Introduction au modèle linéaire via le test de l'effet d'une co-variable qualitative

J.N. Bacro
master ESDB, Université Montpellier

2023/2024

Le modèle linéaire

Objectif général du cours :

on dispose de n observations ${}^t(y_1, \ldots, y_n)$ d'une variable aléatoire Y d'intérêt et on cherche à **expliciter et quantifier l'effet de p co-variables $\mathbf{X_1}, \ldots, \mathbf{X_p}$**, connues par leurs réalisations $(x_{1j}, \ldots, x_{pj})$ (variables quantitatives) ou par la valeur de leurs modalités (variables qualitatives) pour le $j$ième individu, $j = 1, \ldots, n$, **sur l'espérance de Y.**

Une écriture générale pourrait être :

$$Y_j = f(X_{1j}, \ldots, X_{pj}) + \varepsilon_j \, , \; j = 1, \ldots, n$$

où $f(\cdot)$ fonction déterministe à préciser, dont **l'expression va dépendre du statut des co-variables** (quantitatives et/ou qualitatives), et ε un *résidu aléatoire* ...

- $f(\cdot)$?

$\rightsquigarrow$ *expression* **linéaire en les paramètres** ... par exemple du type suivant pour des co-variables quantitatives :

$$f(x_{1j}, \ldots, x_{pj}) = \theta_0 + \theta_1 x_{1j} + \ldots + \theta_p x_{pj}$$

avec $\theta_0, \ldots, \theta_p$ paramètres inconnus à déterminer.

Le modèle linéaire

Selon la nature des co-variables $X_1, \ldots, X_p$, on distingue généralement trois approches

1. variables quantitatives : *régression*.

   $X_1, \ldots, X_p$ sont alors dit *régresseurs* ;

2. variables qualitatives : *analyse de variance*.

   $X_1, \ldots, X_p$ sont alors dit *facteurs* ;

3. variables qualitatives et quantitatives : *analyse de covariance*.

Ces différentes approches peuvent être présentées de façon individuelle mais on verra que l'on dispose d'une **écriture unifiée** offrant un cadre général pour traiter indifféremment chacune de ces approches (d'où une notion de modèle linéaire *général*)

Exemples

1. Examen corrigé par 3 correcteurs : en moyenne, les correcteurs ont-ils noté différemment (*effet correcteur*) ?
2. La consommation moyenne d'essence de voitures peut-elle (correctement) s'expliquer selon le poids des voitures considérées ? selon le poids et la puissance des voitures considérées ?
3. Le poids moyen de grains de colza peut-il s'expliquer selon la dose d'engrais et le type de rotation de culture utilisé dans les parcelles ?
4. Peut-on juger de l'effet d'un médicament contre la tachycardie sur la base d'une injection à deux groupes (dont un témoin) de patients d'âge moyen différent ?

Exemple typique : effet d'une variable qualitative

On considère les résultats à un examen noté sur 70 et 3 correcteurs.

Question : les correcteurs ont-ils notés en moyenne différemment ?

On considère 3 échantillons de 5 copies prises au hasard dans chaque paquet corrigé : $(y_{ij})_{ij}$, $i = 1, \ldots, 5$, $j = 1, 2, 3,$;

moyenne correcteur sur échantillon : $\bar{y}_j$, $j = 1, 2, 3$;
espérance des notes du correcteur j : μ_j, $j = 1, 2, 3$;

$\rightsquigarrow$ *les différences observées entre les $\bar{y}_j$ sont-elles imputables aux fluctuations aléatoires liées à l'échantillonnage ou sont-elles caractéristiques d'une différence significative entre les μ_j, donc d'un effet correcteur ?*

Pour répondre, deux approches possibles (et équivalentes) :

- une approche "directe" (type tests statistiques) ;
- une approche modélisation . . . **cette dernière permettra de géneraliser le raisonnement à des contextes plus complexes que l'exemple présenté.**

ANOVA 1 facteur : *l'approche directe*

Les différences entre les $\bar{y}_j$ sont-elles *suffisamment grandes* pour ne pas être raisonnablement imputées aux fluctuations aléatoires ?

$H_0 : \mu_1 = \mu_2 = \mu_3$
H_1 ?

Données :

$C1$	$C2$	$C3$
47	55	54
53	54	50
49	58	51
50	61	51
46	52	49

D'où : $\bar{y}_1 \equiv y_{\cdot 1} = 49$; $y_{\cdot 2} = 56$; $y_{\cdot 3} = 51$ et $\bar{\bar{y}} \equiv y_{\cdot\cdot} = 52$

On a aussi : $\sum_{j=1}^{3}(y_{\cdot j} - y_{\cdot\cdot})^2 = 26$ (intérêt de ce calcul ?) et $\sum_{j=1}^{3}(y_{\cdot j} - y_{\cdot\cdot}) = 0$!

Estimation de la variance entre les moyennes de correcteurs (groupes) :
$s^2_{moy} = \frac{1}{\text{nbre groupes-1}} \sum_{j=1}^{\text{nbre groupes}}(y_{\cdot j} - y_{\cdot\cdot})^2 = 13$

$\rightsquigarrow$ *résumés de l'ensemble des mesures :* moyenne générale ; moyenne des groupes ; variance des moyennes de groupes . . . *Caractérisent toutes les sources potentielles de variation ?*

ANOVA 1 facteur : *l'approche directe*

On pourrait obtenir les mêmes résumés que précédemment avec des valeurs autres à *l'intérieur* des groupes ! Par exemple :

$C1$	$C2$	$C3$
(47)50	(55)48	(54)57
(53)42	(54)57	(50)59
(49)53	(58)65	(51)48
(50)45	(61)59	(51)46
(46)55	(52)51	(49)45

mêmes résumés . . . **mais avec des variations intra-groupes beaucoup plus fortes !** (ex : écart groupe 1 passe de 6 à 13 ; groupe 2 de 9 à 17 . . . !)

Rôle de cette variabilité intra-groupe ?
selon l'importance de la variabilité dans les groupes, les mêmes différences moyennes observées entre les groupes peuvent conduire à des conclusions différentes quant à H_0

Quel graphe pour illustrer cette affirmation ?

$C1$	$C2$	$C3$
(47)50	(55)48	(54)57
(53)42	(54)57	(50)59
(49)53	(58)65	(51)48
(50)45	(61)59	(51)46
(46)55	(52)51	(49)45

$y_{.1} = 49$; $y_{.2} = 56$; $y_{.3} = 51$ et $\bar{\bar{y}} \equiv y_{..} = 52$

ANOVA 1 facteur : *l'approche directe*

En d'autres termes : **la variabilité des moyennes échantillonnales est-elle grande par rapport aux fluctuations échantillonnales ?**

Caractériser les fluctuations échantillonnales ?

sous l'hypothèse d'une variance des mesures individuelles commune aux groupes considérés (la variable aléatoire "note de l'examen" a une même variance quelque soit le correcteur) $\rightsquigarrow$ moyenne pondérée des variabilités intra-groupes, i.e.

$$s^2_{mes} = \frac{(n_1 - 1)s_1^2 + \ldots + (n_p - 1)s_p^2}{(n_1 - 1) + \ldots + (n_p - 1)} = \frac{(n_1 - 1)s_1^2 + \ldots + (n_p - 1)s_p^2}{N - p}$$

Sous l'hypothèse précédente, s^2_{mes} *est une estimation de* σ^2 ...

Pour simplifier, supposons n mesures dans chacun des J groupes.

Question : s^2_{moy} est elle grande par rapport à s^2_{mes} ?

On considère la statistique : $F = \frac{nS^2_{moy}}{S^2_{mes}}$; comme $\sigma^2_{moy} = \frac{\sigma^2}{n}$, sous H_0, on s'attend à une fluctuation autour de 1 car rapport de deux estimations distinctes de la même quantité σ^2 !

ANOVA 1 facteur : *l'approche directe*

Si H_0 fausse, S^2_{moy} va être "gonflée" car les moyennes seront dispersées alors que S^2_{mes} continuera d'estimer le σ^2 commun ... d'où F significativemennt > 1.

Si J désigne le nombre de groupes, on montrera que sous H_0,

$$F \sim Fish(J-1, \sum_j n_j - J).$$

Ici, $J = 3$ et donc $Fish(2, 12)$ à considérer.
Après calcul, F prend la valeur 8.3 et le quantile à 99% de la loi $Fish(2, 12)$ est 6.93.

Conclusion ?

ANOVA : égalité fondamentale

Supposons n mesures dans chacun des J groupes. On a :

$$\sum_{j=1}^{J}\sum_{i=1}^{n}(y_{ij}-y_{..})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n}(y_{ij}-y_{.j})^2 + \sum_{j=1}^{J}\sum_{i=1}^{n}(y_{.j}-y_{..})^2$$

$$SCT_c = SC_{intra} + SC_{inter}$$

Présentation standard *(table d'analyse de variance)* :

Source variation	Degrés de liberté	Somme des carrés	Carrés moyens	F
Inter	J - 1	SC_{inter}	$\frac{SC_{inter}}{J-1} \equiv V_A$	$\frac{V_A}{V_R}$
Intra	J(n-1)	SC_{intra}	$\frac{SC_{intra}}{J(n-1)} \equiv V_R$	
Totale (cor)	nJ - 1	SCT_c		

On a bien :

1. $s^2_{mes} = \frac{\sum_j \sum_i (y_{ij}-y_{.j})^2}{J(n-1)} = \frac{SC_{intra}}{J(n-1)}$
2. $ns^2_{moy} = \frac{\sum_j \sum_i (y_{.j}-y_{..})^2}{J-1} = n\frac{\sum_j (y_{.j}-y_{..})^2}{J-1} = \frac{SC_{inter}}{J-1}$

Rmq :
si tailles d'échantillons n_j différentes, pas de gros changements !
$\rightsquigarrow \sum_{j=1}^{J} n_j(y_{.j}-y_{..})^2$; $\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-y_{..})^2$; $\sum_{i=1}^{J}(n_i-1)$; $\sum_{i=1}^{J} n_i - 1$.

ANOVA 1 facteur : *l'approche modélisation*

La co-variable *Correcteur* considérée est qualititative à 3 niveaux et dire que les notes des correcteurs ont une espérance commune μ peut s'écrire sous la forme :

$$y_{ij} = \mu + e_{ij},\ i = 1, \ldots, 5\ , j = 1, 2, 3$$

où μ joue le rôle d'une quantité commune à toutes les notes et e_{ij} la fluctuation échantillonnale de la $i$ème note du groupe j autour de μ *(résidu aléatoire après prise en compte d'un effet systématique)*.

Dire que les notes des groupes diffèrent systématiquement d'un groupe à l'autre peut s'écrire :

$$y_{ij} = \mu_j + e_{ij},\ i = 1, \ldots, 5\ , j = 1, 2, 3$$

où μ_j représente une quantité caractéristique des notes du correcteur j (effet correcteur) et e_{ij} la fluctuation échantillonnale de la $i$ème note du groupe j autour de μ_j.

Rejeter H_0 (égalité des espérances) revient à dire que les espérances de notes peuvent changer d'un correcteur à l'autre et que donc une structuration des notes selon le correcteur doit être envisagée : il y a un effet correcteur !

ANOVA 1 facteur : *l'approche modélisation*

Les quantités e_{ij} représentent la part de variation de la variable réponse qui n'est pas "expliquée" par le modèle considéré $\rightsquigarrow$ notion de résidus de modèle (ou erreur. . . *mais mal dit !)*

Intuitivement : modèle explique d'autant mieux la réponse que les (e_{ij}) sont *"petits"* . . .

Pour l'exemple considéré,
soit M_0 le modèle : $y_{ij} = \mu + e_{ij}$, $i = 1, \ldots, 5$, $j = 1, 2, 3$;
$\rightarrow$ une estimation de e_{ij} : $\hat{e}_{ij} = y_{ij} - \hat{\mu} = y_{ij} - y_{..}$

soit M_1 le modèle : $y_{ij} = \mu_j + e_{ij}$, $i = 1, \ldots, 5$, $j = 1, 2, 3$;
$\rightarrow$ une estimation de e_{ij} : $\hat{e}_{ij} = y_{ij} - \hat{\mu}_j = y_{ij} - y_{.j}$

ANOVA : *l'approche modélisation*

Meilleur modèle ? critère : $\sum_j \sum_i \hat{e}_{ij}^2$ minimale !

Sous le modèle M_0 : $\sum_j \sum_i \hat{e}_{ij}^2 = \sum_j \sum_i (y_{ij} - y_{..})^2 = SCT_c \equiv SCR_{M_0}$

Sous le modèle M_1 : $\sum_j \sum_i \hat{e}_{ij}^2 = \sum_j \sum_i (y_{ij} - y_{.j})^2 = SC_{intra} \equiv SCR_{M_1}$

Par la relation fondamentale :

$$SC_{inter} = SCR_{M_0} - SCR_{M_1}$$

SC_{inter} : "chute" de la SCR lorsque l'on passe du modèle M_0 (1 paramètre) au modèle M_1 (J paramètres)

Rmqs :

1. on a nécessairement $SCR_{M_1} \leq SCR_{M_0}$!
2. SC_{inter} : SC des variations imputables à la variable "groupe" *(ici : correcteur)*.

Intuitivement : *M_1 meilleur que M_0 si CR_{M_1} significativement $<$ à SCR_{M_0}.*
$\rightsquigarrow$ **comparer la différence $SCR_{M_0} - SCR_{M_1}$ aux fluctuations échantillonnales :**

- si même ordre de grandeur : M_1 ne fait pas mieux que M_0 ;
- si significativement plus grande : M_1 significativement meilleur que M_0 ;

ANOVA : *l'approche modélisation*

Soit $H_0 : \mu_1 = \mu_2 = \mu_3 \equiv \mu$ (pas d'effet correcteur) et H_1 : au moins un des μ_j diffère des autres (il y a effet correcteur), rejeter H_0 au profit de H_1 revient à rejeter M_0 au profit de M_1 *(M_1 meilleur !)*.

On a vu que le rejet de H_0 se décide au travers de la statistique

$$F = \frac{\frac{SC_{inter}}{d.d.l.(SC_{inter})}}{\frac{SC_{intra}}{d.d.l.(SC_{intra})}}$$

c'est-à-dire

$$F = \frac{\frac{SCR_{M_0} - SCR_{M_1}}{d.d.l.(SCR_{M_0} - SCR_{M_1})}}{\frac{SCR_{M_1}}{d.d.l.(SCR_{M_1})}}$$

Loi du rapport sous H_0 (M_0) : loi de Fisher

$$Fish(d.d.l.(SCR_{M_0} - SCR_{M_1}), d.d.l.(SCR_{M_1}))$$

D'où vient cette loi ?

Postulats du modèle linéaire et loi de Fisher

Critère utilisé : *critère de moindres carrés pour les résidus des modèles* . . . il faut donc que les résidus soient comparables et on impose les postulats suivants pour les résidus d'un modèle :

1. $\mathbb{E}(e_{ij}) = 0$;
2. $\mathbb{V}(e_{ij}) = \sigma^2 \equiv Cte$
3. $(e_{ij})_{i,j}$ non-corrélées ;
4. $e_{ij} \sim N(0, \sigma^2)$

Sous les postulats,

• $SCR_{M_1} = \sum_{i,j} e_{ij}$; $\frac{e_{ij}}{\sigma} \sim N(0,1)$ et $\sum_{i,j} \frac{e_{ij}^2}{\sigma^2} \sim \chi^2(d.d.l.(SCR_{M_1}))$ d'où

$$SCR_{M_1} \sim \sigma^2 \chi^2(d.d.l.(SCR_{M_1}))$$

• SCR_{M_1} et $SCR_{M_1} - SCR_{M_0}$ sont indépendantes (on verra qu'elles correspondent à des projections orthogonales sur des sous-espaces orthogonaux)

• $SCR_{M_1} - SCR_{M_0} \sim \sigma^2 \chi^2(d.d.l.(SCR_{M_1} - SCR_{M_0}))$ (car SCR_{M_0} et SCR_{M_1} sont des χ^2 indépendantes)

Le rapport F considéré est alors bien une loi de Fisher avec les d.d.l. spécifiés.

Table d'analyse de variance d'un modèle linéaire

Table d'analyse de variance du modèle linéaire M_1 par rapport M_0 :

Source variation	Degrés de liberté	Somme des carrés	Carrés moyens	F
Modèle	J - 1	SCM_{M_1}	$\frac{SCM_{M_1}}{J-1} \equiv V_A$	$\frac{V_A}{V_R}$
Résiduelle	J(n-1)	SCR_{M_1}	$\frac{SCR_{M_1}}{J(n-1)} \equiv V_R$	
Totale (cor)	nJ - 1	SCR_{M_0}		

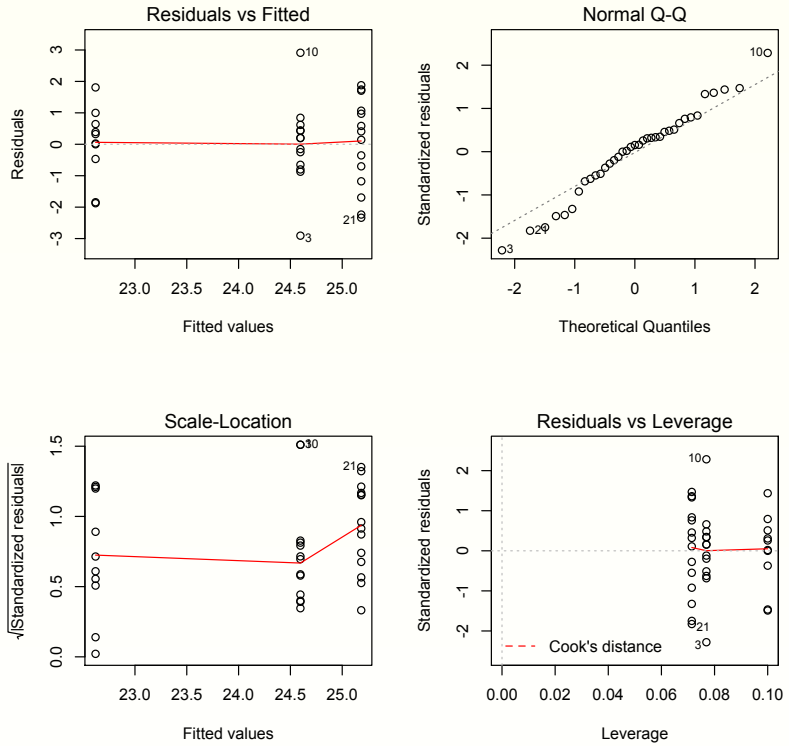
Cette table se généralise à tout modèle linéaire M et sera utilisée pour tester si le modèle M considéré est significativement meilleur que le modèle M_0.

Dans les sommes de carrés, on considérera alors les valeurs $\hat{\mathbf{Y}}$: valeurs de la réponse prédites par le modèle M étudié.

Ici : $SCM_{M1} = \sum_{i,j}(y_{\cdot j} - y_{\cdot\cdot})^2 \equiv \sum_{i,j}(\hat{y}_{ij} - y_{\cdot\cdot})^2$; SCR_{M_1} ? SCR_{M_0} ?

Validation du modèle ?

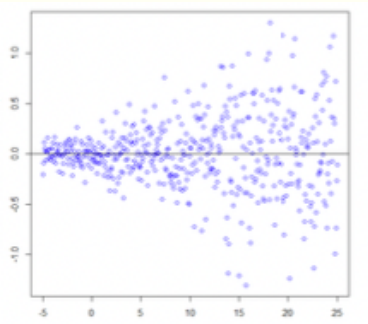
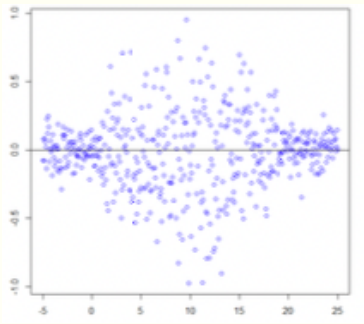
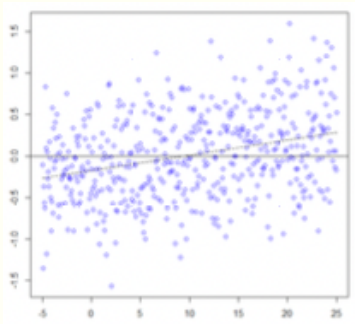
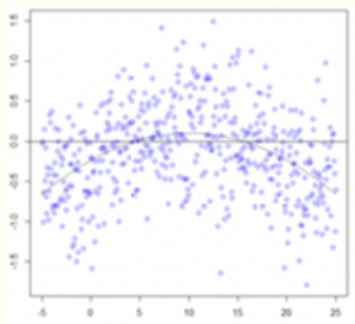
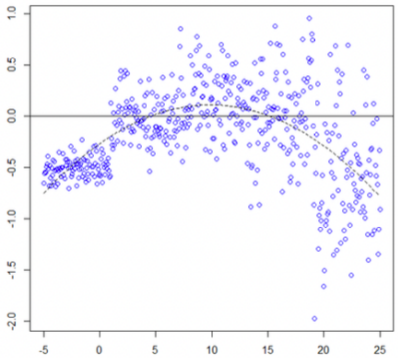
Toutes les approches et interprétations précédemment vues supposent que le modèle considéré est un modèle valide quant aux postulats du ML sur les résidus ! $\rightsquigarrow$ postulats à "valider" ! . . . **pour cela, approche essentiellement graphique !**



Mise en pratique R - Graphes associés - Validation graphique des postulats

- Le premier graphique trace les résidus en fonction des valeurs ajustées avec la tendance moyenne tracée en rouge et les points sortant potentiellement d'une distribution normale indiqués par un numéro. On s'attend à une ligne rouge horizontale (pas de tendance dans les résidus) et pas plus de 5% des points marqués.
- Le graphique 2 (QQ-plot) permet de vérifier graphiquement l'hypothèse de normalité des résidus : si les points sont à peu près alignés en se confondant avec la première bissectrice des axes, on peut dire que les résidus peuvent être distribués selon une loi normale.
- Le troisième répète le premier à une autre échelle.
- Le graphique 4 (Cook's D) permet de repérer les individus pour lesquels le modèle linéaire est mal (ou pas) adapté : pour chaque individu, mesure de la distance entre les estimations des paramètres faites avec et sans l'individu. Des valeurs dans la zone rouge identifient des points qui influencent trop les valeurs des paramètres estimés.

Exemples de graphes (prédits,résidus) que l'on ne veut pas !

source : Jonathan Lenoir, Univ. Picardie

ANOVA 1 facteur : caractériser l'effet d'un facteur

Ecriture alternative permettant de mettre en avant directement l'éventuel effet des modalités d'un facteur :

$$Y_{ij} = \mu + \alpha_j + E_{ij}$$

Différence avec l'écriture précedente $Y_{ij} = \mu_j + E_{ij}$? **Identifiabilité !**

$$Y_{ij} = \mu + \alpha_j + E_{ij} = (\mu - \beta) + (\alpha_j + \beta) + E_{ij} = \mu' + \alpha'_j + E_{ij}$$

$\rightsquigarrow$ infinité de solutions !

$(\alpha_j)_j$ non-estimables . . . *contraintes nécessaires !*

Comment s'en sortir ?

approche matricielle permet une résolution élégante, exploitable pour tous modéles linéaires !

ANOVA 1 facteur : écriture matricielle

- Etape 1 :

$$\mathbf{Y} = X\theta + \mathbf{E}$$

où **Y** vecteur des réponses, **E** vecteur des résidus, θ vecteur des paramètres, X matrice "design" du modèle (matrice caractéristique du modèle considéré).

Exemple :
pour une expérience dans laquelle chaque modalité du facteur considéré est observée 3 fois, écrire la forme matricielle à considérer pour une anova à un facteur ayant 3 modalités (faire le lien avec l'exemple correcteur !).

ANOVA 1 facteur : écriture matricielle

Exemple :

ANOVA 1 facteur : écriture matricielle

• Etape 2 : Chercher θ tel que $\sum_{i,j} E_{i,j}^2$ minimale (moindres carrés résiduels)

$\sum_{i,j} E_{i,j}^2 = {}^t\mathbf{EE} = {}^t(\mathbf{Y} - X\theta)(\mathbf{Y} - X\theta) \rightsquigarrow$ minimiser $||\mathbf{Y} - X\theta||^2$
$\mathbf{Y}$ vecteur de $\mathbb{R}^n$; soit $\mathcal{P}$ l'espace engendré par $X \equiv (X_1, \ldots, X_p)$ où X_j désigne la $j$ème colonne de X.

- $\mathbb{E}(\mathbf{Y}) = X\theta$ (pourquoi ?)
donc $\mathbb{E}(\mathbf{Y}) \in \mathcal{P}$ comme combinaison linéaire des X_j.

- ${}^t\mathbf{EE} = {}^t(\mathbf{Y} - X\theta)(\mathbf{Y} - X\theta)$ minimum si $X\theta$ projection $\perp$ de $\mathbf{Y}$ sur $\mathcal{P}$ (pourquoi ?)

- $\forall i$, on veut $< X_i/\mathbf{Y} - X_i\theta >= 0$ i.e. ${}^tX_i(\mathbf{Y} - X\theta) = 0\ \forall i$;
on en déduit ${}^tX(\mathbf{Y} - X\theta) = \mathbf{0}$ d'où

résoudre : ${}^tXX\theta = {}^tXY$

La résolution de ce système donnera $\hat{\theta}$ . . . dont on déduira $\hat{\mathbf{Y}} \equiv \widehat{\mathbb{E}(\mathbf{Y})}$ valeurs prédites par le modèle, et $\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}}$

2 cas à considérer : ${}^tXX$ inversible ; ${}^tXX$ non-inversible.

ANOVA 1 facteur : écriture matricielle

Exemple : on considère une expérience avec 1 facteur à 2 modalités et 3 répétitions par modalités. Déterminer l'estimimation de $\theta = {}^t(\mu, \alpha_1, \alpha_2)$

1. dans le cas du modèle $y_{ij} = \mu_j + e_{ij}$, $i = 1, \ldots, 3$, $j = 1, 2$.
2. dans le cas du modèle $y_{ij} = \mu + \alpha_j + e_{ij}$, $i = 1, \ldots, 3$, $j = 1, 2$.
   Vous pourrez considérer successivement les contraintes $\alpha_1 + \alpha_2 = 0$ puis $\alpha_2 = 0$

Vérifiez que les estimations obtenues vont dépendre des contraintes . . .

mais que certaines combinaisons linéaires des paramètres admettent une estimation unique (et heureusement !).

ANOVA 1 facteur : écriture matricielle

Cas 1 : $y_{ij} = \mu_j + e_{ij}$

$$\theta = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}; {}^tXX = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}; ({}^tXX)^{-1} = \frac{1}{9}\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$$

et ${}^tXY = \begin{pmatrix} \sum_i y_{i1} \\ \sum_i y_{i2} \end{pmatrix}$. D'où

$${}^tXX\theta = {}^tXY \Leftrightarrow \widehat{\theta} = ({}^tXX)^{-1}\,{}^tXY$$

et

$$\widehat{\theta} = \begin{pmatrix} \frac{\sum_i y_{i1}}{3} \\ \frac{\sum_i y_{i2}}{3} \end{pmatrix} \equiv \begin{pmatrix} y_{\cdot 1} \\ y_{\cdot 2} \end{pmatrix}$$

ANOVA 1 facteur : écriture matricielle

Cas 2 : $y_{ij} = \mu + \alpha_j + e_{ij}$

$$\theta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} \; X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \; {}^tXX = \begin{pmatrix} 6 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{pmatrix};$$

${}^tXX$ est de rang $r = 2 < 3$ donc non-inversible !

${}^tXY = \begin{pmatrix} \sum_i \sum_j y_{ij} \\ \sum_i y_{i1} \\ \sum_i y_{i2} \end{pmatrix}$. D'où

$${}^tXX\widehat{\theta} = {}^tXY \Leftrightarrow (\star) \begin{cases} 6\mu & + & 3\alpha_1 & + & 3\alpha_2 & = & \sum_{i,j} y_{ij} \\ 3\mu & + & 3\alpha_1 & & & = & \sum_i y_{i1} \\ 3\mu & & & + & 3\alpha_2 & = & \sum_i y_{i2} \end{cases}$$

ANOVA 1 facteur : écriture matricielle

On considère la contrainte : $\alpha_1 + \alpha_2 = 0$

$$\text{Système } (\star) \Rightarrow \left\{ \begin{array}{lcll} \alpha_1 & + & \alpha_2 & = & 0 \\ 3\mu & + & 3\alpha_1 & = & \sum_i y_{i1} \\ 3\mu & + & 3\alpha_2 & = & \sum_i y_{i2} \end{array} \right.$$

$$\Leftrightarrow \left\{ \begin{array}{lcl} \alpha_1 & = & -\alpha_2 \\ 6\mu & = & \sum_i y_{i1} + \sum_i y_{i2} \rightarrow \mu = \frac{1}{6} \sum_{i,j} y_{ij} \\ \alpha_1 & = & \frac{1}{3} \left(\sum_i y_{i1} - 3\mu \right) \rightarrow \alpha_1 = \frac{1}{3} \sum_i y_{i1} - \mu \end{array} \right.$$

$$\widehat{\theta} \equiv \left(\begin{array}{c} \widehat{\mu} \\ \widehat{\alpha_1} \\ \widehat{\alpha_2} \end{array} \right) = \left(\begin{array}{c} y_{\cdot\cdot} \\ y_{\cdot 1} - y_{\cdot\cdot} \\ y_{\cdot 2} - y_{\cdot\cdot} \end{array} \right)$$

(en utilisant $2y_{\cdot\cdot} = y_{\cdot 1} + y_{\cdot 2}$ d'où $y_{\cdot\cdot} - y_{\cdot 2} = y_{\cdot 1} - y_{\cdot\cdot}$ pour déduire $\widehat{\alpha_2}$)

ANOVA 1 facteur : écriture matricielle

Autre contrainte possible : $\alpha_2 = 0$

$$\text{Système } (\star) \Rightarrow \left\{ \begin{array}{rcrcl} & & \alpha_2 & = & 0 \\ 3\mu & + & 3\alpha_1 & = & \sum_i y_{i1} \\ 3\mu & + & 3\alpha_2 & = & \sum_i y_{i2} \end{array} \right.$$

$$\Leftrightarrow \left\{ \begin{array}{lcl} \alpha_2 & = & 0 \\ \mu & = & \frac{1}{3}\sum_i y_{i2} \\ \alpha_1 & = & \frac{1}{3}\sum_i y_{i1} - 3\mu \end{array} \right.$$

D'où

$$\widehat{\theta} \equiv \begin{pmatrix} \widehat{\mu} \\ \widehat{\alpha_1} \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} y_{\cdot 2} \\ y_{\cdot 1} - y_{\cdot 2} \\ 0 \end{pmatrix}$$

Les estimations des α_j dépendent des contraintes choisies ... *par contre, certaines combinaisons minéaires, comme $\mu + \alpha_j$ par exemple, ont une estimation unique !*

Modèle linéaire et écriture matricielle

Les différentes modélisations entrant dans le cadre du modèle linéaire peuvent toujours se mettre sous forme matricielle et tout modèle linéaire peut s'écrire sous la forme :

$$\mathbf{Y} = X\theta + \mathbf{E}$$

La forme de la matrice X est directement liée au type de modèle linéaire considéré ...

et comme nous allons le voir dans la suite du cours :

$\hookrightarrow$ les propriétés de X ou de ${}^tXX$ sont déterminantes pour l'étude et les propriétés du modèle concerné.