M1 IMAT, Année 2009-2010

MODELES LINEAIRES

C.Chouquet
Laboratoire de Statistique et Probabilités - Université Paul Sabatier - Toulouse

Table des matières

1 Préambule **1**
- 1.1 Démarche statistique . . . 1
- 1.2 Un exemple introductif pour la modélisation linéaire d'une variable quantitative . . . 2
  - 1.2.1 Description de la population d'étude . . . 2
  - 1.2.2 Relation entre variables quantitatives . . . 3
  - 1.2.3 Relation entre variable quantitative et variables qualitatives . . . 4
  - 1.2.4 Modélisation d'une variable quantitative en fonction de variables quantitatives et qualitatives . . . 5

2 Présentation du modèle linéaire gaussien **6**
- 2.1 Le modèle linéaire . . . 6
- 2.2 Le modèle linéaire gaussien . . . 7
  - 2.2.1 Ecriture générale . . . 7
  - 2.2.2 Le modèle de régression linéaire . . . 8
  - 2.2.3 Le modèle factoriel . . . 8

3 Estimation **9**
- 3.1 Méthodes d'estimation . . . 9
  - 3.1.1 Principe des moindres carrés . . . 9
  - 3.1.2 Principe du Maximum de Vraisemblance . . . 9
- 3.2 Estimation de θ . . . 10
- 3.3 Valeurs ajustées et résidus calculés . . . 10
- 3.4 Estimation de σ^2 . . . 10
- 3.5 Erreurs standard de $\widehat{\theta}_j$, $\widehat{y}_i$, $\widehat{e}_i$ . . . 11
- 3.6 Construction de l'intervalle de confiance de θ_j . . . 12
- 3.7 Décomposition de la variance . . . 12

4 Test de Fisher **13**
- 4.1 Hypothèse testée . . . 13
  - 4.1.1 Principe . . . 13
  - 4.1.2 Calculs sous H_0 . . . 13
- 4.2 Le test de Fisher-Snédécor . . . 13
  - 4.2.1 Principe . . . 13
  - 4.2.2 La statistique de test . . . 14
  - 4.2.3 Fonctionnement du test . . . 14
- 4.3 Cas particulier où q=1 : le test de Student . . . 15

5 La Régression linéaire **16**
- 5.1 Introduction . . . 16
  - 5.1.1 La problématique . . . 16
  - 5.1.2 Le modèle de régression linéaire simple . . . 16
  - 5.1.3 Le modèle de régression linéaire multiple . . . 17
- 5.2 Estimation . . . 17

5.2.1 Résultats généraux . . . 17
5.2.2 Propriétés . . . 18
5.2.3 Le coefficient R^2 . . . 18
5.2.4 Augmentation mécanique du R^2 . . . 19
5.3 Tests et Intervalles de confiance . . . 20
5.3.1 Test de nullité d'un paramètre du modèle . . . 20
5.3.2 Test de nullité de quelques paramètres du modèle . . . 20
5.3.3 Test de nullité de tous les paramètres du modèle . . . 20
5.3.4 Intervalle de confiance de β_j, de $\overline{Y}_i$ et de $\overline{Y}_0$ . . . 21
5.3.5 Intervalle de prédiction . . . 22
5.4 Sélection des variables explicatives . . . 22
5.4.1 Les critères . . . 22
5.4.2 Les méthodes de sélection . . . 23
5.5 Validation du modèle . . . 23
5.5.1 Contrôle de l'ajustement du modèle . . . 23
5.5.2 Etude des colinéarités des variables explicatives . . . 24

6 L'analyse de variance 26
6.1 Introduction . . . 26
6.2 L'analyse de variance à un facteur . . . 26
6.2.1 Notations . . . 26
6.2.2 Le modèle . . . 26
6.2.3 Paramétrage centré . . . 27
6.2.4 Estimation . . . 27
6.2.5 Propriétés . . . 28
6.2.6 Intervalles de confiance et tests d'hypothèses sur l'effet facteur . . . 29
6.2.7 Comparaisons multiples : Méthode de Bonferroni . . . 29
6.3 Analyse de variance à deux facteurs croisés . . . 30
6.3.1 Notations . . . 30
6.3.2 Le modèle . . . 30
6.3.3 La paramétrisation centrée . . . 31
6.3.4 Estimations des paramètres . . . 31
6.3.5 Le diagramme d'interactions . . . 32
6.3.6 Tests d'hypothèses . . . 32
6.3.7 Tableau d'analyse de la variance à deux facteurs croisés dans le cas d'un plan équilibré . . . 34

7 Analyse de covariance 35
7.1 Les données . . . 35
7.2 Le modèle . . . 35
7.3 La seconde paramétrisation . . . 35
7.4 Tests d'hypothèses . . . 36

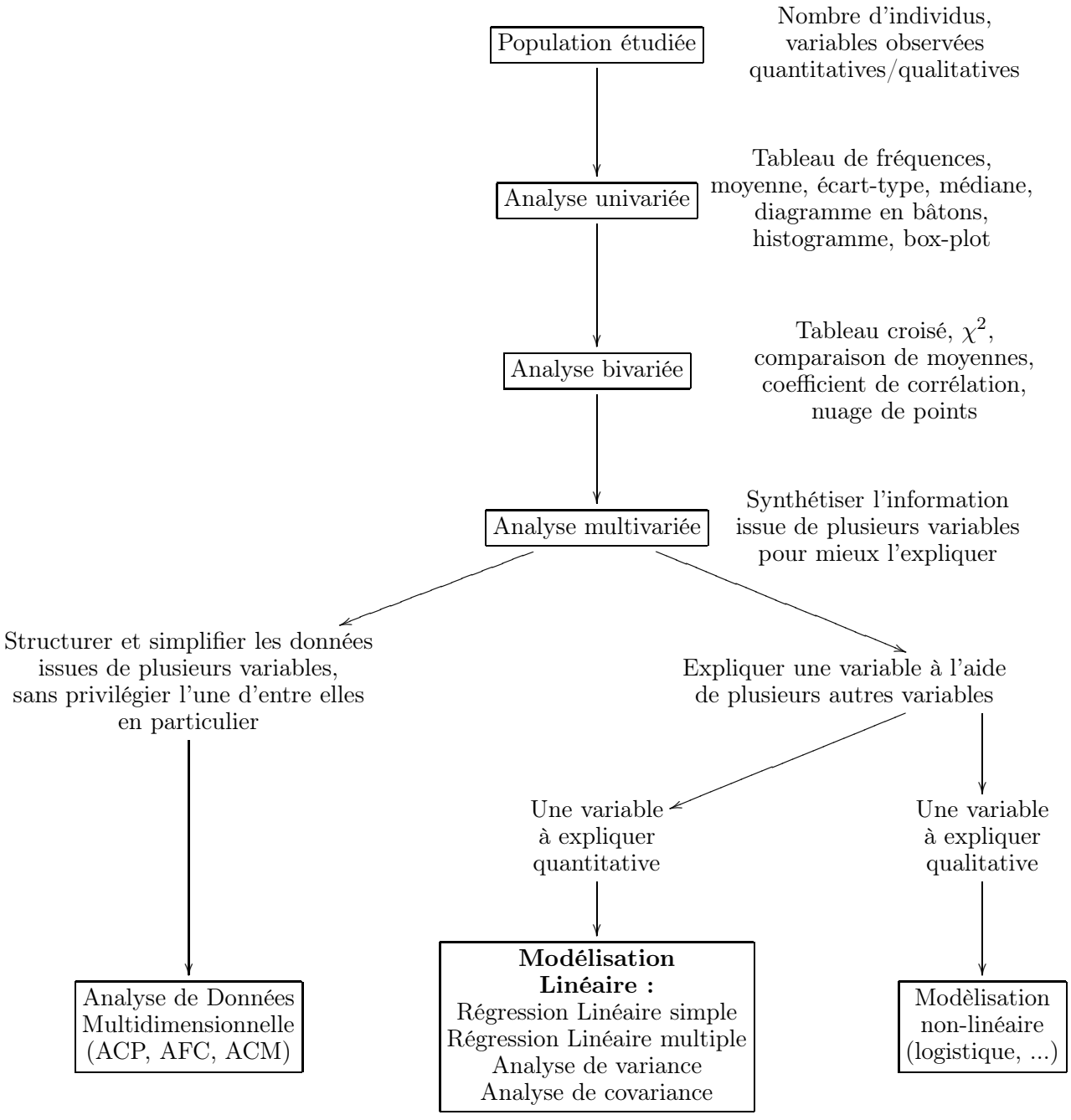
8 Quelques rappels de Statistique et de Probabilités 38
8.1 Généralités . . . 38
8.2 Indicateurs statistiques pour variables quantitatives . . . 39
8.2.1 Moyenne empirique d'une variable . . . 39
8.2.2 La covariance empirique . . . 39
8.2.3 Variance empirique et écart-type empirique . . . 40
8.2.4 Cœfficient de corrélation linéaire empirique . . . 40
8.2.5 Interprétation géométrique de quelques indices statistiques . . . 40
8.2.6 Expressions matricielles . . . 41
8.3 Rappels sur quelques lois de probabilité . . . 42
8.3.1 La distribution Normale $N(\mu, \sigma^2)$ . . . 42

8.3.2 La distribution n-Normale $N_n(\mu, \Gamma)$ 42
8.3.3 La distribution de χ^2 43
8.3.4 La distribution de Student 43
8.3.5 La distribution de Fisher-Snédécor 44
8.4 Rappels de statistique inférentielle 44
8.4.1 Estimation ponctuelle, estimation par intervalle de confiance 44
8.4.2 Notions générales sur la théorie des tests paramétriques 44

Chapitre 1

Préambule

1.1 Démarche statistique



1.2 Un exemple introductif pour la modélisation linéaire d'une variable quantitative

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

Dans une entreprise, on a relevé les salaires des 32 employés (mensuel en euros, noté sal), ainsi que certaines caractéristiques socio-démographiques telles que l'ancienneté dans l'entreprise (en années, notée anc), le nombre d'années d'études après le bac (noté apbac), le sexe ($1 = F/2 = M$, noté sex), le type d'emplois occupés (en 3 catégories codées de 1 à 3, noté emp). Un extrait des données est présenté ci-dessous :

num	anc	sal	sex	apbac	emp
1	7	1231	1	3	2
2	15	1550	1	3	2
...	...	...	...	...	...
33	12	1539	2	2	1
34	13	1587	2	2	2

L'objectif principal de cette étude est d'évaluer l'effet éventuel des caractéristiques socio-démographiques sur le salaire des employés.

1.2.1 Description de la population d'étude

Les variables sont analysées différemment selon leur nature : quantitative ou qualitative. Les variables quantitatives sont résumées sous forme d'indicateurs (moyenne, écart-type, ...), comme dans le tableau ci-dessous, et sont présentées graphiquement sous forme d'histogramme et de boîtes à moustache ou box-plot (Figure 1).

Variable	n	Moyenne	Ecart-type	Médiane	Minimum	Maximum
Ancienneté	32	10.0	6.1	12	1.0	20.0
Salaire	32	1365.4	308.0	1357	926.0	2024.0
Nombre d'années d'études	32	2.3	1.5	2.0	0.0	5.0

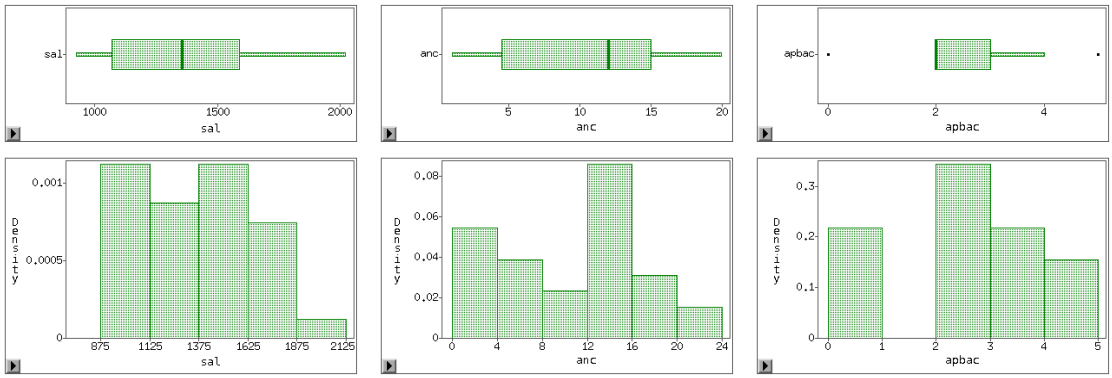


FIG. 1.1 – *Box-plot et histogramme représentant la distribution des variables quantitatives : le salaire, l'ancienneté dans l'entreprise et le nombre d'années d'études après le bac*

Pour les variables qualitatives, on résume les données sous forme de tableau de fréquences (comme ci-dessous) et on les présente graphiquement par des diagrammes en bâtons (Figure 2).

Variable	Modalités	Effectif	Fréquence(%)
Sexe	Féminin (1)	21	65.6%
	Masculin (2)	11	34.4%
Type d'emplois	1	10	31.3%
	2	17	53.1%
	3	5	15.6%

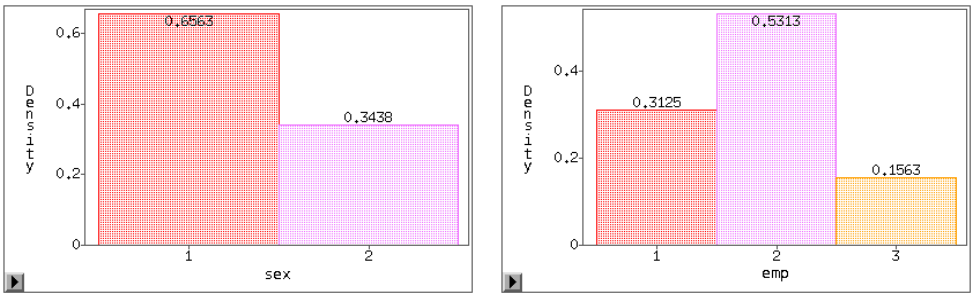


FIG. 1.2 – *Diagramme en bâtons représentant la distribution des variables qualitatives : le sexe (1=F, 2=M) et le type d'emplois occupés (1, 2 ou 3)*

1.2.2 Relation entre variables quantitatives

Etant donné l'objectif de l'étude, nous allons nous intéresser dans cette partie aux relations entre le salaire et les autres variables renseignées. Là encore, selon la nature des variables, les méthodes d'analyse sont différentes.

Pour étudier la relation entre deux variables quantitatives (par exemple, entre le salaire et l'ancienneté, et entre le salaire et le nombre d'année d'études), on peut tracer un nuage de points (Figure 3) et calculer le cœfficient de corrélation linéaire entre ces deux variables :

```
Pearson Correlation Coefficients, N = 32
      Prob > |r| under H0: Rho=0

                          anc          apbac

sal                   0.85559        0.42206
                       <.0001         0.0161
```

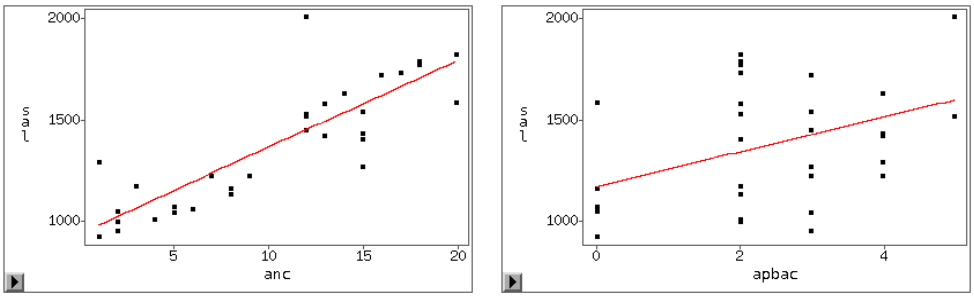


FIG. 1.3 – *Nuage de points représentant la relation entre le salaire et les deux autres variables quantitatives : l'ancienneté et le nombre d'années après le bac*

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

$$sal_i = \underbrace{934.5}_{\text{constante à l'origine}} + \underbrace{42.9}_{\text{pente du salaire sur l'ancienneté}} \times anc_i + e_i$$

La constante à l'origine correspond au salaire moyen des employés au moment de l'entrée dans l'entreprise. La pente représente la variation moyenne de salaire par année d'ancienneté. La pente égale à 42.9 est significativement différente de 0, montrant que le salaire et l'ancienneté sont liés de façon significative. Il en est de même pour la régression linéaire du salaire sur le nombre d'année d'études. Dans cet enseignement, on verra comment estimer les paramètres du modèle et tester leur nullité.

Il peut être également intéressant de modéliser une variable en fonction de plusieurs autres variables, par un modèle de **régression linéaire multiple**. Par exemple, on peut modéliser le salaire en fonction de l'ancienneté et du nombre d'années d'études, ce qui donne l'équation suivante :

$$sal_i = 858.9 + 40.2 \times anc_i + 45.3 \times apbac_i + e_i$$

1.2.3 Relation entre variable quantitative et variables qualitatives

Il est possible d'étudier la relation entre une variable quantitative et une variable qualitative, par exemple entre le salaire et le sexe, ou entre le salaire et le type d'emplois. Cette relation est représentée graphiquement par des box-plots parallèles (Figure 4).

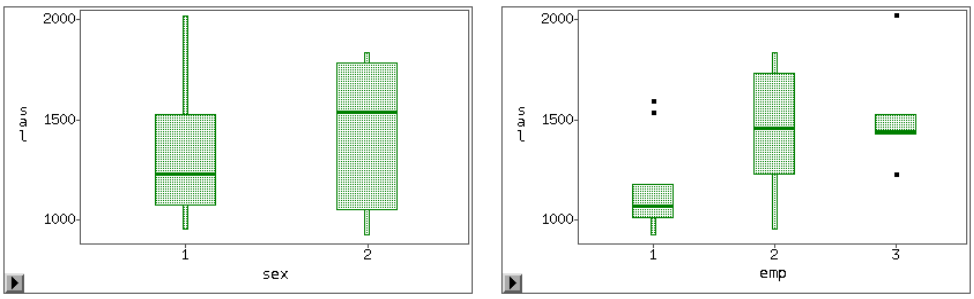


FIG. 1.4 – *Box-plots parallèles représentant la relation entre le salaire et les deux variables qualitatives : le sexe (1=F, 2=M) et le type d'emplois occupés (1, 2 ou 3)*

Intuitivement, pour comparer le salaire des hommes et celui des femmes, on va calculer le salaire moyen -entre autre- pour chaque groupe. De la même façon pour étudier les différences éventuelles entre les trois types d'emplois au niveau du salaire, on peut calculer le salaire moyen pour chaque type d'emplois.
Statistiquement, on modélise le salaire en fonction du sexe en mettant en œuvre un **modèle d'analyse de variance à un facteur** qui s'écrit sous la forme :

$$sal_i = \underbrace{1315.7}_{\text{salaire moyen des femmes}} \times \mathbb{1}_{sexe_i=1} + \underbrace{1460.3}_{\text{salaire moyen des hommes}} \times \mathbb{1}_{sexe_i=2} + e_i$$

Il est également possible d'étudier l'effet conjoint du sexe et du type d'emplois sur le salaire. Intuitivement, on peut étudier les moyennes par classe, en croisant les deux variables qualitatives,

comme dans le tableau ci-dessous :

	Sexe	F	M	Tous sexes confondus
Type d'emplois	1	1182.3	1111.2	1153.9
	2	1312.8	1750.4	1441.5
	3	1593.7	1433.0	1529.4
Tous types confondus		1315.7	1460.3	

Pour étudier l'effet combiné du sexe et du type d'emplois sur le salaire, on met en œuvre un **modèle d'analyse de variance à deux facteurs croisés**. Ce modèle nous permettra d'étudier l'effet de chaque facteur (sexe et type d'emplois) sur le salaire, mais aussi de détecter des combinaisons entre le sexe et le type d'emplois qui donneraient un salaire particulièrement différent des autres classes.

1.2.4 Modélisation d'une variable quantitative en fonction de variables quantitatives et qualitatives

Sur notre exemple, on peut tenter d'expliquer le salaire selon l'ancienneté (variable quantitative) et le sexe (variable qualitative). Dans ce cas, on peut représenter deux nuages de points entre le salaire et l'ancienneté, l'un pour les femmes et l'autre pour les hommes, comme le montre la figure 5.

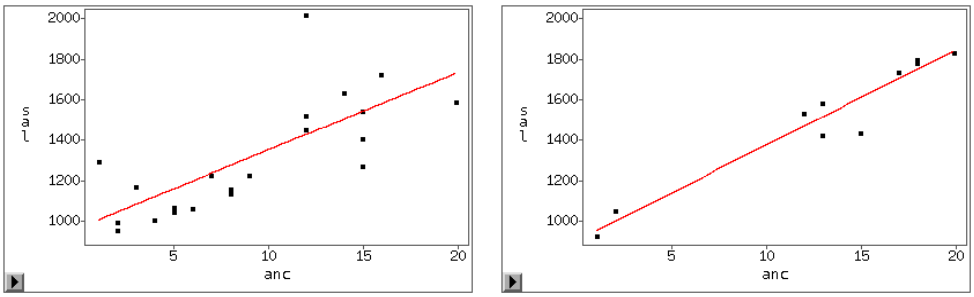


FIG. 1.5 – *Nuages de points représentant la relation entre le salaire et l'ancienneté selon le sexe*

On peut ainsi comparer l'effet de l'ancienneté sur le salaire, selon le sexe. Cela nous amène à mettre en œuvre un **modèle d'analyse de la covariance** permettant de modéliser le salaire en fonction de l'ancienneté et du sexe.

Chapitre 2

Présentation du modèle linéaire gaussien

2.1 Le modèle linéaire

• *Définition :*
On appelle *modèle linéaire* un modèle statistique qui peut s'écrire sous la forme

$$\boxed{Y = \sum_{j=1}^{k} \theta_j X^j + E}$$

On définit les quantités qui interviennent dans ce modèle :

- Y est une v.a.r. que l'on observe et que l'on souhaite expliquer et/ou prédire ; on l'appelle *variable à expliquer* ou *variable réponse* ; on suppose que la variance de Y est constante : c'est ce qu'on appelle l'hypothèse d'homoscédasticité.
- Les k variables $X^1, ..., X^k$ sont des variables réelles ou dichotomiques, non aléatoires et également observées ; l'écriture de ce modèle suppose que l'ensemble des X^j est censé expliquer Y par une relation de cause à effet ; les variables X^j sont appelées *variables explicatives* ou *prédicteurs*.
- Les θ_j $(j = 1, ..., k)$ sont les paramètres du modèle, non observés et donc à estimer par des techniques statistiques appropriées.
- E est le terme d'erreur dans le modèle ; c'est une v.a.r. non observée pour laquelle on pose les hypothèses suivantes :

  $$E(E) = 0 \ ; \ Var(E) = \sigma^2 > 0$$

  où σ^2 est un paramètre inconnu, à estimer.
- Les hypothèses posées sur E impliquent les caractéristiques suivantes sur Y :

  $$E(Y) = \sum_{j=1}^{k} \theta_j X^j \ ; \ Var(Y) = \sigma^2$$

  En moyenne, Y s'écrit donc comme une combinaison linéaire des X^j : la liaison entre les X^j et Y est de nature linéaire. C'est la raison pour laquelle ce modèle est appelé *modèle linéaire*.

L'estimation des paramètres de ce modèle est basée sur n observations simultanées des variables X^j et Y réalisées sur n individus supposés indépendants. Pour la i-ème observation, les valeurs observées des variables sont notées y_i, x_i^1, ..., x_i^k, de sorte que le modèle s'écrit :

$$y_i = \sum_{j=1}^{k} \theta_j x_i^j + e_i$$

Introduisons maintenant :
- $\mathbf{y}$ le vecteur de $\mathbb{R}^n$ composé des valeurs $y_1, ...y_n$,
- $\mathbf{X}$ la matrice (n,k) de rang k, contenant les valeurs observées des k variables explicatives disposées en colonnes,
- θ le vecteur de $\mathbb{R}^k$ contenant les k paramètres du modèle,
- $\mathbf{e}$ le vecteur de $\mathbb{R}^n$ des erreurs du modèle.

On peut donc écrire le modèle sous forme matricielle :

$$\mathbf{y} = X\theta + \mathbf{e}$$

Selon la forme de la matrice X, on est dans le cas de la régression linéaire (X est alors composée de la variable constante $\mathbf{1}$ et des p variables explicatives) ou dans le cas du modèle factoriel (X est composée des variables indicatrices associées aux niveaux du (ou des) facteur(s)).

2.2 Le modèle linéaire gaussien

On reprend la définition précédente du modèle linéaire en ajoutant une hypothèse de normalité des résidus. L'idée sous-jacente réside dans le fait qu'il existe une vraie valeur inconnue θ. Quand on réalise une série d'expériences, on obtient, comme pour les moyennes, les proportions ou les répartitions, une estimation $\widehat{\theta}$, c'est-à-dire une valeur approchée de la vraie valeur θ. Cette estimation de θ est différente selon les échantillons obtenus. D'après le Théorème Centrale Limite, cette estimation tend en moyenne vers la vraie valeur de θ. $\widehat{\theta}$ est donc une variable aléatoire dont on va chercher la distribution. Une fois posée la distribution de $\widehat{\theta}$, la question est de savoir si l'approximation obtenue est bonne ? Peut-on déterminer un intervalle du type $[\widehat{\theta}_j - \varepsilon_j; \widehat{\theta}_j + \varepsilon_j]$ qui contienne très probablement (avec un risque d'erreur petit) la vraie valeur θ_j ?

L'hypothèse de normalité des résidus revient à poser que les n composantes $e_1, ..., e_n$ du vecteur $\mathbf{e}$ sont des observations indépendantes d'une variable aléatoire E distribuée selon une loi $N(0, \sigma^2)$, avec σ^2 inconnu.

2.2.1 Ecriture générale

On appelle modèle linéaire gaussien la donnée d'un vecteur $\mathbf{y}$ de $\mathbb{R}^n$ tel que :

$\mathbf{y} = X\theta + \mathbf{e}$	où	X est une matrice (n, k) de rang k, θ est un vecteur inconnu de $\mathbb{R}^k$, **e est un vecteur de n réalisations indépendantes d'une v.a. normale de moyenne 0 et de variance σ^2 inconnue**.

Cette nouvelle formulation du modèle linéaire a pour conséquences :
- $\mathbf{e}$ est une réalisation d'une variable aléatoire E de distribution $N_n(0, \sigma^2 I_n)$; on peut dire aussi que e_i est une observation de la v.a. E_i distribuée selon une loi $N(0, \sigma^2)$ et les n v.a. réelles E_i sont indépendantes.
- $\mathbf{y}$ est une observation de $Y = X\theta + E$ de distribution $N_n(X\theta, \sigma^2 I_n)$: y_i est l'observation de Y_i de distribution $N((X\theta)_i, \sigma^2)$ et ces n variables aléatoires sont indépendantes.

En faisant intervenir les v.a. Y et E, le modèle linéaire gaussien peut aussi s'écrire sous la forme :

$Y = X\theta + E$ avec $E \sim N_n(0, \sigma^2 I_n)$	où	$Y \in \mathbb{R}^n$, $X \in M_{(n,k)}$, connue, déterministe, de rang k, $\theta \in \mathbb{R}^k$, inconnu, $\sigma^2 \in \mathbb{R}^{*+}$, inconnue.

Il en découle la normalité de Y :

$$Y \sim N_n(X\theta, \sigma^2 I_n)$$

L'hypothèse de normalité des résidus peut se justifier :

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

2.2.2 Le modèle de régression linéaire

On cherche à modéliser une variable quantitative Y en fonction de variables explicatives quantitatives $\mathbf{x}^1$, $\mathbf{x}^2$, ..., $\mathbf{x}^p$. Sous l'hypothèse gaussienne, le modèle de régression linéaire s'écrit :

$$y_i = \beta_0 + \beta_1 x_i^1 + ... + \beta_p x_i^p + e_i$$

avec β_0, β_1, ..., β_p inconnus, et e_1, ..., e_n n observations indépendantes d'une loi $N(0, \sigma^2)$ avec σ^2 inconnue.

2.2.3 Le modèle factoriel

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

$$y_{ij} = \mu_i + e_{ij} \qquad i = 1, ...I \ ; \ j = 1, ..., n_i$$

avec μ_1, ..., μ_I inconnus, et e_{11}, ..., e_{In_I} n observations indépendantes d'une loi $N(0, \sigma^2)$ avec σ^2 inconnue.

Chapitre 3

Estimation

θ est le vecteur des paramètres à estimer. Dans le cas général que nous étudions dans ce chapitre, θ est un vecteurs à k composantes : θ_1, θ_2, ..., θ_k. On note :
Y la variable aléatoire à expliquer,
$\mathbf{y}$ une réalisation de cette v.a. Y,
θ la vraie valeur théorique du vecteur des paramètres du modèle,
$\widehat{\theta}$ l'estimateur de θ,
$\widehat{\theta}(y)$ une réalisation de la v.a. $\widehat{\theta}$ (ou une estimation de θ à partir des données observées).

3.1 Méthodes d'estimation

3.1.1 Principe des moindres carrés

La méthode des moindres carrés consiste à estimer θ en minimisant la somme des carrés des résidus (SSR), telle que

$$\varphi(\widehat{\theta}(y)) = min \sum_{i=1}^{n} (\widehat{e_i})^2 = min \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

Le critère des moindres carrés peut s'écrire aussi de la façon suivante :

$$||\widehat{\mathbf{e}}||^2 = ||\mathbf{y} - X\widehat{\theta}||^2 = Inf_{\theta \in \mathbb{R}^k} ||\mathbf{y} - X\theta||^2$$

Cette méthode d'estimation ne nécessite pas que l'on pose l'hypothèse de normalité des résidus.

3.1.2 Principe du Maximum de Vraisemblance

L'estimation par maximum de vraisemblance est basée sur la vraisemblance du modèle linéaire gaussien :

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

où $f(y_i; \theta)$ est la densité de la loi Normale sur Y.

Pour obtenir l'estimateur $\widehat{\theta}$ du maximum de vraisemblance, on maximise sa log-vraisemblance selon θ en résolvant le système d'équations du maximum de vraisemblance :

$$\frac{\partial}{\partial \theta_j} ln\, L(\theta_1, ..., \theta_k; y) = 0 \text{ pour } j = 1, ..., k.$$

dont $\widehat{\theta}(y)$ est solution, sous réserve que la condition de seconde ordre soit vérifiée. On pourra également obtenir l'estimateur du MV de σ^2 en maximisant la log-vraisemblance selon σ^2.

Remarque : Les estimateurs du Maximum de Vraisemblance de θ sont équivalents aux estimateurs des Moindres Carrés de θ. On pourra le montrer dans le cas de la régression linéaire. En revanche, certaines propriétés ne sont possibles que sous l'hypothèse de normalité des résidus.

3.2 Estimation de θ

Si $\mathbf{y}$ est la réalisation de Y, l'estimation de θ, $\widehat{\theta}(y)$, est l'unique élément de $\mathbb{R}^k$ tel que

$$X\widehat{\theta}(y) = \widehat{\mathbf{y}}.$$

On a donc

$$\boxed{\widehat{\theta}(y) = (X'X)^{-1}X'\mathbf{y}}$$

$\widehat{\theta}(y)$ est l'observation de la v.a. $\widehat{\theta} = (X'X)^{-1}X'Y$: $\widehat{\theta}$ est la transformée de Y par l'a.l. $(X'X)^{-1}X'$.

Propriétés

• $\widehat{\theta}$ est un estimateur sans biais de θ.

• $\widehat{\theta}$ a pour matrice de variance-covariance $\Gamma_{\widehat{\theta}} = \sigma^2(X'X)^{-1}$.

• $\widehat{\theta}$ suit une loi Gaussienne dans $\mathbb{R}^k$.

On peut donc écrire que : $\boxed{\widehat{\theta} \sim N_k(\theta; \sigma^2(X'X)^{-1})}$

3.3 Valeurs ajustées et résidus calculés

Les $\widehat{y}_i$ s'appellent les *valeurs ajustées* ou *valeurs prédites* par le modèle : $\widehat{y}_i$ est une valeur approchée de y_i. On estime également les *résidus* $\widehat{e}_i$.

$$\boxed{\begin{array}{c}\widehat{\mathbf{y}} = X(X'X)^{-1}X'\mathbf{y} \\ \widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}}\end{array}}$$

• $\widehat{\mathbf{y}} = X\widehat{\theta}(y)$ est le vecteur des valeurs ajustées.
$\widehat{\mathbf{y}}$ est l'observation de la v.a. $\widehat{Y} = \underbrace{X(X'X)^{-1}X'}_{H}Y$ avec $\widehat{Y} \sim N_n(X\theta; \sigma^2 H)$.
$H = X(X'X)^{-1}X'$ est appelée la "matrice chapeau" ou "Hat Matrix".

• $\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}}$ est le vecteur des résidus calculés.
$\widehat{\mathbf{e}}$ est l'observation de la variable aléatoire $\widehat{E} = Y - \widehat{Y} = (I_n - H)Y$ avec $\widehat{E} \sim N_n(0; \sigma^2(I_n - H))$.

• Propriétés : $\widehat{Y}$ et $\widehat{E}$ sont deux v.a. indépendantes ; $\widehat{E}$ et $\widehat{\theta}$ sont deux v.a. indépendantes.

3.4 Estimation de σ^2

On note :
σ^2 la vraie valeur théorique de la variance des résidus,
$\widehat{\sigma^2}$ l'estimateur de σ^2,
et $\widehat{\sigma^2}(y)$ la réalisation de la v.a. $\widehat{\sigma^2}$ (ou une estimation de σ^2 à partir des données observées).

Définition
σ^2 est la variance "théorique" des résidus, on l'appelle *variance résiduelle*. Une autre définition de σ^2 est donnée par la variance de Y pour X fixé, c'est-à-dire la variance de Y autour de la droite de régression théorique. Cette définition de σ^2 suggère que son estimation est calculée à partir

des écarts entre les valeurs observées $\mathbf{y}$ et les valeurs ajustées $\widehat{\mathbf{y}}$.
L'estimateur de σ^2 est :

$$\boxed{\widehat{\sigma^2} = \frac{1}{n-k}||\widehat{E}||^2 = \frac{1}{n-k}||Y - \widehat{Y}||^2}$$

L'estimation de σ^2 est donc

$$\widehat{\sigma^2}(y) = \frac{1}{n-k}||\widehat{\mathbf{e}}||^2 = \frac{1}{n-k}||\mathbf{y} - \widehat{\mathbf{y}}||^2 = \frac{||\mathbf{y}||^2 - ||\widehat{\mathbf{y}}||^2}{n-k}$$

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

Rappelons que : $||\mathbf{y}||^2 = \sum_{i=1}^{n} y_i^2$ et que $||\widehat{\mathbf{y}}||^2 = \widehat{\theta}(y)'(X'y)$.

Propriétés

- $\dfrac{(n-k)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-k}$ (Somme des carrés de n v.a. $N(0,1)$ qui vérifient k relations linéaires).
- $\widehat{\sigma^2}$ est un estimateur sans biais de σ^2 et de variance $\dfrac{2\sigma^4}{n-k}$.
- $\widehat{Y}$ et $\widehat{\sigma^2}$ sont deux v.a. indépendantes ; $\widehat{\theta}$ et $\widehat{\sigma^2}$ sont deux v.a. indépendantes.

3.5 Erreurs standard de $\widehat{\theta}_j$, $\widehat{y}_i$, $\widehat{e}_i$

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

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

$Var(\widehat{\theta}_j)$ est donc estimée par $\widehat{\sigma^2}(X'X)^{-1}_{jj}$.
L'erreur standard de $\widehat{\theta}_j(y)$ notée se_j est donc :

$$\boxed{se_j = \sqrt{\widehat{\sigma^2}(y)(X'X)^{-1}_{jj}}}$$

Remarque : L'estimation de la matrice de variance-covariance $\widehat{\sigma^2}(y)(X'X)^{-1}$ est notée `cov b` par SAS.

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

$$r(\widehat{\theta}_j(y), \widehat{\theta}_{j'}(y)) = \frac{\widehat{\sigma^2}(y)(X'X)^{-1}_{jj'}}{se_j \times se'_j} = \frac{(X'X)^{-1}_{jj'}}{\sqrt{(X'X)^{-1}_{jj}(X'X)^{-1}_{j'j'}}}$$

Remarque : L'estimation de la matrice des corrélations de $\widehat{\theta}$ est notée `cor b` par SAS.

- $Var(\widehat{Y}) = \sigma^2 H$ est estimée par $\widehat{\sigma^2}(y)H$.
$\sqrt{\widehat{\sigma^2}(y)H_{ii}}$ est l'erreur standard de $\widehat{y}_i$.

- $\sqrt{\widehat{\sigma^2}(y)(1-H_{ii})}$ est l'erreur standard de $\widehat{e}_i$.
$\dfrac{\widehat{e}_i}{\sqrt{\widehat{\sigma^2}(y)}}$ est le résidu standardisé.
$\dfrac{\widehat{e}_i}{\sqrt{\widehat{\sigma^2}(y)(1-H_{ii})}}$ est le résidu studentisé.

3.6 Construction de l'intervalle de confiance de θ_j

Selon les propriétés de $\widehat{\theta}$, on a écrit que : $\widehat{\theta} \sim N_k(\theta; \sigma^2(X'X)^{-1})$ soit $\widehat{\theta}_j \sim N(\theta_j; \sigma^2(X'X)^{-1}_{jj})$

La v.a. $\dfrac{\widehat{\theta}_j - \theta_j}{\sqrt{\sigma^2(X'X)^{-1}_{jj}}}$ est distribuée selon une loi $N(0;1)$ et la v.a. $\dfrac{(n-k)\widehat{\sigma^2}}{\sigma^2}$ est distribuée selon une loi χ^2_{n-k}.

Ces deux v.a. étant indépendantes, on peut écrire que :

$$T = \frac{\widehat{\theta}_j - \theta_j}{\sqrt{\sigma^2(X'X)^{-1}_{jj}}} \Big/ \sqrt{\frac{(n-k)\widehat{\sigma^2}}{(n-k)\sigma^2}} = \frac{\widehat{\theta}_j - \theta_j}{\sqrt{\widehat{\sigma^2}(X'X)^{-1}_{jj}}} \sim Student(n-k)$$

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

$$\boxed{IC_{1-\alpha}(\theta_j) = \left[\widehat{\theta}_j(y) \pm t_{(1-\frac{\alpha}{2})}\sqrt{\widehat{\sigma^2}(y)(X'X)^{-1}_{jj}}\right] = \left[\widehat{\theta}_j(y) \pm t_{(1-\frac{\alpha}{2})} se_j\right]}$$

3.7 Décomposition de la variance

La mise en œuvre d'un modèle linéaire a pour objectif d'expliquer la variabilité d'une variable y par d'autres variables.

On note :

- $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 = n.Var(y)$ la variabilité totale de y,
- $SSL = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = n.Var(\widehat{y})$ la variabilité de y expliquée par le modèle, c'est-à-dire par les prédicteurs,
- $SSR = \sum_{i=1}^{n}(\widehat{e}_i)^2 = n.Var(\widehat{e})$ la variabilité résiduelle non expliquée par le modèle.

La variance totale de y admet la décomposition :

$$\boxed{Var(y) = Var(\widehat{y}) + Var(\widehat{e})}$$

soit :

$$SST = SSL + SSR$$

On verra par la suite que selon le modèle étudié (régression linéaire ou analyse de variance), cette décomposition amène à des définitions spécifiques à chaque modèle.

D'après le critère des moindres carrés utilisé pour estimer les paramètres, on cherche à minimiser la Somme des Carrés des résidus SSR, et donc à maximimser la Somme des Carrés expliquée par le modèle SSL.

Pour juger de la qualité d'ajustement du modèle aux données, on définit le critère R^2 qui représente la part de variance de y expliquée par le modèle :

$$\boxed{R^2 = SSL/SST = Var(\widehat{y})/Var(y)}$$

Chapitre 4

Test de Fisher

4.1 Hypothèse testée

4.1.1 Principe

On considère un modèle linéaire gaussien

$$Y = X\theta + E \text{ avec } E \sim N_n(0, \sigma^2 I_n)$$

et on s'intéresse à examiner la nullité de certaines composantes de θ ou de certaines combinaisons linéaires des composantes de θ, telles que : $\theta_j = 0$; $\theta_j = \theta_k = 0$ ou $\theta_j = \theta_k$. Ces hypothèses reposent sur la notion de modèles emboîtés : deux modèles sont dits "emboîtés" si l'un peut être considéré comme un cas particulier de l'autre. Cela revient à comparer un modèle de référence à un modèle réduit ou contraint.

Pour spécifier la nullité de certaines composantes de θ, on introduit la matrice Q d'ordre (q, k) où k est le nombre de paramètres dans le modèle de référence et q le nombre de contraintes linéaires testées $(1 \leqslant q \leqslant k)$ telle que :

$$H_0 \ : \ Q \in M_{(q,k)} \mid Q\theta = 0$$

Par exemple, supposons un modèle à $k = 3$ paramètres

- Tester l'hypothèse $H_0 : \theta_2 = 0$ revient à poser $Q\theta = 0$ avec $Q = (0\ 0\ 1)$, $q = 1$.
- $H_0 : \theta_1 = \theta_2 <=> Q\theta = 0$ avec $Q = (0\ -1\ 1)$ ou $Q = (0\ 1\ -1)$, $q = 1$.
- $H_0 : \theta_1 = \theta_2 = 0 <=> Q\theta = 0$ avec $Q = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $q = 2$.

4.1.2 Calculs sous H_0

L'hypothèse nulle étant définie, on a donc posé un modèle contraint que l'on va estimer en supposant H_0 vraie.

On a noté $\widehat{\theta}$ l'estimateur de θ correspondant au modèle de référence. On note $\widehat{\theta}_0$ l'estimateur de θ sous H_0, pour le modèle contraint. On peut obtenir, sous H_0, les valeurs prédites $\widehat{\mathbf{y}}_0$ et les résidus estimés $\widehat{\mathbf{e}}_0$. Le test de Fisher consiste à comparer les estimations du modèle de référence et celles sous H_0.

4.2 Le test de Fisher-Snédécor

4.2.1 Principe

Le test de Fisher-Sénédécor ou test de Fisher est la règle de décision qui permet de décider si on rejette ou ne rejette pas H_0 : "$Q\theta = 0$" :

– Rejeter H_0, c'est décider que $Q\theta \neq 0$, c'est-à-dire que certaines composantes de $Q\theta$ ne sont pas nulles.
– Ne pas rejeter H_0, c'est ne pas exclure que toutes les composantes de $Q\theta$ sont nulles.

On suppose que H_0 est vraie, c'est-à-dire que $Q\theta = 0$. On ré-estime θ par $\widehat{\theta}_0$ caractérisant le modèle contraint (noté M_0). Le vecteur des valeurs ajustées est $\widehat{\mathbf{y}}_0$ et le vecteur des résidus est $\widehat{\mathbf{e}}_0 = \mathbf{y} - \widehat{\mathbf{y}}_0$.

4.2.2 La statistique de test

On utilise la statistique de test suivante :

$$F_{cal} = \frac{(||\widehat{\mathbf{e}}_0||^2 - ||\widehat{\mathbf{e}}||^2)/q}{||\widehat{\mathbf{e}}||^2/(n-k)}$$

Le numérateur représente l'erreur commise en supposant H_0 vraie, sachant que de façon évidente : $||\widehat{\mathbf{e}}_0||^2 \geqslant ||\widehat{\mathbf{e}}||^2$. F_{cal} est donc l'erreur relative due à H_0. Si F_{cal} est grand, on peut rejeter H_0.
Une notation usuelle pour la somme des carrés des résidus est SSR. Dans ce cas, on définit :

$$SSR_0 = ||\widehat{\mathbf{e}}_0||^2 \text{ et } SSR_1 = ||\widehat{\mathbf{e}}||^2$$

d'où l'expression de F_{cal} :

$$F_{cal} = \frac{SSR_0 - SSR_1}{SSR_1} \times \frac{(n-k)}{q} \sim F(q, n-k)$$

On peut également montrer que

$$F_{cal} = \frac{(||\widehat{\mathbf{y}}||^2 - ||\widehat{\mathbf{y}}_0||^2)/q}{\widehat{\sigma^2}(y)}$$

On peut écrire la statistique du test de Fisher-Snédécor sous une autre forme :

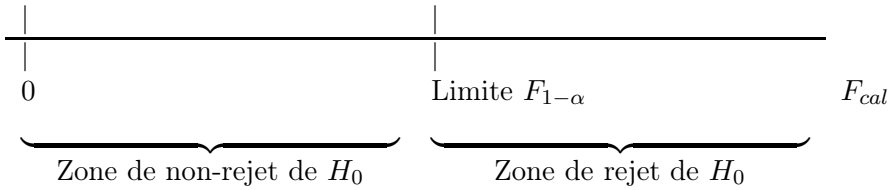
$$F_{cal} = \frac{\widehat{\theta}' Q' (Q(X'X)^{-1}Q')^{-1} Q\widehat{\theta}}{q\ \widehat{\sigma^2}} \sim F(q, n-k)$$

permettant de tester H_0 :"$Q\theta = 0$" contre H_1 :"$Q\theta \neq 0$". Cette expression a l'avantage de ne pas nécessiter l'estimation du modèle contraint.

4.2.3 Fonctionnement du test

Il faut définir une valeur limite $F_{1-\alpha}$ au dessus de laquelle F_{cal} sera considéré comme grand. Dans ce cas, la limite $F_{1-\alpha}$ est le $(1-\alpha)$-quantile de la distribution de Fisher de degrés de liberté q et $n-k$:

$$P[\text{v.a. de Fisher} < F_{1-\alpha}] = 1 - \alpha$$
$$P[\text{v.a. de Fisher} \geqslant F_{1-\alpha}] = \alpha$$



Le risque de première espèce du test de Fisher c'est-à-dire la probabilité de rejeter H_0 alors que H_0 est vraie, vaut α :

$$P[\text{Rejeter } H_0 \mid H_0 \text{ vraie}] = \alpha$$

En effet, on a montré que F_{cal} est distribué selon une loi $F(q, n-k)$ donc selon la règle de décision, la probabilité de rejeter H_0 est la probabilité que $F_{cal} \geqslant F_{1-\alpha}$ si $F_{cal} \sim F(q, n-k)$.

		La réalité (la vérité)	
		H_0 vraie	H_0 fausse
Décision	H_0 non rejetée	Bonne décision	Mauvaise décision (risque de 2ème espèce)
	H_0 rejetée	Mauvaise décision (risque de 1ère espèce)	Bonne décision

4.3 Cas particulier où q=1 : le test de Student

Dans le cas particulier où l'on teste la nullité d'une seule combinaison linéaire des composantes de θ (q=1), la matrice Q est d'ordre $(1, k)$ et l'hypothèse nulle s'écrit :

$$H_0 : \text{“}c'\theta = 0'' \Leftrightarrow \text{“}\varphi = 0''\text{avec } Q = c'$$

.
On a donc $Q(X'X)^{-1}Q' = c'(X'X)^{-1}c = l(c)$ (cf 2.6.2)
$\Rightarrow F_{cal} = \dfrac{\widehat{\varphi}(y)^2}{\widehat{\sigma^2}(y)l(c)} \sim F(1, n-k)$

Or une propriété de la distribution de la loi de Fisher-Snédécor est qu'une distribution de Fisher-Snédécor à 1 et m_2 degrés de liberté est le carré d'une distribution de Student à m_2 degrés de liberté (cf §1.3.5) :

$$P[F(1, n-k) > F_{1-\alpha}] = \alpha = P[(T(n-k))^2 > F_{1-\alpha}] \Rightarrow F_{1-\alpha} = t^2_{1-\alpha/2}$$

On rejette H_0 si $F_{cal} > F_{1-\alpha}$
$\Leftrightarrow |\widehat{\varphi}(y)| > t_{1-\alpha/2}\sqrt{\widehat{\sigma^2}(y)l(c)}$
$\Leftrightarrow -t_{1-\alpha/2}\sqrt{\widehat{\sigma^2}(y)l(c)} < \widehat{\varphi}(y) < +t_{1-\alpha/2}\sqrt{\widehat{\sigma^2}(y)l(c)}.$
Or l'intervalle de confiance de φ (défini au §2.6.3) est

$$[\widehat{\varphi}(y) \pm t_{1-\alpha/2}\sqrt{\widehat{\sigma^2}(y)l(c)}]$$

Le test consiste donc à rejeter H_0 ssi 0 n'appartient à l'intervalle de confiance de φ.

Chapitre 5

La Régression linéaire

5.1 Introduction

5.1.1 La problématique

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

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

- une variable quantitative y prenant la valeur y_i pour l'individu i $\{i = 1, ...n\}$, appelée **variable à expliquer** ou **variable réponse**,
- p variables quantitatives $x^1, x^2, ..., x^p$ prenant respectivement les valeurs $x_i^1, x_i^2, ...x_i^p$ pour l'individu i, appelées **variables explicatives** ou **prédicteurs** ; si $p = 1$, on est dans le cas de la régression simple ; lorsque les valeurs prises par une variable explicative sont choisies par l'expérimentateur, on dit que la variable explicative est *contrôlée*.

Considèrons un couple de variables quantitatives (X, Y). S'il existe une liaison entre ces deux variables, la connaissance de la valeur prise par X change notre incertitude concernant la réalisation de Y. Si l'on admet qu'il existe une relation de cause à effet entre X et Y, le phénomène aléatoire représenté par X peut donc servir à prédire celui représenté par Y et la liaison s'écrit sous la forme $y = f(x)$. On dit que l'on fait de la régression de y sur x.
Dans le cas d'une régression multiple de y sur $x^1, x^2, ..., x^p$, la liaison s'écrit $y = f(x^1, x^2, ..., x^p)$.

Dans les cas les plus fréquents, on choisit l'ensemble des fonctions affines (du type $f(x) = ax + b$ ou $f(x^1, x^2, ...x^p) = a_0 + a_1x^1 + a_2x^2 + ... + a_px^p$) et on parle alors de **régression linéaire**.

5.1.2 Le modèle de régression linéaire simple

Soit un échantillon de n individus. Pour un individu i ($i = 1, ...n$), on a observé

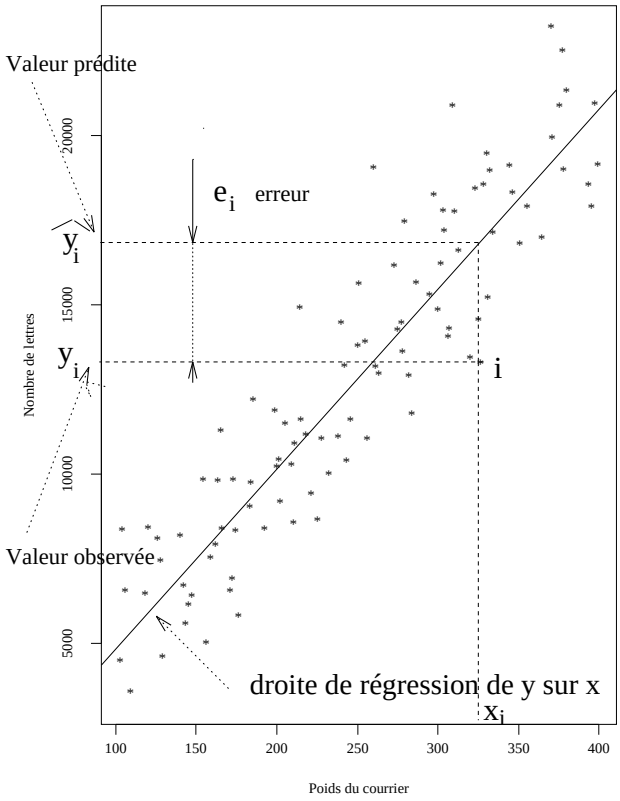
- y_i la valeur de la variable quantitative **y**,
- x_i la valeur de la variable quantitative **x**.

On veut étudier la relation entre ces deux variables, et en particulier, l'effet de **x** (*variable explicative*) sur **y** (*variable réponse*). Dans un premier temps, on peut représenter graphiquement cette relation en tracant le nuage des n points de coordonnées (x_i, y_i). Dans le cas où le nuage de points est de forme "linéaire", on cherchera à ajuster ce nuage de points par une droite.

La relation entre y_i et x_i s'écrit alors sous la forme d'un modèle de régression linéaire simple :

$$\boxed{y_i = \beta_0 + \beta_1 x_i + e_i} \qquad \forall i = \{1, ..., n\} \tag{5.1}$$

où e_i est une réalisation de $E_i \sim N(0, \sigma^2)$, et les n v.a. E_i sont indépendantes.
La première partie du modèle $\beta_0 + \beta_1 x_i$ représente la moyenne de y_i sachant x_i, et la seconde partie e_i, la différence entre cette moyenne et la valeur observée y_i. Le nuage de points est résumé par la droite d'équation $y = \beta_0 + \beta_1 x$.



Pour un x_i donné, correspondent donc y_i la valeur observée et $\beta_0 + \beta_1 x_i$ la valeur prédite par la droite.

5.1.3 Le modèle de régression linéaire multiple

On dispose d'un échantillon de n individus pour chacun desquels on a observé
- y_i, la valeur de la variable réponse $\mathbf{y}$ quantitative,
- $x_i^1, ..., x_i^p$, les valeurs de p autres variables quantitatives $\mathbf{x}^1, ..., \mathbf{x}^p$,

pour $i = \{1, ..., n\}$.
On veut expliquer une variable quantitative $\mathbf{y}$ par p variables quantitatives $\mathbf{x}^1, ..., \mathbf{x}^p$.
Le modèle s'écrit :

$$\boxed{y_i = \beta_0 + \beta_1 x_i^1 + ... + \beta_p x_i^p + e_i} \qquad \forall i = \{1, ..., n\} \tag{5.2}$$

où e_i est une réalisation de $E_i \sim N(0, \sigma^2)$ et où les n v.a. E_i sont indépendantes.

5.2 Estimation

5.2.1 Résultats généraux

• Les paramètres du modèle de régression linéaire sont estimés par :

$$\widehat{\beta}(y) = (X'X)^{-1} X'\mathbf{y}$$

Dans le cas de la régression linéaire simple sous la forme $y_i = \beta_0 + \beta_1 x_i + e_i$, on peut estimer β_0 et β_1 en utilisant aussi les formules suivantes :

$$\widehat{\beta}_0(y) = \overline{y} - \widehat{\beta}_1(y) \times \overline{x} \qquad\qquad \widehat{\beta}_1(y) = \frac{cov(x,y)}{var(x)}$$

On sait que $\widehat{\beta} \sim N_{p+1}(\beta, \sigma^2(X'X)^{-1})$.

• $\widehat{y}_i = \widehat{\beta}_0(y) + \sum_{j=1}^{p} \widehat{\beta}_j(y) x_i^j$ est la valeur ajustée de y_i.

$\widehat{e}_i = y_i - \widehat{y}_i$ est le résidu calculé.

• Une estimation de σ^2 est :

$$\widehat{\sigma^2}(y) = \frac{\sum_{i=1}^{n} (\widehat{e}_i)^2}{n-p-1}$$

On déduit les erreurs standard des paramètres estimés $\widehat{\beta}_0(y), ..., \widehat{\beta}_p(y)$, des valeurs ajustées et des résidus calculés :

- erreur standard de $\widehat{\beta}_j(y)$: *se* de $\widehat{\beta}_j(y) = \sqrt{\widehat{\sigma^2}(y)(X'X)^{-1}_{j+1,j+1}}$
- erreur standard de $\widehat{y}_i$: *se* de $\widehat{y}_i = \sqrt{\widehat{\sigma^2}(y)(X(X'X)^{-1}X')_{ii}} = \sqrt{\widehat{\sigma^2}(y)H_{ii}}$
- erreur standard de $\widehat{e}_i$: *se* de $\widehat{e}_i = \sqrt{\widehat{\sigma^2}(y)(1-H_{ii})}$

5.2.2 Propriétés

1. $\overline{\widehat{e}} = 0$,
2. $\overline{\widehat{y}} = \overline{y}$,
3. La droite de régression passe par le point de coordonnées $(\overline{x}, \overline{y})$
4. Le vecteur des résidus n'est pas corrélé avec la variable explicative : $cov(\mathbf{x}, \widehat{\mathbf{e}}) = 0$
5. Le vecteur des résidus n'est pas corrélé avec la variable ajustée Y : $cov(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = 0$
6. La variance de Y admet la décomposition :

$$var(\mathbf{y}) = var(\widehat{\mathbf{y}}) + var(\widehat{\mathbf{e}}). \tag{5.3}$$

7. Le carré du coefficient de corrélation de $\mathbf{x}$ et de $\mathbf{y}$ s'écrit sous les formes suivantes :

$$\boxed{r^2(\mathbf{x},\mathbf{y}) = \frac{var(\widehat{\mathbf{y}})}{var(\mathbf{y})} = 1 - \frac{var(\widehat{\mathbf{e}})}{var(\mathbf{y})}}$$

On en déduit que la variance empirique de $\mathbf{y}$ se décompose en somme d'une part de *variance expliquée* ($var(\widehat{\mathbf{y}})$) et d'une *variance résiduelle* ($var(\widehat{\mathbf{e}})$), et que le carré de $r(\mathbf{x},\mathbf{y})$ est le rapport de la variance expliquée sur la variance de la variable à expliquer.

5.2.3 Le coefficient R^2

On déduit de cette décomposition que le coefficient R^2, défini comme le carré du coefficient de corrélation de $\mathbf{x}$ et $\mathbf{y}$ est une mesure de qualité de l'ajustement, égale au rapport de la variance effectivement expliquée sur la variance à expliquer :

$$\boxed{R^2 = r^2(\mathbf{x},\mathbf{y}) = \frac{var(\widehat{\mathbf{y}})}{var(\mathbf{y})} \qquad 0 \leqslant R^2 \leqslant 1}$$

Le R^2 est la *proportion de variance expliquée par la régression.*

Pour calculer le R^2, on utilise également les expressions :

$$\boxed{R^2 = 1 - \frac{var(\widehat{\mathbf{e}})}{var(\mathbf{y})} = 1 - \frac{SSR}{n\ var(\mathbf{y})}}$$

La plupart des logiciels n'utilise pas la décomposition (5.3), mais plutôt la décomposition obtenue en multipliant cette expression par n :

$$\boxed{SST = SSL + SSR}$$

où :

- $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$ est la somme totale des carrés corrigés de $\mathbf{y}$,
- $SSL = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$ est la somme des carrés expliquée par le modèle,
- $SSR = \sum_{i=1}^{n}(\widehat{e}_i)^2$ est la somme des carrés des résidus.

La propriété (5) ci-dessus montre que la variance de la variable à expliquer (ou totale) se décompose en somme de la variance expliquée par le modèle ($var(\widehat{\mathbf{y}})$) et de la variance résiduelle ($var(\widehat{\mathbf{e}})$). On note encore R^2 le rapport de la variance expliquée sur la variance totale, soit :

$$R^2 = \frac{var(\widehat{\mathbf{y}})}{var(\mathbf{y})} = 1 - \frac{var(\widehat{\mathbf{e}})}{var(\mathbf{y})}$$

• *Définition :* On appelle *coefficient de corrélation multiple* de $\mathbf{y}$ avec $\mathbf{x}^1, ..., \mathbf{x}^p$, et on note $r(\mathbf{y}, (\mathbf{x}^1, ..., \mathbf{x}^p))$ le coefficient de corrélation linéaire empirique de $\mathbf{y}$ avec $\widehat{\mathbf{y}}$:

$$r(\mathbf{y}, (\mathbf{x}^1, ..., \mathbf{x}^p)) = r(\mathbf{y}, \widehat{\mathbf{y}})$$

• *Propriété :* Le coefficient R^2 de la régression multiple est égal au carré du coefficient de corrélation linéaire empirique $r(\mathbf{y}, (\mathbf{x}^1, ..., \mathbf{x}^p))$.

5.2.4 Augmentation mécanique du R^2

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

$$y_i = \beta_0 + \beta_1 x_i^1 + \ldots + \beta_j x_i^j + \ldots + \beta_{p-1} x_i^{p-1} + e_i,$$

alors les coefficients $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_j, \ldots, \widehat{\beta}_{p-1})$ estimés minimisent

$$\varphi(\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i^1 + \ldots + \beta_j x_i^j + \ldots + \beta_{p-1} x_i^{p-1})\Big)^2.$$

Si on rajoute une nouvelle variable explicative (la variable $\mathbf{x}^p$) au modèle, on obtient

$$y_i = \beta_0 + \beta_1 x_i^1 + \ldots + \beta_j x_i^j + \ldots + \beta_{p-1} x_i^{p-1} + \beta_p x_i^p + e_i,$$

et les coefficients estimés, notés $(\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_j, \ldots, \tilde{\beta}_{p-1}, \tilde{\beta}_p)$ minimisent la fonction :

$$\psi(\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_{p-1}, \beta_p) = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i^1 + \ldots + \beta_j x_i^j + \ldots + \beta_{p-1} x_i^{p-1} + \beta_p x_i^p)\Big)^2,$$

qui est par construction telle que

$$\psi(\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_{p-1}, 0) = \varphi(\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_{p-1}).$$

D'où l'inégalité :

$$\psi(\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_j, \ldots, \tilde{\beta}_{p-1}, \tilde{\beta}_p) \leqslant \psi(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_j, \ldots, \widehat{\beta}_{p-1}, 0) = \varphi(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_j, \ldots, \widehat{\beta}_{p-1}).$$

d'où le résultat. On verra par la suite qu'augmenter ainsi "mécaniquement" le R^2 n'est pas forcément synonyme d'amélioration de modèle.

5.3 Tests et Intervalles de confiance

5.3.1 Test de nullité d'un paramètre du modèle

On étudie l'effet de la présence d'une variable explicative X^j dans le modèle en testant l'hypothèse nulle :

$$H_0 : \beta_j = 0$$

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 la statistique de Student :

$$\boxed{T_{cal} = \frac{\widehat{\beta_j}}{\text{se de } \widehat{\beta_j}} \sim Student(n-p-1)}$$

à comparer avec la valeur limite $t_{(n-p-1),(1-\frac{\alpha}{2})}$.
Si $|T_{cal}| \geqslant t_{(n-p-1),(1-\frac{\alpha}{2})}$ alors on rejette H_0.
Si $|T_{cal}| < t_{(n-p-1),(1-\frac{\alpha}{2})}$ alors on ne peut pas rejeter H_0.

5.3.2 Test de nullité de quelques paramètres du modèle

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

$$H_0 : \beta_1 = \beta_2 = ... = \beta_q = 0 \text{ avec } q \leqslant p$$

Sous l'hypothèse alternative, au moins un des paramètres β_1, ..., β_q est non-nul.

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

$$y_i = \beta_0 + \beta_1 x_i^1 + ... + \beta_p x_i^p + e_i \text{ sous } H_1$$

$$\text{versus } y_i = \beta_0 + \beta_{q+1} x_i^{q+1} + ... + \beta_p x_i^p + e_i \text{ sous } H_0$$

L'hypothèse H_0 peut être testée au moyen de la statistique :

$$\boxed{F_{cal} = \frac{SSR_0 - SSR_1}{SSR_1} \times \frac{n-p-1}{q} \sim F(q, n-p-1)}$$

où SSR_0 est la somme des carrés des résidus du modèle "réduit" sous H_0 et SSR_1 est la somme des carrés des résidus du modèle de référence.

On compare F_{cal} à la valeur limite $F_{1-\alpha}(q, n-p-1)$: si $F_{cal} \geqslant F_{1-\alpha}(q, n-p-1)$ alors on rejette H_0.

Remarque : dans le cas où q=1, on teste la nullité d'un seul paramètre du modèle. Etant la propriété selon laquelle une v.a. distribuée selon une loi $F(1, m_2)$ est le carré d'une v.a. de Student à m degrés de liberté (cf §1.1.5), le test de Fisher-Snédécor ci-dessus et le test de Student (vu au paragraphe précédent) donnent les mêmes conclusions.

5.3.3 Test de nullité de tous les paramètres du modèle

Tester l'hypothèse de nullité de tous les paramètres du modèle (associés aux variables explicatives) :

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

revient à comparer la qualité d'ajustement du modèle de référence à celle du modèle blanc.
Cette hypothèse composée de p contraintes signifie que les p paramètres associés aux p variables explicatives sont nuls, c'est-à-dire qu'aucune variable explicative présente dans le modèle ne permet d'expliquer Y.
Sous H_0, le modèle s'écrit :

$$y_i = \beta_0 + e_i \text{ avec } \widehat{\beta}_0 = \overline{y}$$

et la somme des carrés des résidus (SSR_0) est égale à la somme des carrés totale (SST).

La statistique de Fisher-Snédécor permettant de **tester la nullité des p paramètres du modèle** peut donc s'écrire :

$$\boxed{F_{cal} = \frac{SSL_1}{SSR_1} \times \frac{n-p-1}{p} = \frac{R^2}{1-R^2} \times \frac{n-p-1}{p} \sim F(p, n-p-1)}$$

où SSL_1 est la somme des carrés du modèle de référence avec $SST = SSL_1 + SSR_1$, et R^2 est le critère d'ajustement du modèle de référence.

On compare F_{cal} à la valeur limite $F_{1-\alpha}(p, n-p-1)$: si $F_{cal} \geqslant F_{1-\alpha}(p, n-p-1)$ alors on rejette H_0 et on conclut qu'il existe au moins un paramètre non nul dans le modèle.

5.3.4 Intervalle de confiance de β_j, de $\overline{Y}_i$ et de $\overline{Y}_0$

• L'intervalle de confiance du paramètre β_j au risque α (ou de sécurité $1-\alpha$) est de la forme :

$$\boxed{IC_{1-\alpha}(\beta_j) = \left[\widehat{\beta}_j(y) \pm t_{n-p-1,1-\alpha/2} \times \text{ se de } \widehat{\beta}_j(y)\right]}$$

• On note $\overline{Y}_i$ la réponse moyenne de Y_i associée au jeu de valeurs $(x_i^1, x_i^2, ..., x_i^p)$ des variables explicatives : $\overline{Y}_i = (X\beta)_i = \beta_0 + \sum_j \beta_j x_i^j$. On l'estime par : $\widehat{y}_i = \widehat{\beta}_0(y) + \sum_j \widehat{\beta}'_j(y) x_i^j$.
L'intervalle de confiance de $\overline{Y}_i$ au risque α est :

$$\boxed{IC_{1-\alpha}(\widehat{Y}_i) = \left[\widehat{y}_i \pm t_{n-p-1,1-\alpha/2} \times \text{ se de } \widehat{y}_i\right]}$$

• Pour des valeurs données x_0^1, x_0^2, ..., x_0^p des variables explicatives, la réponse moyenne est :

$$\overline{Y}_0 = \beta_0 + \beta_1 x_0^1 + ... + \beta_p x_0^p = X_0'\beta \text{ où } X_0' = (1\ x_0^1\ x_0^2\ ...\ x_0^p)$$

L'estimateur de $\overline{Y}_0$ est :

$$\widehat{\beta}_0 = X_0'\widehat{\beta}$$

et la variance de cet estimateur est :

$$Var(\overline{Y}_0) = Var(X_0'\widehat{\beta}) = \sigma^2 X_0'(X'X)^{-1})X_0$$

L'estimation de $\overline{Y}_0$ est $\widehat{y}_0 = X_0'\widehat{\beta}(y)$ d'où on déduit l'intervalle de confiance de $\overline{Y}_0$ au risque α :

$$\boxed{IC_{1-\alpha}(\overline{Y}_0) = \left[\widehat{y}_0 \pm t_{n-p-1,1-\alpha/2} \times \sqrt{\widehat{\sigma^2}(y)(X_0'(X'X)^{-1}X_0)}\ \right]}$$

5.3.5 Intervalle de prédiction

Avant toute chose, il est important de comprendre la différence entre l'intervalle de confiance de $\widehat{Y}_0$ et l'intervalle de prédiction. Dans les deux cas, on suppose un jeu de valeurs données des variables explicatives. Dans le premier cas, on veut prédire une réponse moyenne correspondant à ces variables explicatives alors que dans le second cas, on cherche à prédire une nouvelle valeur "individuelle". Par exemple, si on étudie la liaison entre le poids et l'âge d'un animal, on peut prédire la valeur du poids à 20 jours soit comme le poids moyen d'animaux à 20 jours, soit comme le poids à 20 jours d'un nouvel animal. Pour le nouvel animal, on doit prendre en compte la variabilité individuelle, ce qui augmente la variance de l'estimateur et donc la largeur de l'intervalle.

La prédiction est à nouveau donnée par $\widehat{y}_0 = X_0'\widehat{\beta}(y)$. En revanche, la variance de la prédiction devient :

$$Var(\overline{Y}_0) + Var(E_0) = \sigma^2(1 + X_0'(X'X)^{-1}X_0)$$

L'intervalle de prédiction de sécurité $1-\alpha$ est donné par :

$$\boxed{\left[\widehat{y}_0 \pm t_{n-p-1,1-\alpha/2} \times \sqrt{\widehat{\sigma^2}(y)(1 + X_0'(X'X)^{-1}X_0)}\right]}$$

5.4 Sélection des variables explicatives

En présence de p variables explicatives dont on ignore celles qui sont réellement influentes, on doit rechercher un modèle d'explication de Y à la fois performant (résidus les plus petits possibles) et économique (le moins possible de variables explicatives).

5.4.1 Les critères

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

- choisir, parmi tous les modèles, le modèle pour lequel $\widehat{\sigma^2}(y)$ est minimum ;
- choisir, parmi tous les modèles, celui pour lequel le R^2 ajusté est maximum avec

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

  où p est le nombre de variables explicatives dans le modèle ;
- choisir le modèle pour lequel C_p de Mallows est minimum avec

  $$C_p = \frac{\sum(\widehat{e}_i)^2}{\widehat{\sigma^2}(y)} + 2p - n$$

- choisir le modèle pour lequel le critère PRESS (Prediction Sum of Squares) de Allen est minimum :

  $$PRESS = \sum_i (y_i - \tilde{y}_i)^2$$

  où $\tilde{y}_i$ est obtenu de la façon suivante :
  - on retire l'observation i du jeu de données,
  - β est alors estimé par $\widehat{\beta}^{(-i)}$,
  - $\tilde{y}_i$ est la prédiction de y_i d'après cette estimation de β.

5.4.2 Les méthodes de sélection

Toutes les méthodes de sélection nécessitent la donnée d'un des critères cités précédemment qui permet de comparer des modèles ayant des nombres de paramètres différents. On choisit donc un critère de qualité à optimiser, la variable à expliquer $\mathbf{y}$ et un ensemble de p variables candidates à l'explication de $\mathbf{y}$. Pour k fixé, on cherche le groupe de k variables, qui, parmi les p variables, explique le mieux $\mathbf{y}$. Comme la recherche du maximum du R^2 sur tous les ensembles de k variables prises parmi p peut prendre trop longtemps (ils sont au nombre de C_p^k) et peut amener à des artéfacts (un "bon" résultat qui n'en est pas un), on utilise souvent des méthodes pas à pas, qui sont soit ascendantes, descendantes ou "stepwise" :

1. Les méthodes ascendantes : On cherche d'abord la variable qui explique le mieux $\mathbf{y}$ au sens du R^2 (R^2 maximum), puis on cherche celle qui, ajoutée à la première, augmente le plus le R^2, etc. Un critère d'arrêt de la procédure peut-être obtenu en utilisant des critères du type R^2 ajusté, C_p de Mallows ou critère AIC : par exemple, on arrête le processus lorsque le R^2 ajusté commence à décroître.
2. Les méthodes descendantes : On part du modèle utilisant les p variables explicatives et on cherche, parmi les p variables, celle qui peut être supprimée en occasionnant la plus forte croissance du critère. Cette variable étant supprimée, on itère le processus tant que le R^2 ajusté ne décroît pas.
3. Les Méthodes stepwise : Partant d'un modèle donné, on opère une sélection d'une nouvelle variable (comme avec une méthode ascendante), puis on cherche si on peut éliminer une des variables du modèle (comme pour une méthode descendante) et ainsi de suite. Il faut définir pour une telle méthode un critère d'entrée et un critère de sortie.
4. On peut citer la méthode des "s best subsets" (ou "s meilleurs sous-ensembles") : on cherche de façon exhaustive parmi les sous-ensembles de s variables, les s meilleurs, au sens du critère considéré.

5.5 Validation du modèle

5.5.1 Contrôle de l'ajustement du modèle

Une fois le modèle mis en œuvre, on doit vérifier *a posteriori* le "bien-fondé statistique" de ce modèle du point de vue de la normalité des résidus et de l'adéquation de la valeur ajustée $\widehat{y_i}$ à la valeur observée y_i et de l'absence de données aberrantes. Pour se faire un idée sur ces questions, on peut étudier :

1. les résidus "standardisés" : $r_i = \dfrac{\widehat{e_i}}{\sqrt{\widehat{\sigma^2}(y)}}$.
2. les résidus "studentisés" : $t_i = \dfrac{\widehat{e_i}}{se \text{ de } \widehat{e_i}}$ dont on compare la répartition à la distribution $N(0;1)$ (tout en étant conscient que les n résidus ne sont pas indépendants mais liés par $p+1$ relations linéaires) en tracant le P-P Plot ou le Q-Q Plot (droite de Henry) et en comparant la proportion des résidus compris entre -1 et $+1$, entre -2 et $+2$, entre -2.6 et $+2.6$ respectivement à 70%, 95% et 99%. De grands résidus signalent plutôt des valeurs atypiques de la variable à expliquer.
3. le graphe des n points $(y_i, \widehat{y_i})$ qui "doivent" être à peu près alignés selon la droite de pente 1.
4. le graphe des n points $(\widehat{e_i}, \widehat{y_i})$ qui doit correspondre à celui de deux variables non-corrélées.
5. l'effet levier par les éléments diagonaux de la matrice H. En effet, l'estimation des paramètres est très sensible à la présence de points extrêmes pouvant modifier de façon substantielle les résultats. Une observation est influente si l'élément diagonal de la matrice H correspondant à cette observation est grand. L'effet levier apparait principalement pour des observations dont les valeurs prises par les variables explicatives sont éloignées de la moyenne.

6. les mesures d'influence peuvent aussi permettre de déceler des points "atypiques" avec la distance de Cook D_i pour l'individu i : $(\widehat{\beta} - \widehat{\beta}^{(-i)})'T'T(\widehat{\beta} - \widehat{\beta}^{(-i)})$ où T est le vecteur des résidus studentisés. Cette distance conclut à une infuence de l'observation i lorsque la valeur de D_i dépasse 1.

5.5.2 Etude des colinéarités des variables explicatives

Le problème

L'estimation des paramètres et de leurs variances nécessite le calcul de l'inverse de la matrice $(X'X)$. On dit que $(X'X)$ est mal conditionnée si son déterminant est proche de 0. La matrice $(X'X)^{-1}$ sera alors très grande. Cette situation se produit lorsque les variables explicatives sont très corrélées entre-elles. On parle alors de multi-colinéarité et cela conduit à des estimations biaisées des paramètres avec des variances importantes.

Remarque : Dans le cas extrême où certaines variables explicatives sont des constantes ou sont des combinaisons linéaires des autres, alors les colonnes de la matrice X sont des vecteurs linéairement liées et $X'X$ est singulière. Dans ce cas, `SAS` élimine certaines variables en leur affectant d'autorité un cœfficient nul.

Les critères de diagnostic

Il s'agit de diagnostiquer ces situations critiques puis d'y remédier. Une des techniques (la plus simple, mais pas la plus rapide) est de détecter les fortes liaisons entre variables explicatives en faisant la régression de chaque variable explicative sur les autres variables explicatives et en mesurant les liaisons par le R^2 de chacune des ces régressions. Un autre critère de diagnostic permet de détecter les problèmes de multi-colinéarité entre variables : le facteur d'inflation de la variance (VIF).
Soit $\tilde{X}$ la matrice des données observées centrées (c'est-à-dire la matrice X privée de la colonne $\mathbb{1}$ et centrée) et S la matrice diagonale contenant les écart-types empiriques des variables X^j, on peut définir R la matrice des corrélations sous la forme :

$$R = \frac{1}{n} S^{-1} \tilde{X}' \tilde{X} S^{-1}$$

On note $\tilde{\beta}$ le vecteur des paramètres associées aux p variables explicatives centrées. On peut montrer que $\widehat{\tilde{\beta}}$ et $Var(\widehat{\tilde{\beta}})$ peuvent s'exprimer en fonction de $\tilde{X}$:

$$\widehat{\tilde{\beta}} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y \text{ et } Var(\widehat{\tilde{\beta}}) = (\tilde{X}'\tilde{X})^{-1}\sigma^2$$

et on peut en déduire une nouvelle expression de $Var(\widehat{\tilde{\beta}})$:

$$Var(\widehat{\tilde{\beta}}) = \frac{\sigma^2}{n} S^{-1} R^{-1} S^{-1}$$

Si on note $Var(\widehat{\tilde{\beta}}_j)$ le jème élément diagonal de la matrice de variance-covariances de $\widehat{\tilde{\beta}}$ et V_j le jème élément diagonal de la matrice R^{-1} alors

$$Var(\widehat{\tilde{\beta}}_j) = \frac{\sigma^2}{n} \frac{V_j}{Var(X_j)}.$$

V_j est appelé *facteur d'inflation de la variance* (VIF) : plus V_j est grand, plus la variance de $\widehat{\beta}_j$ est grande. V_j peut s'exprimer comme :

$$V_j = \frac{1}{1 - R_j^2}$$

où R_j est le cœfficient de corrélation multiple obtenu en régressant X_j sur les $p-1$ autres variables explicatives. On appelle *tolérance* $1 - R_j^2$. Une tolérance et un facteur d'inflation de la variance qui tendent vers 1 signifient une absence de multicolinéarité entre les variables explicatives. En revanche, si la tolérance tend vers 0 et le facteur d'inflation de la variance vers ∞, alors on détecte un problème de multicolinéarité entre les variables explicatives.

Une première solution : la régression "ridge"

Une façon d'éviter ce problème d'inversibilité et donc de réduire les inconvénients de variables explicatives fortement corrélées est de remplacer $\widehat{\beta}$ par

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

où c est une réel choisi par l'utilisateur de la façon suivante : $\tilde{\beta}$ n'est plus un estimateur sans biais de β, mais il est de variance plus petite que $\widehat{\beta}$. On calcule l'erreur quadratique de $\tilde{\beta}$ ($variance + biais^2$) et on choisit c de façon que l'erreur quadratique de $\tilde{\beta}$ soit minimum.

Une seconde solution : la régression sur composantes principales

C'est une autre façon de "gérer" les colinéarités des variables explicatives :

- on fait l'A.C.P. des variables explicatives et on considère les composantes principales ; on note C la matrice des composantes principales : $C = (x^1|x^2|...|x^p)M$;
- on remplace les variables explicatives par les composantes principales qui sont non corrélées de variances décroissantes : on écrit donc le modèle sous la forme $y = \tilde{X}\gamma + e$ avec $\tilde{X} = (\mathbb{1}|C) = XB$ donc $\beta = B\gamma$;
- on estime γ par $\widehat{\gamma} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}y$. Comme les dernières composantes de $\widehat{\gamma}$ sont petites et de grandes se, on les remplace par 0 ce qui donne $\tilde{\gamma}$ qui est un estimateur biaisé de γ, donc $\tilde{\beta} = B\tilde{\gamma}$ est un estimateur biaisé de β de plus petite variance que $\widehat{\beta}$. On calcule l'erreur quadratique de $\tilde{\beta}$ et on choisit le nombre de composantes principales que l'on néglige de façon à minimiser l'erreur quadratique de $\tilde{\beta}$.

Chapitre 6

L'analyse de variance

6.1 Introduction

On applique des modèles factoriels quand on dispose :
- d'une variable quantitative à expliquer,
- d'une ou de plusieurs variables *qualitatives* explicatives, appelées *facteurs*.

• *Définition d'un facteur*

1. Un facteur est dit *contrôlé* si ses valeurs ne sont pas observées mais fixées par l'expérimentateur.
2. Les modalités des variables qualitatives explicatives sont appelées *niveaux* du facteur.

• *Définition d'un plan d'expérience*

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

6.2 L'analyse de variance à un facteur

6.2.1 Notations

On appelle plan à un facteur un plan d'expériences défini par un seul facteur ; on dispose donc d'une variable quantitative à expliquer et d'une seul facteur explicatif. On note
- i l'indice du groupe ou de la "cellule", définie par le facteur explicatif,
- I le nombre de groupes ($i = 1, \ldots, I$),
- n_i le nombre d'expériences dans le groupe i,
- $j = 1, \ldots, n_i$ l'indice de l'expérience dans le groupe i,
- enfin $n = \sum_{i=1}^{I} n_i$ le nombre total d'expériences.

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

6.2.2 Le modèle

On modélise une variable quantitative en fonction d'un facteur à I niveaux. $\mathbf{y}$ est la variable à expliquer qui prend la valeur y_{ij} pour l'individu j du niveau i du facteur. Le modèle s'écrit :

$$\boxed{y_{ij} = \mu_i + e_{ij}} \text{ avec } i = 1, ..., I \,;\, j = 1, ..., n_i \text{ et } n = \sum_{i=1}^{I} n_i$$

où e_{ij} est une réalisation de $E_{ij} \sim N(0, \sigma^2)$ et où les n v.a. E_{ij} sont indépendantes.

Le modèle peut également s'écrire sous la forme :

$$\mathbf{y} = (\mathbb{1}_1|\mathbb{1}_2|...|\mathbb{1}_I) \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{pmatrix} + \mathbf{e} \text{ avec } E \sim N(0, \sigma^2 I_n)$$

où $\mathbb{1}_i$ est l'indicatrice du niveau i. Ce modèle contient I paramètres à estimer.

6.2.3 Paramétrage centré

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

Introduisons quelques nouvelles notations : $\mu = \dfrac{\sum_i \mu_i}{I} = \overline{\mu}_{.}$ l'effet moyen général et $\alpha_i = \mu_i - \mu$ l'effet différentiel (centré) du niveau i. Le modèle initial peut s'écrire sous la forme :

$$\boxed{y_{ij} = \mu + \alpha_i + e_{ij} \text{ avec } \sum_{i=1}^{I} \alpha_i = 0}$$

ou bien :

$$\mathbf{y} = \mu \mathbb{1} + \sum_{i=1}^{I-1} \alpha_i (\mathbb{1}_i - \mathbb{1}_I) + \mathbf{e}$$

6.2.4 Estimation

On note $\overline{y}_{i.}$ la moyenne des observations y_{ij} dans la cellule i :

$$\overline{y}_{i.} = \frac{1}{n_i} \sum_{i=1}^{n_i} y_{ij}$$

Les coefficients μ_i sont estimés par les moyennes $\overline{y}_{i.}$ des observations dans les cellules :

$$\widehat{\mu}_i(y) = \overline{y}_{i.} = \frac{\sum_{j=1}^{I} y_{ij}}{n_i}$$

On les appelle les *effets principaux des facteurs*. Leur variance est estimée par :

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

Pour les deux autres paramétrisations : $\widehat{\mu}(y) = \dfrac{\sum_{i=1}^{n} \overline{y}_{i.}}{I} = \overline{y}_{..}$; $\widehat{\alpha}_i = \overline{y}_{i.} - \overline{y}_{..}$

Les valeurs ajustées $\widehat{y}_{ij}$ dans la cellule i sont constantes et sont égales aux moyennes $\overline{y}_{i.}$ des observations dans la cellule i :

$$\widehat{y}_{ij} = \overline{y}_{i.}$$

dont on déduit les résidus estimés :

$$\widehat{e}_{ij} = y_{ij} - \overline{y}_{i.}$$

L'estimation de σ^2 est donnée par :

$$\widehat{\sigma^2}(y) = \frac{\sum_i \sum_j (y_{ij} - \overline{y}_{i.})^2}{n - I}$$

6.2.5 Propriétés

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

1. La moyenne des résidus par cellule est nulle : pour tout $i = 1, \ldots, I$, $\overline{\widehat{e}}_{i.} = 0$
2. La moyenne générale des résidus est nulle : $\overline{\widehat{e}} = 0$
3. La moyenne des valeurs ajustées est égale à la moyenne des valeurs observées : $\overline{\widehat{y}} = \overline{y}$
4. $cov(\widehat{\mathbf{e}}, \widehat{\mathbf{y}}) = 0$
5. $var(\mathbf{y}) = var(\widehat{\mathbf{y}}) + var(\widehat{\mathbf{e}})$

La dernière propriété nous amène à définir les quantités suivantes :

- On appelle *variance inter-groupe* la quantité $var(\widehat{\mathbf{y}})$, qui s'écrit encore :

$$var(\widehat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{I} n_i (\overline{y}_{i.} - \bar{y})^2$$

  C'est la *variance des moyennes par cellule*, pondérées pour les poids des cellules n_i/n.
- On appelle *variance intra-groupe*, ou variance résiduelle, la quantité $var(\widehat{\mathbf{e}})$, qui s'écrit encore :

$$var(\widehat{\mathbf{e}}) = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i.})^2 = \frac{1}{n} \sum_{i=1}^{I} n_i Var_i(y)$$

  où $Var_i(y)$ est la variance des valeurs observées dans le niveau i : $Var_i(y) = \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i.})^2$

  $var(\widehat{\mathbf{e}})$ est donc la moyenne des variances des observations dans les cellules.
- La relation $var(\mathbf{y}) = var(\widehat{\mathbf{y}}) + var(\widehat{\mathbf{e}})$ s'écrit ici

$$\text{Variance totale} = \text{variance inter} + \text{variance intra}$$

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

$$R^2 = \frac{var(\widehat{\mathbf{y}})}{var(\mathbf{y})} = 1 - \frac{var(\widehat{\mathbf{e}})}{var(\mathbf{y}))}$$

On l'appelle rapport de corrélation empirique entre la variable quantitative $\mathbf{y}$ et le facteur considéré. C'est une *mesure de liaison entre une variable qualitative et une variable quantitative.* On peut mentionner les deux cas particuliers suivants :

$R^2 = 1 \iff \widehat{\mathbf{e}} = 0 \iff y_{ij} = \overline{y}_{i.} \forall i,\ \forall j = 1, \ldots, n_i$
$\mathbf{y}$ est constante dans chaque cellule

$R^2 = 0 \iff var(\widehat{\mathbf{y}}) = 0 \iff \overline{y}_{i.} = \overline{y}\ \forall i = 1, \ldots, I,$
La moyenne de $\mathbf{y}$ est la même dans chaque cellule

6.2.6 Intervalles de confiance et tests d'hypothèses sur l'effet facteur

Dans le cadre général du modèle gaussien, on a montré que les estimateurs des paramètres du modèle sont distribués selon une loi gaussienne. Cette propriété peut s'appliquer au modèle à un facteur pour lequel on a posé l'hypothèse de normalité des résidus.
On a montré précédemment que :

$$E(\widehat{\mu}_i) = \mu_i \text{ et } Var(\widehat{\mu}_i) = \frac{\sigma^2}{n_i}$$

d'où on déduit :

$$\widehat{\mu}_i \sim N(\mu_i; \frac{\sigma^2}{n_i})$$

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

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

• On veut étudier l'effet du facteur sur la variable **y** en posant l'hypothèse d'égalité de tous les paramètres du modèle :

$$H_0 : \mu_1 = \mu_2 = ... = \mu_I = \mu \Leftrightarrow \forall i \ \alpha_i = 0$$

$$\text{vs } H_1 : \exists (i, i') \text{ tel que } \mu_i \neq \mu_{i'}$$

Sous H_0, tous les paramètres μ_i sont égaux et le modèle s'écrit :

$$y_{ij} = \mu_0 + e_{ij} \text{ avec } \widehat{\mu}_0(y) = \overline{y} = \frac{1}{n} \sum_i \sum_j y_{ij}$$

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

$$F_{cal} = \frac{\sum_i \sum_j (\overline{y}_{i.} - \overline{y})^2}{\sum_i \sum_j (y_{ij} - \overline{y}_{i.})^2} \times \frac{n-I}{I-1} = \frac{\text{SSL}}{\text{SSR}} \times \frac{n-I}{I-1} \sim F(I-1, n-I)$$

où SSL est la somme des carrés inter-groupes et SSR est la somme des carrés intra-groupes.
Toutes ces estimations peuvent être présentées sous la forme d'un tableau d'analyse de la variance à un facteur :

Source	ddl	Somme des Carrés	Moyenne des Carrés	F_{cal}	$F_{1-\alpha}$
Facteur	$I-1$	$\sum_{i=1}^{I} n_i(\overline{y}_{i.} - \overline{y})^2 = SSL$	$\frac{SSL}{I-1} = MSL$	$\frac{MSL}{\widehat{\sigma^2}(y)}$	$F_{1-\alpha}(I-1, n-I)$
Résiduel	$n-I$	$\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i.})^2 = SSR$	$\frac{SSR}{n-I} = \widehat{\sigma^2}(y)$		
Total	$n-1$	$\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2 = SST$			

6.2.7 Comparaisons multiples : Méthode de Bonferroni

Si on rejette l'hypothèse d'égalité de tous les paramètres μ_i avec le test précédent, on conclut qu'au moins deux paramètres μ_i et $\mu_{i'}$ sont différents. On peut donc chercher à identifier les couples (i, i') pour lesquels $\mu_i \neq \mu_{i'}$. Il y a donc $I(I-1)/2$ comparaisons possibles. Pour identifier ces couples, il est possible de tester les hypothèses $\mu_i - \mu_{i'} = 0$ avec un test de Student tel que le

risque de première espèce conjoint soit α.

Ceci consiste donc à déterminer un intervalle de confiance de $\mu_i - \mu_{i'}$ de sécurité $1-\gamma$ avec $\gamma = 2\alpha/I(I-1)$:

$$IC_{1-\gamma}(\mu_i - \mu_{i'}) = \left[(\overline{y}_{i.} - \overline{y}_{i'.}) \pm t_{n-I,1-\gamma/2} \sqrt{\widehat{\sigma^2}(y)(\frac{1}{n_i} + \frac{1}{n_{i'}})} \right]$$

Les couples pour lesquels 0 n'appartient pas à l'intervalle de confiance sont ceux pour lesquels $\mu_i \neq \mu_{i'}$. La sécurité conjointe de ces intervalles est au moins égale à $1-\alpha$.

Dans le cas particulier où le facteur est composé de deux niveaux, le problème se résume à la comparaison de deux moyennes μ_1 et μ_2 de deux distributions gaussiennes de même variance. Le test de comparaison de μ_1 et μ_2 est un test de Student basé sur la statistique :

$$T_{cal} = \frac{|\overline{y}_{1.} - \overline{y}_{2.}|}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\widehat{\sigma^2}(y)}} \sim Student(n-2)$$

$$\text{avec } \widehat{\sigma^2}(y) = \frac{1}{n-2} \sum_{i=1}^{2} \sum_{j=1}^{n} (y_{ij} - \overline{y}_{i.})^2$$

Lorsque l'hypothèse de normalité des distributions n'est pas réaliste, mais que l'on peut accepter que les deux distributions sont identiques à un paramètre de position près δ, on peut utiliser le test des rangs pour comparer δ à 0 : le test de Wilcoxon pour le cas de deux échantillons indépendants, le test de Mann-Whitney pour le cas de données appariés. Le test de Kruskall-Wallis est une généralisation du test de Wilcoxon dans le cas de plusieurs échantillons (c'est-à-dire d'un facteur à plus de deux niveaux).

6.3 Analyse de variance à deux facteurs croisés

6.3.1 Notations

On appelle *cellule* une case du tableau, *facteur ligne* le facteur qui définit les lignes du tableau, *facteur colonne* celui qui définit les colonnes du tableau. De plus, on note par :

$i = 1, \ldots, I$	les indices des niveaux du facteurs ligne (le facteur ligne a I niveaux),
$j = 1, \ldots, J$	les indices des niveaux du facteurs colonne (le facteur colonne a J niveaux),
n_{ij}	le nombre d'observations pour le niveau i du facteur ligne et pour le niveau j du facteur colonne (on dit encore nombre d'observations dans la cellule (i,j),
$l = 1, \ldots, n_{ij}$	les indices des observations dans la cellule (i,j),
y_{ijl}	la l-ième observation dans la cellule (i,j),
$\overline{y}_{ij.}$	la moyenne des observations dans la cellule (i,j), $(y_{ij.} = 1/n_{ij} \sum_l y_{ijl})$.

6.3.2 Le modèle

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

$$\boxed{y_{ijl} = \mu_{ij} + e_{ijl}} \text{ avec } i = 1, ..., I; j = 1, ..., J; l = 1, ..., n_{ij}$$

où e_{ijl} est une réalisation de $E_{ijl} \sim N(0, \sigma^2)$, n v.a. indépendantes.
Deux autres paramétrisations permettent de décomposer μ_{ij} afin de définir des quantités, fonctions des μ_{ij}, qui mesurent les effets "séparés" des deux facteurs et les effets "conjoints".

6.3.3 La paramétrisation centrée

Cette première paramétrisation décompose μ_{ij} par rapport à un effet moyen général. On définit ainsi les nouveaux paramètres qui interviennent dans cette décomposition :

$\mu = \frac{1}{IJ}\sum_i \sum_j \mu_{ij} = \overline{\mu}_{..}$ = effet moyen général,

$\overline{\mu}_{i.} = \frac{1}{J}\sum_j \mu_{ij}$ = effet moyen du niveau i du facteur ligne,

$\alpha_i^L = \overline{\mu}_{i.} - \overline{\mu}_{..}$ = effet différentiel du niveau i du facteur ligne,

$\overline{\mu}_{.j} = \frac{1}{I}\sum_i \mu_{ij}$ = effet moyen du niveau j du facteur colonne,

$\alpha_j^C = \overline{\mu}_{.j} - \overline{\mu}_{..}$ = effet différentiel du niveau j du facteur colonne,

$\alpha_{ij} = \mu_{ij} - \overline{\mu}_{i.} - \overline{\mu}_{.j} + \overline{\mu}_{..}$ = interaction du niveau i du facteur ligne et du niveau j du facteur colonne.

Ces paramètres vérifient les conditions suivantes :

$$\sum_i \alpha_i^L = 0 \; ; \; \sum_j \alpha_j^C = 0 \; ; \; \forall i \sum_j \alpha_{ij} = 0 \; ; \; \forall j \sum_i \alpha_{ij} = 0$$

Le modèle complet s'écrit alors sous la forme :

$$\boxed{y_{ijl} = \mu + \alpha_i^L + \alpha_j^C + \alpha_{ij} + e_{ijl}}$$

Les $I.J$ paramètres μ_{ij} sont donc redéfinis en fonction de :

- μ : un paramètre de centrage général,
- α_i^L : $I-1$ paramètres qui caractérisent globalement sur j les I niveaux du facteur ligne,
- α_j^C : $J-1$ paramètres qui caractérisent globalement sur i les J niveaux du facteur colonne,
- α_{ij} : $(I-1)(J-1)$ paramètres qui prennent en compte que les effets des niveaux du facteur ligne varie selon le niveau du facteur colonne.

6.3.4 Estimations des paramètres

• μ_{ij} est estimé par $\widehat{\mu}_{ij}(y) = \frac{1}{n_{ij}}\sum_{l=1}^{n_{ij}} y_{ijl} = \overline{y}_{ij.}$ avec $\widehat{\mu}_{ij} \sim N(\mu_{ij}, \frac{\sigma^2}{n_{ij}})$

• On en déduit $\widehat{\overline{\mu}}_{i.} = \overline{y}_{i..} = \frac{1}{J}\sum_{j=1}^{J} \overline{y}_{ij.}$, $\widehat{\overline{\mu}}_{.j} = \overline{y}_{.j.} = \frac{1}{I}\sum_{i=1}^{I} \overline{y}_{ij.}$ et $\widehat{\mu} = \overline{y}_{...} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} \overline{y}_{ij.}$

• Valeurs ajustées et résidus estimés : $\widehat{y}_{ijl} = \widehat{\mu}_{ij} = \overline{y}_{ij.}$ et $\widehat{e}_{ijl} = y_{ijl} - \overline{y}_{ij.}$

• $\widehat{\sigma^2}(y) = \frac{1}{n-IJ}\sum_{ijk}(\widehat{e}_{ijl})^2 = \frac{1}{n-IJ}\sum_{ijk}(y_{ijl} - \overline{y}_{ij.})^2$ avec $n = \sum_{ij} n_{ij}$

• Des estimations de μ_{ij}, on déduit les estimations de μ, α_i^L, α_j^C et α_{ij} en remplacant μ_{ij} par son estimation dans les définitions de μ, α_i^L, α_j^C et α_{ij}.

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

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}}(y_{ijl} - \overline{y})^2$$

$$SSL = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(\overline{y}_{ij.} - \overline{y})^2$$

$$SSR = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} Var_{ij}(y) \qquad \text{où} \qquad Var_{ij}(y) = \frac{1}{n_{ij}}\sum_{l=1}^{n_{ij}}(y_{ijl})^2 - (\overline{y}_{ij.})^2$$

6.3.5 Le diagramme d'interactions

Le diagramme d'intéractions permet de visualiser graphiquement la présence ou l'absence d'intéractions. Pour chaque j fixé, on représente dans un repère othogonal les points (i, j) de coordonnées (i, μ_{ij}), et on trace les segments joignant les couples de points $((i-1), j)$, (i, j). On obtient ainsi pour chaque j fixé une ligne brisée.

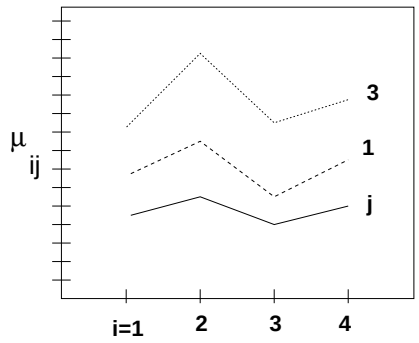


FIG. 6.1 – *Construction d'un diagramme d'interactions*

• *Propriété :* Si l'hypothèse de non-interaction est vraie, alors les lignes brisées dans le diagramme d'interaction sont parallèles.
En, effet, la ligne brisée associée au niveau j joint les points $(1, \mu_{1j})$, $(2, \mu_{2j})$, ..., (I, μ_{Ij}). S'il n'y a pas d'interactions, alors ces points ont pour coordonnées $(1, u_1 + v_j)$, $(2, u_2 + v_j), \ldots, (I, u_I + v_j)$, et les lignes brisées associées au niveaux j et j' se correspondent par une translation verticale d'amplitude $v_{j'} - v_j$.
On lit sur ce diagramme l'effet principal des modalités j (le niveau moyen d'une ligne brisée), l'effet principal des modalités i (la moyenne des ordonnées des points à abcisse fixée). En ce qui concerne les interactions, on obtiendra rarement des lignes brisées strictement parallèles. Le problème sera alors de savoir si leur non-parallèlisme traduit une interaction significative. Ce sera l'un des points de la partie sur le modèle linéaire gaussien.

6.3.6 Tests d'hypothèses

Trois hypothèses sont couramment considérées :

– l'hypothèse d'absence d'interactions entre les deux facteurs ou hypothèse d'additivité des deux facteurs :

$$H_0^{L,C} : \forall i, j,\ \alpha_{ij} = 0$$

qui impose $(I-1)(J-1)$ contraintes ;

– l'hypothèse d'absence d'effet du facteur ligne :

$$H_0^{L} : \forall i,\ \alpha_i^L = 0$$

qui impose $(I-1)$ contraintes ;

– l'hypothèse d'absence d'effet du facteur colonne :

$$H_0^C : \forall j,\ \alpha_j^C = 0$$

qui impose $(J-1)$ contraintes.

Pour ces trois hypothèses, le calcul de la statistique consiste à ré-estimer les paramètres sous la contrainte que l'hypothèse est vraie, à en déduire les nouvelles estimations des μ_{ij}, les valeurs ajustées et les résidus calculés sous cette hypothèse. On en déduit la statistique du test.

Une remarque très importante porte sur la démarche de ces tests d'hypothèses : **S'il existe des interactions entre les deux facteurs, alors les deux facteurs qui constituent cette interaction doivent impérativement être introduits dans le modèle ; dans ce cas, il est donc inutile de tester l'effet de chacun des deux facteurs**. En effet, la présence d'interactions entre les deux facteurs signifie qu'il y a un effet combiné des deux facteurs, et donc un effet de chaque facteur.

• Tester l'**hypothèse de non-interaction entre les deux facteurs** consiste à comparer le modèle complet (avec interactions) et le modèle additif (sans interactions) en utilisant la statistique de Fisher :

$$F_{cal} = \frac{SSR_{L,C} - SSR/(I-1)(J-1)}{SSR/n - IJ} = \frac{SSI}{SSR} \times \frac{n-IJ}{(I-1)(J-1)} \sim F((I-1)(J-1), n-IJ)$$

où $SSR_{l,c}$ est la somme des carrés des résidus du modèle additif, SSR est la somme des carrés des résidus du modèle complet et SSI la somme des carrés corrigés de l'effet d'interaction entre les deux facteurs.

• Tester l'**hypothèse d'absence d'effet du facteur ligne** est intéressant si le test précédent a permis de montrer l'absence d'interactions. En effet, si les termes d'interactions sont introduits dans le modèle, les facteurs qui constituent cette interaction doivent également apparaître dans le modèle. Cette remarque est également valable pour l'hypothèse d'absence d'effet du facteur colonne. Pour étudier l'effet du facteur ligne, on pose l'hypothèse H_0^L ce qui revient à comparer le modèle additif (à $I+J-1$ paramètres)

$$y_{ijl} = \mu + \alpha_i^L + \alpha_j^C + e_{ijl}$$

et le modèle à un facteur (à J paramètres)

$$y_{ijl} = \mu + \alpha_j^C + e_{ijl}$$

Le test est basé sur la statistique de Fisher-Snédécor :

$$F_{cal} = \frac{(SSR_C - SSR_{L,C})/(I-1)}{SSR_{L,C}/n - (I+J-1)} \sim F(I-1, n-(I+J-1))$$

où SSR_C est la somme des carrés des résidus du modèle à un facteur (le facteur colonne) et $SSR_{L,C}$ est la somme des carrés des résidus du modèle additif (à deux facteurs sans interaction).

• Pour étudier l'effet du facteur colonne, on compare le modèle à deux facteurs sans interaction au modèle à un facteur (à I paramètres) :

$$y_{ijl} = \mu + \alpha_i^L + e_{ijl}$$

et on teste l'**hypothèse d'absence d'effet du facteur colonne** H_0^C à partir de la statistique :

$$F_{cal} = \frac{(SSR_L - SSR_{L,C})/(J-1)}{SSR_{L,C}/n - (I+J-1)} \sim F(J-1, n-(I+J-1))$$

où $SSR_{L,C}$ est la somme des carrés des résidus du modèle additif et SSR_L est la somme des carrés des résidus du modèle à un facteur (le facteur ligne).

6.3.7 Tableau d'analyse de la variance à deux facteurs croisés dans le cas d'un plan équilibré

Dans le cas du modèle à deux facteurs croisés, la variance inter-cellule (expliquée par le modèle) peur être décomposée en une variance expliquée par le premier facteur, une variance expliquée par le second facteur et par une variance expliquée par les interactions entre les deux facteurs. Dans le cas d'un plan équilibré à deux facteurs (où $\forall(i,j),\ n_{ij} = n_0$), on définit les quantités suivantes :

- $SS1$, la somme des carrés corrigés de l'effet différentiel du premier facteur (Ligne) :

$$SS1 = n_0 J \sum_{i=1}^{I} (\overline{y}_{i..} - \overline{y}_{...})^2 = n_0 J \sum_{i=1}^{I} (\widehat{\alpha}_i^L)^2$$

- $SS2$, la somme des carrés corrigés de l'effet différentiel du second facteur (Colonne) :

$$SS2 = n_0 I \sum_{j=1}^{J} (\overline{y}_{.j.} - \overline{y}_{...})^2 = n_0 I \sum_{j=1}^{J} (\widehat{\alpha}_j^C)^2$$

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

$$SSI = n_0 \sum_{i=1}^{I} \sum_{j=1}^{J} (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 = n_0 \sum_{i=1}^{I} \sum_{j=1}^{J} (\widehat{\alpha}_{ij})^2$$

On peut montrer que :

$$SSL = SS1 + SS2 + SSI$$

On peut ainsi dresser le tableau d'analyse de la variance d'un plan équilibré à deux facteurs croisés :

Source	ddl	Somme des Carrés	Moyenne des Carrés	F_{cal}	$F_{1-\alpha}$
Ligne	$I-1$	$SS1$	$\frac{SS1}{I-1} = MS1$	$\frac{MS1}{\widehat{\sigma^2}(y)}$	$F_{1-\alpha}(I-1, n-IJ)$
Colonne	$J-1$	$SS2$	$\frac{SS2}{J-1} = MS2$	$\frac{MS2}{\widehat{\sigma^2}(y)}$	$F_{1-\alpha}(J-1, n-IJ)$
Interaction	$(I-1)(J-1)$	SSI	$\frac{SSI}{(I-1)(J-1)} = MSI$	$\frac{MSI}{\widehat{\sigma^2}(y)}$	$F_{1-\alpha}((I-1)(J-1), n-IJ)$
Résiduel	$n-IJ$	SSR	$\frac{SSR}{n-IJ} = \widehat{\sigma^2}(y)$		
Total	$n-1$	SST			

Chapitre 7

Analyse de covariance

7.1 Les données

Sur un échantillon de n individus, on observe deux variables quantitatives $\mathbf{x}$ et $\mathbf{y}$, et une variable qualitative T. La variable quantitative $\mathbf{y}$ est la variable réponse que l'on cherche à expliquer en fonction de la variable quantitative $\mathbf{x}$ et de du facteur T à J niveaux.
Chaque individu de l'échantillon est repéré par un double indice (i, j), j représentant le niveau du facteur T auquel appartient l'individu et i correspondant à l'indice de l'individu dans le niveau j. Pour chaque individu (i, j), on dispose d'une valeur x_{ij} de la variable $\mathbf{x}$ et d'une valeur y_{ij} de la variable $\mathbf{y}$.

Pour chaque niveau j de T (avec $j = 1, ..., J$), on observe n_j valeurs $x_{1j}, ..., x_{n_j j}$ et n_j valeurs $y_{1j}, ..., y_{n_j j}$ de Y.
$n = \sum_{j=1}^{J} n_j$ est le nombre d'observations.

Ces données peuvent être représentées conjointement sur un même graphique permettant de visualiser la relation éventuelle entre $\mathbf{y}$, $\mathbf{x}$ et T. Il s'agit de tracer un nuage de points de coordonnées (x_{ij}, y_{ij}), où tous les points du niveau j (j=1,...,J) sont représentés par le même symbole.

7.2 Le modèle

Le modèle est explicité dans le cas simple où une variable quantitative Y est expliquée par une variable qualitative T à J niveaux et une variable quantitative, appelée covariable X. Le modèle s'écrit :

$$\boxed{y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}} \qquad \text{avec } i = 1, ..., n_j \text{ et } j = 1, ...J.$$

Cela revient à estimer une droite de régression linéaire de Y sur X pour chaque niveau j du facteur T. Pour le niveau j, on estime les paramètres β_{0j}, constante à l'origine de la droite de régression, et β_{1j}, pente de la droite de régression.

7.3 La seconde paramétrisation

Comme pour les modèles factoriels, SAS opère une reparamétrisation faisant apparaître des effets différentiels par rapport à un niveau de référence, en général le dernier niveau du facteur. Le modèle associé à cette nouvelle paramétrisation s'écrit :

$$y_{ij} = \beta_{0J} + \underbrace{(\beta_{0j} - \beta_{0J})}_{\gamma_{0j}} + \beta_{1J} x_{ij} + \underbrace{(\beta_{1j} - \beta_{1J})}_{\gamma_{1j}} x_{ij} + e_{ij} \text{ avec } i = 1, ..., n_j \text{ et } j = 1, ...J - 1.$$

Le dernier niveau est considéré comme le niveau de référence caractérisé par β_{0J} et β_{1J}. Les autres paramètres γ_{0j} et γ_{1j} représentent respectivement, pour chaque niveau j, l'écart entre les

constantes à l'origine des niveaux j et J, et l'écart entre les pentes de régression des niveaux j et J.

Cette paramétrisation permet de faire apparaître :
- un effet d'interaction entre la covariable X et le facteur T (γ_{1j}) ;
- un effet différentiel du facteur T sur la variable Y (γ_{0j}) ;
- un effet différentiel de la covariable X sur la variable Y $(\beta_1 J)$.

7.4 Tests d'hypothèses

Comme pour le modèle factoriel, il est important de suivre une démarche logique dans la mise en place des tests d'hypothèses. La première étape doit consister à tester l'hypothèse de non-interaction entre le facteur T et la covariable X :

$$H_0^i : \beta_{11} = \beta_{12} = ... = \beta_{1J} \Leftrightarrow \gamma_{11} = \gamma_{12} = ... = \gamma_{1J-1} = 0$$

en comparant le modèle dit complet :

$$y_{ij} = \beta_{0J} + \gamma_{0j} + \beta_{1J} x_{ij} + \gamma_{1j} x_{ij} + e_{ij} \text{ avec } i = 1, ..., n_j \text{ et } j = 1, ...J - 1.$$

au modèle sans interaction :

$$(i)\ y_{ij} = \beta_{0J} + \gamma_{0j} + \beta_{1J} x_{ij} + e_{ij}$$

Si on rejette cette hypothèse, on conclut à la présence d'interactions dans le modèle. Il est alors inutile de tester l'absence d'effet du facteur T ou de la covariable X sur Y, car toute variable constituant une interaction doit appaître dans le modèle.
En revanche, si ce premier test montre que l'hypothèse H_0^i est vraisemblable et qu'il n'existe pas d'interaction entre T et X (les J droites de régression partagent la même pente de régression), on peut alors évaluer l'effet de la covariable X sur Y et celui du facteur T sur Y.

On peut tester deux hypothèses en comparant le modèle sans interaction :

$$y_{ij} = \beta_{0J} + \gamma_{0j} + \beta_{1J} x_{ij} + e_{ij} \text{ avec } i = 1, ..., n_j \text{ et } j = 1, ...J - 1.$$

à chacun des modèles réduits suivants :

(ii) $y_{ij} = \beta_{0J} + \gamma_{0j} + e_{ij}$
correspondant à l'hypothèse d'absence d'effet de la covariable X sur Y

$$H_0^{ii} : \beta_{11} = \beta_{12} = ... = \beta_{1J} = 0$$

Seul le facteur T explique Y, on met en place un modèle à un facteur.

(iii) $y_{ij} = \beta_{0J} + \beta_{1J} x_{ij} + e_{ij}$
correspondant à l'hypothèse d'absence d'effet du facteur T sur Y

$$H_0^{iii} : \beta_{01} = \beta_{02} = ... = \beta_{0J} \Leftrightarrow \gamma_{01} = \gamma_{02} = ... = \gamma_{0J-1} = 0$$

Les J droites de régression partagent la même constante à l'origine, seule la covariable X explique Y : on met en place un modèle de régression linéaire simple.

Ces différentes hypothèses sont testées en comparant le modèle complet au modèle réduit par la statistique de Fisher-Snédécor :

$$F_{cal} = \frac{(SSR_0 - SSR_1)/q}{SSR_1/ddl} \sim F(q, ddl)$$

où :

- SSR_1 est la somme des carrés des résidus du modèle complet,
- SSR_0 est la somme des carrés des résidus du modèle contraint,
- q est le nombre de contraintes posées sous l'hypothèse nulle, c'est-à-dire le nombre de paramètres dans le modèle complet - le nombre de paramètres dans le modèle comtraint,
- ddl est le nombre de degrés de liberté associé aux résidus du modèle complet, c'est-à-dire le nombre d'observations - le nombre de paramètres dans le modèle complet.

Cette statistique est à comparer à la valeur limite $F_\alpha(q, ddl)$. Si F_{cal} est supérieure à cette valeur limite, on rejette l'hypothèse nulle.

Chapitre 8

Quelques rappels de Statistique et de Probabilités

8.1 Généralités

• *Définition :* Une *unité statistique* est un *individu* ou *objet* sur lequel on effectue des mesures ou observations. Les unités statistiques sont numérotées de 1 à n ; on note $I = \{1, \ldots, n\}$ cet ensemble d'indices.

L'ensemble des individus pourra être un échantillon (une partie) d'une population plus grande. Sous des hypothèses fondées sur la théorie du calcul des probabilités, il sera possible de déduire d'observations sur l'échantillon des conclusions applicables à l'ensemble de la population. C'est l'objet de la *statistique inférentielle*. On parlera alors de *variable aléatoire*, et une valeur observée sera appelée une *réalisation* de la variable aléatoire.
L'ensemble des observations pourra aussi concerner toute la population. On parle alors de données exhaustives. Dans ce cas, et même dans le cas d'observations partielles, on peut avoir comme seul objectif de décrire les données observées, sans chercher à établir de loi valable pour des cas non observés. C'est le but de la *statistique descriptive*.

• *Définition :* On appelle *variable statistique* (ou simplement *variable*) un ensemble de n observations de même type effectuées sur les n individus.

Typologie des variables statistiques
• On dit qu'une variable est *quantitative* quand elle prend ses valeurs dans l'ensemble des réels.
• Si elle prend ses valeurs dans un ensemble dont le nombre d'éléments est fini, on dit qu'elle est *qualitative* (on dit aussi *catégorielle* ou *nominale*). Pour ce type de variable, dans le cadre du modèle linéaire, on parle de *facteurs*. L'ensemble des valeurs d'une variable qualitative est appelé l'ensemble des *modalités* de la variable ; pour un facteur, on parle de l'ensemble des *niveaux* du facteur. Si l'ensemble des modalités possède une structure d'ordre, on parle de variable *ordinale* ou *qualitative ordonnée*.

Notations des variables quantitatives
On note y_i l'observation relative à l'individu i. La variable quantitative $\mathbf{y}$ est identifiée au vecteur de $\mathbb{R}^n$ de coordonnées y_i. Tous les vecteurs sont par convention représentés en colonne et noté en caractères latin minuscule gras. Un scalaire est désigné par un caractère grec (ou latin) ordinaire, une matrice par une lettre majuscule. On note donc dans la suite :

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \cdots \\ y_i \\ \cdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

une variable observée sur les n individus. On parlera aussi du vecteur des observations. Lorsqu'il y a plusieurs variables, elles sont indicées par j ($j = 1, \ldots, p$, indice placé en haut), et on note $\mathbf{y}^j$ la j-ème variable. Ainsi pour p variables :

$$\mathbf{y}^1 = \begin{pmatrix} y_1^1 \\ \cdots \\ y_i^1 \\ \cdots \\ y_n^1 \end{pmatrix}, \ldots \mathbf{y}^j = \begin{pmatrix} y_1^j \\ \cdots \\ y_i^j \\ \cdots \\ y_n^j \end{pmatrix}, \ldots \mathbf{y}^p = \begin{pmatrix} y_1^p \\ \cdots \\ y_i^p \\ \cdots \\ y_n^p \end{pmatrix}.$$

L'espace $\mathbb{R}^n$ est appelé *espace des variables.*

8.2 Indicateurs statistiques pour variables quantitatives

8.2.1 Moyenne empirique d'une variable

• *Définition :* La moyenne empirique d'une variable $\mathbf{y}$ est définie par :

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

Propriété caractéristique : $a \in \mathbb{R}$ est la moyenne empirique de la variable $\mathbf{y} \iff$

$$\sum_{i=1}^{n} (y_i - a) = 0$$

• *Propriété :* L'application qui à une variable $\mathbf{y}$ de $\mathbb{R}^n$ associe sa moyenne empirique est une *forme linéaire* sur $\mathbb{R}^n$ (application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}$).

• *Définition :*
  • Une variable de moyenne nulle est dite *centrée*;
  • soit $\mathbf{1}_n$ le vecteur de $\mathbb{R}^n$ dont toutes les coordonnées sont égales à 1, alors :

$$\begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_i - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} = \mathbf{y} - \bar{y}\mathbf{1}_n$$

  est appelée *variable centrée* de $\mathbf{y}$. Ses valeurs sont les écarts à la moyenne de la variable $\mathbf{y}$.

8.2.2 La covariance empirique

• *Définition :* La *covariance empirique* de $\mathbf{y}$ et $\mathbf{z}$ s'écrit :

$$cov(\mathbf{y}, \mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(z_i - \bar{z}) = \frac{1}{n} \sum_{i=1}^{n} y_i z_i - \bar{y}\bar{z}.$$

• *Propriété :* La covariance empirique possède les propriétés suivantes :
  • $cov(\mathbf{y}, \mathbf{z}) = cov(\mathbf{z}, \mathbf{y})$
  • $cov(\mathbf{y}, \mathbf{z}) = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{i' \in I} (y_i - y_{i'})(z_i - z_{i'})$
  • La covariance est une forme bilinéaire :
    – linéarité à droite : pour tous réels α, β, pour toutes variables $\mathbf{z}$ et $\mathbf{t}$:

$$cov(\mathbf{y}, \alpha\mathbf{z} + \beta\mathbf{t}) = \alpha cov(\mathbf{y}, \mathbf{z}) + \beta cov(\mathbf{y}, \mathbf{t}),$$

    – linéarité à gauche : s'obtient de la même manière par permutation.
  • La covariance d'une variable avec une constante est nulle.

8.2.3 Variance empirique et écart-type empirique

• *Définition :* La *variance empirique* de $\mathbf{y}$ est :

$$var(\mathbf{y}) = cov(\mathbf{y}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2$$

L'écart-type empirique est $\sigma_y = \sqrt{var(\mathbf{y})}$.

• *Propriété :* La variance possède les propriétés suivantes :
- $var(\mathbf{y}) = 0 \Leftrightarrow \forall i \in I, \quad y_i = \bar{y} \Leftrightarrow \mathbf{y}$ est une variable constante.
- $var(\mathbf{y}) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{i' \in I}(y_i - y_i{}')^2$
- $\forall \alpha$ et $\beta \in \mathbb{R}$:

$$var(\alpha\mathbf{y} + \beta\mathbf{1}_n) = var(\alpha\mathbf{y}) = \alpha^2 var(\mathbf{y})$$

La transformation $\mathbf{y} \to \mathbf{y} + \beta\mathbf{1}_n$ correspond à un changement de l'origine de l'échelle des mesures, et la transformation $\mathbf{y} \to \alpha\mathbf{y}$ correspond à un changement d'unité.

• *Définition :* On appelle variable *centrée réduite* associée à $\mathbf{y}$ la variable $\mathbf{z} = (z_i)_{i=1}^n$ telle que :

$$z_i = \frac{(y_i - \bar{y})}{\sigma_y}$$

• *Propriété :* $\mathbf{z}$ est une variable centrée réduite si et seulement si $\bar{z} = 0$ et $var(\mathbf{z}) = 1$.

8.2.4 Cœfficient de corrélation linéaire empirique

• *Définition :* Le cœfficient de corrélation linéaire empirique de $\mathbf{y}^1$ et $\mathbf{y}^2$ est :

$$r(\mathbf{y}^1, \mathbf{y}^2) = \frac{cov(\mathbf{y}^1, \mathbf{y}^2)}{\sqrt{var(\mathbf{y}^1)var(\mathbf{y}^2)}}$$

• *Propriété :* Le cœfficient de corrélation linéaire vérifie les propriétés suivantes :
- $r(\mathbf{y}, \mathbf{y}) = 1$; $r(\mathbf{y}^1, \mathbf{y}^2) = r(\mathbf{y}^2, \mathbf{y}^1)$; $r(\mathbf{y}^1, \mathbf{y}^2) \in [-1, +1]$
- $r(\mathbf{y}^1, \mathbf{y}^2) = \pm 1 \Leftrightarrow \exists \alpha$ et $\beta \ \mid \ \forall i \in I : y_i^1 = \alpha y_i^2 + \beta$ avec $signe(\alpha) = signe(r)$.
- $\forall \alpha, \beta, \alpha', \beta' \in \mathbb{R}, \ \ r(\alpha\mathbf{y}_1 + \beta\mathbf{1}_n, \alpha'\mathbf{y}_2 + \beta'\mathbf{1}_n) = signe(\alpha\alpha')\, r(\mathbf{y}^1, \mathbf{y}^2)$.

Deux variables de corrélation linéaire nulle sont dites *non corrélées*. Attention, le cœfficient de corrélation linéaire ne mesure la liaison que lorsque celle-ci est de type linéaire. Il suppose aussi une "bonne" répartition des observations. On donne figure 8.1. des exemples de diagrammes de dispersion de paires de variables statistiques : pour chaque paire, on donne la valeur du coefficient de corrélation empirique. En cas de liaison non linéaire, on peut utiliser le cœfficient de corrélation des rangs : on ne s'intéresse alors qu'à l'ordre des observations. Pour le calculer, on remplace dans la formule donnant le cœfficient de corrélation linéaire empirique, les valeurs de chaque variable par les rangs des valeurs observées.

8.2.5 Interprétation géométrique de quelques indices statistiques

On munit l'espace des variables $\mathbb{R}^n$ du produit scalaire défini par :

$$< \mathbf{x}, \mathbf{x}' >= \frac{1}{n}\sum_{i=1}^{n} x_i x_i',$$

et on note $\| \ \|$ la norme associée : $\|\mathbf{x}\| =< \mathbf{x}, \mathbf{x} >^{1/2}$. Si $\mathbf{x}$ est la variable centrée de $\mathbf{y}$, on a :
- $\bar{y} =< \mathbf{y}, \mathbf{1}_n >$
- $\mathbf{y}$ est centré $\Leftrightarrow \mathbf{y} \perp \mathbf{1}_n$
- $\mathbf{x}$ est la projection orthogonale de $\mathbf{y}$ sur le sous-espace vectoriel $\mathbf{1}_n^{\perp}$;

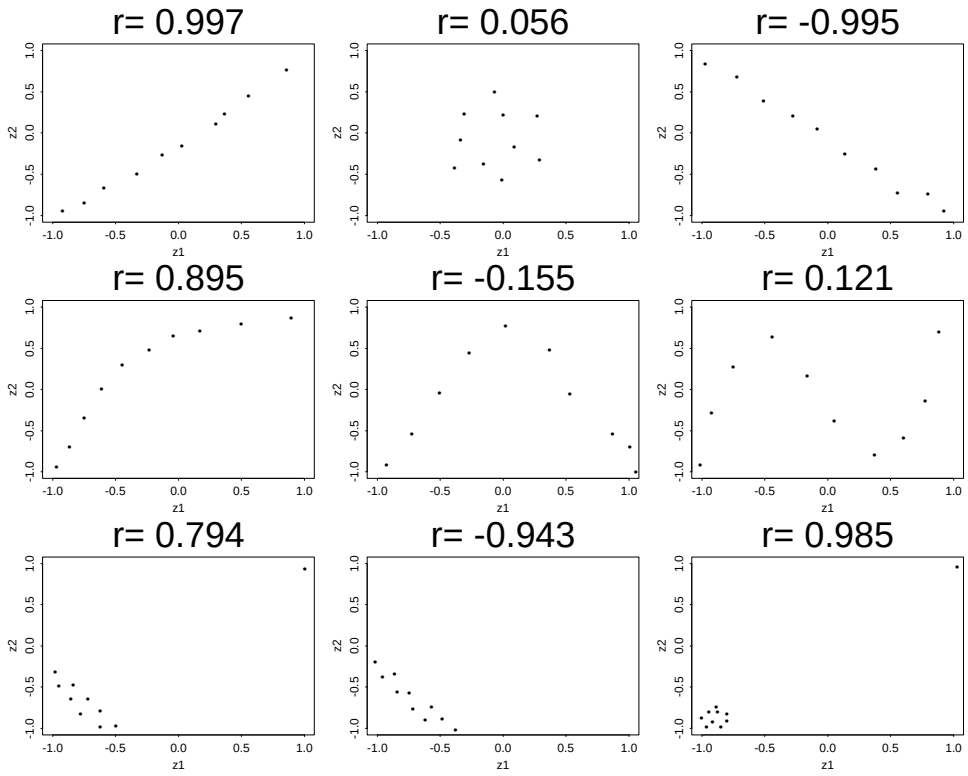


FIG. 8.1 – *Coefficient de corrélation linéaire empirique dans différents cas de figures*

- $var(\mathbf{y}) = \|\mathbf{y} - \bar{y}\mathbf{1}_n\|^2 = \|\mathbf{x}\|^2$;
- $cov(\mathbf{y}^1, \mathbf{y}^2) = < \mathbf{y}^1 - \bar{y}^1\mathbf{1}_n, \mathbf{y}^2 - \bar{y}^2\mathbf{1}_n > = < \mathbf{x}^1, \mathbf{x}^2 >$;
- $r(\mathbf{y}^1, \mathbf{y}^2) = \dfrac{< \mathbf{y}^1 - \bar{y^1}\mathbf{1}_n, \mathbf{y}^2 - \bar{y^2}\mathbf{1}_n >}{\|\mathbf{y}^1 - \bar{y^1}\mathbf{1}_n\|\|\mathbf{y}^2 - \bar{y^2}\mathbf{1}_n\|} = \dfrac{< \mathbf{x}^1, \mathbf{x}^2 >}{\|\mathbf{x}^1\|\|\mathbf{x}^2\|} = cov(\mathbf{x}^1, \mathbf{x}^2)$

8.2.6 Expressions matricielles

Soit p variables quantitatives $\{\mathbf{y}^j, j = 1, \ldots, p\}$ où $\mathbf{y}^j$ est le vecteur colonne de $\mathbb{R}^n$ d'éléments y_i^j. Enfin Y et X sont les matrices $n \times p$ de colonnes $\mathbf{y}^j$ et $\mathbf{x}^j$ respectivement.

• *Définition :* La matrice $n \times p$ contenant les variables $\mathbf{y}^j$ en colonne est appelée **tableau de données** ; la matrice X, contenant en colonne les variables centrées, est le *tableau centré*.

• *Définition :* La matrice Γ définie par $\Gamma_{jk} = cov(\mathbf{y}^j, \mathbf{y}^k)$ est appelée *matrice de variance-covariance empirique* des variables $\mathbf{y}^1, \ldots, \mathbf{y}^p$.

On a comme expressions matricielles :

- $\bar{y^j} = \frac{1}{n}(\mathbf{y}^j)'\mathbf{1}_n = \frac{1}{n}\mathbf{1}'_n\mathbf{y}^j$;
- $cov(\mathbf{y}^j, \mathbf{y}^k) = \frac{1}{n}(\mathbf{y}^j - \bar{y^j}\mathbf{1}_n)'(\mathbf{y}^k - \bar{y^k}\mathbf{1}_n) = \frac{1}{n}(\mathbf{x}^j)'\mathbf{x}^k$
- $var(\mathbf{y}^j) = \frac{1}{n}(\mathbf{y}^j - \bar{y^j}\mathbf{1}_n)'(\mathbf{y}^j - \bar{y^j}\mathbf{1}_n) = \frac{1}{n}(\mathbf{x}^j)'\mathbf{x}^j$
- $\Gamma = \frac{1}{n}X'X$
- Soit $a = (a_j)_{j\in J}$ et $b = (b_j)_{j\in J}$ deux vecteurs de $\mathbb{R}^p$. Alors $Ya = \sum_{j\in J} a_j\mathbf{y}^j$, $Yb = \sum_{k\in J} b_k\mathbf{y}^k$ et $cov(\sum_{j\in J} a_j\mathbf{y}^j, \sum_{k\in J} b_k\mathbf{y}^k) = a'\Gamma b$

• *Propriété :* La matrice Γ est *symétrique* ($\Gamma' = \Gamma$) et *positive* (pour tout vecteur u de $\mathbb{R}^p$, $u'\Gamma u \geqslant 0$).

• *Définition :* La matrice de corrélation empirique d'un ensemble de p variables $\mathbf{y}^j$ $(j = 1, \ldots, p)$ est une matrice carrée d'ordre p dont l'élément (i, j) est la corrélation $r(\mathbf{y}^i, \mathbf{y}^j)$.

On a des propriétés analogues à celle de la matrice de covariance. En particulier si X désigne la matrice ayant en colonnes les variables centrées et résuites, alors la matrice des corrélations est

$R = X'DX$. De plus, cette matrice est un résumé de l'ensemble des liaisons entre les variables deux à deux. On peut comparer la matrice de corrélation à la matrice des diagrammmes de dispersion.

8.3 Rappels sur quelques lois de probabilité

8.3.1 La distribution Normale $N(\mu, \sigma^2)$

• *Définition :*
Une v.a. V est distribuée normalement de moyenne μ et de variance σ^2, notée $N(\mu, \sigma^2)$, si sa densité est définie par :

$$f(v) = \frac{1}{\sigma\sqrt{2\pi}} exp \frac{-(v-\mu)^2}{2\sigma^2} \; ; \; v \in \mathbb{R}$$

Si a et b sont deux scalaires, $aV + b$ est distribuée selon une loi Normale $N(a\mu + b, a^2\sigma^2)$.

• *Définition :*
Une v.a. Z est distribuée selon une loi Normale centrée réduite si elle est définie comme :

$$Z = \frac{V - \mu}{\sigma} \sim N(0, 1)$$

et sa densité est :

$$f(z) = \frac{1}{\sqrt{2\pi}} exp(-\frac{z^2}{2})$$

• *Propriétés :*
$E(Z^3) = 0$ et $E(Z^4) = 3$.
De plus, un avantage est que cette loi est tabulée.
Les variables normales possèdent la propriété d'additivité :

- La somme de deux variables X_1 et X_2 indépendantes suivant des lois normales $N(m_1, \sigma_1)$ et $N(m_2, \sigma_2)$ respectivement, est une variable $N(m_1 + m_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.
- Cependant, toute combinaison linéaire de p variables normales non indépendantes est normale à condition que le vecteur des p variables normales suive une loi normale à p dimensions.

8.3.2 La distribution n-Normale $N_n(\mu, \Gamma)$

• *Définition :*
Soit $V = (V_1, V_2, ..., V_n)$ un n-uplet, V est distribuée selon une loi n-Normale $N_n(\mu, \Gamma)$ où $\mu \in \mathbb{R}^n$ et Γ est une matrice (n, n) définie positive, si sa densité f est définie par

$$f(v) = \frac{1}{(2\pi)^{n/2} det(\Gamma)} exp(-\frac{(v-\mu)'\Gamma^{-1}(v-\mu)}{2})$$

• *Propriétés :*

- Γ est la matrice de variance-covariance de V de dimension (n, n) :

  $$\text{Si } V = (V_1, V_2, ..., V_n) \text{ alors } \Gamma_{jk} = cov(V_j, V_k).$$

- Si $\Gamma = diag(\lambda_1, \lambda_2, ..., \lambda_n)$, $V_1, V_2, ..., V_n$ sont indépendantes.
- Si $V_1 \sim N(\mu_1, \sigma_1^2)$, $V_2 \sim N(\mu_2, \sigma_2^2)$, et V_1 et V_2 indépendantes alors

  $$\begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Gamma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right)$$

  Cette propriété reste vraie en n dimensions : Si $V \sim N_n(\mu, \sigma^2 I_n)$, dans toute base orthonormée de $\mathbb{R}^n$, les composantes de V V_1, V_2, ..., V_n sont indépendantes, gaussiennes, de variance σ^2 et $E(V_i)$ est la i-ème composante de μ dans cette base.

– Si V_1 et V_2 ne sont pas indépendantes, $\begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$ ne suit pas en général une loi Binormale.

– Si V est $N_n(\mu, \Gamma)$ et si A est une matrice (m, n) de rang m, alors

$$AV \sim N_m(A\mu, A\Gamma A')$$

– La distribution marginale de chaque composante du vecteur V est une loi Normale.

8.3.3 La distribution de χ^2

• *Définition :*
La distribution du χ^2 est la distribution de la somme des carrés de m v.a. gaussiennes centrées, réduites et indépendantes :

$$\chi^2 = \sum_{i=1}^{m} V_i^2 \sim \chi^2_m \qquad \text{avec } V_i \sim N(0,1) \text{ et indépendantes.}$$

$E(V_i^4) = 3$ et $Var(V_i) = E(V_i^2) = 1$ donc $E(\chi^2) = m$ et $Var(\chi^2) = 2m$.
m est le degré de liberté.
La densité de la distribution de χ^2_m est f définie par :

$$f(v) = \begin{cases} \dfrac{v^{\frac{m}{2}-1} exp(-\frac{v}{2})}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} & si\ v \geqslant 0 \\ 0 & si\ v < 0 \end{cases}$$

• *Propriétés :*

– Si deux variables de χ^2 de degrés m_1 et m_2 sont indépendantes, leur somme est un χ^2 de degré de liberté $m_1 + m_2$.
– Pour n grand, on peut approcher la distribution χ^2_m par la distribution $N(m, 2m)$.
– Si les variables aléatoires V_i ne sont pas indépendantes, mais vérifient k relations linéaires, le nombre de degrés de liberté diminue de k.
– On rencontre quelquefois une distribution de χ^2 *décentrée* qui est la distribution de la somme des carrés de m v.a. gaussiennes $V_1, V_2, ..., V_m$ indépendantes avec $E(V_i) = \mu_i$ et $Var(V_i) = 1$; son degré de liberté est m ; son paramètre de décentrage est $\lambda = \sum_{i=1}^{m} \mu_i^2$. On montre que $E(\chi^2_{m,\lambda}) = m + \lambda$ et $Var(\chi^2_{m,\lambda}) = 2m + 4\lambda$.

8.3.4 La distribution de Student

• *Définition :*
La distribution de Student est la distribution de la v.a. T définie par :

$$T = \frac{V_1}{\sqrt{V_2/m}} \sim Student(m)$$

$$\text{avec } V_1 \sim N(0,1),\ V_2 \sim \chi^2_m, \text{ et } V_1 \text{ et } V_2 \text{ indépendantes.}$$

Son degré de liberté est m. Sa densité est symétrique par rapport à 0 et est définie par :

$$f(v) = \frac{1}{(1 + \frac{v^2}{m})^{\frac{m+1}{2}} \beta(\frac{1}{2}, \frac{m}{2})\sqrt{m}}$$

La moyenne est 0 et la variance $\dfrac{m}{m-2}$ pour $m \geqslant 3$.

Pour $m = 5$, $P[-2.57 \leqslant T \leqslant 2.57] = 0.95$ et $P[-4.03 \leqslant T \leqslant 4.03] = 0.99$.
Pour $m = 10$, $P[-2.23 \leqslant T \leqslant 2.23] = 0.95$ et $P[-3.17 \leqslant T \leqslant 3.17] = 0.99$.
Pour m grand, on peut approcher la distribution de Student par la distribution $N(0,1)$.

8.3.5 La distribution de Fisher-Snédécor

• *Définition :*
La distribution de Fisher-Snédécor est la distribution de la v.a. F définie par :

$$F = \frac{V_1}{V_2} \times \frac{m_2}{m_1} \sim F(m_1, m_2)$$

$$\text{où } V_1 \sim \chi^2_{m_1},\ V_2 \sim \chi^2_{m_2},\ V_1 \text{ et } V_2 \text{ indépendantes.}$$

m_1 est le degré de liberté attaché à V_1 (*degré de liberté du numérateur*) et m_2 est le degré de liberté attaché à V_2 (*degré de liberté du dénominateur*). Sa moyenne est $\frac{m_2}{m_2 - 2}$. Sa densité est nulle sur $\mathbb{R}^2$.

• *Propriétés :*
- $V \sim F(m_1, m_2)$ alors $\frac{1}{V} \sim F(m_2, m_1)$.
- Si $m_1 = 1$ alors V_1 est le carré d'une v.a. $N(0,1)$:

$$F = \frac{(N(0,1))^2}{V_2/m_2} = \left(\frac{N(0,1)}{\sqrt{V_2/m_2}} \right)^2$$

  F est donc le carré d'une v.a. de Student de degré de liberté m_2 : $F = T^2$ où $T \sim Student(m_2)$.

8.4 Rappels de statistique inférentielle

Soit X une variable aléatoire de loi P_θ où θ désigne un paramètre inconnu à estimer. L'objectif de la statistique inférentielle est de produire une inférence sur θ (estimation ou test) sur la base d'un échantillon de n observations $(x_1, ..., x_i, ..., x_n)$ où x_i est la réalisation de X_i ; les X_i étant supposés en général *i.i.d.* (c'est à dire indépendants et identiquement distribués) de loi P_θ. Dans cette section, θ désigne un paramètre réel, à valeurs dans $\Theta \subseteq \mathbb{R}$.

8.4.1 Estimation ponctuelle, estimation par intervalle de confiance

Estimateur et estimation

On appelle estimateur de θ toute fonction des X_i à valeurs dans Θ. Un estimateur de θ est souvent noté $\widehat{\theta_n}$. La première qualité d'un estimateur est d'être convergent : ce qui signifie que $\widehat{\theta_n}$ converge (en probabilité) vers θ quand $n \to \infty$. Il est également souhaitable d'utiliser des estimateurs sans biais, c'est à dire tels que $E(\widehat{\theta_n}) = \theta$. Si $\widehat{\theta_n}$ est noté $T_n(X_1, \ldots, X_n)$ alors $T_n(x_1, \ldots, x_n)$ s'appelle une estimation ponctuelle de θ. Il est important de réaliser qu'une estimation de θ est une grandeur numérique alors qu'un estimateur de θ est une variable aléatoire.

Définition d'un intervalle de confiance

On appelle intervalle de confiance d'un paramètre θ associé à un n-échantillon $(X_1, X_2, ..., X_n)$, un intervalle $I = [A, B]$ dont les bornes A et B sont des fonctions des X_i, et tel que $P(I \ni \theta) = 1 - \alpha$; $1 - \alpha$ s'appelle le niveau de confiance de l'intervalle de confiance. Il est important de noter que I est un intervalle aléatoire au sens où les bornes A et B sont aléatoires.

8.4.2 Notions générales sur la théorie des tests paramétriques

La théorie des tests paramétriques consiste à formuler des hypothèses particulières sur le paramètre θ de la loi P_θ ; puis à apporter un jugement sur ces hypothèses (plus particulièrement, à trancher entre deux hypothèses). Ce jugement est basé, d'une part, sur les résultats obtenus sur un ou plusieurs échantillons extraits de la population étudiée et d'autre part, sur l'acceptation d'un

certain risque dans la prise de décision. A titre indicatif, les tests peuvent être classés en différentes catégories :

- test sur une hypothèse relative à la valeur particulière d'un paramètre,
- test de conformité de deux distributions ou test d'ajustement entre une distribution théorique et une distribution expérimentale,
- test de comparaison de deux populations,
- test d'indépendance de deux caractères dans un tableau de contingence.

Formulation des hypothèses

On veut tester une hypothèse, que l'on appelera hypothèse nulle notée H_0 à savoir :

$$H_0 : \theta = \theta_0$$

contre une hypothèse alternative notée H_1. Cette hypothèse H_1 peut se formuler de différentes façons :

$$\theta \neq \theta_0 \text{ ou } \theta > \theta_0 \text{ ou } \theta < \theta_0 \text{ ou } \theta = \theta_1.$$

La décision (cad : choisir soit H_0, soit H_1) se faisant sur la base des observations $x_1, \ldots, x_n$.

Risques et probabilités d'erreur

Pour des événements dans lesquels le hasard intervient, toute décision prise comporte un certain risque que cette décision soit erronée. On peut par exemple accepter un risque égal à 5% de rejeter l'hypothèse H_0 alors qu'elle est vraie ; c'est aussi le risque d'accepter à tort l'hypothèse H_1. Ce risque, noté α, est le risque de rejeter à tort l'hypothèse H_0 alors qu'elle est vraie. On l'appelle risque de première espèce :

$$\alpha = P(rejeter\ H_0\ /\ H_0\ vraie) = P(choisir\ H_1/H_0\ vraie)$$

On appelle région critique qu'on note en général W (ou parfois R_c), l'ensemble des valeurs de la v.a. de décision (appelée statistique de test) qui conduisent à écarter H_0 au profit de H_1. La région complémentaire $\overline{W}$ représente la région d'acceptation de H_0. La règle de décision peut se formuler ainsi : si la valeur de la statistique de test considérée appartient à la région d'acceptation $\overline{W}$, on choisit H_0 ; si elle appartient à la région critique W, on choisit H_1.

Il existe un deuxième risque d'erreur appelé risque de deuxième espèce et noté risque β. C'est le risque de ne pas rejeter H_0 alors que H_1 est vraie :

$$\beta = P(\text{ne pas rejeter } H_0 \text{ / } H_1 \text{ vraie}) = P(\text{choisir } H_0 \text{ / } H_1 \text{ vraie})$$

En introduisant la région critique, on peut aussi écrire :

$$P(\overline{W}/H_0) = 1 - \alpha \quad P(W/H_1) = 1 - \beta$$

La quantité $1 - \beta$ s'appelle la puissance du test. Elle représente la probabilité de ne pas rejeter H_0 alors que H_1 est vraie. Ces différentes situations sont résumées dans le tableau suivant :

Vérité / Décision	H_0	H_1
$H_0(\overline{W})$	$1-\alpha$	β
$H_1(W)$	α	$1-\beta$

Comme indiqué précédemment, l'hypothèse alternative H_1 peut se formuler de différentes façons. On peut visualiser ces différentes hypothèses H_1 et montrer ainsi la forme de la région critique :

Test unilatéral à droite	Test unilatéral à gauche
$H_0 : \theta = \theta_0$	$H_0 : \theta = \theta_0$
$H_1 : \theta > \theta_0$	$H_0 : \theta < \theta_0$

Test bilatéral
$H_0 : \theta = \theta_0$
$H_1 : \theta \neq \theta_0$

Démarche de construction d'un test

Pour élaborer un test statistique portant sur θ, il faut suivre la démarche suivante :

1. Formuler de façon précise l'hypothèse nulle H_0 et l'hypothèse alternative H_1.
2. Fixer, avant l'expérience, le risque de première espèce α, c'est-à-dire le risque de rejeter à tort l'hypothèse nulle alors qu'elle est vraie.
3. Préciser les conditions d'application du test : forme de la loi de probabilité de l'estimateur du paramètre d'intérêt, taille de l'échantillon, variance connue ou inconnue, ...
4. Choisir une statistique de test, c'est-à-dire une fonction de $(X_1, ...X_n)$ égale à T_n ou intimement liée à T_n, et donner sa loi de probabilité sous les hypothèses nulle et alternative.
5. Déterminer la région critique ou région de rejet de l'hypothèse nulle H_0 compte tenu de H_1 et en déduire la règle de décision.
   W : région critique conduisant au rejet de H_0 : $P(W/H_0) = \alpha$
   $\overline{W}$: région de non-rejet (ou d'acceptation) de H_0 : $P(\overline{W}/H_0) = 1 - \alpha$
6. Calculer la valeur numérique de la statistique de test en utilisant les données de l'échantillon.
7. Donner les conclusions du test :
   – Si cette valeur appartient à W, on rejette H_0 au profit de H_1 ;
   – Si cette valeur appartient à $\overline{W}$, on ne peut pas rejeter H_0.