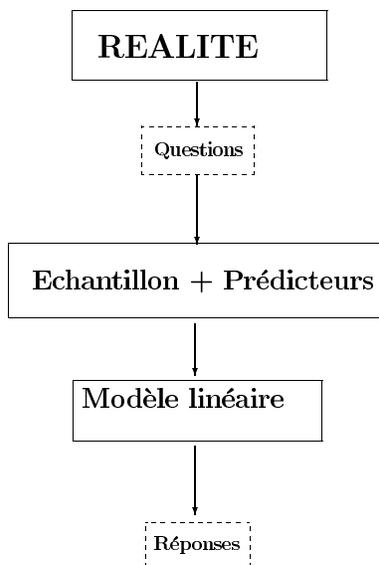


Plan du Chapitre 2

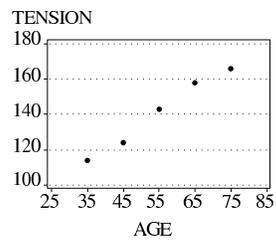
1. Utilité d'un modèle
2. Modèle et Sous-modèle
3. Test d'Hypothèses : comparaison de Modèles

Utilité d'un modèle



Réalité	inatteignable	
modèle choisi	$\mu_n = \sum x_{pn} \theta_p$	paramètres inconnus
modèle estimé	$\hat{\mu}_n = \sum x_{pn} \hat{\theta}_p$	paramètres estimés

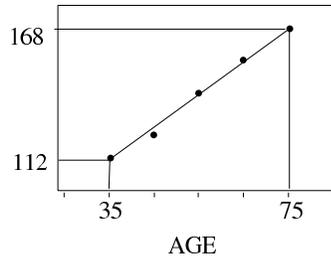
Exemple : Etude de la relation entre la tension artérielle et l'âge



Modèle choisi ?

Modèle estimé ?

TENSION



$$\hat{\beta} = \frac{168 - 112}{75 - 35} \quad 168 = \hat{\alpha} + 1.4 \times 75$$

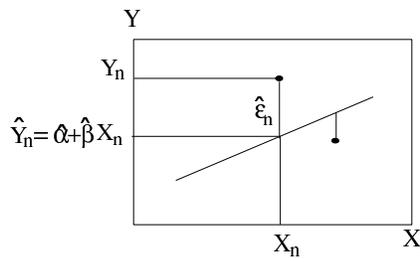
$$\hat{\beta} = 1.4 \quad \hat{\alpha} = 63$$

$$\hat{Y} = 63 + 1.4 \times X$$

Le meilleur modèle estimé ?

Quantifier l'utilité d'un modèle

Pour quantifier l'utilité d'un modèle, on définit une **distance** entre la valeur *prédite* \hat{Y}_n et la valeur observée Y_n sur tout l'échantillon.



Parmi les distances possibles, on trouve

- $\sum_n |Y_n - \hat{Y}_n|$
- $\sum_n (Y_n - \hat{Y}_n)^2$
- $\max |Y_n - \hat{Y}_n|$

La distance choisie $\sum_n (Y_n - \hat{Y}_n)^2$ correspond à la **Somme des Carrés des Écarts Résiduelle** et est notée SCE_R .

modèle choisi $Y_n = \sum_p x_{np} \theta_n + \epsilon_n$

modèle estimé $Y_n = \sum_p x_{np} \hat{\theta}_n + \hat{\epsilon}_n$

Soit \hat{Y}_n la meilleure prédiction

de Y_n dans le modèle estimé

$$\hat{Y}_n = \sum_p x_{np} \hat{\theta}_n$$

Critère de choix du "meilleur" modèle estimé, celui qui minimise la **distance**

$$\sum_n (Y_n - \hat{Y}_n)^2$$

⇕

Trouver le "meilleur" modèle estimé, c'est donc rechercher les θ_p tels que

$$\sum_n \left(Y_n - \sum_p x_{pn} \theta_p \right)^2 \text{ soit minimale}$$

c'est le critère des **Moindres Carrés**

Synthèse sur l'utilité d'un modèle

L'utilité d'un modèle est donc quantifiée par

$$\sum_n (Y_n - \hat{Y}_n)^2$$

qui est la Somme des Carrés des Écarts Résiduelle.

On peut aussi vouloir le modèle demandant le moins de paramètres (principe de parcimonie)

le modèle le plus utile sera celui qui pour "une même SCE_R" aura le moins de paramètres

définition

Modèle

$$\mu_n = \sum_p x_{np} \theta_n$$

Sous-modèle
emboîté

on a ôté un ou des paramètres au
modèle (soit en les annulant, soit en
les égalisant)

Exemples

Exemple : Étude des forêts à travers la hauteur d'arbres

Modèle	$\mu_n = \mu + \alpha_i$
--------	--------------------------

↓

$\alpha_1 = \alpha_2 = \alpha_3$

↓

Sous-modèle emboîté	$\mu_n = \mu$
---------------------	---------------

Exemple : Étude du rendement de blé en fonction de doses de fertilisants

Modèle	$\mu_n = 1.\theta_0 + Z1_n.\theta_1 + Z2_n.\theta_2 + Z3_n.\theta_3$
--------	----------------------------------------------------------------------

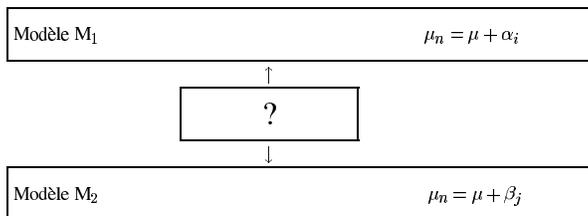
↓

$\theta_1 = 0$

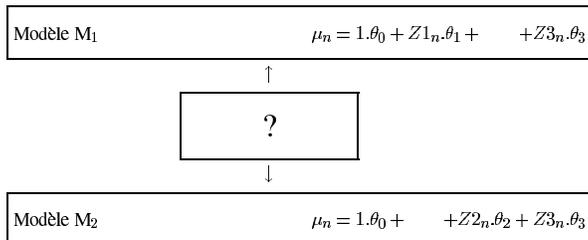
↓

Sous-modèle emboîté	$\mu_n = 1.\theta_0 + \quad + Z2_n.\theta_2 + Z3_n.\theta_3$
---------------------	--------------------------------------------------------------

Exemple : Étude de la teneur en huile de population de tournesol



Exemple : Étude du rendement de blé en fonction de doses de fertilisants



Modèle irréductible

définition

Un modèle est sous **forme irréductible**

⇔ **tous ses paramètres** sont importants

⇔ il n'existe pas un modèle moins paramétré

Exemple : Étude des forêts à travers la hauteur d'arbres

$$\mu_n = \mu + \alpha_i \quad i = 1 \dots I = 3$$

Exemple : Étude du rendement de blé en fonction de doses de fertilisants

$$\mu_n = 1.\theta_0 + Z1_n.\theta_1 + Z2_n.\theta_2 + Z3_n.\theta_3$$

Exemple : Niveau scolaire

$$\mu_n = 1.\theta_0 + Z1_n.\theta_1 + Z2_n.\theta_2 + Z3_n.\theta_3 + Z4_n.\theta_4$$

μ_n : note moyenne espérée au lycée

$Z1_n$: note moyenne au collège

$Z2_n$: note moyenne en mathématique au lycée

$Z3_n$: note moyenne en philosophie au lycée

$Z4_n$: notes moyennes au lycée ($Z2_n + Z3_n$)

Test par rapport au modèle le plus simple

Modèle M_1	$\mu_n = \theta_0 + \sum_{p=1}^{P-1} x_{np} \theta_p$
Nbre de paramètres irréductibles	P
SCE_{M_1}	$\sum_{n=1}^N (Y_n - \hat{Y}_n)$

|

$H_0 = \{\theta_p = 0 \quad p = 1 \dots P - 1\}$
$H_1 = \{\exists p / \theta_p \neq 0\}$

↓

Modèle M_0	$\mu_n = \theta_0$
Nbre de paramètres irréductibles	1
SCE_{M_0}	$\sum_{n=1}^N (Y_n - Y.)$

$$\begin{array}{ccccccc}
 SCE_{M_0} & = & (SCE_{M_0} - SCE_{M_1}) & + & SCE_{M_1} \\
 \downarrow & & \downarrow & & \downarrow \\
 ddl & & N-1 & = & P-1 & + & N-P
 \end{array}$$

$$SCE_{M_0} = \sum_{n=1}^N (Y_n - \bar{Y})^2$$

$$SCE_{M_1} = \sum_{n=1}^N (Y_n - \hat{Y}_n)^2$$

Statistique de test

Statistique :

$$F = \frac{(SCE_{M_0} - SCE_{M_1}) / (P - 1)}{SCE_{M_1} / (N - P)}$$

$$H_0 = \{\theta_p = 0 \quad p = 1 \dots P - 1\}$$

$$H_1 = \{\exists p / \theta_p \neq 0\}$$

$$F \sim F(P - 1, N - P) \text{ sous } H_0$$

Source	ddl	SC	CM	F	Pr > F
Modèle	P-1	$SCE_{M_0} - SCE_{M_1}$	CM_M	F	$\Pr(F(P-1, N-P) > F)$
Résiduelle	N-P	SCE_{M_1}	CM_R		
Totale	N-1	SCE_{M_0}			

$$CM_M = \frac{(SCE_{M_0} - SCE_{M_1})}{P-1}$$

$$CM_R = \frac{SCE_{M_1}}{N-P} = \hat{\sigma}^2$$

$$F = \frac{(SCE_{M_0} - SCE_{M_1})/(P-1)}{SCE_{M_1}/(N-P)}$$

Modèle M_2 **Test M_1 contre M_2**

Source	ddl	SC	CM	F
Modèle	P-1	$SCE_{M_0} - SCE_{M_2}$	CM_{M_2}	$\frac{CM_{M_2}}{CM_{R_2}}$
Résiduelle	N-P	SCE_{M_2}	CM_{R_2}	
Totale	N-1	SCE_{M_0}		

|

$H_0 =$ Les paramètres non communs à M_1 et M_2 sont tous nuls

$H_1 =$ Au moins un des paramètres non communs à M_1 et M_2 est non nul

↓

Sous Modèle emboîté M_1

Source	ddl	SC	CM	F
Modèle	Q-1	$SCE_{M_0} - SCE_{M_1}$	CM_{M_1}	$\frac{CM_{M_1}}{CM_{R_1}}$
Résiduelle	N-Q	SCE_{M_1}	CM_{R_1}	
Totale	N-1	SCE_{M_0}		

Statistique : $F = \frac{(SCE_{M_1} - SCE_{M_2})/(P-Q)}{SCE_{M_2}/(N-P)}$

Statistique :

$$F = \frac{(SCE_{M_1} - SCE_{M_2}) / (P - Q)}{SCE_{M_2} / (N - P)}$$

$H_0 = \{ \text{Les paramètres non communs à } M_1 \text{ et } M_2 \text{ sont tous nuls} \}$

$H_1 = \{ \text{Au moins un des paramètres non communs à } M_1 \text{ et } M_2 \text{ est non nul} \}$

$F \sim F(P - Q, N - P) \quad \text{sous } H_0$