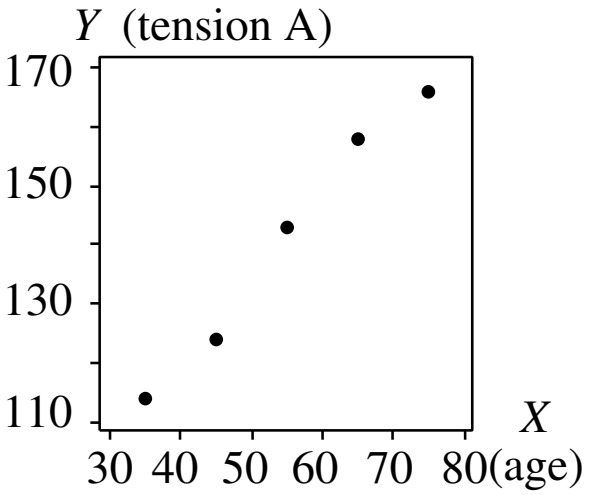
INTRODUCTION AU MODELE LINEAIRE
CAS D'UNE REGRESSION SIMPLE

Etude de la tension artérielle en fonction de l'âge des individus

tension Y	114	124	143	158	166
âge X	35	45	55	65	75



$$Y_i = \alpha + \beta\, X_i + \varepsilon_i$$

$$\widehat{\alpha} = 65.1 \qquad \widehat{\beta} = 1.38$$

- On repart de l'ensemble traité la veille (régression).

- Montrer, dans un premier temps,
  - le tableau des couples de données (X, Y)
  - la représentation graphique associée
  - la modélisation mathématique

- Rappeler, dans un second temps, les estimations des paramètres (calculés la veille).

CAS D'UNE REGRESSION SIMPLE

$$Y_i = \alpha + \beta\, X_i + \varepsilon_i$$

$$\begin{aligned}
114 &= \alpha + \beta \times 35 + \varepsilon_1 \\
124 &= \alpha + \beta \times 45 + \varepsilon_2 \\
143 &= \alpha + \beta \times 55 + \varepsilon_3 \\
158 &= \alpha + \beta \times 65 + \varepsilon_4 \\
166 &= \alpha + \beta \times 75 + \varepsilon_5
\end{aligned}$$

$$\begin{pmatrix} 114 \\ 124 \\ 143 \\ 158 \\ 166 \end{pmatrix} = \alpha \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \beta \begin{pmatrix} 35 \\ 45 \\ 55 \\ 65 \\ 75 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$$

$$(\, Y_i \,) \quad = \alpha \; (\, 1 \,) \; + \; \beta \;\; (\, X_i \,) + \; (\, \varepsilon_i \,)$$

**L'ensemble des équations qui suivent doit être construit progressivement au tableau (le transparent peut être montré à la fin pour une écriture propre et bien alignée).*

- Rappeler que si A est un vecteur (ou une matrice), $\lambda(A) = (\lambda A)$: tous les termes du vecteur (ou de la matrice) sont multipliés par λ.

- Ecrire les equations relatives à tous les couples (Y_i, X_i).

- Ecrire ensuite le système d'équations précédent sous forme d'une combinaison linéaire des vecteurs colonnes: $(1), (X_i), (\varepsilon_i)$.

- Rappeler, à partir de deux matrices simples, A_{2X2} et B_{2X3} , comment on réalise le produit matriciel

$$A \times B \rightarrow C_{2X3}$$

(pour préparer la forme synthétique donnée au transparent suivant).

CAS D'UNE REGRESSION SIMPLE

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ 1 & X_4 \\ 1 & X_5 \end{pmatrix} \times \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$$

$$\begin{array}{ccccccc} \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \boxed{Y} & = & X & & \theta & + & \varepsilon \end{array}$$

Dans le cas où n couples (X_i, Y_i) sont observés :

- Y est un vecteur de dimension n
- θ est un vecteur de dimension p, si p est le nombre de paramètres à estimer
- ε est un vecteur de dimension n
- X est une matrice de dimension (n, p)

$$
\underset{(n,1)}{Y} = \underset{(n,2)}{X} \underset{(2,1)}{\theta} + \underset{(n,1)}{\varepsilon}
$$

**A continuer, comme précédemment, au tableau.*

- Après avoir rappelé comment on réalisait un produit matriciel, on donnera la forme condensée ci-contre. (On vérifiera lentement le produit des deux matrices pour bien montrer que c'est les mêmes équations que précédemment).

- On introduira ensuite une forme synthétique de l'équation précédente: $Y = X\theta + \varepsilon$. On insistera bien sur le fait qu'elle correspond à une formalisation très générale de tout modèle linéaire.

- On rappellera les dimensions des vecteurs et matrices de l'équation générale, et on vérifiera la cohérence de celles-ci au niveau de l'équation.

- On rappellera aussi au tableau que:

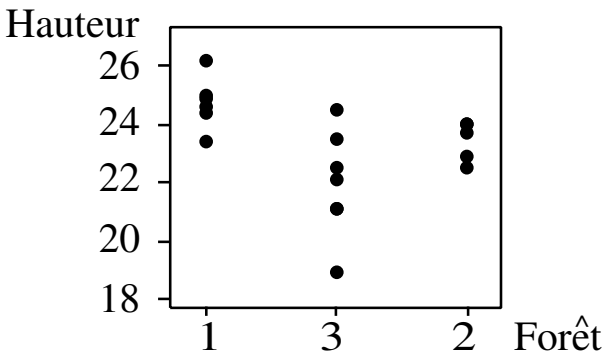
$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_5 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_5 \end{pmatrix} \qquad \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_5 \end{pmatrix}$$

ANALYSE DE LA VARIANCE A 1 FACTEUR

Comparaison des hauteurs d'arbres moyennes dans 3 forêts

Forêt 1	Forêt 2	Forêt 3
23.4	22.5	18.9
24.4	22.9	21.1
24.6	23.7	21.1
24.9	24.0	22.1
25.0	24.0	22.5
26.2		23.5
		24.5
$n_1 = 6$	$n_2 = 5$	$n_3 = 7$

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$


Hauteur
26
24
22
20
18
1
3
2
Forêt

- On repart de l'exemple traité la veille pour l'introduction à l'analyse de la variance

- Montrer dans un premier temps:
  - **le tableau de données**
    les n_i (nombres d'arbres) sont différents d'une forêt à l'autre

  - **la modélisation mathématique simple:**
    * Y_{ij} est la hauteur du $j^{ième}$ arbre de la forêt i
    * μ_i est la moyenne de tous les arbres de la forêt i
    * ε_{ij} est le résidu associé au $j^{ième}$ arbre de la forêt i

  - **une représentation graphique**

ANALYSE DE LA VARIANCE A 1 FACTEUR

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

ou

Y_{ij}	$=$	μ	$+$	α_i	$+ \varepsilon_{ij}$
Hauteur de l'arbre j dans la forêt i		Moyenne des hauteurs de toutes les forêts		Effet moyen de la forêt i	

$$\mu_i = \mu + \alpha_i \;\; ; \;\; \mu = \frac{\mu_1+\mu_2+\mu_3}{3} \text{ et } \sum_i \alpha_i = 0$$

Construction de la table d'analyse de variance

Source	SCE	ddl	CM	F
Variation entre forêts	25.31	2	12.65	7.3 **
Variation résiduelle	26.00	15	1.733	

Conclusion : $\exists$ au moins 2 forêts différentes !

Revenir au tableau.

- Passer d'une modélisation à l'autre,
  avec $\mu_i = \mu + \alpha_i$ et $\mu = \frac{\mu_1+\mu_2+\mu_3}{3}$
  α_i est l'effet moyen de la forêt i.

- Dire qu'on est amené à introduire la contrainte :

$$\sum_i \alpha_i = 0$$

  (par définition, la somme des effets moyens des trois forêts est nulle).

- Rappeler ensuite les résultats de la table d'analyse de la variance calculée la veille (sur transparent).

Y_{11} (23.4)	=	μ	+	α_1				+	ε_{11}
Y_{12} (24.4)	=	μ	+	α_1				+	ε_{12}
Y_{13} (24.6)	=	μ	+	α_1				+	ε_{13}
Y_{14} (24.9)	=	μ	+	α_1				+	ε_{14}
Y_{15} (25.0)	=	μ	+	α_1				+	ε_{15}
Y_{16} (26.2)	=	μ	+	α_1				+	ε_{16}
Y_{21} (22.5)	=	μ	+		α_2			+	ε_{21}
Y_{22} (22.9)	=	μ	+		α_2			+	ε_{22}
Y_{23} (23.7)	=	μ	+		α_2			+	ε_{23}
Y_{24} (24.0)	=	μ	+		α_2			+	ε_{24}
Y_{25} (24.0)	=	μ	+		α_2			+	ε_{25}
Y_{31} (18.9)	=	μ	+				α_3	+	ε_{31}
Y_{32} (21.1)	=	μ	+				α_3	+	ε_{32}
Y_{33} (21.1)	=	μ	+				α_3	+	ε_{33}
Y_{34} (22.1)	=	μ	+				α_3	+	ε_{34}
Y_{35} (22.5)	=	μ	+				α_3	+	ε_{35}
Y_{36} (23.5)	=	μ	+				α_3	+	ε_{36}
Y_{37} (24.5)	=	μ	+				α_3	+	ε_{37}

$$Y_{ij} = \mu \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \alpha_1 \begin{pmatrix} \underline{1} \\ \underline{0} \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} \underline{0} \\ \underline{1} \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} \underline{0} \\ \underline{0} \\ 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{ij} \end{pmatrix}$$

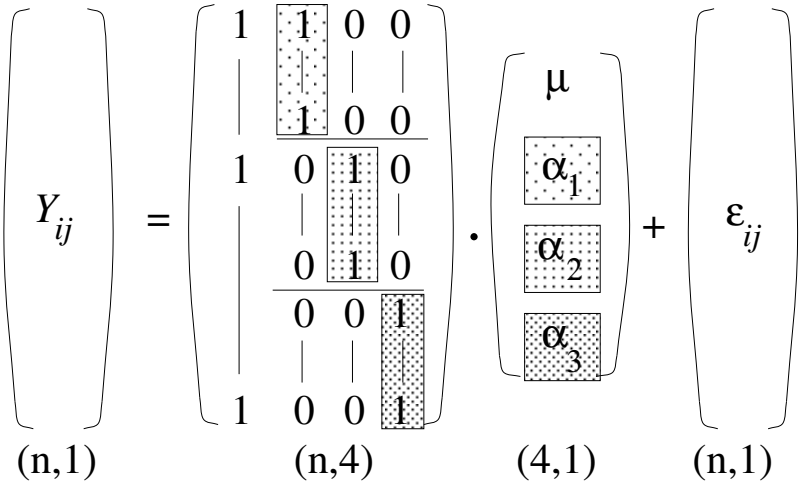
**Les équations qui suivent doivent être progressivement établies au tableau, en veillant au décallage des α_1, α_2 et α_3.*

- On écrira ensuite la forme condensée présentée en bas du transparent : combinaison linéaire de vecteurs colonnes, bien alignés par rapport au système d'équations précédent.

  (Le transparent peut être projeté après l'écriture au tableau pour une écriture très propre).

ANALYSE DE LA VARIANCE A 1 FACTEUR

Formalisation :



$$Y = X\theta + \varepsilon$$

Pour n observations :

- Y et ε sont des vecteurs de dimension n
- θ est un vecteur de dimension p
- X est une matrice de dimension (n, p)
- p est le nombre de paramètres à estimer (ici $p = 4$)

**Toujours au tableau*

- On écrira, sous forme matricielle, la combinaison des vecteurs "colonnes" précédente, en donnant les dimensions des vecteurs et matrices impliqués.

- On remarquera, comme dans l'exemple traité pour la régression, que l'écriture matricielle associée au modèle d'analyse de la variance est de la forme: $Y = X\theta + \varepsilon$.

- Remarquer que le nombre de paramètres à estimer est plus grand que dans le cas de la régression et que les dimensions de X et θ sont différentes de celles de la matrice X et du vecteur θ de l'équation de régression.

  MAIS! LE FORMALISME EST STRICTEMENT IDENTIQUE

REGRESSION LINEAIRE MULTIPLE

Exemple : Régression du rendement moyen des blés d'hiver dans l'Allier (Y_i) en fonction :

- de l'année (X_i)
- du nombre de jours où $\theta_{min} < 6°C$, entre le 1er et le 20 mai (Z_i)

Rendement Y	Année X_i	Nbre Jours Z_i
43	80	12
35	81	8
45	82	11
40	83	4
51	84	12
51	85	10
40	86	2
53	87	10
47	88	2
57	89	11

Les deux exemples qui suivent n'auront pas été introduits la veille ; ils doivent permettre de montrer que le formalisme général du modèle linéaire est toujours le même, lorsque l'on aborde des cas un peu plus complexes que l'analyse de la variance à un facteur ou la régression à un régresseur.

Remarque pour le formateur:
la variable X_i est une variable discrète ; il n'est donc pas évident de la choisir comme régresseur ; néanmoins, l'ensemble est parlant...

Cas d'une régression linéaire multiple à deux régresseurs

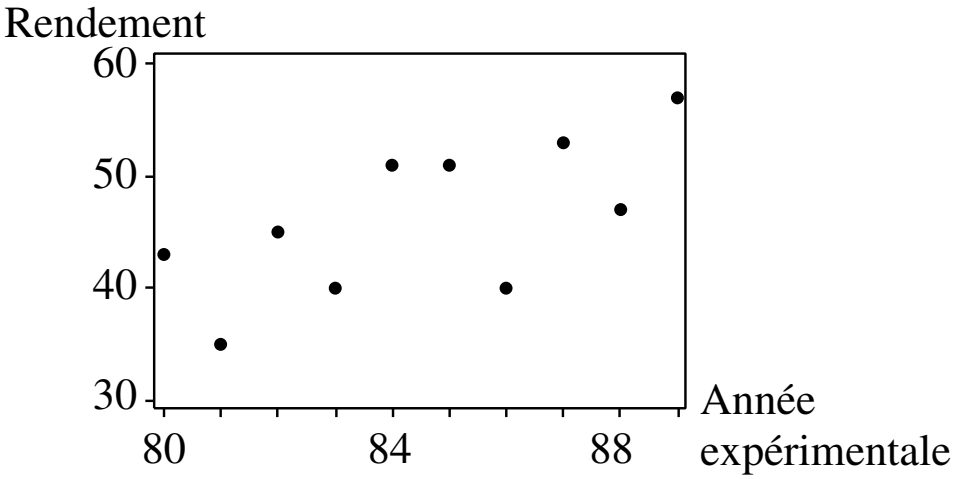
- On cherche à expliquer le rendement de champs de blé dans l'Allier en fonction de deux paramètres : l'année de culture (X_i) et le nombre de jours où la température minimale est descendue au dessous de 6°C, entre le premier et le 20 mai (Z_i).

- On présentera l'ensemble des données (Y_i, X_i, Z_i)

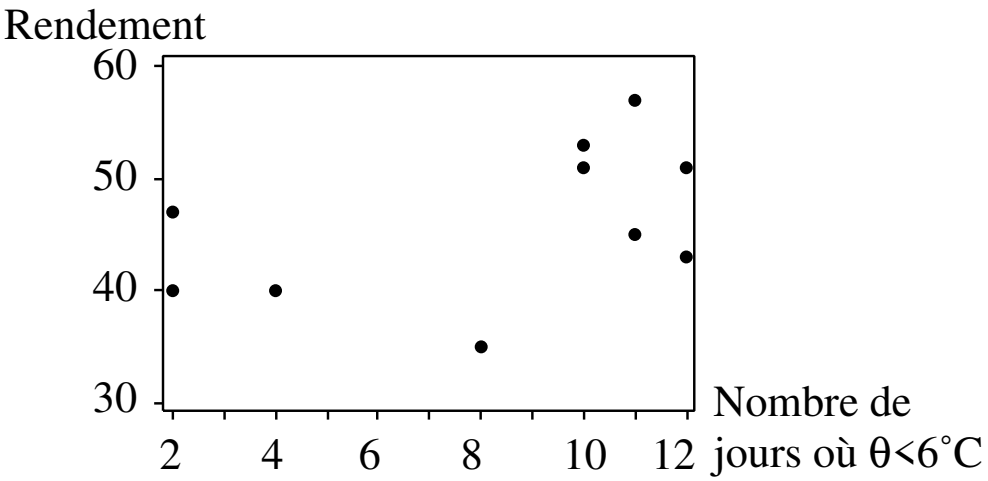
REGRESSION LINEAIRE MULTIPLE

Modélisations et graphiques simples :

$$Y_i = \alpha' + \beta' X_i + \varepsilon'_i$$



$$Y_i = \alpha'' + \beta'' Z_i + \varepsilon''_i$$



$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$

- On présentera les représentations graphiques correspondants aux deux analyses à un seul régresseur : ou X_i ou Z_i, avec les modèles mathématiques correspondants.

- La question que l'on peut poser est alors : "Au lieu de considérer indépendemment les effets des deux régresseurs X_i et Z_i, peut-on construire une combinaison linéaire de ceux-ci qui permette un meilleur ajustement des rendements de blé ? "

  ⟵ D'où le nouveau modèle ci-contre.

REGRESSION LINEAIRE MULTIPLE

Décomposition du problème :

$$
\begin{array}{rcccccc}
43 & = & \alpha & + & \beta_1.80 + & \beta_2.12 & + & \varepsilon_1 \\
35 & = & \alpha & + & \beta_1.81 + & \beta_2.\ 8 & + & \varepsilon_2 \\
45 & = & \alpha & + & \beta_1.82 + & \beta_2.11 & + & \varepsilon_3 \\
 & & & & \vdots & & & \\
57 & = & \alpha & + & \beta_1.89 + & \beta_2.11 & + & \varepsilon_{10}
\end{array}
$$

$$
\begin{pmatrix} Y_i \end{pmatrix} = \alpha. \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1. \begin{pmatrix} X_i \end{pmatrix} + \beta_2. \begin{pmatrix} Z_i \end{pmatrix} + \begin{pmatrix} \varepsilon_i \end{pmatrix}
$$

Ecriture matricielle ?

$$
\begin{array}{c}
\\ \\ \begin{pmatrix} Y_i \end{pmatrix} \end{array} = \begin{array}{c} \begin{matrix} \mathbf{1} & X_i & Z_i \end{matrix} \\ \begin{pmatrix} 1 & 80 & 12 \\ 1 & 81 & 8 \\ \vdots & \vdots & \vdots \\ 1 & 89 & 11 \end{pmatrix} \end{array} . \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \\ \varepsilon_i \\ \\ \end{pmatrix}
$$

**Revenir au tableau.*

- Ecrire toutes les équations relatives à l'ensemble des triplets (Y_i, X_i, Z_i)

- Mettre cette succession d'équations sous forme d'une combinaison linéaire de vecteurs colonnes et **s'arrêter là!**

- Distribuer aux stagiaires la feuille d'exercice n° 1 et les laisser la compléter:
  - écriture matricielle
  - dimension des vecteurs et matrices

- corriger avec eux.

REGRESSION LINEAIRE MULTIPLE

Formalisation :

$$Y = X\theta + \varepsilon$$

- Y est un vecteur de dimension 10
- ε est un vecteur de dimension 10
- θ est un vecteur de dimension 3
- X est une matrice de dimension $(10, 3)$

Conclusion: Tout modèle de régression linéaire relève du même formalisme:

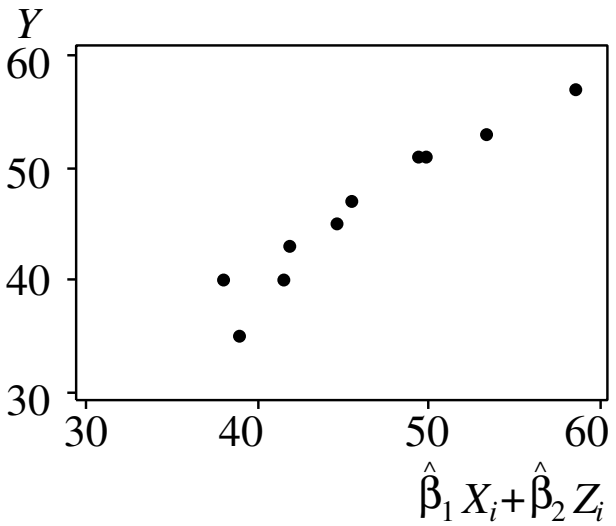
$$Y = X\theta + \varepsilon$$

L'augmentation du nombre de régresseurs conduit à une augmentation des dimensions de la matrice X et du vecteur θ.

REGRESSION LINEAIRE MULTIPLE

Résultats

$$\widehat{\alpha} = -132.31$$
$$\widehat{\beta_1} = \quad 1.99$$
$$\widehat{\beta_2} = \quad 1.23$$



Une combinaison linéaire des variables "Année" et "Nombre de Jours" conduit à une prédiction du Rendement bien meilleure que si chacune d'elles est considérée isolément

On peut terminer sur l'exemple en montrant que le modèle qui considère simultanément les deux régresseurs conduit à un ajustement bien meilleur que chacun des modèles tenant compte d'un seul régresseur.

ANALYSE DE VARIANCE A 2 FACTEURS

Etude de la teneur en huile de populations de tournesol d'origines variées, croisées à 2 ensembles de testeurs, T1 et T2

Origine / Testeur	AFRIQUE	HONGRIE	MAROC
T 1	43.54 45.30	44.25 42.55	47.28 49.40
T2	47.21 47.73	44.34 46.49	47.75 49.47

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

↙ Teneur en huile de la population k d'origine j et croisée au testeur i

↘ Teneur en huile moyenne de toutes les populations d'origine j croisées au testeur i

ANALYSE A DEUX FACTEURS CROISES, SANS INTERACTION:

Présentation d'un tableau de données à deux entrées, où sont notées les teneurs en huile de différentes populations de tournesol d'Afrique, de Hongrie et du Maroc, lorsqu'elles sont croisées à deux ensembles de testeurs : T_1 et T_2.
(Les testeurs sont des variétés de référence, dont les performances moyennes en croisement avec d'autres variétés sont généralement bonnes).

- **Modélisation mathématique simple:**

  - Y_{ijk} est la teneur en huile de la $k^{ième}$ population, d'origine j, croisée à l'ensemble des testeurs i.

  - μ_{ij} est la moyenne des teneurs en huile de toutes les populations d'origine j, croisées à l'ensemble des testeurs i.

ANALYSE DE VARIANCE A 2 FACTEURS

Autre écriture du modèle, sous l'hypothèse d'absence d'interaction entre les facteurs "testeur" et "origine"

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

α_i : testeur, β_j : origine

Origine / Testeur	AFRIQUE	HONGRIE	MAROC	$Y_{i..}$	$Y_{i..} - Y_{...}$
T 1	43.54 45.30	44.25 42.55	47.28 49.40	45.386	-0.89
T2	47.21 47.73	44.34 46.49	47.75 49.47	47.165	+0.89
$Y_{.j.}$	45.945	44.407	48.475	$Y_{...}$ = 46.275	
$Y_{.j.} - Y_{...}$	-0.33	-1.87	+2.20		

avec $\widehat{\mu} = Y_{...}$

$\widehat{\alpha_i} = Y_{i..} - Y_{...}$ et $\sum_i \alpha_i = 0$

$\widehat{\beta_j} = Y_{.j.} - Y_{...}$ et $\sum_j \beta_j = 0$

**Au tableau.*

Sous l'hypothèse d'absence d'interaction entre les facteurs "testeurs" et "origine", on présentera une autre écriture du modèle précédent, en faisant ressortir les effets de chacuns des facteurs.

- On montrera, intuitivement, comment sont estimés les paramètres $\widehat{\mu}$, $\widehat{\alpha_i}$ et $\widehat{\beta_j}$, à partir du tableau de données initial (que l'on aura copié au tableau) : on ajoutera progressivement les colonnes $(Y_{i..})$ et $(Y_{i..} - Y_{...})$ et les lignes $(Y_{.j.})$ et $(Y_{.j.} - Y_{...})$.

- On vérifie que $\Sigma_i \, \alpha_i = 0$ et $\Sigma_j \, \beta_j = 0$, avec les estimations choisies.

ANALYSE DE VARIANCE A 2 FACTEURS

$$\hat{\varepsilon}_{ijk} = Y_{ijk} - \widehat{\mu} - \widehat{\alpha_i} - \widehat{\beta_j}$$

$\hat{\varepsilon}_{ijk}$			
T 1	-1.515 0.245	0.735 -0.965	-0.305 1.815
T2	0.375 0.895	-0.955 1.195	-1.615 0.105

Construction de la table d'analyse de la variance

Source de variation	SCE	ddl	CM	F
Origine	33.745	2	16.87	10.3 **
Testeur	9.487	1	9.587	5.8 *
Résiduelle	13.11	8	1.639	

- On montrera aussi comment sont estimées les erreurs, et on donnera le tableau des $\widehat{\varepsilon_{ijk}}$.

- On peut alors, comme dans le cas de l'analyse à un facteur, construire une table d'analyse de la variance intégrant les effets de chacun des deux facteurs étudiés.

Conclusion: Il existe au moins une origine différente des autres; de plus, les deux ensembles de testeurs sont différents.

ANALYSE DE VARIANCE A 2 FACTEURS

TESTEUR 1	AFR	Y_{111}	$=\mu$	$+\alpha_1$		$+\beta_1$			$+\varepsilon_{111}$
		Y_{112}	$=\mu$	$+\alpha_1$		$+\beta_1$			$+\varepsilon_{112}$
	HON	Y_{121}	$=\mu$	$+\alpha_1$		$+$	β_2		$+\varepsilon_{121}$
		Y_{122}	$=\mu$	$+\alpha_1$		$+$	β_2		$+\varepsilon_{122}$
	MAR	Y_{131}	$=\mu$	$+\alpha_1$		$+$		β_3	$+\varepsilon_{131}$
		Y_{132}	$=\mu$	$+\alpha_1$		$+$		β_3	$+\varepsilon_{132}$
TESTEUR 2	AFR	Y_{211}	$=\mu$	$+$	α_2	$+\beta_1$			$+\varepsilon_{211}$
		Y_{212}	$=\mu$	$+$	α_2	$+\beta_1$			$+\varepsilon_{212}$
	HON	Y_{221}	$=\mu$	$+$	α_2	$+$	β_2		$+\varepsilon_{221}$
		Y_{222}	$=\mu$	$+$	α_2	$+$	β_2		$+\varepsilon_{222}$
	MAR	Y_{231}	$=\mu$	$+$	α_2	$+$		β_3	$+\varepsilon_{231}$
		Y_{232}	$=\mu$	$+$	α_2	$+$		β_3	$+\varepsilon_{232}$

$$\begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \end{pmatrix} = \mu. \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \alpha_1. \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \alpha_2. \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \beta_1. \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \beta_2. \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \beta_3. \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{131} \\ \varepsilon_{132} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \\ \varepsilon_{231} \\ \varepsilon_{232} \end{pmatrix}$$

**Au tableau.*
(On utilisera éventuellement le transparent pour une synthèse "propre" où tout est bien aligné.)

- Mettre les Y_{ijk} en ligne au tableau (à gauche pour pouvoir ensuite écrire les équations complètes).

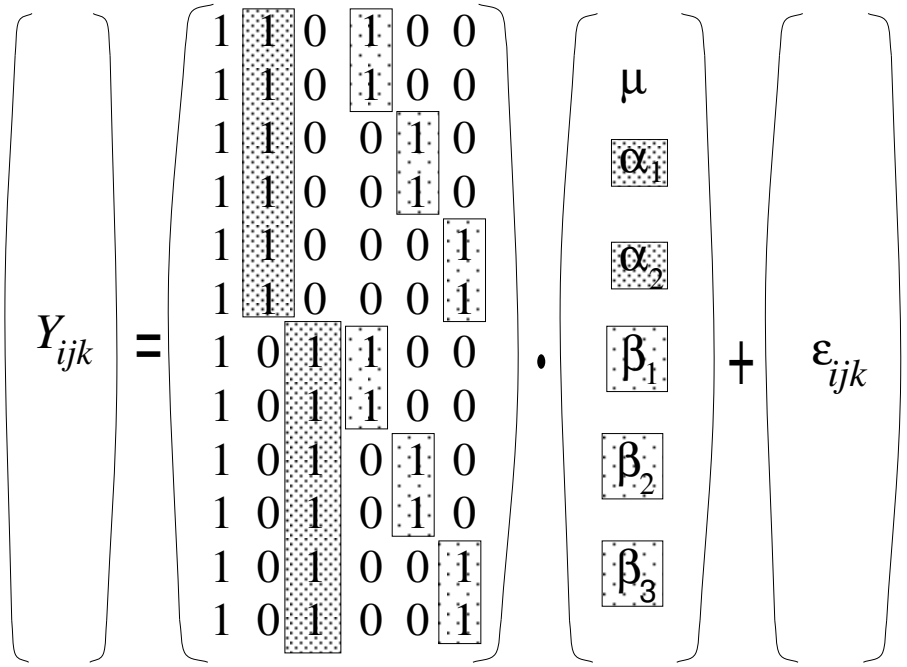
- Les décomposer selon la modélisation mathématique précédente qui fait apparaître les effets de chacun des deux facteurs : on décalera nettement les deux modalités du premier facteur (α_1, α_2) et puis celles du second (β_1 , β_2 ,β_3).

- Montrer alors l'écriture correspondant à une combinaison linéaire de vecteurs colonnes et EN RESTER LA!

- Distribuer ensuite les feuilles relatives à l'exercice n°2 et laisser aux stagiaires le temps de le résoudre.

ANALYSE DE VARIANCE A 2 FACTEURS

Ecriture matricielle



Formalisation

$$Y = X\,\theta + \varepsilon$$

Pour n observations :

- Y et ε sont des vecteurs de dimension 12
- θ est un vecteur de dimension 6
- X est une matrice de dimension $(12, 6)$

- On montrera (après l'avoir exposé au tableau) le corrigé de l'exercice en insistant sur le fait que tous les modèles d'analyse de la variance relèvent du même formalisme, les dimensions des vecteurs et matrices étant fonction du nombre de données de départ et de la modélisation choisie (c'est-à-dire du nombre de paramètres à estimer).

Remarque sur la définition d'un modèle linéaire

$$\mathbf{y}_1 = \alpha_1 + \beta_1.\mathbf{x}^2 + \gamma_1.\mathbf{x}^3 + \varepsilon_1 \quad (1)$$

$$\mathbf{y}_2 = \alpha_2 + \beta_2.\mathbf{log}\,\mathbf{x} + \gamma_2.\mathbf{e}^{-\mathbf{z}} + \varepsilon_2 \quad (2)$$

$$\mathbf{y}_3 = \alpha_3 + \beta_3.\mathbf{x}^{\gamma_3} + \varepsilon_3 \quad (3)$$

$$\mathbf{y}_4 = \alpha_4.\left(\frac{1}{\beta_4+\mathbf{x}^2}\right) + \gamma_4.\mathbf{e}^{-\delta_4.\mathbf{z}} + \varepsilon_4 \quad (4)$$

Les modèles (3) et (4) sont "non linéaires" en les paramètres $(\alpha, \beta, \gamma, \delta)$

Un modèle sera dit "linéaire", s'il est linéaire en les paramètres estimés.

- On montrera **l'opposition** entre les **modèles (1),(2)** (linéaires en α, β et γ) et **(3),(4)** (non linéaires en α, β et γ).

- les modèles (1) et (2) sont linéaires en les paramètres, même s'ils comprennent des fonctions "puissance", "logarithme" et "exponentielle".

INTRODUCTION AU MODELE LINEAIRE
FORMALISME GENERAL

Tout modèle, *linéaire en les paramètres à estimer*, sera exprimé selon le formalisme suivant :

$$\underset{(n,1)}{Y} = \underset{(n,p)}{X}\ \underset{(p,1)}{\theta} + \underset{(n,1)}{\varepsilon}$$

Tous les modèles de régression et d'analyse de variance relèvent du même formalisme !

Conclusions

1. Tout modèle linéaire relève du même formalisme.

2. Rappel des dimensions des vecteurs et matrices de l'équation:

$$Y = X\theta + \varepsilon$$

3. Tous les modèles de régression linéaire et d'analyse de la variance relèvent du même formalisme.

Pause conseillée avant d'envisager la résolution du système:

$$Y = X\theta + \varepsilon$$