

Modèles Linéaires

Correction TD 6 : ANOVA Licence 3 MIASHS


Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Résumé

L'objectif ambitionné dans cette fiche est d'étudier l'Analyse de Variance d'un point de vue géométrique à l'aide des outils d'algèbre linéaire en partant d'un exemple pratique. Plus précisément

- étudier un cas pratique d'Analyse de Variance,
- construction du test d'hypothèses à l'aide de l'algèbre linéaire,
- interprétation des résultats,
- effectuer une analyse à l'aide des fonctions de .

1 Exposition des à un produit toxique dans une usine

On s'intéresse à un problème de santé en milieu industriel et plus précisément à l'exposition à un produit toxique dans une usine fabriquant des produits chimiques. Cette usine dispose de 4 ateliers et le plan proposé pour étudier l'exposition à la toxicité des produits est le suivant : on analyse un certain indicateur sur 20 personnes prises au hasard dans chaque atelier. Le résultat des analyses révèle un résultat entier noté x qui indique le degré d'exposition auquel la personne a été soumise au cours de son travail. Les résultats sont résumés dans le tableau suivant.

	Atelier 1	Atelier 2	Atelier 3	Atelier 4
y=7	5	0	1	7
y=8	4	4	8	5
y=9	3	4	3	8
y=10	4	5	5	0
y=11	4	7	3	0

La première question qui vient à l'esprit est bien sûr de se demander si les ateliers sont équivalents du point de vue du risque d'exposition de ses employés. Plusieurs techniques sont possibles pour décider si tel est bien le cas. C'est ce que nous allons étudier dans la suite.

2 Hypothèse gaussienne et analyse de la variance

On peut commencer comme on le fait souvent par supposer que les résultats x sont des réalisations de variables approximativement gaussiennes Y_{ij} où i est le numéro de l'employé i dans l'atelier j . Une façon de tester si tel est le cas est le test du χ^2 par exemple ou le test de Kolmogorov-Smirnov ou encore le test de Shapiro-Wilks. Le modèle gaussien qui va nous permettre de mettre en évidence une différence entre les ateliers est le suivant.

On suppose que la loi de Y_{ij} est gaussienne $\mathcal{N}(\mu_j, \sigma^2)$ où la variance ne dépend ni de l'atelier ni de l'employé. L'idée est donc de tester l'hypothèse suivante :

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{v.s.} \quad H_1 : \exists j \text{ s.t. } \mu_i \neq \mu_j.$$

On va chercher une variable pivotale dont la loi sera connue et qui nous permettra, si elle est trop grande par exemple, de conclure que l'hypothèse H_0 est fausse avec un certain risque prescrit à l'avance.

On appelle \mathbf{y} le vecteur des observations, obtenu en mettant les ateliers "les uns sous les autres", c'est à dire

$$\mathbf{y} = [y_{1,1}, \dots, y_{20,1}, y_{1,2}, \dots, y_{20,2}, y_{1,3}, \dots, y_{20,3}, y_{1,4}, \dots, y_{20,4}]^\top$$

où \cdot^\top désigne la transposition. Sans supposer H_0 , notre modèle nous impose donc que

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{1}$$

ou, plus précisément

$$\mathbf{y} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} + \boldsymbol{\varepsilon}$$

où $\boldsymbol{\varepsilon}$ est un vecteur de composantes indépendantes gaussiennes $\mathcal{N}(0, \sigma^2)$. Dans le cas où H_0 est satisfaite, si on note $\mu = \mu_1 = \dots = \mu_4$, on a

$$\mathbf{y} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \boldsymbol{\varepsilon}.$$

Introduisons les notations suivantes

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \quad \text{et} \quad \mathbf{e} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Dans le cas général où on ne suppose pas H_0 , on a

$$\mathbb{E}[Y] = X\mu$$

ce qui montre donc que $\mathbb{E}[Y]$ est dans l'espace image de dimension 4 de la matrice \mathbf{X} que nous noterons V . De même, si H_0 est vérifiée,

$$\mathbb{E}[Y] = \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix} \mu$$

et donc, $\mathbb{E}[Y]$ est dans l'espace image de dimension 1 de la matrice constituée d'une seule colonne de 1, *i.e.*, de notre vecteur *unité*.

3 Construction théorique du test de Fisher

Cette section ne contient pas de question mais explique pourquoi on applique, au final, un test de Fisher. Elle vous montre aussi l'importance de l'algèbre linéaire pour ce type d'analyse.

Rappelons que V est l'espace engendré par les colonnes de la matrice de design \mathbf{X} . Notons U l'espace image de dimension 1 de la matrice constituée d'une seule colonne de 1. Notons également W le **supplémentaire orthogonal** de U dans V , c'est à dire tous les vecteurs de V qui sont orthogonaux à tous les vecteurs de U .

L'idée est alors d'étudier les projections de \mathbf{y} sur U , sur W et sur V^\perp , le supplémentaire orthogonal à V dans l'espace total \mathbb{R}^{80} , et de construire une variable pivotale à partir de ce qu'on va découvrir sur la loi de ces projections. Le résultat suivant est la clé de l'étude.

Notations. Si H est un sous espace vectoriel de \mathbb{R}^n et \mathbf{z} un vecteur de \mathbb{R}^n , on notera dans toute la suite par $\mathbf{p}_H(\mathbf{z})$ la projection orthogonale de \mathbf{z} sur H . On notera par \mathbf{I} la matrice identité.

3.1 Un résultat sur les projections des vecteurs Gaussiens

Il y a un résultat vraiment très utile et qui relie l'algèbre linéaire avec les probabilités dans le cas des vecteurs dont les composantes sont *i.i.d.* $\mathcal{N}(0, 1)$: c'est le théorème de Cochran [[Cochran, 1934](#)].

Théorème 3.1: Théorème de Cochran

Soit Z un vecteur gaussien dans \mathbb{R}^n de loi $\mathcal{N}(0, I)$ et soit H un sous-espace vectoriel de \mathbb{R}^n de dimension d . Alors

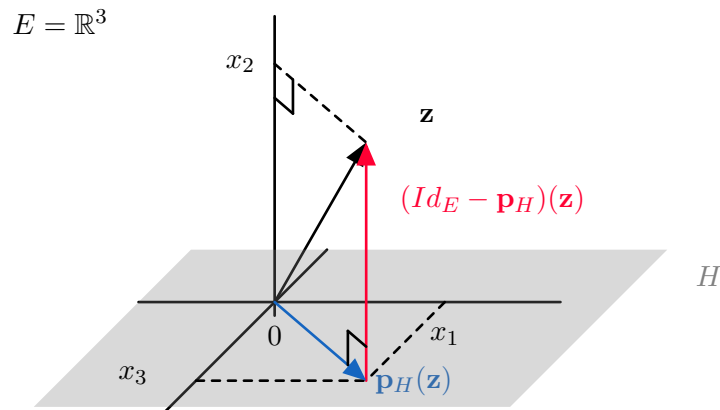
$$\|\mathbf{p}_H(Z)\|^2 \text{ suit une loi } \chi^2(d).$$

Maintenant, si H_1 et H_2 sont deux sous-espaces vectoriels orthogonaux de dimensions respectives d_1 et d_2 . Alors $\mathbf{p}_{H_1}(Z)$ et $\mathbf{p}_{H_2}(Z)$ sont indépendants.

3.2 Formule explicite de la projection sur H

La projection sur H s'exprime de manière très explicite en utilisant une base de H comme le suit : soit $\mathbf{A} \in \mathbb{R}^{n \times d}$ une matrice dont les colonnes forment une base de H . Alors, la matrice $\mathbf{A}^\top \mathbf{A}$ est inversible et la projection orthogonale $\mathbf{p}_H(\mathbf{z})$ de \mathbf{z} sur H est donnée par

$$\mathbf{p}_H(\mathbf{z}) = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{z}.$$



En effet, calculer la projection de \mathbf{z} sur H , connaissant une base de H peut s'exprimer comme le problème de trouver un vecteur w de la forme

$$w = \mathbf{A}\beta$$

c'est à dire combinaison linéaire des colonnes de \mathbf{A} , et qui soit le plus proche possible de \mathbf{z} : cela revient à résoudre

$$\min_{\beta} \|\mathbf{z} - \mathbf{A}\beta\|_2^2$$

On connaît déjà la solution à ce problème :

$$\hat{\boldsymbol{\beta}} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{z}$$

et donc le meilleur \mathbf{w} est $\mathbf{A}\hat{\boldsymbol{\beta}}$, c'est à dire $\mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{z}$.

Dans le cas de notre problème, nous avons les propriétés suivantes qui permettront de construire des estimateurs utiles et d'appliquer le Théorème 3.1. En effet, sous les hypothèses du modèle (1) on a

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

donc $\mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\mu}$ et ainsi $\mathbb{E}[\mathbf{y}] \in V$, l'espace engendré par les colonnes de \mathbf{X} . nous avons ainsi,

$$\mathbf{p}_V(\mathbf{y}) = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

et donc

$$\mathbb{E}[\mathbf{p}_V(\mathbf{y})] = \mathbb{E}[\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}] = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{y}] = \mathbf{p}_V(\mathbb{E}[\mathbf{y}]).$$

En résumé

$$\mathbb{E}[\mathbf{p}_V(\mathbf{y})] = \mathbf{p}_V(\mathbb{E}[\mathbf{y}]) = \mathbb{E}[\mathbf{y}].$$

Mais, si l'hypothèse nulle est vraie, on a

$$\mathbb{E}[\mathbf{p}_U(\mathbf{y})] = \mathbb{E}[\mathbf{y}].$$

On en déduit que

- $\mathbf{p}_V(\mathbf{y})$ est un estimateur non biaisé de $\mathbb{E}[\mathbf{y}]$
- sous l'hypothèse H_0 , $\mathbf{p}_U(\mathbf{y})$ est un estimateur non biaisé de $\mathbb{E}[\mathbf{y}]$.

3.3 Test de Fisher d'appartenance à un sous-espace

Construisons maintenant le test. La projection de \mathbf{y} sur V^\top est

$$\mathbf{p}_{V^\top}(\mathbf{y}) = \mathbf{y} - \mathbf{p}_V(\mathbf{y}).$$

De plus, par le lemme précédent,

$$\mathbb{E}[\mathbf{p}_{V^c}(\mathbf{y})] = \mathbb{E}[\mathbf{y} - \mathbf{p}_V(\mathbf{y})] = \mathbb{E}[\mathbf{y}] - \mathbb{E}[\mathbf{p}_V(\mathbf{y})] = 0$$

et donc le Théorème 3.1 s'applique. Ainsi, comme

$$\begin{aligned} \mathbf{y} - \mathbf{p}_V(\mathbf{y}) &= \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon} - \mathbf{p}_V(\mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}), \\ &= \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon} - P_V(\mathbf{X}\boldsymbol{\mu}) - \mathbf{p}_V(\boldsymbol{\varepsilon}), \\ &\quad \downarrow \text{comme } \mathbf{p}_V(\mathbf{X}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\mu} \\ &= \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{\mu} - \mathbf{p}_V(\boldsymbol{\varepsilon}), \\ &= \boldsymbol{\varepsilon} - \mathbf{p}_V(\boldsymbol{\varepsilon}), \\ \mathbf{y} - \mathbf{p}_V(\mathbf{y}) &= \mathbf{p}_{V^\perp}(\boldsymbol{\varepsilon}) \end{aligned}$$

et donc

$$\mathbf{y} - \mathbf{p}_V(\mathbf{y}) = \boldsymbol{\varepsilon} - \mathbf{p}_V(\boldsymbol{\varepsilon}) = \mathbf{p}_{V^\perp}(\boldsymbol{\varepsilon}).$$

ce qui donne au final, comme $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$,

$$\frac{\|\mathbf{y} - \mathbf{p}_V(\mathbf{y})\|_2^2}{\sigma^2} = \|\mathbf{p}_{V^\perp}(\boldsymbol{\varepsilon}/\sigma)\|_2^2 \sim \chi^2(n-4).$$

D'autre part, la projection de \mathbf{y} sur W est $\mathbf{p}_V(\mathbf{y}) - \mathbf{p}_U(\mathbf{y})$.

Sous l'hypothèse H_0 , cette projection a une espérance nulle et le Théorème 3.1 s'applique encore. On obtient alors

$$\begin{aligned} \frac{\|\mathbf{p}_V(\mathbf{y}) - \mathbf{p}_U(\mathbf{y})\|_2^2}{\sigma^2} &= \frac{\|\mathbf{p}_V(\mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}) - \mathbf{p}_U(\mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon})\|_2^2}{\sigma^2}, \\ &= \frac{\|\mathbf{p}_V(\mathbf{X}\boldsymbol{\mu}) + \mathbf{p}_V(\boldsymbol{\varepsilon}) - \mathbf{p}_U(\mathbf{X}\boldsymbol{\mu}) - \mathbf{p}_U(\boldsymbol{\varepsilon})\|_2^2}{\sigma^2}, \\ &= \frac{\|\mathbf{X}\boldsymbol{\mu} + \mathbf{p}_V(\boldsymbol{\varepsilon}) - \mathbf{X}\boldsymbol{\mu} - \mathbf{p}_U(\boldsymbol{\varepsilon})\|_2^2}{\sigma^2}, \\ &= \frac{\|\mathbf{p}_V(\boldsymbol{\varepsilon}) - \mathbf{p}_U(\boldsymbol{\varepsilon})\|_2^2}{\sigma^2}, \\ &= \|\mathbf{p}_W(\boldsymbol{\varepsilon}/\sigma)\|_2^2 \sim \chi^2(4-1). \end{aligned}$$

De plus, comme W et V^\perp sont orthogonaux, le Théorème 3.1 implique que les vecteurs $\mathbf{y} - \mathbf{p}_V(\mathbf{y})$ et $\mathbf{p}_V(\mathbf{y}) - \mathbf{p}_U(\mathbf{y})$ sont indépendants.

Ceci nous permet donc de conclure au résultat suivant

Proposition 3.1. *Test de Fisher*

Dans les conditions du modèle, sous l'hypothèse H_0 on a

$$\frac{\|\mathbf{y} - \mathbf{p}_V(\mathbf{y})\|^2/(n-4)}{\|\mathbf{p}_V(\mathbf{y}) - \mathbf{p}_U(\mathbf{y})\|^2/(4-1)} \sim F(n-4, 4-1)$$

où $F(d_1, d_2)$ est le symbole de la loi de Fisher à d_1 et d_2 degrés de liberté.

Nous avons donc obtenu notre variables pivotale : le rapport des normes au carrées des projections de \mathbf{y} sur V^\perp et W . Nous pouvons faire ensuite faire le test permettant de valider H_0 ou de la contredire. On considère la réalisation \mathbf{y} du vecteur à notre disposition. On choisit alors le risque de première espèce α puis on cherche le quantile $f_{1-\alpha}$ de la loi de Fisher $F(n-4, 4-1)$. On calcule alors le rapport

$$r = \frac{\|\mathbf{y} - \mathbf{p}_V(\mathbf{y})\|^2/(n-4)}{\|\mathbf{p}_V(\mathbf{y}) - \mathbf{p}_U(\mathbf{y})\|^2/(4-1)}$$

Maintenant, le test de Fisher est donné par

Test de Fisher sur l'hypothèse $H_0 : \mu_1 = \dots = \mu_4 :$

- si $r > f_{1-\alpha}$, on rejette H_0 ,
- sinon H_0 n'est pas contredite.

Ce test est appelé l'analyse de la variance. Cette dénomination est à première vue paradoxale car le but est tout de même de savoir si les espérances sont égales. Le calcul explicite du rapport montre que l'on est en fait en train de comparer des variances, ce qui résout ce paradoxe apparent.

4 Mise en pratique

4.1 Calcul explicite de r

Rappelons que

$$r = \frac{\|\mathbf{y} - \mathbf{p}_V(\mathbf{y})\|^2/(n-4)}{\|\mathbf{p}_V(\mathbf{y}) - \mathbf{p}_U(\mathbf{y})\|^2/(4-1)}$$

et que

- V est l'image de la matrice \mathbf{X} et la projection de \mathbf{y} sur l'image de \mathbf{X} est donnée par

$$\mathbf{p}_V(\mathbf{y}) = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- U est l'image du vecteur $\mathbf{X} \cdot \mathbf{e}$ où \mathbf{e} est le vecteur dont les composantes sont des 1 et la projection est donnée par

$$P_V(\mathbf{y}) = \mathbf{e}(\mathbf{e}^\top \mathbf{e})^{-1} \mathbf{e}^\top \mathbf{y}.$$

L'objectif est de calculer les formules explicites des projections $\mathbf{p}_V(\mathbf{y})$ et $\mathbf{p}_U(\mathbf{y})$.

1. Montrer que

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{pmatrix} n_1^{-1} & 0 & 0 & 0 \\ 0 & n_2^{-1} & 0 & 0 \\ 0 & 0 & n_3^{-1} & 0 \\ 0 & 0 & 0 & n_4^{-1} \end{pmatrix}$$

et que

$$\mathbf{X}^\top \mathbf{y} = \begin{pmatrix} y_1 + \dots + y_{n_1} \\ y_{n_1+1} + \dots + y_{n_1+n_2} \\ y_{n_1+n_2+1} + \dots + y_{n_1+n_2+n_3} \\ y_{n_1+n_2+n_3+1} + \dots + y_{n_1+n_2+n_3+n_4} \end{pmatrix}$$

Réponse : on a

$$X^t X = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

et cela donne

$$X^t X = \begin{pmatrix} n_1 & 0 & 0 & 0 \\ 0 & n_2 & 0 & 0 \\ 0 & 0 & n_3 & 0 \\ 0 & 0 & 0 & n_4 \end{pmatrix}$$

et donc

$$(X^t X)^{-1} = \begin{pmatrix} n_1^{-1} & 0 & 0 & 0 \\ 0 & n_2^{-1} & 0 & 0 \\ 0 & 0 & n_3^{-1} & 0 \\ 0 & 0 & 0 & n_4^{-1} \end{pmatrix}.$$

Occupons nous aussi de $X^t Y$:

$$X^t Y = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ y_{n_1+n_2+1} \\ \vdots \\ y_{n_1+n_2+n_3} \\ y_{n_1+n_2+n_3+1} \\ \vdots \\ y_{n_1+n_2+n_3+n_4} \end{pmatrix}$$

ce qui donne la formule

$$X^t y = \begin{pmatrix} y_1 + \cdots + y_{n_1} \\ y_{n_1+1} + \cdots + y_{n_1+n_2} \\ y_{n_1+n_2+1} + \cdots + y_{n_1+n_2+n_3} \\ y_{n_1+n_2+n_3+1} + \cdots + y_{n_1+n_2+n_3+n_4} \end{pmatrix}$$

2. Montrer que la projection $\mathbf{p}_V(\mathbf{y})$ est donnée par la formule :

$$\mathbf{p}_V(\mathbf{y}) = \mathbf{X} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \bar{y}_4 \end{pmatrix}$$

et donc la projection consiste à remplacer toutes les réponse par la moyenne des réponses dans le groupe auquel elles appartiennent.

Réponse : Rappelons que

$$y = \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{n_1,1} \\ y_{1,2} \\ \vdots \\ y_{n_2,2} \\ y_{1,3} \\ \vdots \\ y_{n_3,3} \\ y_{1,4} \\ \vdots \\ y_{n_4,4} \end{pmatrix}$$

ce qui donne

$$X^t Y = \begin{pmatrix} \sum_{i=1}^{n_1} y_{i,1} \\ \sum_{i=1}^{n_2} y_{i,2} \\ \sum_{i=1}^{n_3} y_{i,3} \\ \sum_{i=1}^{n_4} y_{i,4} \end{pmatrix}$$

et donc

$$X^t Y = \begin{pmatrix} n_1 \bar{y}_1 \\ n_2 \bar{y}_2 \\ n_3 \bar{y}_3 \\ n_4 \bar{y}_4 \end{pmatrix}$$

Maintenant, comme la projection sur l'image de X , dénotée V est obtenue par la formule

$$\mathbf{p}_V(y) = X(X^t X)^{-1} X^t y$$

et en utilisant les formules de $(X^t X)^{-1}$ et de $X^t y$ que l'on vient de calculer, on obtient

$$\mathbf{p}_V(y) = X \begin{pmatrix} n_1^{-1} & 0 & 0 & 0 \\ 0 & n_2^{-1} & 0 & 0 \\ 0 & 0 & n_3^{-1} & 0 \\ 0 & 0 & 0 & n_4^{-1} \end{pmatrix} \begin{pmatrix} n_1 \bar{y}_1 \\ n_2 \bar{y}_2 \\ n_3 \bar{y}_3 \\ n_4 \bar{y}_4 \end{pmatrix}$$

ce qui donne

$$\mathbf{p}_V(y) = X \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \bar{y}_4 \end{pmatrix}$$

ce qui est égal à

$$\mathbf{p}_V(y) = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_2 \\ \bar{y}_3 \\ \vdots \\ \bar{y}_3 \\ \bar{y}_4 \\ \vdots \\ \bar{y}_4 \end{pmatrix}$$

3. Montrer que la projection $\mathbf{p}_U(\mathbf{y})$ est donnée par la formule :

$$\mathbf{p}_U(\mathbf{y}) = \bar{\mathbf{y}}\mathbf{e},$$

avec

$$\bar{\mathbf{y}} = \frac{1}{n_1 + n_2 + n_3 + n_4} \sum_{j=1}^4 \sum_{i=1}^{n_j} y_{ij}.$$

Ce qui signifie que la projection sur U correspond juste à remplacer toutes les données par la moyenne générale sur toutes les données.

Réponse : Comme U est l'image du vecteur e , la projection de y sur U est donnée par

$$\mathbf{p}_U(y) = e(e^t e)^{-1} e^t y$$

et comme

$$e^t y = y_1 + \cdots + y_{n_1+n_2+n_3+n_4}$$

et $e^t e = n_1 + n_2 + n_3 + n_4 = n$.
Comme

$$\begin{aligned} y_1 + \cdots + y_{n_1+n_2+n_3+n_4} &= y_{1,1} + \cdots + y_{n_1,1} + y_{1,2} + \cdots + y_{n_2,2} \\ &\quad + y_{1,3} + \cdots + y_{n_3,3} + y_{1,4} + \cdots + y_{n_4,4}, \\ &= \sum_{j=1}^4 \sum_{i=1}^{n_j} y_{i,j}, \end{aligned}$$

cela donne

$$\mathbf{p}_U(y) = e \cdot \frac{1}{n_1 + n_2 + n_3 + n_4} \sum_{j=1}^4 \sum_{i=1}^{n_j} y_{i,j}.$$

4. Montrer que

$$\|\mathbf{y} - \mathbf{p}_V(\mathbf{y})\|^2 = \sum_{j=1}^4 \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

Réponse : On a

$$\mathbf{y} - \mathbf{p}_V(\mathbf{y}) = \begin{pmatrix} y_1 - \bar{y}_1 \\ \vdots \\ y_{n_1} - \bar{y}_1 \\ y_{n_1+1} - \bar{y}_2 \\ \vdots \\ y_{n_1+n_2} - \bar{y}_2 \\ y_{n_1+n_2+1} - \bar{y}_3 \\ \vdots \\ y_{n_1+n_2+n_3} - \bar{y}_3 \\ y_{n_1+n_2+n_3+1} - \bar{y}_4 \\ \vdots \\ y_{n_1+n_2+n_3+n_4} - \bar{y}_4 \end{pmatrix} \quad (2)$$

ce qui donne, étant données les valeurs des composantes de y

$$y - \mathbf{p}_V(y) = \begin{pmatrix} y_{1,1} - \bar{y}_1 \\ \vdots \\ y_{n_1,1} - \bar{y}_1 \\ y_{1,2} - \bar{y}_2 \\ \vdots \\ y_{n_2,2} - \bar{y}_2 \\ y_{1,3} - \bar{y}_3 \\ \vdots \\ y_{n_3,3} - \bar{y}_3 \\ y_{1,4} - \bar{y}_4 \\ \vdots \\ y_{n_4,4} - \bar{y}_4 \end{pmatrix} \quad (3)$$

et finalement, comme $\|z\|^2 = z_1^2 + z_2^2 + \dots$, on a

$$\|y - P_V(y)\|^2 = \sum_{i=1}^{n_1} (y_{i,1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{i,2} - \bar{y}_2)^2 + \sum_{i=1}^{n_3} (y_{i,3} - \bar{y}_3)^2 + \sum_{i=1}^{n_4} (y_{i,4} - \bar{y}_4)^2$$

5. Montrer que

$$\|\mathbf{p}_V(\mathbf{y}) - \mathbf{p}_U(\mathbf{y})\|^2 = \sum_{j=1}^4 n_j (\bar{y}_j - \bar{y})^2$$

Réponse : on a

$$\mathbf{p}_V(y) - \mathbf{p}_U(y) = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_2 \\ \bar{y}_3 \\ \vdots \\ \bar{y}_3 \\ \bar{y}_4 \\ \vdots \\ \bar{y}_4 \end{pmatrix} - \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 - \bar{y} \\ \vdots \\ \bar{y}_1 - \bar{y} \\ \bar{y}_2 - \bar{y} \\ \vdots \\ \bar{y}_2 - \bar{y} \\ \bar{y}_3 - \bar{y} \\ \vdots \\ \bar{y}_3 - \bar{y} \\ \bar{y}_4 - \bar{y} \\ \vdots \\ \bar{y}_4 - \bar{y} \end{pmatrix}.$$

Ainsi

$$\|\mathbf{p}_V(y) - \mathbf{p}_U(y)\|^2 = n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + n_3(\bar{y}_3 - \bar{y})^2 + n_4(\bar{y}_4 - \bar{y})^2,$$

ce qui donne le résultat escompté (écrit différemment avec le signe Σ).
On obtient donc à l'aide des calculs que

$$r = \frac{\sum_{j=1}^4 \sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2 / (n - 4)}{\sum_{j=1}^4 n_j (\bar{y}_j - \bar{y})^2 / (4 - 1)},$$

qui est donc la "somme des carrés des résidus intra-groupe" divisé par la "somme des carrés des résidus inter-groupes".

6. Vérifier statistiquement si les 4 ateliers sont équivalents en terme d'exposition au produit toxique étudié en mettant en oeuvre le test de Fisher

```
# Création des objets nécessaires à l'étude

# La matrice de design
X = matrix(0, nrow=80, ncol=4)
X[1:20,1] = rep(1,20)
X[21:40,2] = rep(1,20)
X[41:60,3] = rep(1,20)
X[61:80,4] = rep(1,20)

# Le vecteur des réponses
a1 = c(rep(7,5),rep(8,4),rep(9,3),rep(10,4),rep(11,4))
a2 = c(rep(7,0),rep(8,4),rep(9,4),rep(10,5),rep(11,7))
a3 = c(rep(7,1),rep(8,8),rep(9,3),rep(10,5),rep(11,3))
a4 = c(rep(7,7),rep(8,5),rep(9,8),rep(10,0),rep(11,0))
y = c(a1,a2,a3,a4)

# Notre vecteur unité
e = rep(1,80)
```

On doit ensuite calculer les différentes projections sur les espaces U et V , ce qui nous donne

```
# Projection sur l'espace V
Pv = X%%solve(t(X)%*%X)%*%t(X)%*%y

# Projection sur l'espace U
Pu = e%%solve(t(e)%*%e)%*%t(e)%*%y
```

Il nous reste à calculer le numérateur et le dénominateur qui servent à définir

notre statistique de test r pour la déterminer complètement

```
# Norme du numérateur
num = sum((y-Pv)^2)

# Norme du dénominateur
denom = sum((Pv-Pu)^2)

# Statistique de test
r = (num/76)/(denom/3)
r

## [1] 0.1531694
```

Pour conclure au rejet ou non de l'hypothèse H_0 , on va regarder si la valeur de la statistique de test se trouve ou non dans l'intervalle de confiance de niveau $1 - \alpha$, $I_{1-\alpha}$ de la loi de Fisher à 76 et 3 degrés de libertés, *i.e.*, regardons si

$$r \in [f_{\alpha/2}, f_{\alpha/2}],$$

où $\alpha = 0.05$.

Or ces quantiles sont données par

```
# Borne inférieure de l'intervalle de confiance
borne_inf = qf(0.025,76,3)
# Borne supérieure de l'intervalle de confiance
borne_sup = qf(0.975,76,3)

# Il reste à faire le test
ifelse((r<borne_inf)|(r>borne_sup),"On rejette H0",
"On ne rejette pas H0")

## [1] "On ne rejette H0"
```

Le test nous conduit donc à rejeter l'hypothèse H_0 . On peut donc en conclure qu'au moins deux ateliers sont exposés à des degrés différents.

5 Autres exercices

Tout fonctionne de la même façon dans le cas où le nombre de groupe est différent de 4 groupes : il suffit de remplacer "4" par le nombre de groupes présent dans l'étude.

On veut étudier l'impact d'une ancienne mine d'arsenic sur les composantes hydrochimiques et hydrobiologiques d'un réseau hydrographique de Corse. Les mesures ont été faites sur 3 stations : B2, B3 (sur la Bravona) et P2 (sur un affluent la Presa) où est située la mine d'arsenic. Les tableaux ci-dessous résument la bio-accumulation de l'arsenic (en $\mu g/g$) sur les branchies des truites capturées pour chaque station.

1. Ecrire le modèle d'analyse de variance. Les conditions d'une analyse de variance sont-elles vérifiées dans la Table ??? A l'aide du test de Bartlett, vérifier l'égalité des variances dans ce cas.
2. On propose dans la Table ?? de transformer les données à l'aide de la fonction $x \mapsto \sqrt{x}$. Les conditions d'une analyse de variance sont-elles vérifiées dans la Table ???
3. Si les conditions sont vérifiées, réaliser le tableau d'analyse de variance et proposer un test pour vérifier que la station est un facteur significatif pour la bio-accumulation d'arsenic. Sinon, comment conclure ?

6 Travaux Pratiques sous R

On veut étudier l'effet de la direction du vent sur les pics d'ozone. Pour cela, on va considérer le jeu de données *Ozone* et expliquer la variable **maxO3** par la variable qualitative **vent**.


1. Importer les données et résumer les variables d'intérêts, ici **maxO3** et **vent** et commenter les résultats.

On pourra utiliser la commande suivante

```
# data désigne le jeu de données importé
summary(data[,c("maxO3", "vent")])
```

2. Représenter les données à l'aide des boîtes à moustaches pour illustrer l'effet du vent sur les pics d'ozone :

```
boxplot(maxO3~ozone, data = data, pch=15, cex=.5)
```

3. Réaliser l'analyse de variance pour estimer les paramètres du modèle, à l'aide de la fonction **lm** de .

```
regaov<-lm(maxO3~vent, data=data)
summary(regaov)
```

4. Que représente la ligne *Intercept* et la colonne *Estimate* ? Quelle contrainte a-t-on implicitement imposé sur les paramètres ?

5. Retrouver ces résultats à la main en calculant dans ce cas $(\mathbf{X}^\top \mathbf{X})^{-1}$.
6. Tester à présent la significativité du modèle, à l'aide du tableau d'analyse de variance et conclure.

```
anova(regaov)
```


7. On veut à présent imposer la contrainte $\mu = 0$ d'effet moyen nul. Pour cela, il suffit de spécifier au logiciel un modèle sans constante :

```
regaov2<-lm(maxO3 ~ -1+vent,data=data)
summary(regaov2)
```

8. Que représente dans ce cas la colonne *Estimate* de la matrice *Coefficient*. Retrouver les valeurs des estimateurs à la main.
9. On veut tester l'influence du vent sur le pic d'ozone grâce à un tableau d'analyse de variance. On propose d'utiliser la commande suivante :

```
anova(regaov2)
```

Ce tableau d'analyse de variance est faux. Quand la constante ne fait pas partie du modèle, tester $H_0 : \alpha_1 = \dots = \alpha_I = 0$ n'a pas de sens pour illustrer l'effet du facteur.

10. Pour des raisons particulières, on peut choisir une cellule témoin spécifique ( choisit par défaut la première par ordre alphabétique, ici *Est*) avec la commande suivante :

```
regaov3<-lm(maxO3 ~ C(vent,base=2),data=ozone)
```

Retrouver les résultats de la première analyse de variance, seul l'ordre des coefficients étant modifié.

11. On peut aussi choisir la contrainte $\sum_{i=1}^p \alpha_i = 0$ grâce à la commande :

```
regaov3<-lm(maxO3 ~ C(vent,sum),data=ozone)
```

Interpréter les entrées de la matrice Coefficients dans ce cas. Comment estimer l'effet du vent du Sud ?

Références

- [Cochran, 1934] Cochran, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, 30(2) :178–191.