QUALITÉ ET SOUS-MODÈLES

Plan du Chapitre 2

1. Utilité d'un modèle

2. Modèle et Sous-modèle

3. Test d'Hypothèses : comparaison de Modèles

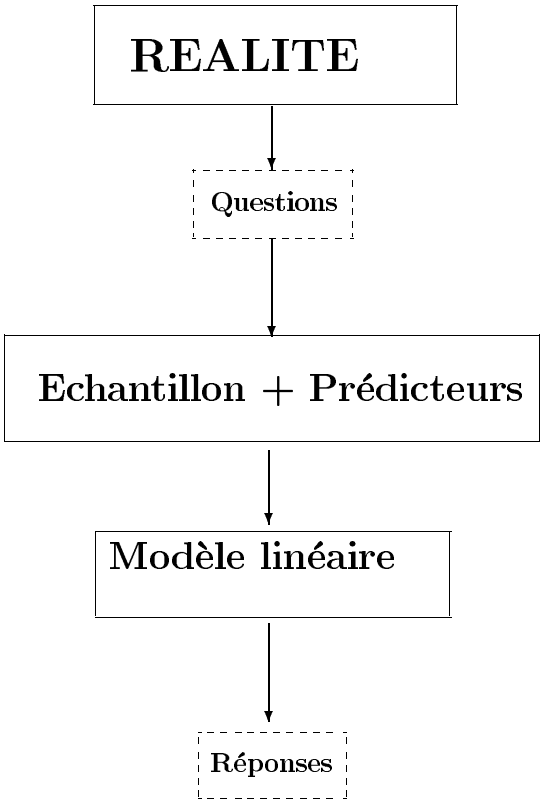
Les objectifs liés aux différents points sont les suivants :

— **Utilité** pourquoi utiliser un modèle ? quelles sont les simplifications qui sont faites ? Comment quantifier l'utilité d'un modèle ?

— **Modèle et sous modèle** définir la notion de sous modèle emboîté, la notion de modèle irréductible.

— **Test d'hypothèse** définir la notion de comparaison de modèles.

Utilité d'un modèle



Pourquoi utiliser un modèle ?

Schéma : Un modèle est une simplification de la réalité.

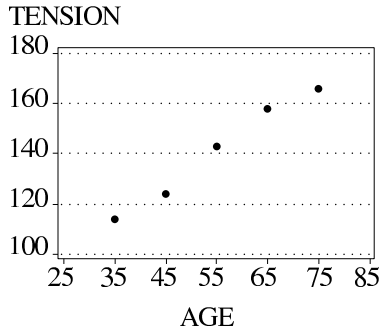
Un modèle est toujours une simplification d'une réalité complexe et non connue. Il résulte de connaissances que l'on a et des données disponibles.

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

Pour rendre compte d'une réalité sur laquelle on se pose une ou des questions, on choisit un modèle dont les paramètres sont inconnus mais que l'on suppose rendre compte de la réalité. Puis à partir d'un échantillon et de prédicteurs, on va estimer les paramètres du modèle choisi.

FPSTAT INRA 1997 II-Qualité et Sous-Modèles-3

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



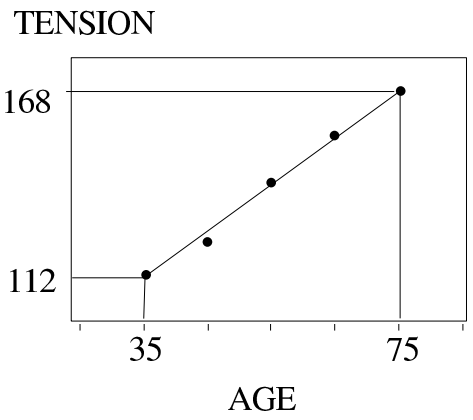
Modèle choisi ?

Modèle estimé ?

Exercice :

On propose un modèle de la forme : $Y = \alpha + \beta X$. C'est le modèle que l'on a choisi. Graphiquement, "à l'œil", chaque stagiaire donne les estimations $\hat{\alpha}$ et $\hat{\beta}$, de la droite qu'il a tracée.

Un exemple de modèle estimé



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

$\widehat{\alpha} = 63$

$\widehat{\beta} = 1.4$

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

Le meilleur modèle estimé ?

On a tracé la droite passant par les points extrêmes : on obtient comme modèle (droite) estimé $\widehat{Y} = 63 + 1.4 \times X$

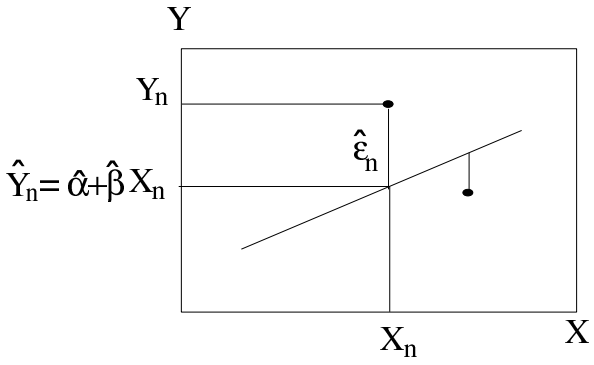
Problème : il y a autant de modèles proposés que de stagiaires. Quel est le "meilleur" modèle estimé ?

Remarque : Peut-on conclure que la tension moyenne à la naissance est 63 ? Non, cela revient à extrapoler et à supposer que la relation entre l'âge et la tension est linéaire entre 0 et 75 ans. En réalité, la courbe est sigmoïdale.

FPSTAT INRA 1997 II-Qualité et Sous-Modèles-5

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

— $\sum_n \left| Y_n - \hat{Y}_n \right|$

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

— $\max \left| Y_n - \hat{Y}_n \right|$

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

— une distance **raisonnable** pour le choix entre les différents modèles estimés est de vouloir que $\sum_n \left(Y_n - \hat{Y}_n \right)^2$ soit minimum

c.a.d. la norme carrée de la différence entre Y et de son meilleur prédicteur ; mais on aurait pu choisir d'autres distances par ex :
$\max \mid Y_n - \hat{Y}_n \mid$ soit minimum ; distance qui a le gros inconvénient de donner beaucoup de poids aux données aberrantes

— Par contre

$$\sum_n \left(Y_n - \hat{Y}_n \right)^2 \quad \text{minimum}$$

$$\Updownarrow$$

$$\sum_n \hat{\epsilon}_n^2 \quad \text{minimum}$$

c'est donc déclarer le "meilleur", celui qui minimise la variabilité de l'erreur résiduelle, c.a.d. la part de variabilité de Y qui ne peut pas être expliquée par le modèle choisi.

Remarque : La somme des carrés des écarts résiduelle divisée par le nombre de degrés de liberté correspondant est une estimation sans biais de la variance σ^2.

Critère d'estimation

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**

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

$$\Updownarrow$$

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**

Il faut un critère pour choisir entre les différents modèles estimés, ce critère est lié à la distance choisie.

Sous les postulats du Modèle Linéaire, la meilleure prédiction de Y_n dans un modèle estimé, c'est

$$\hat{Y}_n = \hat{\mu}_n = \sum_p x_{pn}\hat{\theta}_p$$

Remarque : Les paramètres ainsi estimés et les valeurs ajustées sont sans biais et de variance minimale comparativement aux autres estimateurs combinaisons linéaires des observations.

Remarque : Dans ce cas, l'estimation par le maximum de vraisemblance conduit aux mêmes estimations $\hat{\theta}_p$.

Synthèse sur l'utilité d'un modèle

L'utilité d'un modèle est donc quantifiée par

$$\sum_{n} \left(Y_n - \hat{Y}_n \right)^2$$

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

FPSTAT INRA 1997 II-Qualité et Sous-Modèles-8

Modèle et Sous-modèle

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)

FPSTAT INRA 1997 II-Qualité et Sous-Modèles-9

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

1. les moyennes sont identiques dans les 3 forêts $\mu_1 = \mu_2 = \mu_3$
2. on pense que le régresseur Z1 (fertilisant AZ) n'agit pas sur le rendement ; on testera alors le modèle annulant le paramètre θ_1.

FPSTAT INRA 1997 II-Qualité et Sous-Modèles-10

Contre-exemples

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

Modèle M_1	$\mu_n = \mu + \alpha_i$
↑	
?	
↓	
Modèle M_2	$\mu_n = \mu + \beta_j$

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

Modèle M_1	$\mu_n = 1.\theta_0 + Z1_n.\theta_1 + \quad + Z3_n.\theta_3$
↑	
?	
↓	
Modèle M_2	$\mu_n = 1.\theta_0 + \quad + Z2_n.\theta_2 + Z3_n.\theta_3$

Les sous-modèles ne sont pas emboîtés car on ne peut pas passer du modèle au sous-modèle en annulant des paramètres ou en les égalisant.

Question : Pourquoi un sous modèle emboîté ?

C'est indispensable pour comparer deux modèles.

Modèle irréductible

définition

Un modèle est sous **forme irréductible**

$\Longleftrightarrow$ **tous ses paramètres** sont importants

$\Longleftrightarrow$ il n'existe pas un modèle moins paramétré

FPSTAT INRA 1997 II-Qualité et Sous-Modèles-12

Modèle irréductible ? (oui/non)

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

$\mu_n = \mu + \alpha_i \qquad i = 1 \cdots I = 3$

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 ($Z2_n + Z3_n$)

Si le modèle est réductible, on donnera au tableau l'écriture du modèle irréductible ainsi que le nombre de paramètres irréductibles.

1. le modèle est réductible

   — le modèle irréductible : $\mu_n = \mu_i \quad i = 1 \cdots 3$

   — le nombre de paramètres irréductibles est 3

2. le modèle est irréductible

   — le nombre de paramètres irréductibles est 4

3. le modèle est réductible

   — le modèle irréductible est :

   $\mu_n = 1.\theta_0 + Z1_n.\theta_1 + Z2_n.\theta'_2 + Z3_n.\theta'_3$
   $(\theta'_2 = \theta_2 + \theta_4 \; et \; \theta'_3 = \theta_3 + \theta_4)$

   — le nombre de paramètres irréductibles est 4.

Question : Pourquoi un modèle irréductible ?

C'est pour faciliter le calcul des degrés de liberté pour le test de comparaison de 2 modèles (nécessairement emboîtés).

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}	$\sum_{n=1}^{N} \left(Y_n - \widehat{Y}_n\right)$

|

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

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

↓

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

Exemples de modèle linéaire le plus simple :

en régression linéaire simple :

$\mu_n = \theta_0$ (nullité de la pente)

en analyse de variance à un facteur :

$\mu_n = \theta_0 = \mu$ (égalité des groupes, dans l'exemple forêt $\mu_1 = \mu_2 = \mu_3$)

Remarque : Sur ces exemples, on pourra visualiser la notion de sous modèles emboîtés et l'hypothèse nulle associée.

Comparer 2 modèles emboîtés signifie tester l'apport d'explication du modèle M_1 par rapport au modèle le plus simple M_0 vis à vis du surcoût expérimental occasionné par le nombre de paramètres supplémentaires dans le modèle M_1.
A chaque modèle, on associe son nombre de paramètres irréductibles et la somme des carrés des écarts (de l'erreur) résiduelle.

Somme des carrés écarts résiduelle de M_1 :

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

Somme des carrés écarts résiduelle de M_0 :

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

Remarque : La somme des carrés des écarts résiduelle SCE est indépendante de la forme irréductible ou non du modèle (intrinsèque).

Décomposition de la variabilité

$$SCE_{M_0} = (SCE_{M_0} - SCE_{M_1}) + SCE_{M_1}$$
$$\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$

ddl N-1 = P-1 + N-P

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

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

	Observations	Y_n	
Variation résiduelle	Modèle M_1	$\widehat{Y}_n$	$SCE_{M_1} = \sum \left(Y_n - \widehat{Y}_n\right)^2$
Variation totale	Modèle M_0	$Y. = \widehat{\theta}_0$	$SCE_{M_0} = \sum (Y_n - Y.)^2$

variation totale = variation expliquée + variation résiduelle

Calcul des degrés de liberté

(Le degré de liberté représente le nombre de relations indépendantes).

Pour la variation totale, on a N observations et une information estimée : la moyenne $Y.$ et ainsi on a N-1 ddl.

Pour la variation résiduelle, on a N observations et P informations estimées par le modèle M_1, et ainsi on a N-P ddl.

Pour la variation expliquée, on a (N-1) - (N-P) = P-1 ddl.

Statistique de test

Statistique :

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

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

$H_1 = \{\exists_P \ / \ \theta_p \neq 0\}$

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

Le test F compare le carré moyen (quotient de la somme des écarts résiduelle SCE par le nombre de degré de liberté) de la variation expliquée par le modèle au carré moyen de la variation résiduelle du modèle choisi. On teste si l'information, l'explication apportée par le modèle est significative.

— sous H_0

$$SCE_{M_1} \sim \sigma^2 \chi^2(N-P)$$
$$SCE_{M_0} - SCE_{M_1} \sim \sigma^2 \chi^2(P-1)$$

Ces deux statistiques sont de plus indépendantes d'où la loi de Fisher

$F(P-1, N-P)$ pour le test F

Table d'analyse de la variance

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

$$\mathrm{CM_M} = \frac{(\mathrm{SCE_{M_0}} - \mathrm{SCE_{M1}})}{\mathrm{P}-1}$$

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

$$\mathrm{F} = \frac{(\mathrm{SCE_{M_0}} - \mathrm{SCE_{M1}})/(\mathrm{P}-1)}{\mathrm{SCE_{M_1}}/(\mathrm{N}-\mathrm{P})}$$

Résumé des informations dans la table d'analyse de variance, que le modèle soit une régression, une analyse de la variance ou de covariance.

Remarque : $\mathrm{SCE_{M_0}} - \mathrm{SCE_{M_1}}$ représente la partie de la résiduelle de $\mathrm{M_0}$ expliquée par le passage du modèle $\mathrm{M_0}$ au modèle $\mathrm{M_1}$.

FPSTAT INRA 1997 II-Qualité et Sous-Modèles-17

Test M_1 contre M_2

Modèle 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_1et M_2 sont tous nuls

H_1 = Au moins un des paramètres non communs à M_1et 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 : $F = \frac{(SCE_{M_1} - SCE_{M_2})/(P-Q)}{SCE_{M_2}/(N-P)}$

On a :

— Modèle M_2 à P paramètres irréductibles

— Sous Modèle Emboîté M_1 à Q paramètres irréductibles (Q < P).

On veut tester la qualité d'un modèle M_1 contre un modèle M_2 pour réduire le nombre d'estimations de paramètres.

Pour chaque modèle M_1 et M_2 , on décrit la table d'analyse de variance correspondant au test par rapport à M_0

Remarque : $SCE_{M_1} - SCE_{M_2}$ représente la partie de la résiduelle expliquée par le passage du "petit modèle" M_1 au "grand modèle" M_2 .

Le test de Fisher compare le carré moyen de la variation de la résiduelle expliquée par le passage du "petit modèle" M_1 au "grand modèle" M_2 au carré moyen de la résiduelle du "grand modèle"M_2 . On teste si l'explication apportée par le modèle M_2 par rapport au modèle M_1 est significative.

Statistique de test

Statistique :

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

$H_0 = \{$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$\}$

$F \sim F(P - Q, N - P)$ sous H_0