STATISTIQUES MULTIVARIEES
Analyse de données, ANOVA, RLM

A. BOURHATTAS
EISTI - Département Mathématiques

2011-2012

L'appellation "**Statistiques multivariées**" signifie que l'on s'intéresse à plusieurs variables en même temps. Il s'agit de dégager les éventuelles différences, ressemblances ou relations entre ces variables. Le cours peut être divisé en deux parties :
- Analyse de données en particulier l'anlyse en composantes principales (ACP) et l'analyse factorielle des correspondances (AFC).
- Analyse de variance (ANOVA) et régression linéaire multiple (RLM).

On appelle "**Analyse des Données**" l'ensemble des techniques permettant de décrire statistiquement de grands tableaux (nombre de lignes allant de quelques dizaines à quelques milliers, nombre de colonnes allant de quelques unités à quelques dizaines).
Ce sont des méthodes qui se sont développées grâce à l'utilisation de l'ordinateur, et qui ont une base théorique plus liée à la géométrie euclidienne qu'aux probabilités et statistiques.
Un de leur principaux objectifs est de donner des représentatons graphiques lisibles et permettant de résumer les relations entre les différentes composantes.

Nous verrons en particulier les méthodes multifactorielles qui forment un ensemble de méthodes, avec une même base théorique et plusieurs déclinaisons selon les types de tableaux, à savoir les deux types principaux :
- ACP : Analyse en Composantes Principales, pour les tableaux de variables quantitatives.
- AFC : Analyse Factorielle des Correspondances, pour les tableaux de contingence.

L'**analyse de variance** et la **régression linéaire multiple** quant à elles nous ramèneront dans l'univers des statistiques puisqu'il s'agit en fait d'outils pour tester des hypothèses, d'influence d'une ou plusieurs variables qualitatives sur une variable quantitative dans le cas de l'ANOVA, ou de relations linéaires entre variables quantitatives dans le cas de la RLM.

Table des matières

Chapitre 1

Analyse en composantes principales ACP

1.1 Rappels de statistiques descriptives

1.1.1 Une variable

Pour un **caractère quantitatif** X sur une **population** constituée de n **individus**, La moyenne est donnée par :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La variance, indicateur de dispersion de la variable statistique est :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \overline{x}^2$$

s_x est alors l'écart-type.

1.1.2 Deux variables

Lorsqu'on veut étudier la relation entre deux variables X et Y, on est amené à calculer la covariance de ces deux variables donnée par :

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

On montre alors que :
La droite de régression D de y en x a pour équation $y = ax + b$ avec :
$a = \dfrac{s_{xy}}{{s_x}^2}$ et $b = \bar{y} - a\bar{x}$ La droite D passe par le point moyen $G(\bar{x}; \bar{y})$ du nuage.
Pour apprécier la qualité d'un ajustement linéaire entre ces deux variables, on introduit le coefficient de corrélation linéaire entre x et y qui est défini par :

$$r = \frac{s_{xy}}{s_x \times s_y}$$

1.2 Exemple introductif pour l'ACP

Commençons par un exemple simplifié. Comment représenter géométriquement et avec un maximum d'information le relevé de notes de 5 matières des 10 élèves suivants :

	math.	phys.	info.	fran.	ang.
élève 1	8	9	12	10	11
élève 2	14	13	16	12	13
élève 3	4	6	7	11	10
élève 4	13	15	12	7	6
élève 5	8	7	10	9	8
élève 6	1	2	5	8	7
élève 7	10	8	12	13	11
élève 8	7	10	14	12	15
élève 9	13	13	8	5	4
élève 10	3	6	8	8	9

On sait représenter chaque variable et déterminer ses caractéristiques. On sait également représenter chaque couple de variable et étudier les corrélations correspondantes, mais on n'a pas, jusqu'ici, de moyen de représenter le tableau en entier.
En effet, chaque variable (note de matière) est ici représentée par un point de $\mathbb{R}^{10}$. Ce qui forme un nuage de 5 points de $\mathbb{R}^{10}$.
Et chaque individu (élève) est représenté par un point de $\mathbb{R}^5$. D'où un nuage 10 points de $\mathbb{R}^5$.

L'utilisation d'un logiciel statistique comme SAS ou R donne les représentation suivantes :

Les différentes variables (ici les notes de chaque matière) sont représentées dans un plan.
Le premier axe de ce plan représente une variable virtuelle qui est corrélée positivement avec chaque matière. C'est la première composante principale de la méthode ACP. Sur cet axe, chaque élève aura une abscisse d'autant plus grande que ses notes dans les différents matières seront élevées.
Le deuxième axe représente la deuxième composante principale. Celle-ci est corrélée positivement avec les maths et la physique et négativement avec le français et l'anglais. La deuxième composante distingue donc le profil scientifique du profil littéraire.
On obtient également un graphique pour les individus.

On remarque rapidement que l'ordre des élèves sur le premier axe est celui de leurs résultats globaux. Le premier axe indique donc le résultat général.
Le deuxième axe distingue de manière éclatante les élèves 8, 3 et 7 d'un côté et les élèves 9 et 4 de l'autre. Les 3 premiers ont des notes notablement différenciées entre les matières en faveur des matières littéraires, les deux autres idem mais

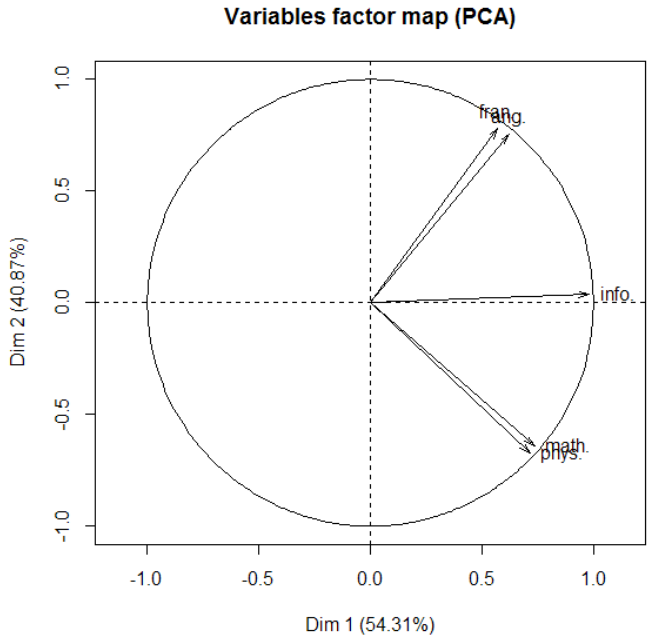

FIGURE 1.1 – ACP variables

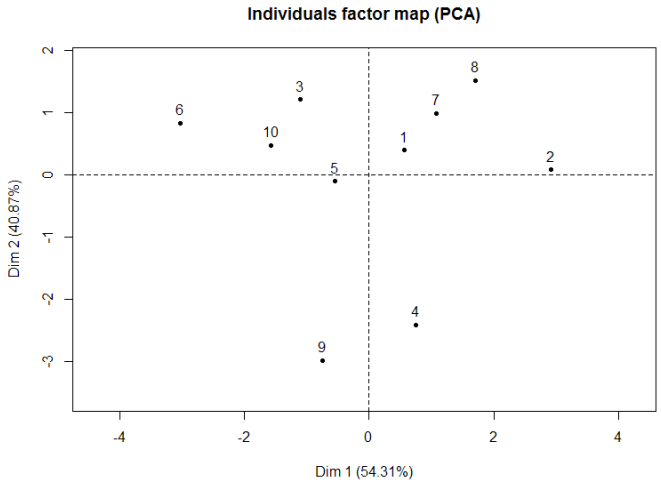

FIGURE 1.2 – ACP individus

en faveur des matières scientifiques. Les élèves dont l'ordonnée est proche de 0 sont ceux dont les notes sont les moins dispersées.

1.3 Cas général

On part d'une matrice X rectangulaire de taille $n \times p$. Celle-ci correspond à n observations ou individus sur lesquels sont mesurées p variables quantitatives.

$$X = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^j & \cdots & x_1^p \\ \vdots & \vdots & & \vdots & & \vdots \\ x_i^1 & x_i^2 & \cdots & x_i^j & \cdots & x_i^p \\ \vdots & \vdots & & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^j & \cdots & x_n^p \end{pmatrix}$$

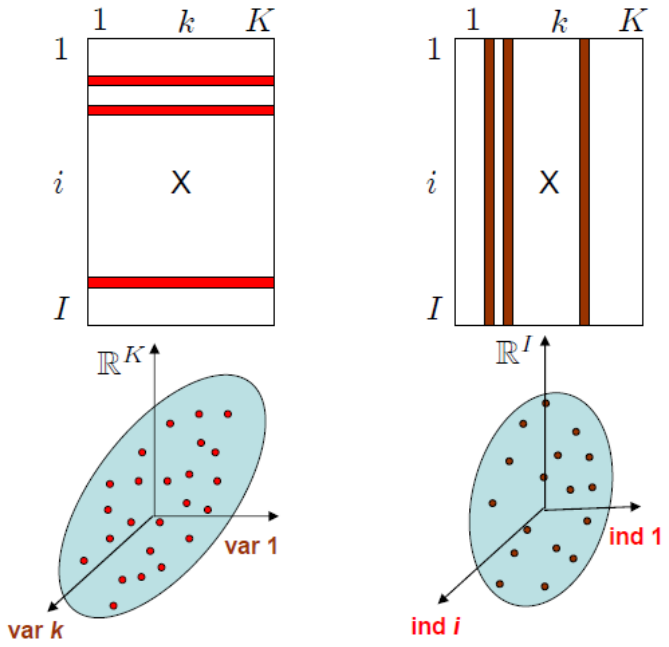

FIGURE 1.3 – les 2 représentations du tableau

Le tableau X peut admettre deux représentations :
- l'une dans un espace vectoriel $\mathbb{R}^p$ avec un nuage de n points correspondant chacun à une ligne X_i(ou individu) ;
- l'autre dans un espace vectoriel $\mathbb{R}^n$ avec un nuage de p points correspondant chacun à une colonne X^j(ou variable).

On commence par **centrer le nuage** autour du point moyen $G = \begin{pmatrix} \overline{x^1} \\ \vdots \\ \overline{x^j} \\ \vdots \\ \overline{x^p} \end{pmatrix}$

On obtient une **nouvelle matrice** $Y = (y_i^j)$ avec $y_{ij} = x_i^j - \overline{x^j}$ pour $1 \leq i \leq n$ et $1 \leq j \leq p$.
La **matrice des variances-covariances** est alors obtenue par :

$$V = \frac{1}{n} Y^t \, Y = (v_{jk})_{1 \leq j,k \leq p}$$

avec $\begin{cases} v_{jk} = s_{jk} = \text{covariance de } X^j \text{ et } X^k \text{ si } k \neq j \\ \mathrm{v}_{jj} = s_j^2 = \text{variance de } X^j \end{cases}$

Notons que si, en plus de centrer les variables, on les réduisait (en divisant chacune par son écart-type), la matrice V donnerait les différents coefficients de corrélation. Nous verrons que ce procédé peut être utile pour réduire l'effet des variables les plus dispersées ou pour effacer l'effet des unités de mesure de chacune.

1.4 Dispersion du nuage

Nous munissons l'espace des individus $\mathbb{R}^p$ de la distance euclidienne associée au produit scalaire usuel : $\langle X; Y \rangle = \sum_i x_i y_i$.
La distance entre deux individus est alors donnée par :
$\|X_i - X_j\|^2 = \sum_{k=1}^{p} (x_i^k - x_j^k)^2$.
En prenant la représentation centrée Y_i des individus, la dispersion du nuage peut être mesurée par :

$$\sum_{1 \leq i,j \leq n} d^2(Y_i, Y_j) = \sum_{1 \leq i,j \leq n} \| \, Y_i\vec{Y}_j \, \|^2 = \sum_{1 \leq i,j \leq n} \| \, O\vec{Y}_i \, \|^2 + \| \, O\vec{Y}_j \, \|^2 - 2 \left\langle O\vec{Y}_i; O\vec{Y}_j \right\rangle = 2n \sum_{j=1}^{n} \| \, O\vec{Y}_j \, \|^2$$

Définition 1.4.1
On appelle **Inertie** du nuage le nombre :

$$I = \frac{1}{n} \sum_{j=1}^{n} \| \, O\vec{Y}_j \, \|^2$$

On sait qu'il s'agit tout simplement de la somme des variances, ou encore de la trace de la matrice V.
Le but de l'ACP est de déterminer un sous-espace H de $\mathbb{R}^p$ de dimension $k < p$, en général $k = 2$ de sorte que la projection orthogonale du nuage sur H soit la moins modifiée possible par rapport au nuage original.
Si on note $P_H Y_j$ le projeté orthogonal de Y_j sur H, le problème revient, de manière équivalente, à :

- Minimiser l'inertie par rapport à H du nuage
  $\min J_H = \frac{1}{n} \sum_{j=1}^{n} \| \, O\vec{Y}_j - O\vec{P_H Y_j} \, \|^2$
- Maximiser l'inertie dans H du projeté orthogonal du nuage
  $\max I_H = \frac{1}{n} \sum_{j=1}^{n} \| \, O\vec{P_H Y_j} \, \|^2$

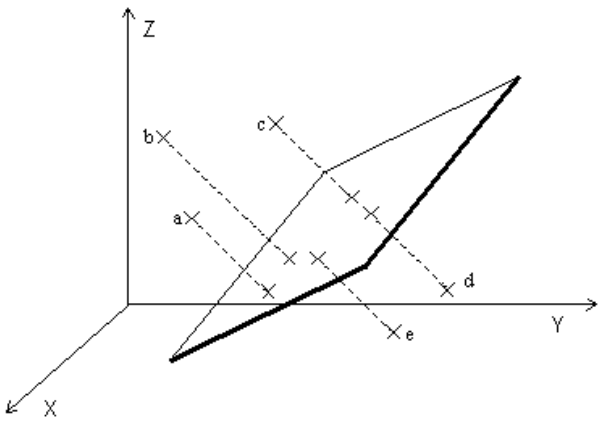

FIGURE 1.4 – Projection sur un plan

1.5 Meilleure projection

Pour répondre au problème précédent, on commence par rechercher la droite qui maximise l'inertie projetée, c à d un vecteur unitaire $\vec{u}$ solution de :

$\max I_u = \frac{1}{n}\sum_{j=1}^{n}\left\langle O\vec{Y}_j;\vec{u}\right\rangle^2$ sous la contrainte $\| \vec{u} \|^2 = 1$

Ensuite, on cherche un vecteur orthogonal à $\vec{u}$ qui répond au même problème, ce qui permet d'obtenir un plan. On continue ainsi jusqu'à la dimension k.
Un peu de géométrie euclidienne et d'algèbre linéaire (voir TD 1) permet d'établir le théorème suivant :
Théorème 1.4.2 :
Le sous-espace H_k de dimension k recherché est obtenu de la manière suivante :

- **Si u_k est le vecteur propre unitaire de V associée à la k-ième plus grande valeur propre. Alors $H_k = Vect(u_1, ..., u_k)$.**
- **L'inertie projetée sur le k-ième axe propre est donnée par $I_{u_k} = \lambda_k$.**
- **l'inertie sur H_k est la somme des inerties moyennes sur les k axes propres principaux : $I_{H_k} = \sum_{i=1}^{k} \lambda_i$.**

Il faut noter que la matrice V, étant symétrique semi-définie positive, est diagonalisable avec des valeurs propres réelles positives. De plus, les vecteurs propres forment une base orthonormée.
Les droites portées par les vecteurs propres ainsi obtenus sont appelés **axes principaux**.

1.6 Composantes principales

Les vecteurs u_l obtenus par diagonalisation de V nous permettent de définir de nouvelles variables C^l combinaisons linéaires des Y^j.

Ces variables, appelées **facteurs principaux**, sont définies par :

$$C^l = Yu_l = \sum_{j=1}^{n} u_l^j Y^j$$

Elles expliquent mieux que les originales, la dispersion du nuage. Elles ont entre autres qualités d'être non corrélées. Plus précisément :

$\text{cov}(C^l, C^k) = \begin{cases} \lambda_l & \text{si } l = k \\ 0 & si \quad l \neq k \end{cases}$

Pour exprimer les relations entre les composantes principales et les variables d'origine on utilise les coefficients de corrélation.
la covariance est donnée par : $cov(C^l, Y^j) = \lambda_l u_l^j$.
Ce qui donne un coefficient de corrélation entre C^l et Y^j égal à :

$$r(C^l, Y^j) = \frac{cov(C^l, Y^j)}{\sqrt{var(C^l) var(Y^j)}} = \sqrt{\lambda_l} \frac{u_l^j}{s_j}.$$

Dans la représentation graphique des variables donnée par l'ACP on se place dans le cercle unité et on représente chaque variable par le couple de ses coefficients de corrélation avec les 2 premières composantes principales.

Ainsi, Y^j sera représentée par le point de coordonnées $(\sqrt{\lambda_1}\frac{u_1^j}{s_j}; \sqrt{\lambda_2}\frac{u_2^j}{s_j})$.

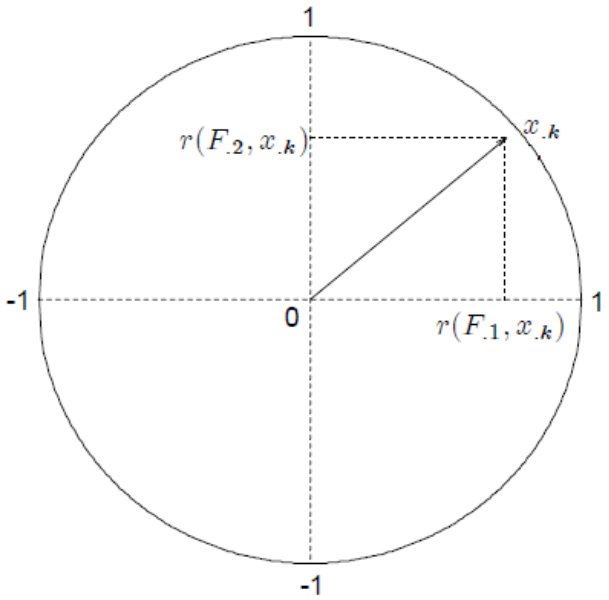

FIGURE 1.5 – Représentation des variables

Les **composantes principales** sont les nombres : $C_i^l = \langle Y_i; u_l \rangle$ qui permettent de placer l'individu i dans l'axe factoriel u_l. Ce sont les coordonnées de la projection des individus sur les axes définis par les u_l.

1.7 Choix du nombre de composantes

Il y a plusieurs stratégies pour le choix du nombre de composantes principales à retenir.
Rappelons que la somme des variances de toutes les variables, ou encore l'inertie du nuage, est égale à la trace de la matrice V.
Elle est donc égale à la somme de ses valeurs propres. La contribution en termes de variance de chaque composante C^l est donc égale à : $\frac{\lambda_l}{\sum_{i=1}^{p} \lambda_i}$.

- Une stratégie consiste à choisir un nombre de composantes garantissant un pourcentage minimal de contribution à la variance totale. On retient en général 80% ou 90%.
- Une deuxième règle consiste à retenir toutes les composantes dont la contribution est supérieure à $\frac{100}{p}\%$.
- Une troisième méthode est basée sur la représentation graphique de la contribution de chaque composante. On recherche un "coude" dans la courbe affine par morceaux qui relie les contributions par ordre décroissant.

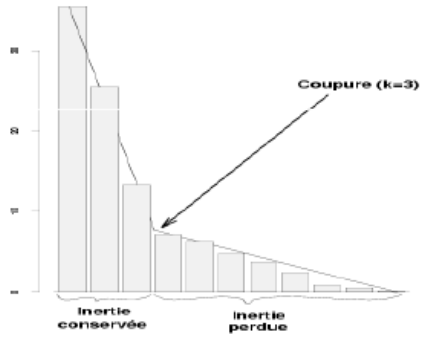

FIGURE 1.6 – Choix du nombre d'axes

Nous utiliserons principalement la première méthode.
Lorsque nous choisissons de retenir un nombre k de composantes, nous remplaçons les individus Y_i par $\hat{Y}_i = \sum_{l=1}^{k} C_i^l u_l$.
Si on note $C(k)$ la matrices des k composantes principales retenues et $U(k)$ la matrice des k vecteurs propres correspondants de V, ceci revient à remplacer la matrice Y par $\hat{Y} = C(k)U(k)$.
L'erreur commise est alors $E(k) = \frac{1}{n} \sum_{i=1}^{n} \| Y_i - \hat{Y}_i \|^2 = \sum_{i=k+1}^{p} \lambda_i$.

1.8 Aide à l'interprétation

1. *Interprétation des axes*
   Pour interpréter le sens d'un axe, on commence par regarder quels sont les variables et les individus qui participent le plus à la formation de l'axe.

Pour celà, on regarde la contribution de chaque individu et chaque variable à l'inertie de cet axe. On retient ceux et celles dont la contribution est supérieure à la moyenne.
La contribution de l'individu i à l'inertie de l'axe k est donnée par :

$$CTR_k(Y_i) = \frac{(C_i^k)^2}{\lambda_k}$$

Il faut noter que la sommes des contributions des individus à un axe est égale à 1.
La règle est de retenir les individus dont la contribution est supérieure à $\frac{1}{n}$, càd lorsque $C_i^k > \sqrt{\lambda_k}$. La contribution de la variable j à l'inertie de l'axe k est quant à elle :

$$CTR_k(Y^j) = (u_{jk})^2$$

En pratique : On retient les variables dont la contribution est supérieure à la contribution moyenne ($> \frac{1}{\sqrt{p}}$).

2. *Etude des proximités entre points*

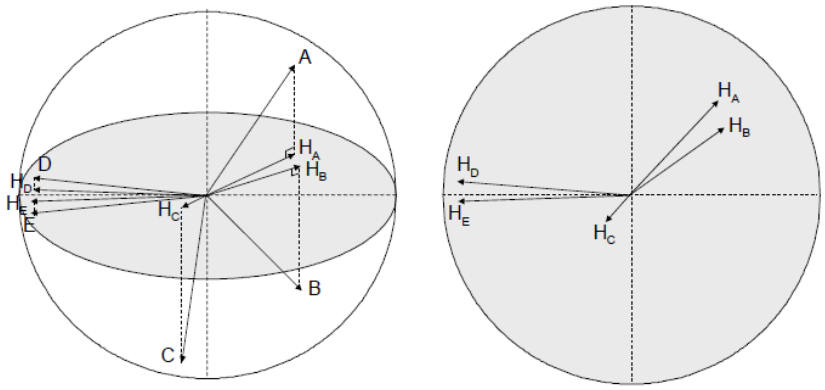

FIGURE 1.7 – Proximité réelle ou virtuelle

Sur un graphique issu d'une ACP, il arrive qu'une proximité entre points (individus ou variables) soit plus due à la projection qu'à une réelle proximité des points.
Voilà pourquoi il faut s'intéresser à la qualité de représentation d'un individu i sur un axe principal k.
Pour évaluer cette qualité, on utilise $\cos^2(\theta_{ik})$ où θ_{ik} représente l'angle entre l'individu et sa projection.
Plus cet angle sera proche de 0, plus l'individu sera bien représenté, le cosinus carré sera proche de 1.
On a le même résultat avec les variables et leur $\cos^2$. Une variable est d'autant mieux représentée sur un axe qu'elle est proche du bord du cercle des corrélations et de l'axe, d'autant plus mal représentée qu'elle est proche de l'origine.

1.9 Variables et individus supplémentaires

Il arrive qu'on ait besoin de placer des individus ou des variables supplémentaires sur les plans de projection. Soit parce que ces variables ou individus sont atypiques auraient biaisé l'ACP si on les avait intégrés. Soit parce qu'ils sont plus faciles à appréhender et que leur projection pourraient nous aider à mieux interpréter les axes supplémentaires.
On les appelle dans ce cas des variables ou individus supplémentaires.
La projection d'un individu supplémentaire Y_{n+1} représenté par les modalités des variables initiales centrées

$$Y_{n+1} = (Y_{n+1}^1, Y_{n+1}^2, ..., Y_{n+1}^p)$$

sur un axe principal dirigé par le vecteur propre u_l se fait comme pour les autres individus en calculant sa l-ème composante principale :

$$C_{n+1}^l = \langle Y_{n+1}; u_l \rangle .$$

Pour une variable supplémentaire centrée Y^{p+1}, il faut calculer ses coefficients de corrélations avec les facteurs principaux.
Pour celà, on commence par calculer ses coefs de corrélation avec les $(Y^j)_{1 \leq j \leq p}$ grâce au produit matriciel :

$$^tYY^{p+1}$$

Ensuite, on utilise les composantes de u_l comme coefficients de combinaison linéaire pour obtenir la corrélation avec l'axe principal correspondant.

$$r(C^l, Y^{p+1}) =^t Y^{p+1} Y u_l$$

1.10 Dualité variable-individu

Nous avons jusque là travaillé avec le nuage des individus pour lequel nous avons cherché les meilleures directions de projection afin d'expliquer l'inertie totale.
Nous aurions pu faire le choix de travailler sur le nuage des variables. La solution du problème d'optimisation de la projection se trouve alors être définie par les vecteurs propres da la matrice $W = \frac{1}{n} Y \ Y^t$.
Les valeurs propres sont les mêmes et les vecteurs propres v^l sont liés aux u_l par :

$$\left\langle Y^j; u_k \right\rangle = \sqrt{\lambda_k} v_j^k$$

1.11 ACP centrée réduite

Tous les logiciels statistiques appliquent, par défaut, l'ACP sur des variables centrées réduites. Cela signifie que toutes les variances sont égales à 1 et que les covariances sont en fait les coefficients de corrélation.

$$\forall i, j \quad s_j = 1 \quad et \quad s_{ij} = r_{ij}$$

La matrice V que l'on diagonalise se trouve donc être la matrice des corrélations. Ses coefficients diagonaux sont égaux tous à 1, ce qui explique que la somme des valeurs propres soit égale au nombre de variables indépendantes.
Inutile donc de rappeler qu'il faut centrer et réduire tout individu ou variable avant toute tentative d'interprétation.

Chapitre 2

Analyse factorielle des correspondances AFC

2.1 Dans quel cas utilise-t-on l'AFC

L'AFC est une technique d'analyse qui a été développée pour les tableaux de contingence, c à d les tableaux croisés de répartition d'une population entre les modalités de deux caractères quantitatifs ou discrets (ou encore continus discrétisés par regroupement en classe).
Dans ces tableaux, les lignes et les colonnes jouent des rôles symétriques (on peut tout à fait les échanger). Par extension, elle s'applique à tout tableau positif dont les sommes des lignes ainsi que les sommes des colonnes ont une signification précise et ou les lignes et les colonnes jouent des rôles symétriques.

2.2 Rappels sur contingence et indépendance

Le tableau de contingence est un moyen particulier de représenter simultanément deux caractères observés sur une même population, qu'ils soient qualitatifs, discrets ou bien continus et regroupés en classes.
Si on note X et Y les deux caractères, et n la taille de l'échantillon. Les modalités ou classes de X étant notées $c_1, \ldots, c_r$, celles de Y étant notées $d_1, \ldots, d_s$. On définit :
- n_{hk} l'effectif conjoint de c_h et d_k : c'est le nombre d'individus pour lesquels X prend la valeur c_h et Y la valeur d_k,
- $n_{h\bullet} = \sum_{k=1}^{s} n_{hk}$ l'effectif marginal de c_h : c'est le nombre d'individus pour lesquels X prend la valeur c_h,
- $n_{\bullet k} = \sum_{h=1}^{r} n_{hk}$ l'effectif marginal de d_k : c'est le nombre d'individus pour lesquels Y prend la valeur d_k.

On représente ces valeurs dans un tableau à double entrée, dit tableau de contingence.

Chaque ligne et chaque colonne correspond à un sous-échantillon particulier.

X \ Y	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
⋮	⋮		⋮		⋮	⋮
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
⋮	⋮		⋮		⋮	⋮
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

FIGURE 2.1 – tableau de contingence

La ligne d'indice h est la répartition sur $d_1, \ldots, d_s$ des individus pour lesquels le caractère x prend la valeur c_h.
La colonne d'indice k est la répartition sur $c_1, \ldots, c_r$ des individus pour lesquels le caractère y prend la valeur d_k.
En divisant les lignes et les colonnes par leurs sommes, on obtient sur chacune des distributions empiriques constituées de fréquences conditionnelles. Pour $h = 1, \ldots, r$ et $k = 1, \ldots, s$, on les notera :
$f_{k|h} = \dfrac{n_{hk}}{n_{h\bullet}}$ et $\quad f_{h|k} = \dfrac{n_{hk}}{n_{\bullet k}}$.
Ces distributions empiriques conditionnelles s'appellent les profils-lignes et profils-colonnes.

Les deux caractères sont indépendants si la valeur de l'un n'influe pas sur les distributions des valeurs de l'autre. Lorsque c'est le cas, les profils-lignes sont tous peu différents de la distribution marginale de Y, et les profils-colonnes de celle de X :
$f_{k|h} = \dfrac{n_{hk}}{n_{h\bullet}} \approx f_{\bullet k} = \dfrac{n_{\bullet k}}{n}$ et $\quad f_{h|k} = \dfrac{n_{hk}}{n_{\bullet k}} \approx f_{h\bullet} = \dfrac{n_{h\bullet}}{n}$.
Ceci est équivalent à dire que les fréquences conjointes sont proches des produits de fréquences marginales.
$f_{hk} = \dfrac{n_{hk}}{n} \approx f_{h\bullet}\, f_{\bullet k} = \dfrac{n_{h\bullet}}{n}\, \dfrac{n_{\bullet k}}{n}$.
On se retrouve donc avec deux distributions de probabilité sur l'ensemble produit $\{c_1, \ldots, c_r\} \times \{d_1, \ldots, d_s\}$:
– Les fréquences conjointes réalisées par l'échantillon d'une part (les f_{hk}).
– Les fréquences théoriques en cas d'indépendance et qui sont égales aux produits de fréquences marginales d'autre part, (les $f_{\text{théo}} = f_{h\bullet} f_{\bullet k}$).
Un des moyens de quantifier leur proximité est de calculer la distance du chi-deux de l'une par rapport à l'autre. Dans ce cas particulier, on parle de chi-deux de contingence .
Définition : La distance du chi-deux de contingence de la distribution empirique (f_{hk}) à la distribution théorique $(f_{h\bullet} f_{\bullet k})$ vaut :

$$\begin{aligned} D^2_{\chi^2} &= \sum_{h=1}^{r} \sum_{k=1}^{s} \frac{(f_{hk} - f_{h\bullet} f_{\bullet k})^2}{f_{h\bullet} f_{\bullet k}} \\ &= -1 + \sum_{h=1}^{r} \sum_{k=1}^{s} \frac{n_{hk}^2}{n_{h\bullet}\, n_{\bullet k}} \ . \end{aligned}$$

La distance du chi-deux vaut 0 si les deux caractères sont indépendants.
On montre que, sous l'hypothèse $(\mathcal{H}_0)$ d'indépendance, et pour n assez grand, $nD^2_{\chi^2}$ suit une loi du chi-deux, dont le nombre de degrés de liberté est donné par : $(r-1)(s-1)$.
C'est cette distance qui va nous permettre de définir la distance entre deux profils lignes ou deux profils colonnes et ainsi de définir l'inertie du nuage de chaque type de profil.

2.3 Nuages et profils

On part d'un tableau de contingence $N = (n_{ij})_{1\leq i\leq r, 1\leq j\leq s}$ d'effectif total n.
Prenons par exemple la répartition de 115 étudiants du cycle ingénieur de l'EISTI en fonction, d'une part de leur origine, et d'autre part de l'aspect dominant de leurs options.

	MATHS	INFO	MANAGT
CPGE	25	32	3
CPI	3	10	7
AUTRES	2	8	25

Soit

$$N = \begin{pmatrix} 25 & 32 & 3 \\ 3 & 10 & 7 \\ 2 & 8 & 25 \end{pmatrix}$$

La distribution marginale en ligne est donnée par les : $n_{i\bullet} = \sum_{j=1}^{r} n_{ij}$.

Celle des colonnes par les : $n_{\bullet j} = \sum_{i=1}^{s} n_{ij}$.

	MATHS	INFO	MANAGT	TOTAL
CPGE	25	32	3	60
CPI	3	10	7	20
AUTRES	2	8	25	35
TOTAL	30	50	35	115

On introduit alors les matrices diagonales :
$D_L = \text{diag}(n_{1\bullet}, n_{2\bullet}, ..., n_{s\bullet})$ et $D_C = \text{diag}(n_{\bullet 1}, n_{\bullet 2}, ..., n_{\bullet r})$

$$D_L = \begin{pmatrix} 60 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 35 \end{pmatrix} \text{ et } \mathrm{D}_C = \begin{pmatrix} 30 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 35 \end{pmatrix}$$

Le tableau des profils-lignes est alors donné par : $X = D_L^{-1}N$, dont les coefficients sont les fréquences conditionnelles $\dfrac{n_{ij}}{n_{i\bullet}}$.
Matrice des profils lignes

$$X = D_L^{-1}N = \begin{pmatrix} \frac{25}{60} & \frac{32}{60} & \frac{3}{60} \\ \frac{3}{20} & \frac{10}{20} & \frac{7}{20} \\ \frac{2}{35} & \frac{8}{35} & \frac{25}{35} \end{pmatrix} = \begin{pmatrix} 0,417 & 0,533 & 0,05 \\ 0,15 & 0,5 & 0,35 \\ 0,057 & 0,229 & 0,714 \end{pmatrix}$$

Le tableau des profils-colonnes est quant à lui donné par : $X = ND_C^{-1}$, et ses les coefficients sont les fréquences conditionnelles $\frac{n_{ij}}{n_{\bullet j}}$.
Matrice des profils colonnes

$$X = ND_C^{-1} = \begin{pmatrix} \frac{25}{30} & \frac{32}{50} & \frac{3}{35} \\ \frac{3}{30} & \frac{10}{50} & \frac{7}{35} \\ \frac{2}{30} & \frac{8}{50} & \frac{25}{35} \end{pmatrix} = \begin{pmatrix} 0,83 & 0,64 & 0,086 \\ 0,1 & 0,2 & 0,2 \\ 0,07 & 0,16 & 0,714 \end{pmatrix}$$

Avant de parler du nuage associé par exemple aux profils-lignes, il faut noter que toutes les lignes n'ont pas le même poids selon la distribution marginale des lignes.
Par conséquent, le profil-ligne (i), qui correspond à la i-ème modalité du premier caractère, sera affecté du poids $f_{i\bullet} = \frac{n_{i\bullet}}{n}$.
Ceci revient à introduire une matrice des poids égale à

$$\frac{1}{n}D_L = \begin{pmatrix} \frac{60}{115} & 0 & 0 \\ 0 & \frac{20}{115} & 0 \\ 0 & 0 & \frac{35}{115} \end{pmatrix} = \begin{pmatrix} 0,522 & 0 & 0 \\ 0 & 0,174 & 0 \\ 0 & 0 & 0,304 \end{pmatrix}$$

pour les profils-lignes.
Les profils-colonnes héritent eux de la matrice poids :

$$\frac{1}{n}D_C = \begin{pmatrix} \frac{30}{115} & 0 & 0 \\ 0 & \frac{50}{115} & 0 \\ 0 & 0 & \frac{35}{115} \end{pmatrix} = \begin{pmatrix} 0,261 & 0 & 0 \\ 0 & 0,435 & 0 \\ 0 & 0 & 0,304 \end{pmatrix}$$

.

Nous nous limiterons dans la suite aux profils-lignes, tous les résultats pouvant être extrapolés pour les profils-colonnes.
Le nuage des profils-lignes s'obtient en associant à chaque ligne du tableau X un point de l'espace $\mathbb{R}^s$.
Il faut noter que le point moyen de ce nuage est le profil marginal des lignes $(f_{\bullet 1}, f_{\bullet 2}, ..., f_{\bullet s})$.
Notez aussi que, du fait que la somme des coordonnées de chaque profil vaut 1, tous les points du nuage se trouvent sur l'hyperplan d'équation : $x_1 + x_2 + ... + x_s = 1$.

2.4 Métrique du χ^2

La métrique utilisée pour mesurer la proximité entre profils-lignes est celle du χ^2, définie par :

$$d^2_{\chi^2}(\ell_i, \ell_k) = \sum_{j=1}^{s} \frac{n}{n_{\bullet j}} \left(\frac{n_{ij}}{n_{i\bullet}} - \frac{n_{kj}}{n_{k\bullet}} \right)^2 = \sum_{j=1}^{s} \frac{n}{n_{\bullet j}} \left(\ell_{ij} - \ell_{kj} \right)^2 = \sum_{j=1}^{s} \frac{1}{f_{\bullet j}} \left(x_{ij} - x_{kj} \right)^2$$

On peut par exemple calculer la distance entre le profil-ligne 1 et le profil-ligne 3 de la manière suivante :

$$d^2_{\chi^2}(\ell_1, \ell_3) = \sum_{j=1}^{3} \frac{n}{n_{\bullet j}} \left(\frac{n_{1j}}{n_{1\bullet}} - \frac{n_{3j}}{n_{3\bullet}} \right)^2 = \sum_{j=1}^{3} \frac{1}{f_{\bullet j}} \left(x_{1j} - x_{3j} \right)^2$$

$$= \frac{1}{0,261}(0,417-0,057)^2 + \frac{1}{0,435}(0,533-0,229)^2 + \frac{1}{0,304}(0,05-0,714)^2$$

C'est donc la distance associée au produit scalaire de matrice

$$M = nD_C^{-1} = \begin{pmatrix} \frac{1}{0,261} & 0 & 0 \\ 0 & \frac{1}{0,435} & 0 \\ 0 & 0 & \frac{1}{0,304} \end{pmatrix}$$

Le choix de cette métrique se justifie par ses propriétés remarquables et notamment :

- Lorsque deux colonnes ont le même profil, on peut les regrouper sans que celà change les distances, au sens du χ^2, entre les lignes.
- Les coefficients $\frac{n}{n_{\bullet j}} = \frac{1}{f_{\bullet j}}$ qui apparaissent dans la matrice M reviennent à privilégier les écarts lorsqu'ils concernent des facteurs globalement rares. Ceci traduit l'idée que ce qui est rare est plus caractéristique et permet de mieux distinguer.

Une fois la métrique choisie, on peut définir l'inertie du nuage comme pour l'ACP, puis rechercher le meilleur sous-espace de projection avec le moins de perte d'inertie possible.

2.5 AFC et lien avec l'ACP

Faire une analyse factorielle des correspondances (AFC) revient alors à faire deux ACP, l'une sur les profils-lignes et l'autre sur les profils colonnes.
Dans le premier cas on utilise comme donnée la matrice $X = D_L^{-1}N$ avec la métrique $M = nD_C^{-1}$ et la matrice des poids $\frac{1}{n}D_L$.

Dans le second on utilise comme donnée la matrice $X = D_C^{-1}N^t$ avec la métrique $M = nD_L^{-1}$ et la matrice des poids $\frac{1}{n}D_L C$.

On peut montrer aisément que les facteurs principaux sont obtenus comme vecteurs propres de la matrice $D_C^{-1}N^t\ D_L^{-1}N$ pour les profils-lignes et de la matrice $D_L^{-1}N\ D_C^{-1}N^t$ pour les profils-colonnes.
Il faut aussi noter que les points moyens sont des vecteurs propres triviaux associés à la valeur propre 1. Tous les logicils statistiques ignorent l'axe principal correspondant.
On obtient donc les mêmes données de sortie que pour l'ACP sauf que la représentation graphique superpose les plans principaux des deux analyses.

2.6 Interprétation d'une AFC

- On commence toujours par commenter les profils marginaux qui synthétisent la répartition globale de chaque caractère.

– Ensuite on doit faire le choix du nombre d'axes à retenir en suivant exactement la même méthode que pour l'ACP.

– L'interprétation des axes se fait de la même manière, en commençant par repérer les points-lignes qui ont la plus grande contribution et en distinguant, d'après les coordonnées les positives des négatives. Il ne faut pas oublier de vérifier la qualité de représentation de chaque point grâce au $\cos^2$. On fait de même pour les points-colonnes.
– Sur le graphique un point-ligne se trouve représenté à la position moyenne (ou barycentre) des points colonnes pondérés par leur importance relative dans la ligne considérée. Idem pour les points-colonnes.
– Une proximité entre un point-ligne i et un point colonne j peut être interprétée comme attraction dans la mesure où la part de i dans la colonne j est importante. Mais cette proximité graphique peut aussi être due à l'attraction d'autres points-lignes. Le seul cas où cette attraction est sans équivoque c'est lorsque les points sont proches du bord.

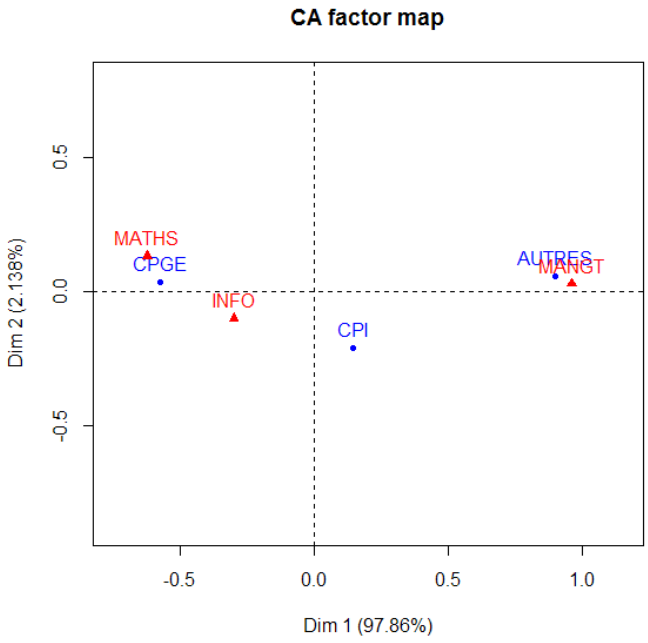

FIGURE 2.2 – AFC - EISTI

Le graphique donnant les résultats de l'AFC appliquée à notre exemple simplifié indique une attraction des étudiants CPGE vers les options MATHS et une similaire des étudiants AUTRES vers le MANAGT.
Les options INFO semblent réparties entre les CPGE et les CPI.
Les modalités CPI et INFO sont plus proches de l'origine et donc proches du profil moyen.

Chapitre 3

Analyse de variance ANOVA

L'analyse de variance permet d'évaluer l'influence d'un ou plusieurs facteurs qualitatifs sur une variable quantitative.
On part d'une variable aléatoire quantitative X à expliquer par des variables qualitatives A_i appelées variables explicatives.
On dispose de plusieurs échantillons de X correspondant aux différentes modalités des A_i.
On suppose que l'influence éventuelle des A_i se limite à la moyenne et on note $\mu_1, \mu_2, ..., \mu_k$ les moyennes des différents échantillons. On suppose donc que tous les échantillons sont tirés d'une population qui a la même variance σ^2.
L'analyse de variance revient à tester l'hypothèse : $(\mathcal{H}_0) : \mu_1 = \mu_2 = ... = \mu_k$ contre $(\mathcal{H}_1) : \exists\, i, j, \quad \mu_i \neq \mu_j$.
On distingue l'analyse à un facteur lorsqu'on a une seule variable explicative de l'analyse à deux facteurs ou de l'analyse multifactorielle dans les autres cas.

3.1 Analyse de variance à un facteur

3.1.1 Présentation générale

On souhaite tester les effets des k modalités d'une variable qualitative A, appelée facteur, sur une variable quantitative X pour laquelle on dispose donc de k échantillons formés de $n_1, \ldots, n_k$ individus.

Dans le modèle probabiliste, à chaque modalité A_j de A correspond un échantillon :

$(X_1^{(j)}, \ldots, X_{n_j}^{(j)})$ de X pour le j-ème groupe.

les variables aléatoires modélisant les données des différents groupes sont supposées indépendantes et de même loi $\mathcal{N}(\mu_j, \sigma^2)$. On suppose donc en particulier que la variance σ^2 est constante indépendante de A, hypothèse qu'un test aura validé auparavant.

On cherche à savoir si la variabilité observée dans les données est uniquement due au hasard, ou s'il existe effectivement des différences significatives entre les groupes, imputables au facteur.
Pour cela, on va comparer les variances empiriques de chaque échantillon, à la variance de l'échantillon global, de taille $n_1 + \cdots + n_k = n$.

La moyenne des variances de chaque groupe (pondérée par les effectifs) résume la variabilité à l'intérieur des classes, d'où le nom de variance intra-classes, ou variance résiduelle.
La variance des moyennes des différents groupes décrit les différences entre classes qui peuvent être dues au facteur, d'où le nom de variance inter-classes, ou variance expliquée.

Si les moyennes sont significativement différentes, on s'attend à ce que la variance expliquée soit grande, comparée à la variance résiduelle. La décomposition de la variance de l'échantillon global en variance expliquée et variance résiduelle est explicitée dans le résultat suivant.

Proposition
Si on note :
- $\overline{X}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i^{(j)}$ la moyenne empirique de la j-ième classe,
- $V^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} (X_i^{(j)} - \overline{X^{(j)}})^2$ la variance empirique de la j-ième classe,
- $\overline{X}$ la moyenne de l'échantillon global,
- $V_{intra} = \sum_{j=1}^{k} \frac{n_j}{n} V^{(j)}$ la moyenne des variances (variance intra-classes),
- $V_{inter} = \sum_{j=1}^{k} \frac{n_j}{n} (\overline{X}^{(j)} - \overline{X})^2$ la variance des moyennes (variance inter-classes),
- S^2 la variance de l'échantillon global.

Alors :
$S^2 = V_{intra} + V_{inter}$.

Démonstration : Ecrivons :

$$\begin{aligned}
S^2 &= \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_i^{(j)} - \overline{X})^2 \\
&= \sum_{j=1}^{k} \frac{1}{n} \sum_{i=1}^{n_j} (X_i^{(j)} - \overline{X^{(j)}} + \overline{X^{(j)}} - \overline{X})^2 \\
&= \sum_{j=1}^{k} \frac{n_j}{n} \frac{1}{n_j} \sum_{i=1}^{n_j} (X_i^{(j)} - \overline{X^{(j)}})^2 + \frac{1}{n} \sum_{j=1}^{k} n_j (\overline{X^{(j)}} - \overline{X})^2 + 2 \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_i^{(j)} - \overline{X^(j)})(\overline{X^{(j)}} - \overline{X}) \\
&= \sum_{j=1}^{k} \frac{n_j}{n} V^{(j)} + \sum_{j=1}^{k} \frac{n_j}{n} (\overline{X}^{(j)} - \overline{X})^2 + 0 \\
&= V_{intra} + V_{inter} \qquad \square.
\end{aligned}$$

La comparaison entre V_{inter} et V_{intra} va nous permettre de tester l'hypothèse d'égalité des espérances. :

$$\mathcal{H}_0 \ : \ \mu_1 = \cdots = \mu_k \ .$$

Pour déterminer la variable de décision et la loi qu'elle suit, on rappelle que la somme de deux variables indépendantes suivant des chi-deux suit encore une loi du chi-deux, et que leur rapport pondéré suit une loi de Fisher. Plus précisément :

Proposition
Sous l'hypothèse $\mathcal{H}_0$:

- $n\frac{V_{intra}}{\sigma^2}$ suit la loi du chi-deux $\mathcal{X}^2(n-k)$.
- $n\frac{V_{inter}}{\sigma^2}$ suit la loi du chi-deux $\mathcal{X}^2(k-1)$.
- $\frac{V_{inter}/(k-1)}{V_{intra}/(n-k)}$ suit la loi de Fisher $\mathcal{F}(k-1, n-k)$.

Le test ANOVA consiste donc à rejeter l'égalité des moyennes (reconnaître un effet du facteur A) quand le rapport pondéré de la variance expliquée (inter-classes) à la variance résiduelle (intra-classes) est significativement trop grand par rapport aux quantiles de la loi $\mathcal{F}(k-1, n-k)$.

3.1.2 Exemple

On veut savoir si la quantité de nitrates, prélevées le long d'une rivière, varie d'une station à l'autre. Pour cela, on dispose des résultats de 10 prélévements effectués dans 3 stations différentes (k=3).

Station 1	Station 2	Station 3
50,00	162,00	120,00
52,00	350,00	120,00
123,00	125,00	122,00
100,00	320,00	221,00
200,00	112,00	253,00
250,00	200,00	141,00
220,00	40,00	182,00
220,00	162,00	175,00
300,00	160,00	160,00
220,00	250,00	214,00

Le résultat de l'analyse de variance effectuée par Excel donne :

Analyse de variance: un facteur

RAPPORT DÉTAILLÉ

Groupes	*Nombre d'échantillons*	*Somme*	*Moyenne*	*Variance*
Colonne 1	10	1735	173,5	7445,61111
Colonne 2	10	1881	188,1	9048,98889
Colonne 3	10	1708	170,8	2203,73333

ANALYSE DE VARIANCE

Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*	*F*	*Probabilité*	*Valeur critique pour F*
Entre Groupes	1732,466667	2	866,2333333	0,1389803	0,87086372	3,354130829
A l'intérieur des groupes	168285	27	6232,777778			
Total	170017,4667	29				

FIGURE 3.1 – Résultats ANOVA nitrates

La valeur critique indiquée ici pour F correspond à un risque de niveau $\alpha = 0,05$.
La valeur observée dans l'échantillon est inférieure au seuil : $0,13898 < 3,35413$.

On ne peut rejeter l'hypothèse $\mathcal{H}_0$.
Autrement dit il n'y a pas de différence significative entre les moyennes de prélévements des trois stations (qui pourrait signaler un déversement de nitrates entre les stations par exemple).

3.2 Analyse de variance à deux facteurs

3.2.1 Présentation générale

On veut cette fois tester l'effet de deux variables qualitatives A et B sur une variable quantitative X.
Les modalités de A seront indicées par la lettre i, celles de B par j. L'indice k servira à distinguer les valeurs prises par X pour une modalité fixée A_i de A et B_j de B.
L'analyse de variance à deux facteurs va nous permettre de détecter un éventuel effet du facteur A ou du facteur B ou de l'interaction entre A et B.
Le principe est le même que pour l'ANOVA1, on décompose la variance en plusieurs termes correspondant chacun à l'un des effets possibles.
Pour simplifier la présentation, on va travailler avec la somme des carrés des écarts SCE.
On notera $\overline{X}$ la moyenne globale, $\overline{X_{i\bullet}}$ la moyenne correspondant à la modalité A_i, $\overline{X_{\bullet j}}$ la moyenne correspondant à la modalité B_j et $\overline{X_{ij}}$ la moyenne de l'échantillon correspondant à A_i et B_j.
On notera également p le nombre de modalités de A, q le nombre de modalités de B et n_{ij} l'effectif de l'échantillon correspondant à A_i et B_j.
On obtient donc la décomposition suivante :

- $SCET = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(X_{ijk}-\overline{X})^2 = SCEA + SCEB + SCEAB + SCER$
  est la somme des carrés des écarts **totale**,

- $SCEA = \sum_{i=1}^{p} n_{i\bullet}(\overline{X_{i\bullet}}-\overline{X})^2$
  est la somme des carrés des écarts **expliquée par A**,

- $SCEB = \sum_{j=1}^{q} n_{\bullet j}(\overline{X_{\bullet j}}-\overline{X})^2$
  est la somme des carrés des écarts **expliquée par B**,

- $SCEAB = \sum_{i=1}^{p}\sum_{j=1}^{q}(\overline{X_{ij}}-\overline{X_{i\bullet}}-\overline{X_{\bullet j}}+\overline{X})^2$
  est la somme des carrés des écarts qui mesure **l'influence de l'interaction des facteurs A et B** sur la moyenne.

- $SCER = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(X_{ijk}-\overline{X_{ij}})^2$
  est la somme des carrés des écarts **résiduels**.

On teste ensuite l'effet (ou l'absence d'effet) de chaque facteur à l'aide des variables suivantes :

1. **Facteur A** :
   Sous l'hypothèse $(\mathcal{H}_0)$: $\mu_{i\bullet} = \mu$, pour tout $i, 1 \leq i \leq p$, la variable : $F_A = \dfrac{SCEA/(p-1)}{SCER/(n-pq)}$ suit une loi de Fisher-Snedecor à $(p-1, n-pq)$ degrés de liberté : elle permet de tester l'influence du facteur A.
2. **Facteur B** :
   Sous l'hypothèse $(\mathcal{H}_0)$: $\mu_{\bullet j} = \mu$, pour tout $j, 1 \leq j \leq q$, la variable : $F_B = \dfrac{SCEB/(q-1)}{SCER/(n-pq)}$ suit une loi de Fisher-Snedecor à $(q-1, n-pq)$ degrés de liberté : elle permet de tester l'influence du facteur B.
3. **Interaction A et B** :
   Sous l'hypothèse $(\mathcal{H}_0)$: $\mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu = 0$, pour tout i, j, la variable : $F_{AB} = \dfrac{SCEAB/(p-1)(q-1)}{SCER/(n-pq)}$ suit une loi de Fisher-Snedecor à $((p-1)(q-1), n-pq)$ degrés de liberté : elle permet de tester l'influence de l'interaction des facteurs A et B.

3.2.2 Exemple

Méthode	Niv 1	Niv2
Magistral	81,00	70,00
	84,00	56,00
	70,00	81,00
	69,00	78,00
	89,00	88,00
	95,00	45,00
	69,00	83,00
	80,00	77,00
TICE	88,00	89,00
	98,00	94,00
	89,00	95,00
	95,00	98,00
	98,00	87,00
	93,00	85,00
	95,00	87,00
	93,00	93,00
Audiovisuel	87,00	97,00
	94,00	93,00
	95,00	85,00
	92,00	95,00
	86,00	82,00
	87,00	89,00
	100,00	97,00
	96,00	93,00
Autodidacte	85,00	84,00
	92,00	73,00
	63,00	57,00
	69,00	73,00
	65,00	91,00
	74,00	71,00
	56,00	68,00
	85,00	62,00

FIGURE 3.2 – Données pour ANOVA2

Nous disposons des résultats à l'examen final (note sur 100) de 64 étudiants répartis en deux niveaux d'étude et en quatre méthode d'apprentissage (cours magistral, TICE, méthodes audiovisuelles et méthodes autonomes ou d'autodidactes).
Nous désirons savoir si la méthode d'enseignement, le niveau et l'interaction des

deux ont une influence significative sur les résultats moyens des étudiants. L'outil analyse de variance à deux facteurs avec répétition d'expérience d'Excel donne les résultats suivants pour un risque de niveau $\alpha = 0,05$:

ANALYSE DE VARIANCE

Source	*SCE*	*d.d.l.*	*Moy carrés*	*F*	*Proba*	*Val critique F*
Échantillon	5006,625	3	1668,875	19,569073	8,3953E-09	2,769430949
Colonnes	144	1	144	1,6885306	0,19911896	4,012973319
Interaction	109,625	3	36,5416667	0,4284842	0,73337285	2,769430949
A l'intérieur du groupe	4775,75	56	85,28125			
Total	10036	63				

FIGURE 3.3 – Résultats ANOVA2

- La ligne "échantillon", qui correspond à l'influence du facteur A, ici la méthode d'apprentissage, indique une valeur $F_A = 19,569$ pour une valeur critique de $2,769$. On en déduit qu'il faut rejeter l'hypothèse $(\mathcal{H}_0)$; il y a bien influence de la méthode sur les résultats.
- La ligne nommée "colonnes" correspond au facteur B, ici le niveau. $1,6885 < 4,01297$. Au risque de 5 % on peut affirmer qu'il n'y a pas d'effet du niveau sur les résultats.
- La ligne nommée "interaction" correspond à l'effet de l'interaction des deux facteurs. $0,733 < 2,769$. Au risque de 5 % on peut affirmer qu'il n'y a pas d'effet de l'interaction entre méthode d'apprentissage et niveau sur les résultats.

Chapitre 4

Régression linéaire multiple RLM

4.1 Présentation générale

Lorsque l'on sait qu'une variable quantitative Y est liée à un certain nombre de variables quantitatives $X^1, X^2, ..., X^p$, hypothèse que l'on aura justifié à l'aide d'un test par exemple, on peut proposer un modèle de relation entre ces variables.
On parle de régression linéaire lorsqu'on prend un modèle de relation linéaire c'est à dire de type :

$$Y = \alpha_0 + \alpha_1 X^1 + \alpha_2 X^2 + ... + \alpha_p X^p + \varepsilon$$

On parle alors de variable expliquée Y, de variables explicatives $X^1, X^2, ..., X^p$ et de résidu ε.

Dans ce modèle linéaire, on considère que la variable (à expliquer) Y_i mesurée sur un individu i donné est une variable aléatoire, dont la loi dépend des valeurs prises sur cet individu par les caractères explicatifs X_i^j, qui sont déterminés de manière exacte.
Ceci se traduit par la relation :

$$Y_i = \alpha_0 + \alpha_1 X_i^1 + \alpha_2 X_i^2 + ... + \alpha_p X_i^p + \varepsilon_i$$

Les hypothèses esssentielles du modèle sont :
$(\varepsilon_1, \ldots, \varepsilon_n)$ sont des variables aléatoires indépendantes et de même loi avec :
- $E(\varepsilon_i) = 0$.
- $V(\varepsilon_i) = \sigma^2$.
- $COV(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$.
- La plupart du temps, on supposera les résidus gaussiens, càd $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Dans ce modèle on a donc :

$$E(Y) = \alpha_0 + \alpha_1 X^1 + \alpha_2 X^2 + ... + \alpha_p X^p$$

$$V(Y) = \sigma^2 I_n$$

Nous allons voir comment tester la validité d'un tel modèle, trouver des estimateurs des coefficients α_i et évaluer leur qualités.

4.2 Régression linéaire simple

On part d'une variable expliquée Y et d'une variable explicative X liés par la relation :

$$Y_i = \alpha_0 + \alpha_1 X_i + \varepsilon_i$$

L'estimation des paramètres α_0, α_1 se fait en minimisant la somme des carrés des écarts entre observations et modèle (méthode des moindres carrés).

$$\min_{\alpha_0,\alpha_1} \sum_{i=1}^{n} (Y_i - \alpha_0 - \alpha_1 X_i)^2$$

4.2.1 Estimation des paramètres

La solution donne comme estimateurs :

$$\hat{\alpha_1} = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha_0} = \overline{y} - \hat{\alpha_1}\overline{x}$$

Si on suppose de plus que les résidus sont gaussiens, c à d que $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, on obtient les mêmes estimateurs par la méthode du maximum de vraisemblance. On montre que ce sont des estimateurs sans biais et de variance minimale.
Ces estimateurs permettent d'obtenir une valeur estimée ou prédite de Y pour une valeur de X donnée :

$$\hat{Y_i} = \hat{\alpha_0} + \hat{\alpha_1} X_i$$

Les résidus estimés $\hat{e_i} = y_i - \hat{Y_i}$, permettent de définir un estimteur non biaisé de la variance résiduelle :

$$\hat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} \hat{e_i}^2$$

Qui permet à son tour de définir des estimateurs pour les variances des deux coefficients :

$$\hat{\sigma_0^2} = \hat{\sigma^2}(\hat{\alpha_0}) = \hat{\sigma^2} \frac{\sum x_i^2}{n \sum (x_i - \hat{x})^2}$$

$$\hat{\sigma_1^2} = \hat{\sigma^2}(\hat{\alpha_1}) = \frac{\hat{\sigma^2}}{\sum (x_i - \hat{x})^2}$$

4.2.2 Coefficient de déterminantion

On peut donner une interprétation géométrique de la méthode des moindres carrés, qui est la suivante :
On considère les valeurs de Y et de X comme les coordonnées des vecteurs $\vec{Y}$ et $\vec{X}$ de $\mathbb{R}^n$. On note $\mathbb{1}$ le vecteur de $\mathbb{R}^n$ dont toutes les coordonnées valent 1.

Le problème revient alors à chercher le vecteur $\hat{Y} = \alpha_0\mathbb{1} + \alpha_1\vec{X}$ appartenant au plan engendré par $\mathbb{1}$ et $\vec{X}$ qui est le plus proche, au sens de la distance euclidienne, du vecteur $\vec{Y}$.
La réponse étant évidemment la projection orthogonale de $\vec{Y}$ sur ce plan.

Le théorème de Pythagore nous permet alors d'écrire :

$$\left\|\vec{Y} - \bar{y}\mathbb{1}\right\|^2 = \left\|\hat{Y} - \bar{y}\mathbb{1}\right\|^2 + \|\hat{\varepsilon}\|^2$$

$$\sum_{i=1}^{n} |y_i - \bar{y}|^2 = \sum_{i=1}^{n} |\hat{y}_i - \bar{y}|^2 + \sum_{i=1}^{n} |\varepsilon_i|^2$$

On retrouve ainsi la décomposition de la variance en
totale = expliquée + résiduelle ou encore
Somme des Carrés Totaux = Somme des Carrés Expliqués+Somme des Carrés Résiduels :

$$SCT = SCE + SCR$$

On définit alors le coefficient de déterminantion R^2 par :

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

On peut l'interpréter en disant que ($100 \times R^2$ %) de la variance est expliquée par le modèle linéaire de régression.

4.2.3 Lois des estimateurs

En partant de l'hypothèse de normalité des résidus, on montre que :

$$\frac{\hat{\alpha_0} - \alpha_0}{\hat{\sigma_0}} \sim t_{n-2}$$

$$\frac{\hat{\alpha_1} - \alpha_1}{\hat{\sigma_1}} \sim t_{n-2}$$

La loi de Student permet donc de tester si les coefficients sont significativement non nuls, ou d'en donner des intervalles de confiance.

Pour tester la validité globale du modèle, on n'utilise pas le coefficient de détermination R^2, mais sa version corrigée en tenant compte des degrés liberté :

$$F = \frac{R^2}{1 - R^2}(n-2) \sim \mathcal{F}(1, n-2)$$

Notez que une loi de Fisher à 1 et k degrés de libertés est le carré d'une loi de Student k degrés de liberté.

4.3 Régression linéaire multiple

Dans le cas ou on dispose de plusieurs variables explicatives $X^1, X^2, ..., X^p$, on regroupe leurs valeurs dans une matrice X d'ordre $n \times (p+1)$ dont la première colonne est formée de 1 et les autres colonnes correspondent aux X^j.

$$X = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^j & \cdots & x_1^p \\ \vdots & \vdots & & \vdots & & \vdots & \\ 1 & x_i^1 & x_i^2 & \cdots & x_i^j & \cdots & x_i^p \\ \vdots & \vdots & & \vdots & & \vdots & \\ \vdots & \vdots & & \vdots & & \vdots & \\ 1 & x_n^1 & x_n^2 & \cdots & x_n^j & \cdots & x_n^p \end{pmatrix}$$

Le modèle linéaire aura alors la forme matricielle suivante :

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

où Y est un vecteur colonne $(n \times 1)$, β est le vecteur colonne $((p+1) \times 1)$ des coefficients du modèle :

$$\beta =^t (\alpha_0, \alpha_1, ..., \alpha_p)$$

et ε est le vecteur $(n \times 1)$ des résidus.

On rajoute une hypothèse supplémentaire disant que le rang de la matrice doit être maximal, càd $rang(X) = p + 1$.
Ce qui revient à exiger que les colonnes de X soient indépendantes.
Dans le cas contraire, on dit qu'on a un problème de colinéarité, et on le résout en éliminant les colonnes (les variables) qui sont combinaison linéaires des autres.

4.3.1 Estimateurs des moindres carrés, du maximum de vraisemblance

Les paramètres à estimer sont β et σ^2. Si on se place dans le cadre gaussien, la vraisemblance est donnée par :

$$\mathcal{L}(Y, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (X\beta)_i)^2\right)$$

La log-vraisemblance sera donc :

$$\ell(Y, \beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\|Y - X\beta\|^2}{2\sigma^2}$$

Si on fixe la valeur de σ^2, on voit que la vraisemblance est maximale lorsque $J(\beta) = \dfrac{\|Y - X\beta\|^2}{2\sigma^2}$ est minimale.
La méthode des moindres carrés recoupe donc la méthode du maximum de vraisemblance pour le paramètre β.
Pour trouver l'estimateur de β, on résout :

$$\nabla_\beta J(\beta) = -2\ {}^tXY + 2\ {}^tXX\beta = 0$$

Comme ${}^tXX$ est d'ordre $(p+1)\times(p+1)$ et de rang $(p+1)$, elle est inversible et donc :

$$\hat{\beta} = ({}^tXX)^{-1}\ {}^tXY$$

La valeur estimée de Y devient alors :

$$Y_E = X\hat{\beta} = X({}^tXX)^{-1}\ {}^tXY$$

La méthode du maximum de vraisemblance, donne par conséquent comme estimateur de la variance :

$$\hat{\sigma^2} = \frac{\|Y - Y_E\|^2}{n}$$

4.3.2 Lois des estimateurs

En fait on établit un résultat plus général concernant le vecteur aléatoire $\hat{\beta}$:

$$\hat{\beta} \sim \mathcal{N}_{p+1}\left(\beta, \sigma^2({}^tXX)^{-1}\right)$$

Ce qui établit entre autres que $\hat{\beta}$ est un estimateur sans biais β.
Pour la variance on établit de même que :

$$\frac{\|Y - Y_E\|^2}{\sigma^2} \sim \chi^2_{(n-p-1)}$$

L'EMV trouvé pour la variance est donc biaisé, on le remplace traditionnellement par l'estimateur non biaisé :

$$\frac{\|Y - Y_E\|^2}{n-p-1}$$

4.3.3 Tests sur la significativité des coefficients

De la loi suivie par le vecteur $\hat{\beta}$, on déduit, pour le k-ème coefficient le résultat suivant :

$$\frac{\hat{\alpha_k} - \alpha_k}{\sigma\sqrt{({}^tXX)^{-1}_{kk}}} \sim \mathcal{N}_1(0,1)$$

Ceci nous permet, en remplaçant σ par son estimateur, d'établir que :

$$\frac{\hat{\alpha_k} - \alpha_k}{\hat{\sigma}(\hat{\alpha_k})} = \frac{\hat{\alpha_k} - \alpha_k}{\sqrt{({}^tXX)^{-1}_{kk}\dfrac{\|Y - Y_E\|^2}{n-p-1}}} \sim t_{n-p-1}$$

La connaissance de l'estimateur de la variance de $\hat{\alpha_k}$, à savoir $\hat{\sigma}(\hat{\alpha_k}) = ({}^tXX)^{-1}_{kk}\dfrac{\|Y - Y_E\|^2}{n-p-1}$, conjuguée avec la table de la loi de Student à $(n-p-1)$ d.d.l, permet de tester l'hypothèse $(\mathcal{H}_0)$: $\alpha_k = 0$ et aussi de déterminer des intervalles de confiance pour ce coefficient.

4.3.4 Coefficient de déterminantion et test global sur le modèle

Comme dans le cas de la régression linéaire simple, le coefficient de détermination est donné par :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = \frac{\left\|Y_E - \bar{Y}\mathbb{1}\right\|^2}{\left\|Y - \bar{Y}\mathbb{1}\right\|^2}$$

Ce coefficient indique lorsqu'il est proche de 1 que le modèle linéaire est valable avec au moins une variables explicative.
Mais il a le défaut d'augmenter automatiquement lorsqu'on ajoute de nouvelles variables explicatives. C'est pourquoi on introduit le coefficient de détermination ajusté, qui corrigé ce défaut et que l'on note généralement R_a^2 :

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

Pour tester la significativité du modèle, càd la non nullité de tous les coefficients, on introduit :

$$F = \frac{R^2}{1-R^2}\frac{n-p-1}{p} \sim \mathcal{F}(p, n-p-1)$$

Le test de F global concerne les hypothèses :
$(\mathcal{H}_0 : \alpha_1 = \alpha_2 = \alpha_p = 0)$; l'ensemble des p variables explicatives n'apportent pas une information utile pour la prédiction de Y, sous le modèle linéaire.
$(\mathcal{H}_1 : \exists j, \quad \alpha_j \neq 0)$; au moins une des variables explicatives est associée à Y par le modèle linéaire.
La loi de Fisher à p et $(n-p-1)$ d.d.l permet alors de conclure quant à la validité ou non du modèle global pour un niveau de risque donné.

4.4 Pratique de la RLM

4.4.1 Sorties des logiciels statistiques

Les logiciels statistiques (R ou SAS par exemple) donnent en sortie pour chacun des différents paramètres α_k :

- la valeur estimée $\hat{\alpha_k}$.
- l'estimation de l'écart-type (standard error) $\hat{\sigma}(\hat{\alpha_k})$.
- la t-value qui est le rapport des deux précédentes.
- la probabilité Pr $>$ |t| qui représente la probabilité de rejeter à tort l'hypothèse $(\mathcal{H}_0 : \alpha_k = 0)$.

Le paramètre constant α_0 correspond à la ligne ***intercept***.

Pour la validité du modèle linéaire global, vous obtenez en sortie :
- le coefficient de détermination R^2.
- le coefficient de détermination ajusté R_a^2.
- les résultats de l'analyse de variance SCE,SCR et SCT avec les d.d.l. associés.
- la statistique de Fisher F.
- la probabilité Pr $>$ F qui représente la probabilité de rejeter à tort l'hypothèse $(\mathcal{H}_0 : \alpha_1 = \alpha_2 .... = \alpha_p = 0)$.

4.4.2 Analyse des hypothèses du modèle

Avant de passer à l'analyse des résultats de la régression, il faut commencer par vérifier les hypothèses fondamentales du modèle :
- linéarité.
- nullité de la moyenne des écarts.
- valeur constante σ^2 de la variance des écarts, quelque soit l'observation.
- indépendance des erreurs.
- normalité des erreurs.

Il existe un certain nombre de méthodes graphiques permettant de détecter un éventuel défaut de ces hypothèses.
Lorsqu'on trace le graphique des erreurs e_i en fonction des y_i ou des x_i^k, on s'attend à ce que les erreurs soient uniformément réparties dans une bande autour de la valeur 0.
Les exemples suivants montrent quelques exemples de graphiques indiquant des défauts de constance de la variance ou de corrélation des erreurs.

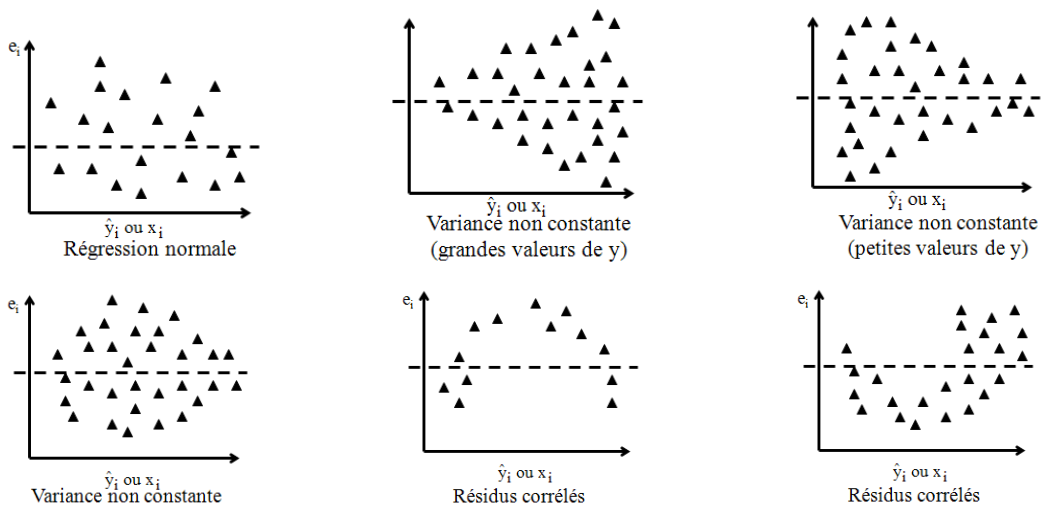

FIGURE 4.1 - Distribution des erreurs

Certains de ces défauts peuvent être corrigés en remplaçant y par $y' = \sqrt{y}$ ou $\log(y)$ ou $\frac{1}{y}$.

Pour vérifier la normalité des erreurs, on utilise un diagramme de Q-Q plot, obtenu en ordonnant les erreurs et les plaçant en fonction des quantiles de la loi normale. Les points obtenus doivent être très proches de la première bissectrice.

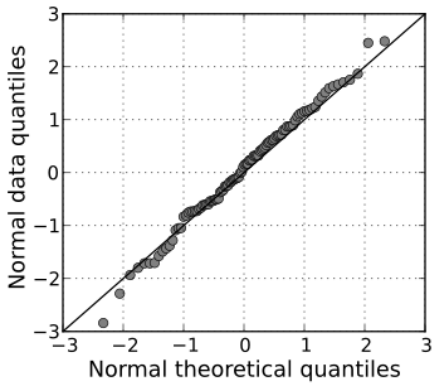

FIGURE 4.2 - Exemple de Q-Q plot

Dans le cas ou le modèle linéaire ne convient pas, la forme du nuage peut suggérer un type de relation entre x et y et permettre ainsi de modifier les variables pour que les relations soient linéaires.
Voici quelques exemples.

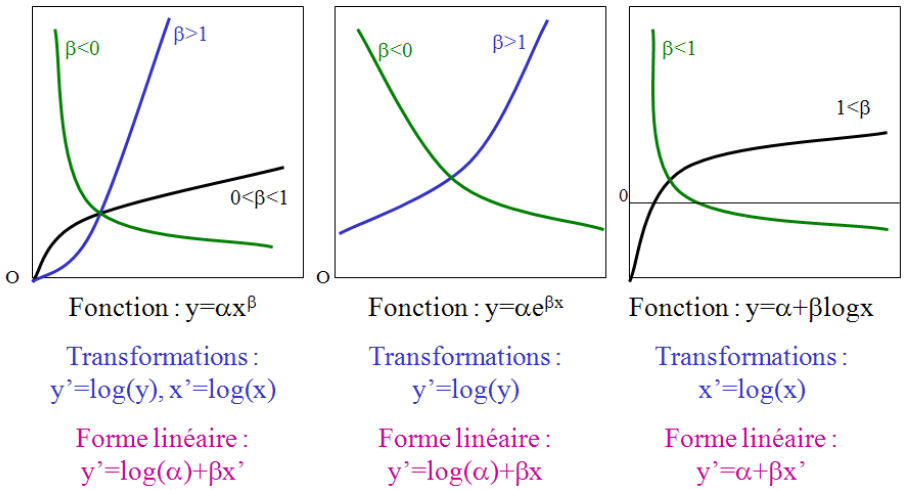

FIGURE 4.3 - Exemples de transformations utiles des variables

4.4.3 Sélection des variables explicatives

Pour choisir les variables X^k à retenir pour expliquer linéairement Y, on peut utiliser une méthodes pas à pas.
Celà consiste à éliminer successivement ou à ajouter successivement des variables explicatives.

- **La méthode descendante** consiste à éliminer la variable la moins significative parmi les p : celle qui a le t de Student le moins significatif. On recalcule alors la régression puis on recommence jusqu'à être satisfait.
- **La méthode ascendante** procède en sens inverse : On part de la

meilleure régression à une seule variable puis on ajoute la variable la plus significative.

4.4.4 Un exemple

On veut vérifier si on peut expliquer le taux de natalité par une relation linéaire avec la densité de population, l'espérance de vie des femmes et le taux d'alphabétisation de certains pays.
Pour celà on part des données suivantes :

PAYS	DENS	ESPVIEF	alphab	natal
Arab.Saoud.	0.80	70.00	62.00	6.67
Argentine	1.20	75.00	95.00	2.80
Belgique	32.90	79.00	99.00	1.70
Bolivie	0.70	64.00	78.00	4.21
Canada	0.30	81.00	97.00	1.80
CoréeS.	44.70	74.00	96.00	1.65
France	10.50	82.00	99.00	1.80
GB	23.70	80.00	99.00	1.83
Indonésie	10.20	65.00	77.00	2.80
Iran	3.90	67.00	54.00	6.33
Liban	34.30	71.00	80.00	3.39
Roumanie	9.60	75.00	96.00	1.82
Russie	0.90	74.00	99.00	1.83
Sénégal	4.30	58.00	38.00	6.10
Somalie	1.00	55.00	24.00	7.25
Suède	1.90	81.00	99.00	2.10
Suisse	17.00	82.00	99.00	1.60
Syrie	7.40	68.00	64.00	6.65
Turquie	7.90	73.00	81.00	3.21
Ukraine	8.70	75.00	97.00	1.82
Vénézuela	2.20	76.00	88.00	3.05

Procédons par méthode descendante. En gardant les trois variables explicatives, le logiciel R donne les résultats suivants :

```
Call:
lm(formula = natal ~ DENS + ESPVIEF + alphab, data = T)

Residuals:
     Min       1Q   Median       3Q      Max
-1.01405 -0.36801 -0.07267  0.28287  1.72874

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.81343    2.42653   3.220  0.00503 **
DENS        -0.01498    0.01361  -1.100  0.28648
ESPVIEF      0.05114    0.04995   1.024  0.32028
alphab      -0.09779    0.01768  -5.533 3.65e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7356 on 17 degrees of freedom
Multiple R-squared: 0.8841,     Adjusted R-squared: 0.8636
F-statistic: 43.22 on 3 and 17 DF,  p-value: 3.605e-08
```

FIGURE 4.4 - Résultats RLM globales

Les valeurs du R^2 et du R_a^2 (plus de $0,8$) indiquent que les variables DENS, ESPVIEF et alphab expliquent bien le taux de natalité.
La valeur du F et la p-value indiquent que le modèle linéaire est significatif.
Pour les paramètres, les résultats indiquent que seuls le facteur constant (intercept) et le taux d'alphabétisation sont significativement non nuls.
Dans l'étape suivante, on enlève la variable qui a la plus mauvaise valeur de $P(>|t|)$, càd ici l'espérance de vie, et on recommence. Les résultats obtenus sont les suivants :

```
Call:
lm(formula = natal ~ DENS + alphab, data = T)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98594 -0.33622 -0.01539  0.25353  1.78009

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.210400   0.638509  15.991 4.40e-12 ***
DENS        -0.015901   0.013597  -1.169    0.257
alphab      -0.081607   0.007912 -10.315 5.53e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7366 on 18 degrees of freedom
Multiple R-squared: 0.8769,     Adjusted R-squared: 0.8633
F-statistic: 64.14 on 2 and 18 DF,  p-value: 6.471e-09
```

FIGURE 4.5 - Résultats RLM $natal = f(DENS, alphab)$

Les R^2 sont meilleurs tout comme la p-value, le modèle est meilleur. Mais la variable densité reste non significative.
On choisit donc le modèle natalité en fonction du taux d'alphabétisation. Les résultats confirment que c'est le meilleur modèle.

```
> model=lm(natal~alphab,data=T)
> summary(model)

Call:
lm(formula = natal ~ alphab, data = T)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0020 -0.3117 -0.1279  0.2084  1.7795

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.280927   0.641773   16.02 1.72e-12 ***
alphab      -0.084538   0.007577  -11.16 8.76e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7437 on 19 degrees of freedom
Multiple R-squared: 0.8676,     Adjusted R-squared: 0.8606
F-statistic: 124.5 on 1 and 19 DF,  p-value: 8.765e-10
```

FIGURE 4.6 - Résultats RLM $natal = f(alphab)$

La meilleur formule du modèle est donc :

$$natal = 10.280927 - 0.084538 * alphab$$