0, 1, 2, 3$, on considère les hypothèses :

$$H_0 : \beta_j = 0 \quad contre \quad H_1 : \beta_j \neq 0$$

On calcul la statistique de test T_{obs} :
Pour

$$\beta_0 : T_{obs} = \frac{\hat{\beta}_0}{\hat{\sigma}(\hat{\beta}_0)} = -0.6156512$$

Pour

$$\beta_1 : T_{obs} = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} = 11.0781957$$

Pour

$$\beta_2 : T_{obs} = \frac{\hat{\beta}_2}{\hat{\sigma}(\hat{\beta}_2)} = 0.5787104$$

Pour L'écriture sous R :

$>$ *t_Obs* <- hatbeta / hatsigmabetas)
$>$ *T_tab*<- qt(m = 1-0.05/2, df = n-m-1))
$>$ ifelse(abs(*t_Obs*) $>$ *T_tab*, "rejet *H*_0", "non rejet *H*_0"

Interprétation du test
Comme la valeur critique est de 0.545, on décide de rejeter l'hypothèse H_0 pour $j = 1$ et de l'accepter pour $j = 2$. Ainsi on comprend que la variable POIDS n'est pas très influente sur le modèle en présence de la variable NICOTINE. Par contre la variable NICOTINE à une grande influence dans le modèle même en présence de la variable POIDS.

6.11 Conclusion

Nous avons commencer par une étude univariée qui nous a donnée une idée sur la moyenne, l'écart type ainsi que les corrélations entre les variables. Puis nous avons enchainé par l'analyse en composante principale qui nous a éclairé sur les variables qui contribuent le plus à la formation des axes et leur qualité de répresentation. Cela nous a permi de réduire les variables car il est plus facile de contrôler un plus petit nombre de paramètres qu'un grand. Enfin nous avons appliquée la régression linéaire multiple sur ces variables restant. En gros on retient que la nocivité du tabac vient essentiellement de la NICOTINE car elle est source d'addiction. Plus une personne est addicte au tabac,plus elle en consomme et plus le CO issu de la combustion intervient dans sa circulation sanguine ce qui peut conduire à des maladies et à la mort

Conclusion et respectives

Ce projet comprenait deux parties essentielles, une partie théorique dans laquelle nous avons rappeler le principe de l'analyse en composante principale ainsi que des généralités sur la regression multiple et logistique. Nous retenons qu'en présence d'une masse impotante de donnée faisant intervenir plusieurs variables, la reduction des données s'impose, ce qui peut facilement se resoudre par l'ACP. Ensuite on peut passer à la régression multiple. Les tests globaux et paramétriques sont là pour évaluer si nombre modèle est intéressant pour expliquer la variable endogène ainsi que l'apport marginal de chaque régresseurs. Ce projet nous a permi de comprendre qu'un grand nombre de données est toujours traitable mais qu'il faut savoir appliquées les bons outis au bon moment.

Annexe A

Rappel d'algèbre

Nous nous occupons ici uniquement des matrices réelles. On note A une matrice et A' sa transposée.

A.1 Quelques définitions

-Une matrice A est inversible s'il existe une matrice B telle que $AB = BA = I$. On note $B = A^{-1}$.[8]
La matrice carrée A est dite symétrique si $A' = A$, autrement dit si $det(A) = 0$; inversible si $det(A) \neq 0$; idempotente si $A^2 = A$; Orthogonale si $A' = A^{-1}$.
Le polynome caractéristique de la matrice A est défini par $P_A(\lambda) = det(\lambda I - A)$. Les valeurs propres sont les solutions de l'équation $P_X(\lambda) = 0$. Le vecteur x est dit vecteur propre associé à la valeur propre λ s'il est non nul et si $Ax = \lambda x$.
Le noyau d'une matrice A de dimensions $I \times J$ est le sous espace de $\mathbb{R}^J$ défini par :

$$Ker(A) = \{u \in \mathbb{R}^J \backslash Au = 0\}.$$

La définition implique que tous les vecteurs de $Ker(A)$ sont orthogonaux à tous les vecteurs lignes contenus dans la matrice A.
L'image d'une matrice B de dimensions $I \times J$ est le sous-espace de $\mathbb{R}^I$ défini par :

$$Im(B) = \{x \in \mathbb{R}^I \backslash \exists u \in \mathbb{R}^J \quad telque \quad Bu = x\}.$$

Le sous-espace $Im(B)$ est l'ensemble des vecteurs qui peuvent s'écrire comme une combinaison linéaire des colonnes de B. L'image de la matrice B est souvent appelé sous-espace engendré par les colonnes de B. La dimension de l'image de B est égale au rang de B.
Remarque A.1.1 Le sous-espace $Im(B)$ est l'orthogonal de $Ker(B')$.

A.2 Quelques propriétés

A.2.1 Les matrices $n \times p$

- $(A + B)' = A' + B'$ et $(AB)' = A'B'$.

- Le rang d'une matrice $A_{n\times p}$ est la plus petite des dimensions des deux sous espaces engendrés respectivement par les lignes et par les colonnes de A.

- $0 \leq rang(A) \leq min(n, p)$.

- $rang(A) = rang(A')$.

- $rang(AB) \leq min(rang(A), rang(B))$.

- $rang(BAC) = rang(A)$ si B et C sont inversibles.

- $rang(AA') = rang(A'A) = rang(A)$.

- Pour $p \leq n$, si A est de rang p, alors $A'A$ est inversible.

A.2.2 Les matrices $n \times n$

Soit A et B des matrices de taille $n \times n$ de termes courants a_{ij} et b_{ij} et $Tr(.)$ est le trace de la matrice $(.)$.

- $Tr(A) = \sum_{i=1}^{n} a_{ii}$.

- $Tr(A + B) = Tr(A) + Tr(B), Tr(AB) = Tr(BA)$ et $Tr(\alpha A) = \alpha Tr(A)$.

- $Tr(AA') = Tr(A'A) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2$.

- $det(AB) = det(A)det(B)$.
- Si $det(A) \neq 0$, la matrice A est inversible, d'inverse botée A^{-1}, vérifiant $(A^{-1})' = (A')^{-1}$ et $det(A^{-1} = \frac{1}{det(A)}$. De plus, si B est inversible, alors $(AB)^{-1} = A^{-1}B^{-1}$.
- La trace et le determinant ne dépendent pas des bases choisies.

A.2.3 Les matrices $n \times n$

Soit A une matrice carrée de taille $n \times n$:
- Les valeurs propres de A sont réelles.
- Les vecteurs propres de A associés à des valeurs propres différentes sont orthogonaux.
- Si une valeur propre λ est de multiplicité k, il existe k vecteurs propres orthogonaux qui lui sont associés.
- La concaténation de l'ensemble des veceturs propres orthonormés forme une matrice orthogonale U. Comme $U' = U^{-1}$, la diagonalisation de A s'écrit simplement $A = U\Delta U'$, où $\Delta = diag(\lambda_1, ..., \lambda_n)$. Pour résumer on dit qu'une matrice symétrique est diagonalisable en base orthonormée.
- $Tr(A) = \sum_{i=1}^{n} \lambda_i$ et $det(A) = \prod_{i=1}^{n} \lambda_i$.
- Les valeurs propres de A^2 sont les carrés des valeurs propres de A et ces deux

matrices ont les même vecteurs propres.

- Les valeurs propres de A^{-1} (si cette matrice existe) sont les inverses des valeurs propres de A et ces 2 matrices ont les mêmes vecteurs propres.

A.2.4 Les matrices semi-définies positives

Soit A une matrice carrée symétrique de taille $n \times n$:

- La matrice A est semi-définie positive (SDP) si $\forall x \in \mathbb{R}^n, x'Ax \geq 0$.
- La matrice A est définie positive (DP) si $\forall x \in \mathbb{R}^n - \{0\}, x'Ax \geq 0$.
- Les valeurs propres d'une matrice SDP sont toutes positives ou nulles (et réciproquement).
- La matrice A est SDP et inversible si et seulement si A est DP.
- Toute matrice A de la forme $A = B'B$ est SDP. En effet $\forall x \in \mathbb{R}^n, x'Ax = x'B'Bx = (Bx)'Bx = ||Bx||^2 \geq 0$, où $||.||$ correspond à la norme euclidienne de n.
- Toute matrice de projecteur orthogonal est SDP. En effet, les valeurs propres d'un projecteur valent 0 ou 1.
- Si B est SDP, alors $A'BA$ est SDP.
- Si A est DP et si B est SDP, alors $A + B$ est inversible et $A^{-1} - (A + B)^{-1}$ est SDP.

A.3 Propriétés des inverses

Soit M une matrice symétrique inversible de taille $p \times p$, soit u et v deux vecteurs de tailles p. Si $u'M^{-1}v \neq -1$ alors nous avons l'inverse suivante :

$$(M + uv')^{-1} = M^{-1} - \frac{M^{-1}uv'M^{-1}}{1 + u'vM^{-1}}.$$

Soit M une matrice inversible telle que :

$$M = \Big(\frac{T}{U}|\frac{U}{W}\Big),$$

avec T inversible, alors $Q = W - VT^{-1}U$ est inversible et l'inverse de M est :

$$M^{-1} = \Big(\frac{T^{-1} + T^{-1}UQ^{-1}T^{-1}}{-Q^{-1}VT^{-1}}|\frac{-T^{-1}UQ^{-1}}{Q^{-1}}\Big)$$

A.4 Propriétés des inveses

A.4.1 Projection

- Soit V un sous-espace vectoriel de $\mathbb{R}^n$, alos tout vecteur $u \in \mathbb{R}^n$ se décompose de manière unique en une somme d'un vecteur de V et d'un vecteur de $V^{\perp}$.
- Soit V un sous-espace de $\mathbb{R}^n$, l'application linéaire qui à un vecteur u fait correspondre un vecteur u^* tel que $u - u^*$ soit orthogonal à V est appelé projection orthogonale

$(u^* \in V)$.

Projection orthogonal dans l'image et le noyau d'une matrice

Le projecteur orthogonal dans l'image d'une matrice X de plein rang de dimension $n \times p$ avec $n \geq p$ est donné par :

$$P_X = X(X'X)^{-1}X'.$$

Le projecteur orthogonal dans le noyau d'une matrice X' de plein rang de dimension $n \times p$ avec $n \geq p$ est donné par :

$$P_X^{\perp} = I - X(X'X)^{-1}X' = I - P_X.$$

A.4.2 Matrice idempotente

Une matrice P est dite idempotente si $PP = P$. Une matrice de projection est idempotente.
Les matrices P_X et $P_X^{\perp}$ sont bien évidemment idempotentes, en effet :

$$\begin{aligned} P_X P_X &= \{X(X'X)^{-1}X'\}\{X(X'X)^{-1}X'\} \\ &= X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= X(X'X)^{-1}X' = P_X. \end{aligned}$$

car $X'X(X'X)^{-1} = I$
De plus

$$P_X^{\perp}P_X^{\perp} = (I - P_X)(I - P_X) = I - 2P_X + P_X P_X = I - 2P_X + P_X = I - P_X = P_X^{\perp}.$$

Le projecteur orthogonal dans le noyau d'une matrice X' de plein rang de demension $n \times p$ est donné par :

$$P_X^{\perp} = I - X(X'X)^{-1}X' = I - P_X.$$

Théorème A.4.1
Toutes les valeurs propres d'une matrice idempotentes valent 1 ou 0.
Propriété A.4.1 La treca d'une matrice idempotente est égale à son rang.
Remarque A.4.1 Le rang et la trace de la matrice $X(X'X)^{-1}X'$ sont égaux au rang de la matrice$(X'X)^{-1}$. Cette matrice est supposée de plein rang(sinon $X'X$ ne serait pas inversible). Le rang de $(X'X)^{-1}$ et donc de $P_X = X(X'X)^{-1}X'$ est égal au nombre de colonne de X. Le rang de P_X est la dimension du sous-espace sur lequel projette P_X.

A.5 Dérivée par rapport à un vecteur

A.5.1 Gradient

Soit f une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ différentiable. Le gradient de f au point x est par définition :

$$\nabla f(x) = grad(f)(x) = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_p}\right].$$

Si f est de classe C^2, le hessien de f au point x est la matrice carrée de dimension $p \times p$, souvent notée $\nabla^2 f(x)$ ou $Hf(x)$, de terme générique $[Hf(x)]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$. Le théorème de Schwarz assure que cette matrice est symétrique.

A.5.2 Dérivation d'une forme linéaire

Soit a un vecteur de $\mathbb{R}^p$, alors

$$\frac{\partial a'X}{\partial X'} = \left(\frac{\partial \sum_{i=1}^p a_i x_i}{\partial x_1}, \ldots, \frac{\partial \sum_{i=1}^p a_i x_i}{\partial x_j}, \ldots, \frac{\partial \sum_{i=1}^p a_i x_i}{\partial x_p} \right) = (a_1, \ldots, a_j, \ldots, a_p) = a'$$

A.5.3 Dérivation d'une application linéaire

Soit A une matrice de dimension $q \times p$, alors

$$AX = \begin{pmatrix} \sum_{j=1}^p a_{1j} x_j \\ \vdots \\ \sum_{j=1}^p a_{pj} x_j \\ \vdots \\ \sum_{j=1}^p a_{qj} x_j \end{pmatrix}$$

On a

$$\frac{\partial AX}{\partial x_j} = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{pj} \\ \vdots \\ a_{qj} \end{pmatrix}$$

Donc,

$$\frac{\partial AX}{\partial X'} = \left(\begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{q1} \end{pmatrix}, \ldots, \begin{pmatrix} a_{1j} \\ \vdots \\ a_{ij} \\ \vdots \\ a_{qj} \end{pmatrix}, \ldots, \begin{pmatrix} a_{1p} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{qp} \end{pmatrix} \right) = \begin{pmatrix} a_{11} & \cdots & a_{1i} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{j1} & \cdots & a_{ji} & \cdots & a_{jp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{q1} & \cdots & a_{qi} & \cdots & a_{qp} \end{pmatrix}$$

A.5.4 Dérivée d'une forme quadratique

Soit A une matrice de dimension $p \times p$, alors

$$X'A = \sum_{i=1}^p \sum_{j=1}^p a_{ij} x_i x_j = \sum_{i=1}^p a_{ii} x_i^2 + \sum_{i=1}^p \sum_{j=1, j \neq i}^p a_{ij} x_i x_j$$

Donc

$$\frac{\partial X'AX}{\partial x_k} = 2a_{kk} x_k + \sum_{j \neq i} a_{kj} x_j + \sum_{i \neq k} a_{kj} x_j + \sum_{i=1}^p a_{ik} x_i,$$

et

$$\frac{\partial X'AX}{\partial x_k} = \begin{pmatrix} \sum_{j=1}^{p} a_{1j}x_j + \sum_{i=1}^{p} a_{i1}x_i \\ \vdots \\ \sum_{j=1}^{p} a_{kj}x_j + \sum_{i=1}^{p} a_{ik}x_i \\ \vdots \\ \sum_{j=1}^{p} a_{pj}x_j + \sum_{i=1}^{p} a_{ip}x_i \end{pmatrix} = AX + A'X$$

Si la matrice A est symétrique, on a $A = A'$ et donc

$$\frac{\partial X'AX}{\partial x_k} = 2AX.$$

Bibliographie

[1] Abdelghani Ben Tahar *Analyse de données. Polycopié du Cours. Licence Mathématiques Appliquées, année académique 2020-2021*

[2] Yves Tillé *Résumé du Cours de modèle de Régression*

[3] I.Adnane & A.Guertit *Mémoire Licence : Analyse de régression, juin 2018*

[4] F. Berkani *Memoire Master : Application de la régression linéaire multiples sur la Balance Commerciale Algérienne, mai 2016*

[5] *https : //fr.wikipedia.org/wiki/R%C3%A9gression_logistique, consulté en juin 2021*

[6] *https : //fr.wikipedia.org/wiki/R%C3%A9gression_polynomiale, consulté en juin 2021*

[7] *https : //fr.wikipedia.org/wiki/Odds_ratio, consulté en juin 2021*

[8] Yves Tillé *Résumé du Cours d'économétrie. 16 decembre 2008. Chapitre 1(Eléments d'algèbre linéaire)*

[9] Arnaud Guyader *Régression linéaire, Cours tiré des quatre premiers chapitres du livre de Pierre-André Cornillon et Eric MatznerLøber, Régression avec R, paru chez Springer en 2010.*

[10] Philippe Besse *Pratique de la modelisation Statistique,Version janvier 2003 mises a jour : ' www.lsp.ups − tlse.fr/Besse*

[11] Bernard Delyon *Régression,* 1^{er} *décembre 2020 Cours de deuxième année de master*

[12] Lien du dataset, Site Unité de Recherche des Universités Lyon 2 et Lyon 1 *http : //eric.univ − lyon2.fr/ ricco/tanagra/fichiers/fr_Tanagra_Regression_Excel.pdf*