Université Claude Bernard Lyon 1
Année universitaire 2011-2012

UE MASS 42

Statistiques

Frédérique BIENVENÜE, frederique.bienvenue@univ-lyon1.fr
http ://math.univ-lyon1.fr/∼duheille/MASS42.html
Polycopié largement inspiré d'une version précédente de Gabriela CIUPERCA

Table des matières

Chapitre 1

INTRODUCTION

Avant tout, quelques notations :
- les variables aléatoires seront notées par des lettres en majuscules (par exemple X), les réalisations de ces variables seront notées en lettres minuscules (par exemple x).
- la loi Normale $\mathcal{N}(m, \sigma^2)$, de densité $t \to e^{-(x-m)^2/(2\sigma^2)}/\sqrt{2\pi\sigma^2}$. Les paramètres $m \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}^{+*}$ représentent respectivement l'espérance et la variance de cette loi.

Soit X une variable aléatoire de function de répartition $F_\theta(x)$ qui dépend du paramètre θ. On considère n variables aléatoires X_1,...,X_n indépendantes et de même fonction de répartition (i.i.d) $F_\theta(x)$.

1 La notion de Modèle

On va essayer de voir d'abord l'importance de la notion de modèle sur laquelle repose toute la statistique inférentielle. Pour ça, commençons par deux exemples.

EXEMPLE 1 : Étude de la taille moyenne d'une population
On dispose de n données $x_1, ...x_n$ qui sont les tailles de n individus d'une population. *Est-ce que les x_i sont des réalisations de variables aléatoires ?* A priori non. A ce stade, il est donc seulement possible de faire de la statistique descriptive (c'est-à-dire, calculer la moyenne, la variance, ... des données observées).
Pour pouvoir faire de la statistique inférentielle, nous somme obligés de faire des hypothèses assez sévères. Par exemple, on peut faire l'hypothèse classique que la taille d'un individu est distribuée Normalement dans une population.
Soit la v.a. X la taille d'un individu. On suppose que $X \sim \mathcal{N}(m, \sigma^2)$ avec σ^2 connu. Dans ce contexte il est possible d'associer aux données $x_1, ..., x_n$ les variables aléatoires $X_1, ..., X_n$ qui sont de même loi et indépendantes :

$$X_i \sim \mathcal{N}(m, \sigma^2) \qquad i = 1, ..., n$$

Chaque x_i est une réalisation de X_i, et plus précisément, on suppose qu'il existe $\omega \in \Omega$ tel que, pour tout i, on a $x_i = X_i(\omega)$.

Disposant maintenant d'un modèle probabiliste sur la taille des individus dans la population, nous allons pouvoir faire de la statistique inférentielle, c'est-à-dire estimer les paramètres, faire des tests... Par exemple :
Quel est un estimateur (ponctuel) du paramètre inconnu m ? Autrement dit, comment estimer

l'espérance m de la loi de X à partir de l'échantillon $(x_i)_{i\leq n}$? On pense bien évidemment à

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

On montre facilement que $\mathbf{E}(\overline{X_n}) = m$ donc, $\overline{X_n}$ est sans biais.
Quelle est la loi de $\overline{X_n}$? $\overline{X_n}$ en tant que combinaison linéaire de v.a. gaussiennes indépendantes, est une variable gaussienne. Par conséquent, puisque $\mathbf{E}(\overline{X_n}) = m$ et $\text{var}\,(X) = \frac{\sigma^2}{n}$ on a :

$$\overline{X_n} \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$$

Pour trouver un estimateur par intervalle de m, au risque $\alpha = 0.05$ signifie de donner un intervalle centré sur $\overline{X_n}$ qui encadre les valeurs possibles de m avec une possibilité de laisser en dehors des valeurs 0.05. Plus précisément, on cherche le réel $a > 0$, tel que

$$\mathbf{P}\left[m \in \left[\overline{X_n} - a, \overline{X_n} + a\right]\right] = 0.95$$

Quand on dispose d'une réalisation de $\overline{X_n}$ ici, $\overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$, on peut remplacer $\overline{X_n}$ par cette réalisation. On obtient alors un intervalle parfaitement déterminé, dit *Intervalle de Confiance* pour le paramètre inconnu m.

En conclusion, que retenir de ce exemple ?
En faisant l'hypothèse sur la distribution Normale de la taille d'un individu dans la population, on a pu construire des estimateurs du paramètre m : un estimateur ponctuel et un estimateur par intervalle. Sans cette hypothèse, nous n'aurions pu faire que de la statistique descriptive. Sans modèle, on peut toujours prendre $\overline{x_n}$ comme estimation ponctuelle de m, mais on sera incapable d'en évaluer la précision. Il n'y a pas d'intervalle de confiance possible.

EXEMPLE 2 : Étude d'un produit polluant
On considère un échantillon de n données $x_1, ..., x_n$ représentant la quantité d'un produit polluant dans un échantillon d'eau tiré de la Méditerranée.
Est-ce que les x_i *sont des réalisations des variables aléatoires* ?
Non. Il est cette fois impossible d'obtenir la distribution continue du produit polluant dans l'eau de la Méditerranée toute entière. Et il paraît d'autre part difficile d'imaginer un modèle. Donc, sur cet exemple, on ne pourra faire que de la statistique descriptive. On se gardera d'extrapoler les résultats obtenus sur l'échantillon à toute la population.
On ne peut pas faire des hypothèses probabilistes sur n'importe quelles données. Il faut tenir compte toujours du contexte pratique du problème.

Conclusion des deux exemples
Il faut retenir l'idée fondamentale de la statistique inférentielle : obtenir des informations sur une population à partir d'observations partielles de cette population. Il faut de plus retenir que cette approche n'est possible qu'à partir du moment où existe un modèle probabiliste sur cette population.
Dans toute étude statistique il faut savoir faire la part des choses entre les situations où existe un modèle naturel (par exemple un modèle de tirage de boule), les situations où l'on peut faire l'hypothèse d'un modèle (exemple sur la taille) et les situations où il est impossible de supposer l'existence d'un modèle (dernier exemple). Dans les deux premiers types de situations, on

pourra faire de la statistique inférentielle à partir des observations et dans la dernière, on ne fera que de la statistique descriptive.

On va s'occuper du deuxième type de modèle.

2 Notions d'échantillonnage

Soit $\mathcal{E}$ une expérience aléatoire sur l'univers $\{\Omega, \mathbf{P}_\theta\}$. Donc on a défini une v.a. X. Pour faire des tests, estimer θ (θ inconnu), une idée est de considérer une suite d'observations (v.a.) indépendantes et de même loi que X.
Définition. On appelle *n-échantillon d'une loi* $\mathbf{P}_\theta$ toute famille $X_1, ..., X_n$ de v.a. indépendantes et de même loi que X.
Puisque les v.a. X_i ont la même loi que X, elles ont les mêmes moments :

$$\mathbf{E}[X_i] = \mathbf{E}[X], \qquad \text{var}\,(X_i) = \text{var}\,(X), \qquad \mathbf{E}[X_i^k] = \mathbf{E}[X^k]$$

$\forall i = 1, ..., n$, $k \in \mathbb{N}$.

Moments empiriques

On considère le cas des v.a. unidimensionnelles ($\Omega \subseteq \mathbb{R}$). Soit un n-échantillon $X_1, ..., X_n$ (X_i est la v.a. réelle correspondant à la i-ème expérience).
Définition. On appelle *moment empirique d'ordre p* ($p \in \mathbb{N}$) la v.a.

$$U_p^n = \frac{1}{n}\sum_{i=1}^{n} X_i^p$$

et on appelle *moment empirique centré d'ordre p*, la v.a.

$$W_p^n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^p$$

$p = 1 : \overline{X_n} = U_1^n$ la moyenne empirique
$p = 2 : W_2^n = S_n^2$ la variance empirique.
Soient $m_p = \mathbf{E}(X^p)$ et $\mu_p = \mathbf{E}[(X - \mu_1)^p]$ les moments centrés et non-centrés d'ordre p de X (s'ils existent). En utilisant la loi des grands nombres, on obtient le résultat suivant :
Théorème 2.1. *Si* $m_p = \mathbf{E}(X^p) < \infty$, *alors*

$$U_p^n \xrightarrow[n\to\infty]{p.s.} m_p, \qquad W_p^n \xrightarrow[n\to\infty]{p.s.} \mu_p$$

Théorème 2.2. *a) Si* $\mathbf{E}(X^p) < \infty$, *alors* $\mathbf{E}(U_p^n) = \mathbf{E}(X^p) = m_p$. *Cas particulier* $p = 1$, $\mathbf{E}(\overline{X_n}) = m_1 = m$.
b) Si $\mathbf{E}(X^{2p}) < \infty$, *alors* $\text{var}\,(U_p^n) = \{\mathbf{E}(X^{2p}) - [\mathbf{E}(X^p)]^2\}/n$. *Cas particulier :* $p = 1$, $\text{var}\,(\overline{X_n}) = \text{var}\,(X)/n = \sigma^2/n$.
c) Si $\text{var}\,(X) < \infty$ *alors* $\mathbf{E}(W_2^n) = \frac{n-1}{n}\text{var}\,(X)$.

3 Notions d'estimation

Soit une v.a. $X \sim (\Omega, \mathbf{P}_\theta)$, de fonction de répartition F_θ, $\theta \in \Theta \subseteq \mathbb{R}^p$. On suppose que la fonction F_θ est connue, mais pas θ. Soit θ_0 la vraie valeur (inconnue). Le but est de trouver des statistiques (ou fonctions du n-échantillon) qui vont approximer le mieux possible, dans un certain sens, θ_0.

Définition. On appelle *estimateur ponctuel du paramètre* θ_0 (en général on dit θ) toute fonction de l'échantillon, prenant ses valeurs dans Θ : $T_n = T(X_1, ..., X_n)$.
La valeur prise par T pour un n-uplet de données $(x_1, ..., x_n)$ est l'*estimation* de θ : $T(x_1, ..., x_n)$.

Exemple 1. On lance une pièce de monnaie et soit la v.a.

$$X = \begin{cases} 0 & \text{si "face"} \\ 1 & \text{si "pile"} \end{cases}$$

alors $X \sim \mathcal{B}(\theta)$, $\theta = p$. On souhaite estimer θ. On lance la pièce 10 fois : $n = 10$. Les variables $X_1, ..., X_{10}$ suivent la loi $\mathcal{B}(\theta)$. Une réalisation de l'échantillon est : 0, 1, 1, 0, 1, 1, 1, 0, 0, 1. Si on prend $\overline{X_n} \in [0, 1]$ comme estimateur, alors $\overline{X_{10}} = 6/10$. Si on répète 10 fois encore l'expérience : on obtiendra peut-être comme échantillon 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, et on aura alors $\overline{X_{10}} = 5/10$. D'autres estimateurs pour θ : $1/2$, $T = X_1$, $T = (X_1 + X_2)/2$.

Exemple 2. $X_1, ..., X_{10} \sim \mathcal{P}(\lambda)$, λ inconnu.
$\overline{X_n}$ estimateur pour λ, mais aussi $\frac{2}{n(n+1)} \sum_{k=1}^{n} kX_k$.

De ces exemples, c'est clair qu'on doit choisir des estimateurs avec des « bonnes qualités ». Par exemple, pour n grand, on souhaite que $T(X_1, ...X_n)$ tende vers θ_0 dans un certain sens. Les valeurs de deux estimations ne doivent pas être non plus « trop différentes ».

Propriétés des estimateurs

Définition. On dit que l'estimateur $T_n = T(X_1, ..., X_n)$ est *faiblement* (resp. *fortement*) *consistant (ou convergent)* si :

$$T_n \xrightarrow[n\to\infty]{\mathbf{P}} \theta_0 : \qquad \forall \varepsilon > 0, \lim_{n\to\infty} \mathbf{P}[|T_n - \theta_0| \geq \varepsilon] = 0$$

respectivement :

$$T_n \xrightarrow[n\to\infty]{p.s.} \theta_0 : \qquad \mathbf{P}[\lim_{n\to\infty} T_n = \theta_0] = 1$$

Exemple 1. Les moments empiriques sont des estimateurs fortement consistants des moments théoriques. En particulier, $\overline{X_n}$ est estimateur consistant pour $m = \mathbf{E}(X)$.
Exemple 2. $X_i \sim \mathcal{B}(\theta)$, $\overline{X_n}$ est un estimateur fortement consistant pour θ, et

$$\frac{1}{n+2}[2 + \sum_{i=1}^{n} X_i] \xrightarrow[n\to\infty]{p.s.} \theta$$

est un autre estimateur consistant de θ. Donc, les estimateurs consistants ne sont pas uniques.

Définition. On appelle *biais* de l'estimateur T_n, la quantité : $B(T_n, \theta) = \mathbf{E}(T_n) - \theta$.
L'*estimateur* est dit *sans biais* si $B(T_n, \theta) = 0$ et il est dit *asymptotiquement sans biais* si $B(T_n, \theta) \xrightarrow[n\to\infty]{} 0$.

Exemples classiques.

1. U_k^n est un estimateur sans biais pour $\mathbf{E}(X^k) = m_k$, $\overline{X_n}$ pour $m = \mathbf{E}(X)$.
2. W_k^n est un estimateur asymptotiquement sans biais pour μ_k.
3. $S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X_n})^2 = \frac{n}{n-1} W_2^n$ est un estimateur sans biais pour $\mathrm{var}\,(X)$.

4 Introduction à la théorie des tests

Supposons qu'une machine produit des objets dont certains sont défecteux. Soit θ la probabilité que l'objet soit défecteux. Le fabricant désire avoir $\theta \leq \theta_0$ avec θ_0 donné, faute de quoi il doit réviser ou changer la machine. Le paramètre θ est inconnu, et on ne peut que travailler avec un échantillon de la production.

Soit la v.a. X d'une certaine loi de probabilité P.
Une *hypothèse statistique* est un énoncé concernant les caractéristiques (valeurs des paramètres, forme de distribution, ...) d'une ou de plusieurs populations.
Le *test statistique (d'hypothèse)* est une démarche qui a pour but de fournir une règle de décision permettant sur la base des résultats de l'échantillon de faire le choix entre deux hypothèses statistiques.
Les hypothèses qui sont envisagées a priori s'appellent, *l'hypothèse nulle* (H_0) et *l'hypothèse alternative* (H_1).
Pour réaliser des tests, on considère un n-échantillon $(X_1, ..., X_n)$ et une réalisation $(x_1, ..., x_n)$.
Pour fournir une règle de décision, on utilise une statistique de test ou fonction de test, i.e. une fonction $\varphi : \mathbb{R}^n \to [0, 1]$ *(suffisamment régulière).*
La fonction φ est un test de l'hypothèse H_0 contre H_1 avec l'erreur de probabilité α (ou : de niveau $1 - \alpha$) si : $\mathbf{E}[\varphi(X_1, ..., X_n)] \leq \alpha$ sous H_0, c'est-à-dire lorsque les (X_i) vérifient H_0.

On considère que la loi de la v.a. X dépend d'un paramètre θ et on veut faire un test sur ce paramètre : on a affaire à des tests paramétriques. On teste :
$H_0 : \theta \in \Theta_0$, appelée *hypothèse nulle* (parce qu'elle s'écrit sous la forme $g(\theta) = 0$)
contre :
$H_1 : \theta \in \Theta_1$ *l'hypothèse alternative*
avec $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 \subseteq \Theta$.

Si Θ_0 est formée d'un seul élément on dit que H_0 est une *hypothèse simple*, sinon elle est *composite.*
Pour faire le test on a besoin d'une régle de décision : soit $T_n = T(X_1, ..., X_n)$ une *statistique de test* et $\mathcal{R}$ un sous-ensemble de valeurs possibles de T, appelée *région de rejet.* Si $T(x_1, ..., x_n) \in \mathcal{R}$ on rejette H_0 et on accepte H_1. La construction de $\mathcal{R}$ est basée sur la connaissance de la loi de T_n sous H_0.

Définitions : **1)** On appelle *risque de première espèce* et on note $\alpha(\theta)$, la probabilité de rejeter H_0 alors qu'elle est vraie :

$$\alpha(\theta) = \mathbf{P}[T \in \mathcal{R} |\ \theta \in \Theta_0]$$

On appelle *niveau*, noté α, la valeur la plus élevée du risque de première espèce quand θ parcourt Θ_0 :

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta)$$

Si H_0 est l'hypothèse $\theta = \theta_0$ alors $\alpha = \alpha(\theta_0)$.
2) On appelle *risque de deuxième espèce*, noté $\beta(\theta)$, la probabilité d'accepter H_0 alors qu'elle est fausse.
3) On appelle *puissance*, noté $\pi(\theta)$, la probabilité de rejeter H_0 alors qu'elle est fausse. On a $\pi(\theta) = 1 - \beta(\theta)$.
4) *Région de rejet* : $\mathcal{R} = \{(x_1, ..., x_n);\ \ H_0 \text{ rejetée}\}$ telle que

$$\alpha(\theta) = \mathbf{P}[(X_1, ..., X_n) \in \mathcal{R} | H_0 \text{ vraie}].$$

Donc $\mathcal{R}$ dépend de α.

Résumons la démarche à suivre pour effectuer un test d'hypothèse :
1. Choisir H_0 et H_1 de sorte que la possibilité d'égalité soit dans H_0 ;
2. Fixer α ;
3. Déterminer la région de rejet $\mathcal{R}$;
4. Regarder si les observations se trouvent ou pas dans $\mathcal{R}$;
5. Conclure au rejet ou au non rejet de H_0.

Exemple : Tests sur la moyenne d'une loi Normale

On considère un n-échantillon $X_1, ..., X_n$ avec $X_i \sim \mathcal{N}(m, \sigma^2)$

Cas σ^2 connue

Notons par u_α le *fractile (quantile)* d'ordre α pour la loi Normale : si $Y \sim \mathcal{N}(0,1)$ alors

$$\mathbf{P}\left[Y < u_\alpha\right] = \alpha$$

1) $H_0 : m = m_0$ contre $m \neq m_0$

Statistique de test :

$$Z = \sqrt{n}\frac{\overline{X_n} - m_0}{\sigma} \sim \mathcal{N}(0,1) \qquad \text{sous } H_0 \tag{1.1}$$

Zone de rejet

$$\mathcal{R} = \left\{(x_1, ..., x_n) \ / \ \sqrt{n}\frac{|\overline{X_n} - m_0|}{\sigma} > u_{1-\alpha/2}\right\}$$

2) $H_0 : m \leq m_0$ contre $m > m_0$ ou $H_0 : m = m_0$ contre $m > m_0$

Statistique de test : (1.1). Zone de rejet :

$$\mathcal{R} = \left\{(x_1, ..., x_n) \ / \ \sqrt{n}\frac{\overline{X_n} - m_0}{\sigma} > u_{1-\alpha}\right\} = \left\{(x_1, ..., x_n) \ / \ \overline{X_n} > m_0 + \frac{\sigma}{\sqrt{n}}u_{1-\alpha}\right\}$$

3) $H_0 : m \geq m_0$ contre $m < m_0$ ou $H_0 : m = m_0$ contre $m < m_0$

Statistique de test : (1.1). Zone de rejet :

$$\mathcal{R} = \left\{(x_1, ..., x_n) \ / \ \sqrt{n}\frac{\overline{X_n} - m_0}{\sigma} < u_{\alpha}\right\} = \left\{(x_1, ..., x_n) \ / \ \overline{X_n} < m_0 + \frac{\sigma}{\sqrt{n}}u_{\alpha}\right\}$$

Cas σ^2 inconnue

Notons par $t_{p,\alpha}$ le *fractile (quantile)* d'ordre α pour la loi Student à p degrés de liberté : si $Y \sim t(p)$ alors

$$\mathbf{P}\left[Y < t_{p,\alpha}\right] = \alpha$$

On remplace σ^2 par son estimateur sans biais

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X_n})^2\right)$$

$\underline{H_0 : m = m_0 \text{ contre } m \neq m_0}$
Statistique de test :

$$Z = \sqrt{n}\frac{\overline{X_n} - m_0}{S^*} \sim t(n-1) \qquad \text{sous } H_0 \tag{1.2}$$

Zone de rejet :

$$\mathcal{R} = \left\{(x_1, ..., x_n) \;/\; \sqrt{n}\frac{|\overline{X_n} - m_0|}{s^*} > t_{n-1,1-\alpha/2}\right\}$$

2) $\underline{H_0 : m \leq m_0 \text{ contre } m > m_0}$ ou $\underline{H_0 : m = m_0 \text{ contre } m > m_0}$

Statistique de test : (1.2). Zone de rejet :

$$\mathcal{R} = \left\{(x_1, ..., x_n) \;/\; \sqrt{n}\frac{\overline{X_n} - m_0}{s^*} > t_{n-1,1-\alpha}\right\} = \left\{(x_1, ..., x_n) \;/\; \overline{X_n} > m_0 + \frac{s^*}{\sqrt{n}}t_{n-1,1-\alpha}\right\}$$

3) $\underline{H_0 : m \geq m_0 \text{ contre } m < m_0}$ ou $\underline{H_0 : m = m_0 \text{ contre } m < m_0}$

Statistique de test : (1.2). Zone de rejet :

$$\mathcal{R} = \left\{(x_1, ..., x_n) \;/\; \sqrt{n}\frac{\overline{X_n} - m_0}{s^*} < t_{n-1,\alpha}\right\} = \left\{(x_1, ..., x_n) \;/\; \overline{X_n} < m_0 + \frac{s^*}{\sqrt{n}}t_{n-1,\alpha}\right\}$$

5 Généralités sur le Modèle Linéaire

Le but de ce cours est de donner quelques notions élémentaires sur le Modèle Linéaire.

Donnons d'abord la forme générale d'un modèle de régression. Soient $Y, X_1, ..., X_p$ des variables. Dans de nombreux problèmes pratiques, on étudie la relation qui peut exister entre Y et $X_1, ..., X_p$:

$$Y = f(X_1, ..., X_p) \tag{1.3}$$

Mais, assez souvent on met en doute le caractère purement déterministe de cette relation

- soit parce qu'il y a des erreurs de mesure
- soit à cause de l'omission volontaire ou non d'éventuelles variables (ce qui est le plus fréquent)

On ajoute un terme d'erreur et on obtient le modèle statistique

$$Y = f(X_1, ..., X_p) + \varepsilon \tag{1.4}$$

Y est la *variable expliquée, dépendante*, $X_1, ..., X_p$, les *variables explicatives, indépendantes.*
Définition. Le modèle (1.4) est dit de *régression linéaire* si la fonction f est une fonction linéaire de $X_1, ..., X_p$

$$f(X_1, ..., X_p) = a_0 + a_1 X_1 + ...a_p X_p \tag{1.5}$$

En ce qui concerne les variables et les paramètres :

	Aléatoire	**Non aléatoire**
Observable	Y	$X_1, ..., X_p$
Non observable	ε	$a_0, a_1, ..., a_p$

$a_0, ..., a_p$ paramètres inconnus à estimer. Pour estimer ces paramètres on dispose de n observations des variables $Y, X_1, ..., X_p$, notées

variable	Y	X_1	X_2	...	X_p
Observation i	y_i	x_{1i}	x_{2i}	...	x_{pi}

Alors, le modèle de régression linéaire s'écrire :

$$Y_i = a_0 + a_1 X_{1i} + \cdots + a_p X_{pi} + \varepsilon_i \qquad i = 1, ..., n \tag{1.6}$$

où :
- Y_i est une v.a., de réalisation y_i,
- X_{1i} une variable (non aléatoire) avec l'observation x_{1i}.

L'étude statistique du modèle linéaire permet
- d'estimer les paramètres $a_0, ..., a_p$ par moindres carrés ;
- de tester l'influence de certaines variables X_j (par test d'hypothèse) ;
- d'en déduire le meilleur modèle (par l'étude des résidus).

Notons que les erreurs $\varepsilon_1, ..., \varepsilon_n$ sont des v.a. indépendantes, donc $Y_1, ..., Y_n$ aussi.

On suppose que les v.a. ε_i suivent une loi Normale :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \qquad i = 1, ..., n$$

Le cas le plus simple de régression linéaire est pour $p = 1$. Le modèle s'écrit alors :

$$Y_i = a_0 + a_1 X_i + \varepsilon_i \qquad i = 1, ..., n \tag{1.7}$$

modèle appelé *régression linéaire simple.*

Exemple de régression simple
Pour une ville on mesure la pollution en ozone et la vitesse maximale du vent (m/s) pendant 10 jours. Écrire un modèle statistique de la pollution en fonction du vent :
- Y : la concentration de l'ozone (en mg/m^3) ;
- X : vitesse (en m/s).

Obs	1	2	3	4	5	6	7	8	9	10
Y	174	188	176	128	116	88	58	120	92	132
X	1	0.5	1	2	2	2.5	3	2	3	2
$\hat{y}_i$	171	195	171	122	122	98	74	122	74	122
e_i	3	-7	5	6	-6	-10	-16	-2	18	10
er_i	0.28	-0.65	0.46	0.55	-0.55	-0.92	-1.47	-0.18	1.66	0.92

Le calcul des estimations des paramètres a_0 et a_1 donne $\widehat{a_0} = 219.5$ et $\widehat{a_1} = -48.5$. On en déduit les estimations données dans les trois dernières lignes du tableau ci-dessus :

- $\hat{y}_i$: valeur de Y_i prédite par le modèle pour la valeur x_i de X ; $\hat{y}_i = a_0 + a_1 x_i$.
- e_i : $y_i - \hat{y}_i$: erreur du modèle ou résidu
- $\hat{\sigma} = 10.84$: variance sans biais des résidus
- er_i : erreur réduite, $er_i = e_i/\hat{\sigma}$.

On peut également :

- tester si vraiment il y a un lien linéaire entre la pollution d'ozone et le vent (c'est-à-dire que le modèle linéaire est bon)
- prévoir la concentration d'ozone, pour un nouveau jour, si on connaît la vitesse maximale du vent (Si par exemple, on prévoit la vitesse du vent par une autre méthode la veille, on peut prévoir pour le lendemain la pollution).

Chapitre 2

RÉGRESSION LINÉAIRE

1 Couples aléatoires

1.1 Lois

Pour ce cours, nous aurons besoin de manipuler les couples et des vecteurs formés de variables aléatoires, le plus souvent à densité.

Définition 1.1. *Soit (U, V) un couple de variables aléatoires. On dit que le couple (U, V) admet une **densité** sur $\mathbb{R}^2$ lorsqu'il existe une fonction positive $f : \mathbb{R}^2 \to \mathbb{R}^+$ telle que*

- *$\int_{\mathbb{R}} \int_{\mathbb{R}} f(u, v)\ du\ dv = 1$*
- *Pour tous intervalles I et J de $\mathbb{R}$, on a*

$$\mathbf{P}(U \in I \text{ et } V \in J) = \int_I \left(\int_J (f(u, v)\ dv)\ \ du \right)$$

On peut déduire de cette définition la proposition suivante

Proposition 1.1. *Soit (U, V) un couple aléatoire de densité $f : \mathbb{R}^2 \to \mathbb{R}^+$. Pour toute fonction $h : \mathbb{R}^2 \to \mathbb{R}$ positive ou bornée*

$$\mathbf{E}(h(U, V)) = \int_{\mathbb{R}} \int_{\mathbb{R}} h(u, v) f(u, v)\ du\ dv.$$

Les variables aléatoires U et V sont à densité et, en notant g_U et g_V leurs densités respectives, on a pour tout réel x,

$$g_U(x) = \int_{v \in \mathbb{R}} f(x, v)\ dv \qquad \text{et} \qquad g_V(x) = \int_{u \in \mathbb{R}} f(u, x)\ du.$$

On dit que la densité $f : \mathbb{R}^2 \to \mathbb{R}^+$ est la **densité conjointe** du couple (U, V) et que les fonctions g_U et g_V sont les **densités marginales** de (U, V).

1.2 Covariance

Définition 1.2. *La covariance d'un couple (U, V) de variables aléatoires de carré intégrable est définie par*

$$\mathrm{cov}(U, V) = \mathbf{E}\left[(U - \mathbf{E}(U))(V - \mathbf{E}(V)) \right].$$

Le coefficient de corrélation est le réel $\mathrm{Corr}(U, V)$ défini par

$$\mathrm{Corr}(U, V) = \frac{\mathrm{cov}(U, V)}{\sqrt{\mathrm{var}\,(U)\mathrm{var}\,(V)}}.$$

Rem : On peut montrer que si U et V sont de carré intégrable, alors la variable aléatoire UV est intégrable et que, si de plus le couple (U, V) admet pour densité $f : \mathbb{R}^2 \to \mathbb{R}$, alors

$$\mathbf{E}(UV) = \int_{\mathbb{R}} \int_{\mathbb{R}} uvf(u, v)\, du\, dv$$

Le coefficient de corrélation est un réel de l'intervalle $[-1, 1]$. Il ne peut être égal à ± 1 que s'il existe un réel λ tel que $U = \lambda V$.

Ces deux résultats sont la conséquence de l'inégalité de Cauchy-Schwarz : si X et Y sont deux variables aléatoires de carré intégrable alors $|\mathbf{E}(XY)| \leq \sqrt{\mathbf{E}(X^2)\mathbf{E}(Y^2)}$.

On peut voir facilement que

Proposition 1.2. *Si U et V sont de carré intégrable,*

- $\mathrm{cov}(U, U) = \mathrm{var}\,(U)$
- $\mathrm{cov}(U, V) = \mathbf{E}(UV) - \mathbf{E}(U)\mathbf{E}(V) = \mathbf{E}[(U - \mathbf{E}(U))V]$
- $\mathrm{var}\,(U + V) = \mathrm{var}\,(U) + \mathrm{var}\,(V) + 2\mathrm{cov}(U, V)$

1.3 Indépendance

Définition 1.3. *Deux variables aléatoires U et V sont dites indépendantes si pour tous intervalles I et J de $\mathbb{R}$, on a*

$$\mathbf{P}(U \in I, V \in J) = \mathbf{P}(U \in I)\mathbf{P}(V \in J).$$

Proposition 1.3.
- *Si le couple (U, V) est un couple à densité dont la densité conjointe f s'écrit comme le produit des densités marginales g_U et g_V c'est-à-dire, vérifiant, pour tout $(u, v) \in \mathbb{R}^2$, $f(u, v) = g_U(u)g_V(v)$ alors les variables aléatoires U et V sont indépendantes.*
- *Réciproquement, si U et V sont indépendantes et de densité respectivement g_U et g_V, alors le couple (U, V) est à densité, de densité $(u, v) \to g_U(u)g_V(v)$.*

On en déduit alors le corollaire suivant :

Corollaire 1.1. *Si U et V sont deux variables aléatoires indépendantes et intégrables, alors la covariance de (U, V) est nulle.*

La réciproque de ce corollaire est **fausse** : considérons par exemple le couple (U, V) de densité $(u, v) \to \frac{1}{\pi}\mathbf{1}_{u^2+y^2\leq 1}$. Il est de covariance nulle mais n'est pas formé de variables aléatoires indépendantes.

1.4 Couples gaussiens

La propriété essentielle des variables aléatoires gaussiennes est leur stabilité en loi : si U suit une loi de Gauss (ou loi Normale) de moyenne m et de variance σ^2, alors, pour tout couple $(a, b) \in \mathbb{R}^* \times \mathbb{R}$, la variable aléatoire $aU + b$ suit la loi Normale d'espérance $am + b$ et de variance $a^2\sigma^2$. En particulier,

Proposition 1.4. *Si U est de loi Normale d'espérance m et de variance σ^2, alors la variable $(U - m)/\sigma$ est Normale, d'espérance nulle et de variance 1 (centrée et réduite).*

Abordons maintenant la notion de couple gaussien, qui interviendra de façon fondamentale dans la suite de ce cours :

Définition 1.4. *Un couple (U, V) de variables aléatoires est dit* **gaussien** *si, pour tous réels a et b, la variable aléatoire $aU+bV$ suit une loi de Gauss (ou loi Normale) sur $\mathbb{R}$ ou est constante.*

Quelques remarques :

- Un couple gaussien n'admet pas nécessairement de densité : par exemple, si U suit une loi de Gauss sur $\mathbb{R}$ le couple (U, U) est un couple gaussien et il n'admet pas de densité sur $\mathbb{R}^2$.
- Si un couple est gaussien, chacune de ses composantes suit une loi de Gauss ou est constante (la réciproque est fausse).
- Si les variables aléatoires U et V sont indépendantes et de loi Normale, alors le couple (U, V) est gaussien.

La propriété suivante n'est (malheureusement) valable que pour les couples gaussiens :

Proposition 1.5. *Si un couple (U, V) est gaussien, alors sa covariance est nulle si et seulement si U et V sont indépendantes.*

Remarque importante : Par définition d'un couple gaussien, toute combinaison linéaire $aU + bV$ des composantes U et V d'un couple gaussien suit une loi Normale. Pour déterminer complètement la loi de $aU + bV$, il reste alors à déterminer l'espérance et la variance de cette variable aléatoire, ce qui se fait en utilisant les règles de calcul de l'espérance et de la variance d'une somme. Ce résultat sera utilisé à de nombreuses reprises pour des combinaisons linéaires de variables aléatoires gaussiennes indépendantes.

2 Régression linéaire simple

2.1 Description des données du modèle

La variable à expliquer est Y et la variable indépendante est X. Le modèle statistique est

$$Y = aX + b + \varepsilon \tag{2.1}$$

Pour estimer les paramètres a et b, nous disposons de n couples d'observations

Var Obs	1	2	...	i	...	n
Y	y_1	y_2	...	y_i	...	y_n
X	x_1	x_2	...	x_i	...	x_n

Alors, le modèle (2.1) peut être écrit

$$Y_i = aX_i + b + \varepsilon_i, \qquad i = 1, ..., n \tag{2.2}$$

On suppose en ce qui concerne les v.a. ε_i que

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \qquad \text{avec} \quad \sigma^2 \text{ inconnu}$$

et pour $i \neq j$, ε_i et ε_j indépendantes, donc $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$

Attention : Les X_i ne sont pas aléatoires. Il faut plutôt les voir comme des paramètres dont dépendent les Y_i observées. Seules les ϵ_i et donc les Y_i sont aléatoires.

Proposition 2.1. *Si les variables (ϵ_i) sont indépendantes et de loi $\mathcal{N}(m, \sigma^2)$ alors*

1. $\mathbf{E}(Y_i) = aX_i + b$
2. $\text{cov}(Y_i, Y_j) = \begin{cases} \sigma^2 & si \quad i = j \\ 0 & si \quad i \neq j \end{cases}$
3. $Y_i \sim \mathcal{N}(aX_i + b, \sigma^2)$

Preuve :

1. $\mathbf{E}(Y_i) = \mathbf{E}\left(aX_i + b + \varepsilon_i\right) = \mathbf{E}\left(aX_i + b\right) + \mathbf{E}(\varepsilon_i) = aX_i + b$
2. $\text{cov}(Y_i, Y_j) = \mathbf{E}\left[\left(Y_i - \mathbf{E}(Y_i)\right)\left(Y_j - \mathbf{E}(Y_j)\right)\right]$
$= \mathbf{E}\left[\left(Y_i - aX_i - b\right)\left(Y_j - aX_j - b\right)\right] = \mathbf{E}(\varepsilon_i\varepsilon_j) = \begin{cases} \sigma^2 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$

■

Proposition 2.2. *Si les variables (ϵ_i) sont indépendantes et de loi $\mathcal{N}(m, \sigma^2)$ alors $\overline{Y_n}$ suit la loi Normale de paramètres*

$$\mathbf{E}(\overline{Y_n}) = a\overline{X_n} + b \quad \text{et} \quad \text{var}\,(\overline{Y_n}) = \frac{\sigma^2}{n}$$

Preuve : La variable aléatoire $\overline{Y_n}$ est une combinaison linéaire du vecteur $(Y_1, \ldots, Y_n)$ qui est gaussien car il est formé de variables aléatoires indépendantes. Donc $\overline{Y_n}$ suit une loi Normale. Déterminons son espérance et sa variance : $\overline{Y_n} = (\sum_{i=1}^n Y_i)/n$ donc

$$\mathbf{E}(\overline{Y_n}) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}(aX_i + b + \varepsilon_i) = \frac{1}{n}\sum_{i=1}^n (aX_i + b) = \frac{a}{n}\sum_{i=1}^n X_i + b = a\overline{X_n} + b$$

et

$$\text{var}\,(\overline{Y_n}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) = \frac{1}{n^2}\text{var}\left(\sum_{i=1}^n Y_i\right) = \frac{1}{n^2}\sum_{i=1}^n \text{var}\,(Y_i) = \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

■

2.2 Estimation des paramètres du modèle

La construction des estimateurs A et B des paramètres réels a et b est basée sur la méthode des moindres carrés.

Définition. Les *estimateurs des moindres carrés* des a et b sont les v.a. A_n et B_n qui minimisent la somme des carrés des termes d'erreur

$$S(A, B) = \sum_{i=1}^n e_i = \sum_{i=1}^n \left[Y_i - (AX_i + B)\right]^2$$

Théorème 2.1. *Il existe un unique couple (A_n, B_n) qui minimise $S(A, B)$ et on a*

$$A_n = \frac{\sum_{i=1}^n (Y_i - \overline{Y_n})(X_i - \overline{X_n})}{\sum_{i=1}^n (X_i - \overline{X_n})^2} \qquad \text{et} \qquad B_n = \overline{Y_n} - A_n\overline{X_n} \tag{2.3}$$

Preuve : Écrivons $Y_i - AX_i - B$ sous la forme

$$Y_i - AX_i - B = (Y_i - \overline{Y_n} - A(X_i - \overline{X_n})) + (\overline{Y_n} - A\overline{X_n} - B)$$

On a

$$\begin{aligned}\sum_i (Y_i - AX_i - B)^2 &= \sum_i (Y_i - \overline{Y_n} - A(X_i - \overline{X_n}))^2 + \sum_i (\overline{Y_n} - A\overline{X_n} - B)^2 \\ &\quad +2\sum_i [(Y_i - \overline{Y_n} - A(X_i - \overline{X_n}))(\overline{Y_n} - A\overline{X_n} - B)] \\ &= \sum_i (Y_i - \overline{Y_n} - A(X_i - \overline{X_n}))^2 + n(\overline{Y_n} - A\overline{X_n} - B)^2 \\ &\quad +2(\overline{Y_n} - A\overline{X_n} - B)\sum_i (Y_i - \overline{Y_n} - A(X_i - \overline{X_n}))\end{aligned}$$

Vérifions que la dernière somme est nulle. On a

$$\begin{aligned}\sum_i (Y_i - \overline{Y_n} - A(X_i - \overline{X_n})) &= \sum_i Y_i - \sum_i \overline{Y_n} - A\sum_i X_i + A\sum_i \overline{X_n} \\ &= \sum_i Y_i - n\overline{Y_n} - A\sum_i X_i + An\overline{X_n}\end{aligned}$$

qui est nulle par définition de $\overline{Y_n}$ et $\overline{X_n}$.

On obtient donc

$$\sum_i (Y_i - AX_i - B)^2 = \sum_i (Y_i - \overline{Y_n} - A(X_i - \overline{X_n}))^2 + n(\overline{Y_n} - A\overline{X_n} - B)^2$$

Une fois A choisi, on pourra toujours choisir B de sorte que $\overline{Y_n} - A\overline{X_n} - B = 0$, ce qui assurera donc que le dernier terme est minimal.

Minimisons maintenant la première somme. Pour cela, notons

$$A_n = \frac{\sum_i (X_i - \overline{X_n})(Y_i - \overline{Y_n})}{\sum_i (X_i - \overline{X_n})^2}$$

On a

$$\begin{aligned}\sum_i (Y_i - \overline{Y_n} - A(X_i - \overline{X_n}))^2 &= \sum_i (Y_i - \overline{Y_n} - A_n(X_i - \overline{X_n}))^2 + \sum_i (A_n - A)^2 (X_i - \overline{X_n})^2 \\ &\quad +2\sum_i (Y_i - \overline{Y_n} - A_n(X_i - \overline{X_n})(A_n - A)(X_i - \overline{X_n})\end{aligned}$$

Montrons que la somme des doubles-produits est nulle :

$$\sum_i (Y_i - \overline{Y_n} - A_n(X_i - \overline{X_n})(X_i - \overline{X_n}) = \sum_i (Y_i - \overline{Y_n}) \sum_i (X_i - \overline{X_n}) - A_n \sum_i (X_i - \overline{X_n})(X_i - \overline{X_n})$$

Par définition de A_n, cette somme est nulle.

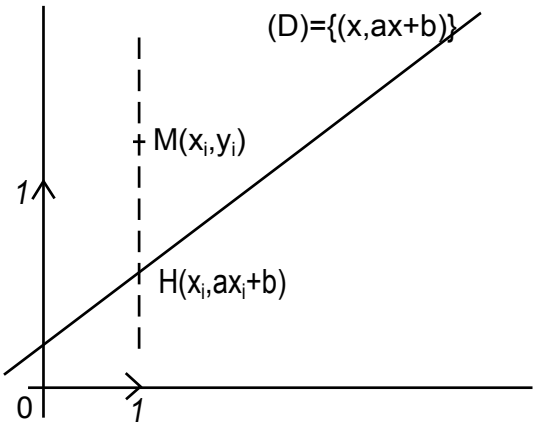
Finalement, on a obtenu :

$$\begin{aligned}\sum_i (Y_i - AX_i - B)^2 &= \sum_i (Y_i - \overline{Y_n} - A_n(X_i - \overline{X_n}))^2 + (A_n - A)^2 \sum_i (X_i - \overline{X_n})^2 + \\ &\quad +n(\overline{Y_n} - A\overline{X_n} - B)^2\end{aligned}$$

Il est maintenant clair que pour minimiser la somme des carrés des erreurs, il faut choisir $A = A_n$ et $B = \overline{Y_n} - A\overline{X_n}$. ■

Définition 2.1. *La droite du plan d'équation* $\{(x, y), y = A_n x + B_n\}$ *est appelée la* **droite de régression linéaire des y en x** *par la méthode des moindres carrés ordinaires (MCO). Le coefficient* A_n *est la pente de cette droite et* B_n *est la valeur à l'origine, également appelée intercept.*

Attention : Il ne s'agit pas de la droite qui minimise la somme des carrés des distances entre le nuage du point et une droite : on s'intéresse à la distance entre un point du nuage et son projeté sur la droite dans la direction verticale. Les rôles des deux caractères x et y ne sont pas symétriques, et l'expression de A_n fait apparaître cette dissymétrie.



Propriétés

Proposition 2.3. *Les v.a.* A_n *et* $\overline{Y_n}$ *ne sont pas corrélées :*

$$\mathrm{Corr}(A_n, \overline{Y_n}) = 0$$

Proposition 2.4. *Les v.a.* A_n *et* B_n *sont des estimateurs sans biais pour les paramètres* a *et* b*, c'est-à-dire que* $\mathbf{E}(A_n) = a$ *et* $\mathbf{E}(B_n) = b$*.*

En ce qui concerne les variances et les covariances de ces estimateurs, on a la proposition suivante :

Proposition 2.5.

$$\mathrm{var}\,(B_n) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X_n}^2}{\sum_{i=1}^n (X_i - \overline{X_n})^2} \right]$$

$$\mathrm{var}\,(A_n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X_n})^2}$$

$$\mathrm{cov}(A_n, B_n) = -\frac{\sigma^2 \overline{X_n}}{\sum_{i=1}^n (X_i - \overline{X_n})^2}$$

En ce qui concerne un estimateur pour la variance σ^2 on a la proposition suivante :

Proposition 2.6. *Un estimateur sans biais pour* σ^2 *est*

$$S_n^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - A_n X_i - B_n)^2$$

Remarque : Une estimation pour σ^2 est donc

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{a}_n x_i - \hat{b}_n\right)^2$$

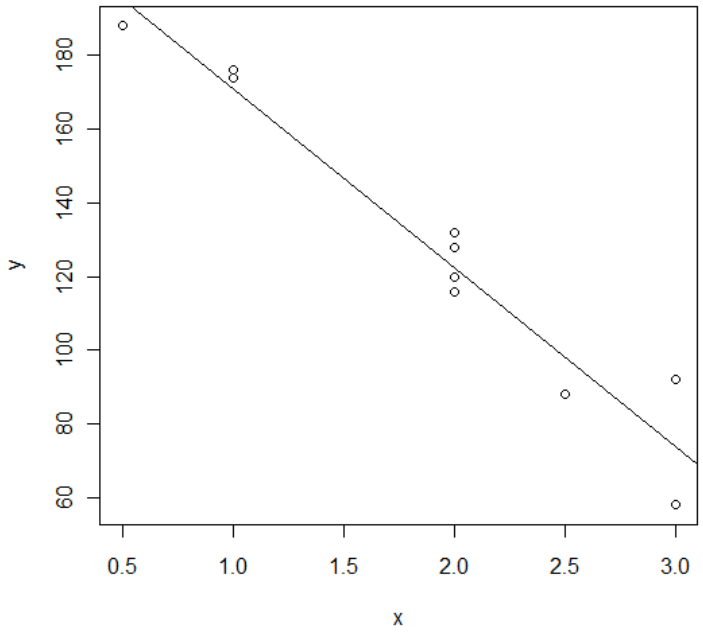
Exemple. Les estimations de a et b sur les données mesurées de l'exemple de taux d'ozone en fonction de la vitesse du vent sont

$$\hat{a}_n = -48.5 \quad \hat{b}_n = 219.5 \quad s_n^2 = 117.5 \quad \hat{\sigma}_n = 10.84$$

$$\widehat{\text{var}\,(A_n)} = 18.4 \quad \widehat{\text{var}\,(B_n)} = 78.05$$

Donc, on peut dire que la pollution d'ozone est liée à la vitesse du vent par la relation linéaire :

$$Y = -48.5X + 219.5$$



Nuage de points et droite de régression (taux d'ozone en fonction de la vitesse du vent)

Les lois des estimateurs

Proposition 2.7. *A_n suit une loi normale d'espérance a et de variance*

$$\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X_n})^2}$$

et B_n suit une loi normale d'espérance b et de variance

$$\sigma^2\left[\frac{1}{n} + \frac{\overline{X_n}^2}{\sum_{i=1}^{n}(X_i - \overline{X_n})^2)}\right]$$

Preuve : On remarque en effet que, puisque les X_i sont déterministes, A_n est une combinaison linéaire des Y_i :

$$A_n = \sum_{i=1}^{n} \frac{X_i - \overline{X_n}}{\sum_{j=1}^{n}(X_j - \overline{X_n})^2} Y_i$$

avec les Y_i de loi normale $\sim \mathcal{N}(aX_i + b, \sigma^2)$, et (Y_i) indépendantes, donc A_n suit une loi normale avec

$$\mathbf{E}(A_n) = a, \operatorname{var}(A_n) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X_n})^2}.$$

Puisque $B_n = \overline{Y_n} - A_n\overline{X_n}$ et que $\overline{Y_n}$ et A_n sont non correlés, on a $\mathbf{E}(B_n) = b$ et

$$\operatorname{var}(B_n) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X_n}^2}{\sum_{i=1}^{n}(X_i - \overline{X_n})^2} \right]$$

A_n et $\overline{Y_n}$ suivent des lois normales d'où

$$B_n \sim \mathcal{N}\left(b, \sigma^2 \left[\frac{1}{n} + \frac{\overline{X_n}^2}{\sum_{i=1}^{n}(X_i - \overline{X_n})^2} \right] \right)$$

■

2.3 Mesure de l'ajustement

On dispose de la forme générale des estimateurs. Pour un ensemble de n couples (x_i, y_i) mesurés, on peut donner une estimation $\hat{a}_n, \hat{b}_n, \hat{\sigma}_n$ (i.e. on peut donner une valeur effective) pour a, b, σ.
Ainsi la *droite de régression* la plus proche du nuage de points (x_i, y_i) est définie par l'équation

$$y = \hat{a}_n x + \hat{b}_n$$

l'estimation de l'observation y_i par le modèle est

$$\hat{y}_i = \hat{a}_n x_i + \hat{b}_n \tag{2.4}$$

qui est une réalisation de la v.a. (estimateur)

$$\hat{Y}_i = A_n X_i + B_n$$

La différence $e_i = y_i - \hat{y}_i$ s'appelle le **résidu** ; et en divisant e_i par $\hat{\sigma}_n$, $\frac{e_i}{\hat{\sigma}_n}$, on a le **résidu réduit**.

Proposition 2.8. *La somme des résidus est nulle.*

Remarque. Il faut faire la différence entre l'erreur :

$$\varepsilon_i = Y_i - aX_i - b$$

avec a, b les vraies valeurs mais inconnues, donc ε_i inconnue, et le résidu :

$$e_i = y_i - \hat{a}_n x_i - \hat{b}_n$$

(avec y_i, x_i mesurées et $\hat{a}_n$ et $\hat{b}_n$ estimées).

Il est souhaitable de donner un indicateur sur la qualité de l'ajustement du modèle $Y_i = aX_i + b + \varepsilon_i$ fourni par l'équation (2.4). Seulement les valeurs des résidus sont insuffisantes

– d'abord, ces différences dépendent de l'unité de mesure
– et ensuite, elles ne donnent pas une indication sur l'ajustement global.

L'indice le plus couramment employé est le coefficient suivant

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y_n})^2}{\sum_{i=1}^{n} (y_i - \overline{y_n})^2}$$

connu sous le nom de **coefficient de détermination**.

Proposition 2.9. $0 \leq R^2 \leq 1$.

Interprétation. Si la valeur de R^2 est proche de 1 on dit que la variable X explique bien la variable Y. Inversement, si R^2 est proche de 0, X n'explique pas bien Y et le modèle de régression linéaire simple consideré n'est pas bon. On va voir plus loin d'où vient cette interprétation.

Exemple de l'ozone : $R^2 = 0.93$.

2.4 Décomposition de la variabilité de Y

Proposition 2.10. *On a*

$$\sum_{i=1}^{n} (y_i - \overline{y_n})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2$$

Dispersion totale de Y = dispersion due au modèle + dispersion résiduelle

$$ST = SM + SR$$

Remarque : ST ne dépend pas du modèle mais des données mesurées et elle s'appelle totale parce qu'elle donne la mesure de variation des données mesurées par rapport à leur moyenne.

Preuve : En effet, pour tout i, on a $y_i - \overline{y_n} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y_n})$, donc

$$\sum_{i=1}^{n} (y_i - \overline{y_n})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \overline{y_n})$$

Or

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} [y_i - (\hat{a}(x_i - \bar{x}) + \bar{y})][(\hat{a}(x_i - \bar{x}) + \bar{y}) - \bar{y}] \\
&= \sum_{i=1}^{n} (y_i - \bar{y} - \hat{a}(x_i - \bar{x}))\hat{a}(x_i - \bar{x}) \\
&= \hat{a}\left[\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \hat{a}\sum_{i=1}^{n} (x_i - \bar{x})^2\right] \\
&= 0
\end{aligned}$$

par définition le résultat de $\hat{a}$, ce qui fournit le résultat souhaité. ■

La régression est résumée dans le tableau ci dessous (appelé tableau d'analyse de variance)

Source de variation	Somme des carrés des écarts	Degrés de liberté	Carré moyen
Régression	$SM = \sum_{i=1}^{n}(\hat{y}_i - \overline{y_n})^2$	1(=2-1)	$SM/1$
Résiduelle	$SR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	n-2	$SR/(n-2)$
Totale	$ST = \sum_{i=1}^{n}(y_i - \overline{y_n})^2$	n-1	

En fait, SM donne une mesure de la variabilité (de l'écart) des estimations $\hat{y}_i$ faites par le modèle par rapport à la moyenne $\overline{y_n}$ des données. SR donne une mesure de la variabilité (de l'écart) entre les estimations $\hat{y}_i$ et les vraies valeurs y_i.

Remarque : Le coefficient R^2 est donc le rapport

$$R^2 = \frac{SM}{ST} = \frac{ST - SR}{ST} = 1 - \frac{SR}{ST}$$

Maintenant on voit mieux d'où vient l'interprétation de R^2 : si R^2 est proche de 1 alors $SR \sim 0$ et la différence entre les valeurs mesurées y_i et celles prédites $\hat{y}_i$ est relativement petite. On divise par ST en fait pour avoir un indicateur qui ne tient pas compte de l'unité de mesure.

Remarque : Revenons à l'estimation de σ^2 :

$$s_n^2 = \hat{\sigma}_n^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a}_n x_i - \hat{b}_n)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{SR}{n-2}$$

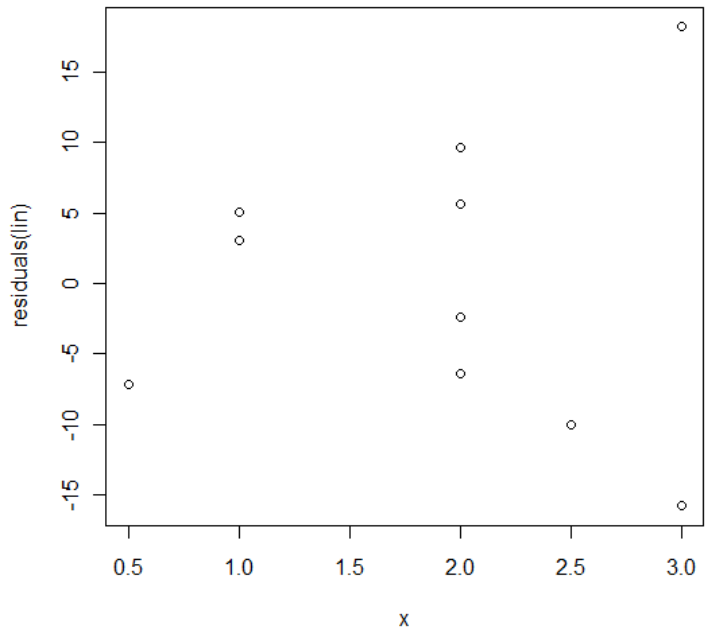
Exemple. Tableau d'analyse de variance pour l'échantillon de mesure de taux d'ozone

Source	S.C.	ddl	C.M.
Modèle	15093	1	15093
Résidu	940	8	117.5
Total	16033	9	

$R^2 = 0.93$.

2.5 Évaluation de l'ajustement

- Jusqu'à présent on a vu que R^2 nous donne une information sur la qualité de l'ajustement. Mais cette quantité est insuffisante pour l'évaluation du modèle.
- On a vu aussi qu'une autre manière simple de détecter les défaillances du modèle consiste à calculer les résidus $e_i = y_i - \hat{y}_i$ et les résidus réduits : $er_i = \frac{e_i}{\hat{\sigma}}$. Puisque les e_i sont des réalisations de la v.a. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, les er_i sont des réalisations d'une v.a. $\mathcal{N}(0, 1)$.
- Un graphique de ces résidus réduits révèle les gros écarts du modèle ; une étude systématique des résidus est un élément essentiel de toute analyse de régression.
  Si le modèle est correct, les résidus réduits doivent se trouver approximativement entre -2 et 2. Ils ne doivent présenter aucune structure particulière. Si jamais ils en présentent une, c'est qu'une structure cachée existe dans les données.



Résidus dans le modèle du taux d'ozone en fonction du vent

2.6 Tests sur les paramètres

On va faire des tests sur les paramètres du modèle. On pourrait tester :
1) L'hypothèse de lien linéaire effectif entre $X_1, ..., X_n$ et les variables aléatoires : $Y_1, ..., Y_n$. En terme de paramètres, ça signifie qu'on testera l'hypothèse

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

équivalent avec

$$H_0 : Y_i = b + \varepsilon_i \qquad \text{contre} \qquad H_1 : Y_i = aX_i + b + \varepsilon_i$$

2) L'hypothèse d'un modèle linéaire spécifié : on testera :

$$H_0 : a = a_0 \quad \text{et} \quad b = b_0 \qquad \Longleftrightarrow \qquad Y_i = a_0 X_i + b_0 + \varepsilon_i$$

contre

$$H_0 : a \neq a_0 \quad \text{ou} \quad b \neq b_0 \qquad \Longleftrightarrow \qquad Y_i = aX_i + b + \varepsilon_i$$

2.6.1 Test du caractère significatif du modèle

Étape 1) : Formulation des hypothèses H_0 et H_1 : L'hypothèse H_0 à tester est « il n'y a pas de lien linéaire entre X et Y » :

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

Étape 2) : Statistique de test : En ce qui concerne la statistique utilisée pour tester H_0, on peut en utiliser deux, qui vont suivre une loi de Student ou une loi de Fisher.

Première méthode : on utilise une v.a. de Student
On sait que

$$Z = \frac{(A_n - a)\sqrt{\sum_{i=1}^n (X_i - \overline{X_n})^2}}{S_n} \sim t(n-2)$$

Sous l'hypothèse H_0 cette hypothèse devient

$$Z = \frac{A_n\sqrt{\sum_{i=1}^n (X_i - \overline{X_n})^2}}{S_n} \sim t(n-2)$$

Étape 3) Construction de la zone d'acceptation. On fixe un risque $\alpha \in (0,1)$. Par la définition de la loi de Student on sait que

$$P\left[|Z| \leq t_{n-2,1-\alpha/2}\right] = 1 - \alpha$$

où $t_{n-2,1-\alpha/2}$ est le fractile d'ordre $1-\alpha/2$ de la loi de Student à $n-2$ degrés de liberté. Alors, la zone d'acceptation est

$$ZA_{H_0,\alpha} = \left[-t_{n-2,1-\alpha/2}, t_{n-2,1-\alpha/2}\right]$$

Étape 4) Calcul de la valeur z de la v.a. Z sur les données observées $(x_i, y_i)_{1\leq i\leq n}$.

$$z = \frac{\hat{a}_n\sqrt{\sum_{i=1}^n (x_i - \overline{x_n})^2}}{\hat{\sigma}_n} = \hat{a}_n\sqrt{\sum_{i=1}^n (x_i - \overline{x_n})^2}\sqrt{\frac{n-2}{\sum_{i=1}^n (y_i - \hat{a}_n x_i - \hat{b}_n)^2}} = \frac{\hat{a}_n}{\sqrt{\text{vâr}\,(A_n)}}$$

Étape 5) Conclusion : On adopte alors la stratégie suivante
- si $z \in ZA_{H_0,\alpha}$ alors on accepte H_0 au risque α (il n'y a pas de lien linéaire entre les deux variables X et Y, avec un risque de α).
- $z \notin ZA_{H_0,\alpha}$ alors on rejette l'hypothèse H_0 et on accepte H_1 (il y a un lien linéaire entre les deux variables)

Deuxième méthode : on utilise une v.a. de Fisher

On sait que

$$\frac{(A_n - a)^2 \sum_{i=1}^n (X_i - \overline{X_n})^2}{S_n^2} \sim F(1, n-2)$$

Alors, sous l'hypothèse H_0

$$Z = \frac{A_n^2 \sum_{i=1}^n (X_i - \overline{X_n})^2}{S_n^2} \sim F(1, n-2)$$

On va écrire cette v.a. sous une autre forme (fonction uniquement de Y) :

Proposition 2.11. *Sous l'hypothèse H_0*

$$Z = (n-2)\frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y_n})^2}{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}$$

Étape 2) La statistique utilisée pour tester H_0 est

$$Z = (n-2)\frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y_n})^2}{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \sim F(1, n-2)$$

Étape 3) Zone d'acceptation. On fixe le risque α.
Par définition de la loi de Fisher

$$P\left(Z \leq f_{1,n-2;1-\alpha}\right) = 1 - \alpha$$

où $f_{1,n-2;1-\alpha}$ est le fractile d'ordre $1-\alpha$ de la loi de la loi de Fisher. Puisque Z ne prend que des valeurs positives, la zone d'acceptation est

$$ZA_{H_0,\alpha} = [0, f_{1,n-2;1-\alpha}]$$

Étape 4) : Calcul de la valeur z de la v.a. Z sur les données (x_i, y_i) observées

$$z = (n-2)\frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y_n})^2}{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Étape 5) : Conclusion
- si $z \in ZA$ alors on accepte H_0
- si $z \notin ZA$ alors on n'accepte pas H_0 et on accepte H_1 au risque α.

Exemple du taux d'ozone : Appliquons les deux méthodes que nous venons de voir pour tester la significativité du modèle linéaire dans l'exemple du taux d'ozone. L'hypothèse à tester est donc

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

Student

2) $Z = \dfrac{A_n\sqrt{\sum_{i=1}^{10}(X_i - \overline{X_n})^2}}{S_n} \sim t(8)$

3) Pour $\alpha = 0.05$, $ZA_{H_0,\alpha} = [-t_{8;0.975}; t_{8;0.975}] = [-2.306; 2.306]$

4) $z = \dfrac{\hat{a}_n\sqrt{\sum_{i=1}^{10}(x_i - \overline{X_n})^2}}{\hat{\sigma}_n} = \dfrac{\hat{a}_n}{\sqrt{\text{var}\,\hat{(}A_n)}} = -\dfrac{48.5}{\sqrt{18.4}} \sim 10$

5) $z \notin ZA \Rightarrow H_0$ rejetée, H_1 acceptée $\Leftrightarrow$ il y a bien une relation linéaire entre la concentration d'ozone et la vitesse du vent.

Fisher

2) $Z = 8\dfrac{\sum^{10}(\hat{Y}_i - \overline{Y_n})^2}{\sum^{10}(\hat{Y}_i - Y_i)^2} \sim F(1, 8)$

3) $ZA_{H_0,\alpha} = [0; f_{1,8;0.95}] = [0; 5.32]$

4) $z = 8\dfrac{\sum^{10}(\hat{y}_i - \overline{y_n})^2}{\sum^{10}(\hat{y}_i - y_i)^2} = 8\dfrac{SM}{SR} = \dfrac{SM}{\hat{\sigma}^2} = \dfrac{15093}{117.5} \sim 12$

5) $z \notin ZA \Rightarrow H_0$ rejetée.

Remarques : **1)** Par les deux méthodes on doit obtenir des résultats concordants.
2) La zone d'acceptation dépend du risque α. Si $\alpha = 0.01$, $[-t_{8;0.92}; t_{8;0.92}] = [-3.355; 3.355]$

2.6.2 Test d'un modèle linéaire spécifié

On veut tester simultanément les deux paramètres a et b. Puisque les estimateurs A_n et B_n des paramètres a et b ne sont pas indépendants, il ne serait pas correct de tester successivement a et puis b.

1) On pose l'hypothèse nulle : $H_0 : a = a_0$ et $b = b_0$
contre l'hypothèse alternative : $H_1 : a \neq a_0$ ou $b \neq b_0$

2) La construction du test repose sur le théorème suivant, que nous ne démontrerons pas :

Théorème 2.2. *Sous l'hypothèse H_0, nous avons*

$$Z = \frac{n-2}{2} \frac{\sum_{i=1}^{n} \left[(A_n - a_0)X_i + (B_n - b_0)\right]^2}{\sum_{i=1}^{n} \left(Y_i - A_n X_i - B_n\right)^2} \sim F(2, n-2)$$

3) *Construction de la zone d'acceptation.* On fixe un risque α et on calcule (en utilisant les tables de la loi de Fisher) $f_{2,n-2;1-\alpha}$ t.q.

$$P\left[Z \leq f_{2,n-2;1-\alpha}\right] = 1 - \alpha$$

La zone d'acceptation est alors

$$ZA_{H_0,\alpha} = [0; f_{2,n-2;1-\alpha}]$$

4) On calcule la valeur z de la v.a. Z sur les données $(x_i, y_i)_{1 \leq i \leq n}$. On a

$$z = \frac{n-2}{2} \frac{\sum_{i=1}^{n} \left[(\hat{a}_n - a_0)x_i + (\hat{b}_n - b_0)\right]^2}{\sum_{i=1}^{n} \left(y_i - \hat{a}_n x_i - \hat{b}_n\right)^2}$$

5) Conclusion
- si $z \in ZA$ alors on accepte H_0
- si $z \notin ZA$ alors on n'accepte pas H_0 et on accepte H_1 au risque α.

Exemple.
1) $H_0 : a = -48 \quad$ et $\quad b = 220$
$H_1 : a \neq -48 \quad$ ou $\quad b \neq 220$
2)

$$Z = 4 \frac{\sum_{i=1}^{10} \left[(A_n + 48)X_i + (B_n - 220)\right]^2}{\sum_{i=1}^{10} \left(Y_i - A_n X_i - B_n\right)^2} \sim^{H_0} F(2, 8)$$

3) $ZA_{H_0,0.05} = [0; f_{2,8;0.95}] = [0; 4,46]$
4)

$$z = 4 \frac{\sum_{i=1}^{10} \left[(-48.5 + 48)x_i + (219.5 - 220)\right]^2}{\sum_{i=1}^{10} \left(y_i + 48.5 x_i - 219.5\right)^2} =$$

5) ....

2.7 Prévision d'une valeur

On est dans la situation suivante : on a n mesures pour les variables X et Y : $(x_i, y_i)_{1\leq i\leq n}$.
Entre les variables Y et X existe un lien linéaire :

$$Y_i = aX_i + b + \epsilon_i \qquad i = 1, ..., n \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

On sait construire des estimateurs A_n et B_n pour les paramètres a et b. Puisqu'on dispose de n données, on peut préciser effectivement quelles sont les valeurs de A_n et B_n : $\hat{a}_n$ et $\hat{b}_n$.
On désire maintenant prévoir la valeur de Y pour une nouvelle valeur de X : x_{n+1}.
La prévision la plus naturelle est

$$\hat{y}_{n+1} = \hat{a}_n x_{n+1} + \hat{b}_n = \hat{a}_n(x_{n+1} - \overline{x_n}) + \overline{y_n}$$

qui est une réalisation de la variable $\hat{Y}_{n+1} = A_n X_{n+1} + B_n$, les estimateurs A_n et B_n étant construits à partir des n premières observations.

Par ailleurs A_n et $\overline{Y_n}$ sont des variables aléatoires normales et non corrélées, donc elles sont indépendantes. On obtient alors aisément l'intervalle de confiance ; la variance de l'estimateur est $\sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x_n})^2}{\sum_{i=1}^n (x_i - \overline{x_n})^2}\right)$

3 Régression linéaire multiple

Exemple.
Supposons que l'on dispose des données suivantes, pour 3 variables :

Obs	y_i	x_{1i}	x_{2i}
1	10	6	28
2	20	12	40
3	17	10	32
4	12	8	36
5	11	9	34

$Corr(Y, X_1) = 0.91 \quad Corr(Y, X_2) = 0.65$
On déduit qu'il peut y avoir un lien linéaire entre Y et X_1, X_2

$$Y_1 = b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i \qquad i = 1, ..., 5 \tag{2.5}$$

avec b_0, b_1, b_2 paramètres inconnus, à estimer.
On définit les matrices et les vecteurs

$$X = \begin{bmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{15} & X_{25} \end{bmatrix} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_5 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_5 \end{bmatrix} \quad \beta = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

Alors, le modèle (2.5) peut être écrit sous la forme matricielle

$$Y = X\beta + \varepsilon \tag{2.6}$$

Démonstration de (2.6)

$$X\beta = \begin{bmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{15} & X_{25} \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_3 \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_{11} + b_2 X_{21} \\ \vdots \\ b_0 + b_1 X_{15} + b_2 X_{25} \end{bmatrix}$$

Donc, la ligne i du vecteur $X\beta + \varepsilon$ est

$$b_0 + b_1 X_{1i} + b_2 X_{2i} = Y_i$$

3.1 Estimation des paramètres

3.1.1 Le cadre du problème

Supposons qu'on a un échantillon de n mesures pour $(p+1)$ variables

$$Y, X_1, ..., X_p$$

avec $p < n$, Y variable aléatoire, X_i variables non-aléatoires. Comme d'habitude on va noter les valeurs mesurées avec des petites lettres

$$y_i, x_{1i}, ..., x_{pi} \qquad i = 1, ..., n$$

Ces données mesurées peuvent être représentées sous la forme d'un tableau

Obs	y	x_1	...	x_j	...	x_p
1	y_1	x_{11}	...	x_{j1}	...	x_{p1}
2	y_2	x_{12}	...	x_{j2}	...	x_{p2}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$
i	y_i	x_{1i}	...	x_{ji}	...	x_{pi}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$
n	y_n	x_{1n}	...	x_{jn}	...	x_{pn}

Donc, x_{ji} est la i-ème observation de la j-ème variable.
On cherche à construire Y comme fonction linéaire des variables $X_1, ..., X_p$.
L'équation modèle pour l'observation i est

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_p X_{pi} + \varepsilon_i \qquad i = 1, ..., n \tag{2.7}$$

On a n équations, une pour chaque observation, et elles peuvent être résumées sous la forme matricielle

$$Y = X\beta + \varepsilon \tag{2.8}$$

avec

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} \qquad X = \begin{bmatrix} 1 & X_{11} & ... & X_{p1} \\ 1 & X_{12} & ... & X_{p2} \\ \vdots & \vdots & \vdots & \\ 1 & X_{1i} & ... & X_{pi} \\ \vdots & \vdots & \vdots & \\ 1 & X_{1n} & ... & X_{pn} \end{bmatrix} \quad \beta = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Pour que le modèle soit complètement spécifié, il faut donner les répartitions des erreurs ε_i. On suppose

$$\mathbf{E}(\varepsilon_i) = 0 \qquad \text{var}\,(\varepsilon_i) = \sigma^2 \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad i = 1, ..., n$$

ε_i et ε_j indép, donc $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$.

Les paramètres du modèle sont : $b_0, b_1, ..., b_p$ et la variance σ^2. Il faut les estimer en connaissant les n observations.

Conséquences

1. $\mathbf{E}(\varepsilon) = 0$ (le vecteur nul de $\mathbb{R}^n$)
2. $\text{var}\,(\varepsilon) = \sigma^2 I_n$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$
3. $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$: les Y_i sont des variables aléatoires indépendantes de loi Normale de variance σ^2 et d'espérance $\mathbf{E}(Y_i) = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_p X_{pi}$.

Commentaires : **1)** La régression linéaire multiple peut être vue comme une extension de la régression simple ($p = 1$).
2) C'est un problème plus difficile : les calculs sont plus difficiles et il est pratiquement impossible de se passer de l'ordinateur.

3.1.2 Estimateurs ponctuels de β et de σ^2

L'estimateur de moindres carrés du vecteur β.

Cet estimateur s'obtient suivant la même procédure que pour la régression simple. C'est le *vecteur aléatoire* qui minimise la fonction :

$$T(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - ... - b_p X_{pi})^2$$

On constate que T peut s'écrire comme la norme du vecteur $Y - X\beta$, où X est la matrice dont les lignes sont formées par les vecteurs $(1, X_{i1}, \ldots, X_{ip})$ des données de la i-éme observation. On a donc

$$T(\beta) = (Y - X\beta)^t.(Y - X\beta) = (Y^t - \beta^t X^t)(Y - X\beta)$$

On peut montrer que si la matrice $(X^t X)$ est inversible, il existe une unique valeur du vecteur β qui minimise T, et que le minimum est atteint en

$$B_n = \left(X^t X\right)^{-1} X^t Y$$

qui est l'**estimateur des moindres carrés du vecteur paramètre** β.

$B_n = \begin{bmatrix} B_{n0} \\ B_{n1} \\ \vdots \\ B_{np} \end{bmatrix}$, B_{ni} est l'estimateur des moindres carrés pour b_i, $i = 0, ..., p$.

Si on a n mesures on obtient une valeur (réalisation) pour la v.a. B_n

$$\hat{\beta}_n = \left(x^t x\right)^{-1} \cdot x^t y$$

où y est le vecteur avec les mesures pour Y et x est la matrice avec les mesures pour $x_1, ..., x_p$, et comportant des « 1 » sur la première colonne. La valeur prédite pour Y par le modèle est

$$\hat{Y} = X\hat{B}_n$$

et $Y - \hat{Y}$ s'appelle résidu.

Exercice : Montrer que pour $p = 1$ on retrouve les valeurs obtenues dans le théorème 2.1.

Propriétés de l'estimateur B_n

1. Estimateur de β est sans biais : $\mathbf{E}(B_n) = \beta$.
   Preuve :

$$\begin{aligned}\mathbf{E}(B_n) &= \mathbf{E}\left[\left(X^tX\right)^{-1}X^tY\right] \\ &= \left(X^tX\right)^{-1}X^t\mathbf{E}(Y) \\ &= \left(X^tX\right)^{-1}X^t(X\beta) = \beta\end{aligned}$$ ■

2. Variance de B_n :

$$\text{var}(B_n) = \begin{bmatrix} \text{var}(B_{n0}) & \text{cov}(B_{n0}, B_{n1}) & ... & \text{cov}(B_{n0}, B_{np}) \\ \text{cov}(B_{n1}, B_{n0}) & \text{var}(B_{n1}) & ... & \text{cov}(B_{n1}, B_{np}) \\ ... & ... & ... & ... \\ \text{cov}(B_{np}, B_{n0}) & \text{cov}(B_{np}, B_{n1}) & ... & \text{var}(B_{np}) \end{bmatrix} = \sigma^2\left(X^tX\right)^{-1}$$

   C'est une matrice $(p+1) \times (p+1)$.
3. Chaque élément B_{nj} composant le vecteur B_n, $j = 0, ..., p$ est une fonction linéaire des variables $Y_1, ..., Y_n$. Cette propriété de linéarité determine les propriétés statistiques de ces estimateurs. En particulier, puisque les Y_i sont indépendantes et de loi Normale, les estimateurs des b_j suivent eux aussi une loi Normale, de variance facilement calculable.
4. Si on note $(X^tX)^{-1} = (c_{ij})_{1 \le i,j \le p+1}$, alors
   - La variance de l'estimateur $B_{n,i-1}$ de b_{i-1} est le i-eme élément diagonal de la matrice $\sigma^2(X^tX)^{-1}$, c'est-à-dire $\sigma^2 c_{ii}$.
   - $\text{cov}(B_{n,i-1}, B_{n,j-1}) = \sigma^2 c_{ij}$ pour $i \neq j$.

Estimateur pour σ^2. On montre que

$$S_n^2 = \frac{(Y - XB_n)^t(Y - XB_n)}{n-p-1} = \frac{(Y - \hat{Y})^t(Y - \hat{Y})}{n-p-1}$$

est un estimateur sans biais de σ^2.
Une estimation de σ^2 est

$$\hat{\sigma}_n^2 = \frac{(y - x\hat{\beta}_n)^t(y - x\hat{\beta}_n)}{n-p-1}$$

Propriétés

1) $(n-p-1)\dfrac{S_n^2}{\sigma^2} \sim \chi^2(n-p-1)$

2) B_i et S_n^2 sont indépendantes pour $\forall i = 0, 1, ..., p$.

3.2 Décomposition de la variabilité de Y

De même que pour la régression linéaire simple, nous avons

$$\sum_{i=1}^{n} (y_i - \overline{y_n})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \overline{y_n})^2$$

$ST = \sum_{i=1}^{n}(y_i - \overline{y_n})^2$ est *la somme des carrés totale* : elle représente la variabilité des observations de Y avant de prendre en compte les effets des variables $X_1, ..., X_p$.
$SR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ est *la somme des carrés résiduelle* (la somme des carrés due aux erreurs) et elle représente la variabilité de Y inexpliquée après que les variables $X_1, ..., X_p$ ont été utilisées dans l'équation de régression pour prédire Y.
$SM = ST - SR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y_n})^2$ *la somme des carrés due au modèle* de régression et mesure la variabilité due aux variables indépendantes $X_1, ..., X_p$ dans l'équation de régression.
On a le tableau de décomposition (ANOVA) :

Source de variation	ddl	S.C.	Carré moyen
Régression	p	SM	SM/p
Résiduelle	$n-p-1$	SR	$SR/(n-p-1)$
Totale	$n-1$	ST	

3.3 Mesure de l'ajustement (empirique)

La mesure de l'ajustement est donnée par le *coefficient de détermination*

$$R^2 = \frac{SM}{ST} \in [0, 1]$$

qui donne une mesure sommaire, quantitative sur la qualité de la prédiction de Y par les variables $X_1, ..., X_p$ dans le modèle de régression linéaire.
Il représente aussi le carré de la corrélation entre Y et $\hat{Y}$.

- Si on a une modélisation parfaite : $Y_i = \hat{Y}_i$ alors $SR = 0$, donc $ST + SM$ donc $R^2 = 1$.
- La valeur de R^2 croît si des nouvelles variables indépendantes sont ajoutées au modèle de régression.
- Similaire à la régression linéaire simple, seulement la valeur de R^2 est insuffisante pour bien caractériser la qualité de l'ajustement.

Exemple.

Obs	Y	X_1	X_2
1	10	6	28
2	20	12	40
3	17	10	32
4	12	8	36
5	11	9	34

$$\hat{\beta} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} = \left(x^t x\right)^{-1} \cdot x^t y = \begin{pmatrix} 2.33 \\ 2.08 \\ -0.208 \end{pmatrix}$$

Le tableau d'analyse de variance :

Source de variation	ddl	S.C.	Carré moyen
Modèle (X_1, X_2)	$p = 2$	62.5	31.25
Résidu	2	11.5	5.75
Totale	4	74	

$$R^2 = \frac{62.5}{74} = 0.85$$

3.4 Tests d'hypothèse

Une fois le modèle de régression multiple fixé et les estimations des paramètres obtenues, on se pose la question de la contribution des variables $X_1, ..., X_p$ sur la prédiction de Y. Un des critères importants dans la sélection d'un modèle est de choisir celui qui, avec le moins de variables, fournit la meilleure description des données étudiées. Dans la cadre de la régression linéaire multiple, p variables peuvent s'avérer superflues et un nombre inférieur q ($q < p$) peut permettre une description aussi bonne.

Il y a deux types de questions que l'on peut se poser :

1. On teste si le groupe entier des variables indépendantes contribue significativement à la prédiction de Y.
2. Test pour ajouter une seule variable, quand les autres variables indépendantes sont déjà dans le modèle.

Test de la significativité du modèle de régression entier

On a le modèle complet

$$Y_i = b_0 + b_1 X_{1i} + ... + b_p X_{pi} + \varepsilon_i$$

L'hypothèse nulle peut se traduire :
H_0 : « Toutes les p variables indépendantes considérées dans le même temps ne produisent pas une variation en Y »
H_0 : « il n'y a pas de régression significative en utilisant les p variables indépendantes dans le modèle », soit :

$$H_0 : b_1 = b_2 = ... = b_p = 0 \qquad \text{contre} \qquad H_1 : \exists j \in \{1, ..., p\} \text{ t.q. } b_j \neq 0$$

Sous l'hypothèse H_0, le modèle réduit est

$$Y_i = b_0 + \varepsilon_i \qquad i = 1, ..., n$$

Pour faire ce test on utilise la statistique

$$Z = \frac{SM(X_1, ..., X_p)/p}{SR(X_1, ..., X_p)/(n-p-1)} = \frac{(ST - SR)/p}{SR/(n-p-1)} \sim F(p, n-p-1)$$

Pour un niveau α fixé, la zone d'acceptation est

$$ZA = [0; f_{p,n-p-1;1-\alpha}]$$

Exemple. $f_{2,2,;0.95} = 19$, $ZA = [0; 19]$, $z = \frac{31.25}{5.75} = 5.4$

Apport d'une seule variable

Si l'ensemble des variables $X_1, ..., X_p$ est significatif dans la prévision de Y, on se pose la question d'effacer les variables qui ne servent pas à la prédiction de Y. Sans réduire la généralité, on suppose que l'on teste l'influence de X_p :
H_0 : X_p ne contribue pas de manière significative à la prédiction de Y si $X_1, ..., X_{p-1}$ sont déjà dans le modèle.
H_1 : X_p contribue de manière significative à la prédiction de Y si $X_1, ..., X_{p-1}$ sont déjà dans le modèle.

$$H_0 : b_p = 0 | b_j \neq 0, j \in \{1, ..., p-1\} \qquad \text{contre} \qquad H_1 : b_p \neq 0 | b_j \neq 0, j \in \{1, ..., p-1\}$$

Modèle complet :

$$Y_i = b_0 + b_1 X_{1i} + ... + b_{p-1} X_{p-1,i} + b_p X_{pi} + \varepsilon_i$$

Modèle réduit

$$Y_i = b_0 + b_1 X_{1i} + ... + b_{p-1} X_{p-1,i} + \varepsilon_i$$

Accepter H_0 signifie que le p^{eme} facteur n'apporte rien de plus après les $p-1$ premières variables. Mais ça ne signifie pas que ce facteur seul n'a pas d'effet sur Y (X_p peur être corrélé avec $X_1, ..., X_{p-1}$).
Pour tester l'hypothèse H_0 on utilise la statistique

$$Z = \frac{SM(X_1, ..., X_p) - SM(X_1, ..., X_{p-1})}{SR(X_1, ..., X_p)/(n-p-1)} \sim f(1, n-p-1)$$

où :
$SM(X_1, ..., X_p)$ est la somme des carrés due au modèle dans le modèle complet
$SM(X_1, ..., X_{p-1})$ est la somme des carrés due au modèle dans le modèle réduit.
Pour un risque α fixé, la zone d'acceptation est

$$ZA_{H_0,\alpha} = [0; f_{1,n-p-1;1-\alpha}]$$

Pour tester H_0, on peut utiliser aussi une statistique qui suit une loi de Student

$$Z = \frac{B_p}{\sqrt{\text{var}\,(B_p)}} \sim t(n-p-1)$$

où B_p est l'estimateur de b_p dans le modèle complet et var (B_p) est la variance de cet estimateur. La zone d'acceptation

$$ZA = \left[-t_{n-p-1;1-\alpha/2}; t_{n-p-1;1-\alpha/2}\right]$$

Autres critères de choix du modèle

Nous donnons ici trois autres critères de choix du nombre de variables à considérer. Pour chacun de ces trois critères, on choisira le modèle pour lequel le paramètre observé est le minimal. On note M_p le modèle complet (où les p variables $X_1, \ldots, X_p$ interviennent) et M_k le modèle de taille k.

– **Cp de Mallows** : On définit $C_p(M_k)$ par

$$C_p(M_k) = \frac{SR(M_k)}{SR(M_p)/(n-p-1)} + 2k - n = \frac{SR(M_k)}{\hat{\sigma}^2} + 2k - n$$

où $\hat{\sigma}^2$ est l'estimateur sans biais de σ^2 dans le modèle complet. Un modèle M_k de taille k sera considéré comme bon si son Cp de Mallows est proche de k. On privilégiera un modèle dont le Cp de Mallows est inférieur à k, avec k petit.

– **AIC** : On définit

$$AIC = n \log\left(\frac{SR(M_k)}{n}\right) + 2k$$

Plus le modèle est de grande taille, plus la somme des carrés résiduels est petite : on pénalise un modèle en fonction du nombre de ses paramètres. On choisit le modèle avec le moins de paramètres parmi ceux dont l'AIC est petit.

– **BIC** : On définit

$$BIC = n \log\left(\frac{SR(M_k)}{n}\right) + k \log(n)$$

Par rapport au critère AIC, on pénalise plus fortement les modèles de grande taille. On choisit le modèle comme avec le critère AIC.

Chapitre 3

ANALYSE DE VARIANCE

1 Analyse de variance à un facteur

1.1 Introduction

Exemple. Les 21 candidats à un oral ont été répartis au hasard entre 3 examinateurs. Le premier examinateur a fait passer l'oral à 6 étudiants, le second à 8 étudiants et le troisième à 7 étudiants. Les notes qu'ils ont eues sont :

Examinateur	**A**	**B**	**C**
	10,11,11,12,13,15	8,11,11,13,14,15,16,16	10,13,14,14,15,16,16
Effectif	6	8	7
Moyenne	12	13	14

On se demande si la variation des moyennes peut être due au hasard ou si elle tient d'un réel "effet d'examinateur".

En général, l'analyse de variance (ANOVA) est une technique statistique utilisée pour étudier l'effet des variables qualitatives sur une variable quantitative Y.

1.2 Terminologie

- *facteur (variable qualitative)* : une variable qui prend un nombre fini de valeurs, pas nécessairement numériques (une valeur constitue une classe). Pour l'exemple, on a le facteur "examinateur" qui prend 3 valeurs : A, B, C.
- *niveau (population)* : les différentes valeurs prises par un facteur.
- *test de l'effet d'un facteur* : tester si les moyennes des populations sont égales.

La variable à modéliser (à prévoir) Y, comme pour la régression linéaire, est une variable qui ne prend que des valeurs numériques.
Pour l'exemple :
- Y : notes ;
- facteur : examinateur ;
- niveaux : A,B, C.

On utilise un vocabulaire particulier, introduit par les agronomes, qui ont été les premiers à s'intéresser à ce type de problème : la variable qualitative susceptible d'influencer sur la distribution de la variable quantitative étudiée est appelée « facteur » et ses valeurs « populations ».

1.3 Données

On suppose qu'on a un seul facteur F et on dispose de k échantillons de tailles respectives $n_1, ..., n_k$, correspondant chacun à un niveau différent du facteur F. On pose

$$n = \sum_{i=1}^{k} n_i$$

À chaque expérience, on mesure la valeur de la variable Y. On peut alors présenter les données à l'aide du tableau suivant :

Niveau (population)	**Nb. obs.**	**Valeurs de Y**
1	n_1	$y_{11}, y_{12}, \ldots, y_{1n_1}$
2	n_2	$y_{21}, y_{22}, \ldots, y_{2n_2}$
$\vdots$	$\vdots$	.....
k	n_k	$y_{k1}, y_{k2}, ..., y_{kn_k}$

On remarque que le nombre d'observations pour chaque population peut ne pas être le même.
Notations : Pour un niveau i :

$$Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}$$

$$\overline{Y_{i\bullet}} = \frac{1}{n_i} Y_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

(la moyenne empirique des Y pour la population i)

$$Y_{\bullet\bullet} = \sum_{i=1}^{k} Y_{i\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

et

$$\overline{Y_{\bullet\bullet}} = \frac{1}{n} Y_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{k} Y_{i\bullet} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

Hypothèse : les k échantillons sont indépendants et de loi Normale. Plus précisément, on suppose que pour tout couple (i, j) les données y_{ij} sont des réalisations de la v.a. $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$ et Y_{ij}, $Y_{i'j'}$ indépendantes pour $i \neq i'$ ou $j \neq j'$.
Autrement dit, pour chaque i, les données $y_{i1}, ..., y_{in_i}$ sont des réalisations des n_i v.a. $Y_{i1}, ..., Y_{in_i}$ indépendantes et de même loi $\mathcal{N}(m_i, \sigma^2)$.
L'objet de cette étude sera de savoir si les moyennes m_i sont toutes égales ou non.

1.4 Modèles statistiques

Puisque $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$, on peut poser :

$$Y_{ij} = m_i + \varepsilon_{ij} \qquad i = 1, ..., k \qquad j = 1, ..., n_i \tag{3.1}$$

avec $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.
Paramètres à estimer : m_i la moyenne de la population i, σ^2 la variance.
Le modèle (3.1) peut être écrit sous une forme équivalente :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1, ..., k \qquad j = 1, ..., n_i \tag{3.2}$$

où :
- μ représente une valeur appelée « effet moyen » ;
- α_i représente l'effet du niveau i du facteur F.

Ainsi, on doit estimer $k + 2$ paramètres : μ et α_i $(i = 1, \ldots, k)$ ainsi que la variance σ^2.
Le modèle écrit sous la forme (3.2) a une indétermination, car $(\mu + \alpha_i)$ peut s'obtenir d'une infinité de manières. On remédie cela, en introduisant une contrainte, qui est en généralement la suivante : $\sum_{i=1}^{k} n_i \alpha_i = 0$.
En utilisant une notation vectorielle, le modèle (3.1) prend la forme :

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{bmatrix} = \begin{bmatrix} 1\ 0\ 0\ ...\ 0 \\ 1\ 0\ 0\ ...\ 0 \\ ..... \\ 1\ 0\ 0\ ...\ 0 \\ 0\ 1\ 0\ ...\ 0 \\ 0\ 1\ 0\ ...\ 0 \\ ..... \\ 0\ 1\ 0\ ...\ 0 \\ ..... \\ ..... \\ 0\ 0\ 0\ ...\ 1 \\ ..... \\ 0\ 0\ 0\ ...\ 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ \vdots \\ m_k \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{bmatrix} \tag{3.3}$$

ou encore

$$Y = X\beta + \varepsilon \tag{3.4}$$

Donc, l'analyse de variance est un modèle linéaire.

1.5 Estimation des paramètres

Pour les modèles (3.1) ou (3.3), il faut trouver les valeurs des m_i qui minimisent la fonction :

$$T(m_i) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - m_i)^2$$

En faisant des calculs, on obtient que : $\hat{m}_i = \bar{Y}_{i\bullet}$. Sous l'hypothèse de normalité et d'indépendance des échantillons, $\bar{Y}_{i\bullet}$ est un estimateur sans biais de m_i et

$$\hat{m}_i = \bar{Y}_{i\bullet} \sim \mathcal{N}\left(m_i, \frac{\sigma^2}{n_i}\right)$$

Pour le modèle (3.2), les paramètres à estimer sont : μ et les α_i, $i = 1, ..., k$.

$$\varepsilon_{ij} = \bar{\varepsilon}_{\bullet\bullet} + (\bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet\bullet}) + (\varepsilon_{ij} - \bar{\varepsilon}_{i\bullet})$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{\varepsilon_{\cdot\cdot}^2} + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{\varepsilon_{i\cdot}} - \overline{\varepsilon_{\cdot\cdot}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\varepsilon_{ij} - \overline{\varepsilon_{i\cdot}})^2 \tag{3.5}$$

On écrit les ε fonction des paramètres à estimer :

$$\varepsilon_{ij} = Y_{ij} - \mu - \alpha_i, \qquad \varepsilon_{i\cdot} = Y_{i\cdot} - n_i\mu - n_i\alpha_i, \qquad \overline{\varepsilon_{i\cdot}} = \overline{Y_{i\cdot}} - \mu - \alpha_i$$

$$\varepsilon_{\cdot\cdot} = Y_{\cdot\cdot} - \sum_{i=1}^{k} n_i\mu - \sum_{i=1}^{k} n_i\alpha_i, \qquad \varepsilon_{\cdot\cdot} = Y_{\cdot\cdot} - n\mu, \qquad \overline{\varepsilon_{\cdot\cdot}} = \overline{Y_{\cdot\cdot}} - \mu$$

alors la relation (3.5) devient :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{\cdot\cdot}} - \mu)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{i\cdot}} - \overline{Y_{\cdot\cdot}} - \alpha_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{i\cdot}})^2 \tag{3.6}$$

Le membre droit de (3.6) est minimisé pour :

$$\hat{\mu} = \overline{Y_{\cdot\cdot}}, \qquad \hat{\alpha}_i = \overline{Y_{i\cdot}} - \overline{Y_{\cdot\cdot}}$$

Il faut vérifier que : $\sum_{i=1}^{k} n_i\hat{\alpha}_i = 0$:

$$\sum_{i=1}^{k} n_i\alpha_i = \sum_{i=1}^{k} n_i\overline{Y_{i\cdot}} - \sum_{i=1}^{k} n_i\overline{Y_{\cdot\cdot}} = \sum_{i=1}^{k} Y_{i\cdot} - n\overline{Y_{\cdot\cdot}} = 0$$

L'estimateur du maximum de vraisemblance modifié pour σ^2 est :

$$S_n^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

1.6 Tests d'hypothèses

Tableau d'analyse de variance

On veut d'abord tester l'hypothèse qu'il n'y a pas k niveaux (populations) différents, mais qu'ils sont tous confondus : les n observations proviennent d'une population unique d'espérance m. Pour le modèle (3.1) ou (3.3), l'hypothèse nulle a la forme :
$H_0 : m_1 = m_2 = ... = m_k = m$
contre
$H_1 : \exists i, j \in \{1, ..., k\}$ tels que $m_i \neq m_j$.
Ou, équivalent pour le modèle (3.2) :
$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$
contre
$H_1 : \exists i \in \{1, ..., k\}$ tel que $\alpha_i \neq 0$

Sous l'hypothèse H_0, le modèle a la forme :

$$M_{reduit} : Y_{ij} = \mu + \varepsilon_{ij}$$

L'estimation pour μ est $\hat{\mu} = \bar{Y}_{\cdot\cdot}$ et la prévision de Y_{ij} est $\hat{Y}_{ij} = \hat{\mu}$. Alors, le résidu, sous l'hypothèse H_0 est : $Y_{ij} - \bar{Y}_{\cdot\cdot}$. La variabilité totale est :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{\cdot\cdot}})^2$$

On peut écrire : $Y_{ij} - \overline{Y_{..}} = (Y_{ij} - \overline{Y_{i.}}) + (\overline{Y_{i.}} - \overline{Y_{..}})$ et par des calculs élémentaires, on obtient :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{..}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{i.}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{i.}} - \overline{Y_{..}})^2$$

(variabilité totale=variabilité résiduelle + variabilité due au modèle) : ST=SR+SM.
On peut résumer cette décomposition par le tableau d'analyse de variance :

Source de variation	ddl	S.C.	Carré moyen
Régression	$k-1$	SM	$SM/(k-1)$
Résiduelle	$n-k$	SR	$SR/(n-k)$
Totale	$n-1$	ST	

Test d'égalité des k effets

Pour tester l'hypothèse H_0 on utilise la statistique :

$$Z = \frac{SM/(k-1)}{SR/(n-k)} \sim F(k-1, n-k) \qquad (\text{sous } H_0)$$

Pour un risque α fixé, la zone d'acceptation est : $ZA_{H_0,\alpha} = [0 \;\; f_{k-1,n-k;1-\alpha}]$

Exemple. Les modèles attachés :

$$Y_{ij} = m_i + \varepsilon_{ij}, \qquad i = 1,2,3, \qquad j = 1,\ldots,n_i, \;\; n_1 = 6 \;\; n_2 = 8 \;\; n_3 = 7 \tag{3.7}$$

ou encore

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1,2,3, \qquad j = 1,\ldots,n_i, \;\; n_1 = 6 \;\; n_2 = 8 \;\; n_3 = 7 \tag{3.8}$$

Les estimations des paramètres : $\hat{\mu} = \bar{y}_{..} = 13.04$, $\hat{\alpha}_1 = \bar{y}_{1.} - \bar{y}_{..} = 12 - 13.04 = -0.96$, $\hat{\alpha}_2 = \bar{y}_{2.} - \bar{y}_{..} = 13 - 13.04 = -0.04$, $\hat{\alpha}_3 = \bar{y}_{3.} - \bar{y}_{..} = 14 - 13.04 = 0.96$. On veut tester s'il y a un effet examinateur : les examinateurs ont-ils le même système de notation ?
$H_0 : m_1 = m_2 = m_3 = m$ contre $H_1 : \exists i \neq j$ tel que $m_i \neq m_j$
$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ contre $H_1 : \exists i \neq j$ tel que $\alpha_i \neq 0$
On obtient : SM=12.95, SR=98 donc $z = (SM/(3-1))/(SR/(21-3)) = 1.19$.
La zone d'acceptation est $ZA_{H_0;1-\alpha} = [0\,; f_{2,18;0.096}] = [0\,; 3.55]$.
Donc, H_0 est acceptée : les examinateurs ont le même système de notation.

Comparaison de moyennes

Le rejet de l'hypothèse d'égalité des moyennes ne signifie pas que tous les m_i sont différents entre eux. On cherche souvent à tester l'égalité entre deux moyennes :
$H_0 : m_h = m_j$ contre $H_1 : m_h \neq m_j$ pour $h \neq j$.
On utilise la statistique de test :

$$Z = \frac{\bar{Y}_{h.} - \bar{Y}_{j.}}{\sqrt{\frac{SR}{n-k}}\sqrt{\frac{1}{n_h} + \frac{1}{n_j}}} \sim t(n-k)$$

La zone d'acceptation $ZA_{H_0,1-\alpha} = [-t_{n-k;1-\alpha/2}\,; t_{n-k;1-\alpha/2}]$.

2 Analyse de variance à deux facteurs

2.1 Introduction

On a vu comment comparer les populations d'un même facteur. Supposons maintenant qu'un expérimentateur souhaite comparer l'influence de trois régimes alimentaires et de deux exploitations sur la production laitière. Les résultats expérimentaux sont dans le tableau suivant.

Expl ↓ R.alim →	**A**	**B**	**C**	**Total**	**Moyenne**
1	7	36	2	45	15
2	13	44	18	75	215
Total	20	80	20	120	
Moyenne	10	40	10		20

2.2 Données

On suppose qu'on a deux facteurs (variables) F1 et F2. Le nombres de niveaux (valeurs possibles) pour F1 est de p et pour F2 est de q. Pour chaque couple (i, j) de niveaux on a $r(\geq 1)$ observations de la variable dépendante Y. Alors, on peut présenter les données à l'aide du tableau suivant :

F1 / F2	**1**	.....	**i**	.....	**p**
1	$y_{111}, \ldots, y_{11r}$	.....	$y_{i11}, \ldots, y_{i1r}$	.....	$y_{p11}, \ldots, y_{p1r}$
⋮	⋮	⋮	⋮	⋮	⋮
j	$y_{1j1}, \ldots, y_{1jr}$	.....	$y_{ij1}, \ldots, y_{ijr}$	.....	$y_{pj1}, \ldots, y_{pjr}$
⋮	⋮	⋮	⋮	⋮	⋮
q	$y_{1q1}, \ldots, y_{1jq}$	.....	$y_{iq1}, \ldots, y_{iqr}$	.....	$y_{pq1}, \ldots, y_{pqr}$

Dans la cellule (i, j) nous avons les valeurs (observations) y_{ijk} : i donne le niveau (population) du facteur $F1$, j le niveau de $F2$ et k la répétition pour un couple (i, j). On a pq cellules et dans chaque cellule il y a r observations.
Notations :

$$\begin{cases} y_{ij\bullet} = \sum_{k=1}^{r} y_{ijk} & \bar{y}_{ij\bullet} = \frac{1}{r} y_{ij\bullet} \\ y_{i\bullet\bullet} = \sum_{j=1}^{q} \sum_{k=1}^{r} y_{ijk} & \bar{y}_{i\bullet\bullet} = \frac{1}{qr} y_{\bullet\bullet} \\ y_{\bullet j\bullet} = \sum_{i=1}^{p} \sum_{k=1}^{r} y_{ijk} & \bar{y}_{\bullet j\bullet} = \frac{1}{pr} y_{\bullet j\bullet} \\ y_{\bullet\bullet\bullet} = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{r} y_{ijk} & \bar{y}_{\bullet\bullet\bullet} = \frac{1}{pqr} y_{\bullet\bullet\bullet} \end{cases}$$

Les observations y_{ijk} sont des réalisations de la v.a. Y_{ijk} sur laquelle on fait les hypothèses :

$$\begin{cases} Y_{ijk} \sim \mathcal{N}(m_{ij}, \sigma^2) & \forall k = 1, \ldots, r \\ Y_{ijk}, Y_{i'j'k'} & \text{indépendantes} \end{cases}$$

En ce qui concerne le nombre r de répétitions on a 2 situations :
- $r > 1$
- $r = 1$. Il n'y a pas de répétition et on va noter $Y_{ij\bullet}$ par Y_{ij}.

Alors, les modèles statistiques considérés seront fonction de ces 2 situations. Les problèmes à traiter seront les mêmes que pour un seul facteur :
- écrire un modèle statistique de Y fonction des facteurs ;
- estimer les effets des niveaux des deux facteurs ;
- test d'hypothèse

2.3 Modèle sans interaction (additif) : r=1

Le modèle le plus simple est d'additionner les effets du facteur $F1$ avec les effets du facteur $F2$:

$$m_{ij} = \mu + \alpha_i + \beta_j \tag{3.9}$$

où :
- μ est l'effet moyen
- α_i est l'effet dû au niveau i du facteur $F1$;
- β_j est l'effet dû au niveau j du facteur $F2$;

Puisque $Y_{ijk} \sim \mathcal{N}(m_{ij}, \sigma^2)$ on peut considérer un modèle :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \tag{3.10}$$

Ce dernier modèle est indéterminé, car on peut obtenir la relation (3.9) par une infinité de manières. On remédie à cela, en introduisant des contraintes, par exemple :

$$\sum_{i=1}^{p} \alpha_i = 0 \qquad \sum_{j=1}^{q} \beta_j = 0$$

Estimation des paramètres

Il faut trouver les valeurs de m_{ij} (ou de μ, α_i, β_j) qui minimisent la fonction :

$$T(m_{ij}) = \sum_{i=1}^{p}\sum_{j=1}^{q} \varepsilon_{ij}^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} (Y_{ij} - m_{ij})^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} (Y_{ij} - \mu - \alpha_i - \beta_j)^2 \tag{3.11}$$

On utilise la même technique que pour l'analyse de variance à un facteur, et on obtient :

$$\hat{\alpha}_i = \overline{Y_{i\cdot}} - \overline{Y_{\cdot\cdot}} \qquad \hat{\beta}_i = \overline{Y_{\cdot j}} - \overline{Y_{\cdot\cdot}} \qquad \hat{\mu} = \overline{Y_{\cdot\cdot}}$$

La valeur prédite pour Y_{ij} est :

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \overline{Y_{i\cdot}} + \overline{Y_{\cdot j}} - \overline{Y_{\cdot\cdot}}$$

Exemple. F1 est le régime alimentaire, qui prend 3 valeurs (A, B, C), donc $p = 3$. $F2$ est l'exploitation, qui prend 2 valeurs (1 et 2), donc $q = 2$. Le modèle statistique est :

$$Y_{ij} = \mu + \alpha_i + \beta_j \qquad i = 1, 2, 3 \quad j = 1, 2$$

où : α_1 est l'effet de l'exploitation n° 1 sur Y, β_1 est l'effet du régime A sur la production laitière... Les estimations des paramètres sont : $\hat{\mu} = \bar{y}_{\cdot\cdot} = 20$, $\hat{\alpha}_1 = \bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot} = 10 - 20 = -10$, $\hat{\alpha}_2 = 20$, $\hat{\alpha}_3 = -10$, $\hat{\beta}_1 = \bar{y}_{\cdot 1} - \bar{y}_{\cdot\cdot} = 15 - 20 = -5$, $\hat{\beta}_2 = 5$. La prévision de Y_{11} (pour le régime alimentaire A et l'exploitation 1) : $\hat{Y}_{11} = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_1 = 20 - 10 - 5 = 5$.

Tableau d'analyse de variance

En partant de l'identité :

$$Y_{ij} - \overline{Y_{\cdot\cdot}} = (Y_{ij} - \overline{Y_{i\cdot}} - \overline{Y_{\cdot j}} + \overline{Y_{\cdot\cdot}}) + (\overline{Y_{i\cdot}} - \overline{Y_{\cdot\cdot}}) + (\overline{Y_{\cdot j}} - \overline{Y_{\cdot\cdot}})$$

On obtient :

$$\sum_{i,j}(Y_{ij}-\overline{Y_{\bullet\bullet}})^2=\sum_{i,j}(Y_{ij}-\overline{Y_{i\bullet}}-\overline{Y_{\bullet j}}+\overline{Y_{\bullet\bullet}})^2+q\sum_{i=1}^{p}(\overline{Y_{i\bullet}}-\overline{Y_{\bullet\bullet}})^2+p\sum_{j=1}^{q}(\overline{Y_{\bullet j}}-\overline{Y_{\bullet\bullet}})^2$$

ou encore $ST = SR + S_{F1} + S_{F2}$. On peut résumer cette décomposition par le tableau d'analyse de variance :

Source de variation	ddl	S.C.	Carré moyen
F1	$p-1$	S_{F1}	$S_{F1}/(p-1)$
F2	$q-1$	S_{F2}	$S_{F2}/(q-1)$
Résidu	$(p-1)(q-1)$	SR	$SR/(p-1)(q-1)$
Totale	$pq-1$	ST	

Test d'hypothèse

On peut tester deux types d'hypothèse : si le modèle significatif ou l'effet de chaque facteur.

Test du modèle. Le modèle n'est pas significatif si aucun des deux facteurs n'influence Y :
$H_0 : \alpha_1 = \cdots = \alpha_p = \beta_1 = \cdots = \beta_q = 0$
contre :
$H_1 : \exists i \in \{1,\ldots,p\}$ ou $\exists j \in \{1,\ldots,q\}$ t.q. $\alpha_i \neq 0$ ou $\beta_i \neq 0$.
Le modèle complet est (3.10) et le modèle réduit : $Y_{ij} = \mu + \varepsilon_{ij}$.
Statistique de test :

$$Z=\frac{(S_{F1}+S_{F2})/(p+q-2)}{SR/(p-1)(q-1)}\sim F(p+q-2,(p-1)(q-1)) \qquad \text{sous } H_0$$

Test d'un facteur. Supposons que l'on veut tester l'effet de F1.
H_0 : F1 n'influe pas Y sachant que F2 est dans le modèle.
$H_0 : \alpha_1 = \cdots = \alpha_p = 0$ contre $H_1 : \exists i \in \{1,\ldots,p\}$ t.q. $\alpha_i \neq 0$.
Le modèle complet est (3.10) et le modèle réduit : $Y_{ij} = \mu + \beta_j + \varepsilon_{ij}$. (modèle à un facteur)
L'hypothèse H_0 peut être traduite sous la forme : la moyenne m_{ij} ne dépend pas de i. Statistique de test :

$$Z=\frac{(S_{F1})/(p-1)}{SR/(p-1)(q-1)}\sim F(p-1,(p-1)(q-1)) \qquad \text{sous } H_0$$

Exemple. Le tableau d'analyse de variance est :

Source de variation	ddl	S.C.	Carré moyen
F1	2	1200	600
F2	1	150	150
Résidu	2	28	14
Totale	5	1378	

On teste si le modèle est significatif : $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = 0$:

$$Z=\frac{(S_{F1}+S_{F2})/(3+2-2)}{SR/2}\sim F(3,2) \qquad \text{sous } H_0$$

$ZA = [0\ ; f_{3,2:0.95}] = [0\ ; 19.2]$, $z = \frac{1350/3}{14} = 32.1 \notin ZA$. Donc H_0 est rejetée et le modèle est significatif.
On teste si le facteur régime alimentaire agit sur la production laitière : $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ sachant que l'exploitation est dans le modèle. L'hypothèse alternative est $H_1 : \exists i \in \{1, 2, 3\}$ t.q. $\alpha_i \neq 0$.
Le modèle sous H_0 est $Y_{ij} = \mu + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, 3,\ j = 1, 2$.
La statistique de test $Z = \frac{S_{F1}/2}{SR/2}$ suit la loi $F(2, 2)$ sous H_0.
$ZA = [0\ ; f_{2,2:0.95}] = [0\ ; 19.0]$, $z = \frac{600}{14} = 42.86 \notin ZA$.
Donc H_0 est rejetée, le régime alimentaire est un facteur influent pour la production laitière.