Choix du modèle en régression linéaire

(C. Preda)

Objectifs : - description du lien entre Y et $\{X_1, \dots, X_p\}$

- estimation des paramètres $\beta_0, \dots, \beta_p$

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$$

(partie commune de deux modèles)

- prévision

utiliser un échantillon test

Critères classiques pour le choix de modèles

I: Le R^2 :

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$$

$$R^2(p) = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Remarque :

1) $R^2(p) \leq R^2(p+1)$

⚠ Ajouter une variable dans le modèle augmente le R^2

2) si $n \leq p+1$ alors $R^2(p) = 1$.

3) $R^2(p) \equiv R^2(n, p)$ alors

$$R^2(n_1, p) \leq R^2(n_2, p) \quad , \quad n_1 > n_2$$

⚠ Il ne faut pas utiliser le R^2 pour comparer des modèles de taille différente (pas le même nombre de variables explicatives).

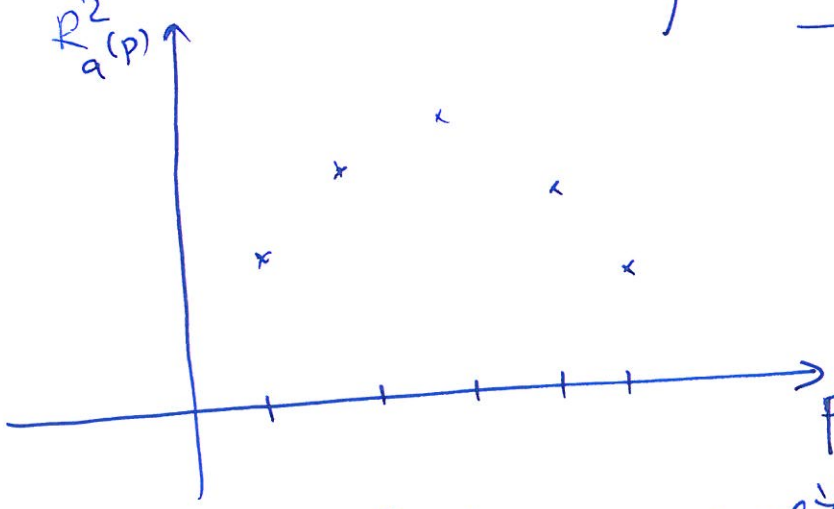
II Le R^2 ajusté

Correction du R^2 corrigeant les "défauts" du R^2

$$R^2_a(p) = 1 - \frac{n-1}{n-p-1}\left(1 - R^2(p)\right)$$

$$= 1 - \frac{n-1}{n-p-1} \cdot \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum (y_i - \bar{y})^2}$$

$\frac{n-1}{n-p-1}$ → facteur de correction



Utilisé pour le choix du modèle afin de décrire les données.

III Le C_p de Mallows

Soit M_q un modèle avec q variables choisies parmi les $\{X_1, \dots X_p\}$.

$$C_p(q) = \frac{\sum_{i=1}^{n} \varepsilon_i^2(q)}{\hat{\sigma}^2} - n + 2(q+1)$$

où $\hat{\sigma}^2$ est l'estimation de la variance résiduelle avec le modèle à p variables (modèle complet).

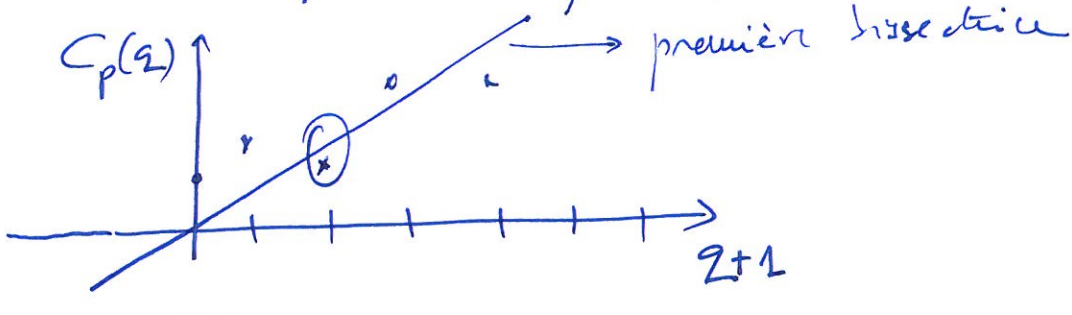
Observation : Si le modèle à q variables est "le bon" alors

$$\sum_{i=1}^{n} \varepsilon_i^2(q) \simeq (n-q-1)\cdot\sigma^2$$

et donc

$$C_p(q) \simeq \frac{(n-q-1)\sigma^2}{\hat{\sigma}^2} - n + 2(q+1) \simeq q+1.$$

Donc, il est recommandé de choisir le modèle avec $C_p(q)$ le plus proche de $q+1$.



IV AIC et BIC

→ AIC : Akaike information Criterion

$$AIC(p) = -2 \log\left(L(\beta_0, \dots \beta_p)\right) + 2(p+1).$$

avec

$$\log\left(L(\beta_0 \dots \beta_p)\right) = -\frac{n}{2} \log\left(\frac{\sum \varepsilon_i^2}{n}\right) - \frac{n}{2}\left(1 + \log \frac{2}{n}\right)$$

On choisit le modèle avec l'AIC le plus petit

→ BIC : Bayesian information criterion

$$BIC(p) = -2 \log\left(L(\beta_0, \dots \beta_p)\right) + (p+1) \log(n)$$

- pour $n > 7$, $\log(n) > 2$ et donc BIC a tendance à sélectionner des modèles "plus petites" que l'AIC.

Tableau :

Critère	Taille du modèle
Bic	faible
Aic	↓
R^2_a	forte

(V) Test des modèles emboités

Soit deux modèles

$M(p_0)$: variables explicatives $X_{i_1}, X_{i_2}, \dots X_{i_{p_0}}$ (p_0 variables parmi p variables $\{X_1, \dots X_p\}$)

$M(p)$: variables $\{X_1 \dots X_p\}$.

$p_0 < p$:

On dit que le modèle $M(p_0)$ est emboité dans le modèle $M(p)$

L'hypothèse nulle est que

$\underline{H_0}$: $E(Y \mid X_1 = x_1, \dots X_p = x_p) = f(x_{i_1}, \dots x_{i_{p_0}})$.

(il n'y a que les variables $X_{i_1}, \dots, X_{i_{p_0}}$ qui sont linéairement liées à Y).

H_1 : $\neg H_0$

Le test de Fisher associé à H_0 est

$$F = \frac{\sum_{i=1}^{n} \varepsilon_i^2(p_0) - \sum_{i=1}^{n} \varepsilon_i^2(p)}{\sum_{i=1}^{n} \varepsilon_i^2(p)} \times \frac{n-p-1}{p-p_0}$$

Sous H_0 : $F \sim \text{Fisher}(p-p_0, n-p-1)$.

Cas particulier : $p_0 = p-1$

$\Rightarrow$ test sur l'influence d'une variable.

Comment construire un bon modèle ?

On a des critères pour comparer des modèles.

Avec p variables $\{X_1 \dots X_p\}$ on peut construire 2^p modèles.

Exemple $p = 3$.

On ne peut pas explorer et comparer tous les modèles pour p grand. La recherche exhaustive est donc limitée.

- Recherche des modèles "optimaux" pas-à-pas (step-by-step)

(1) Méthode pas-à-pas ascendante (forward)

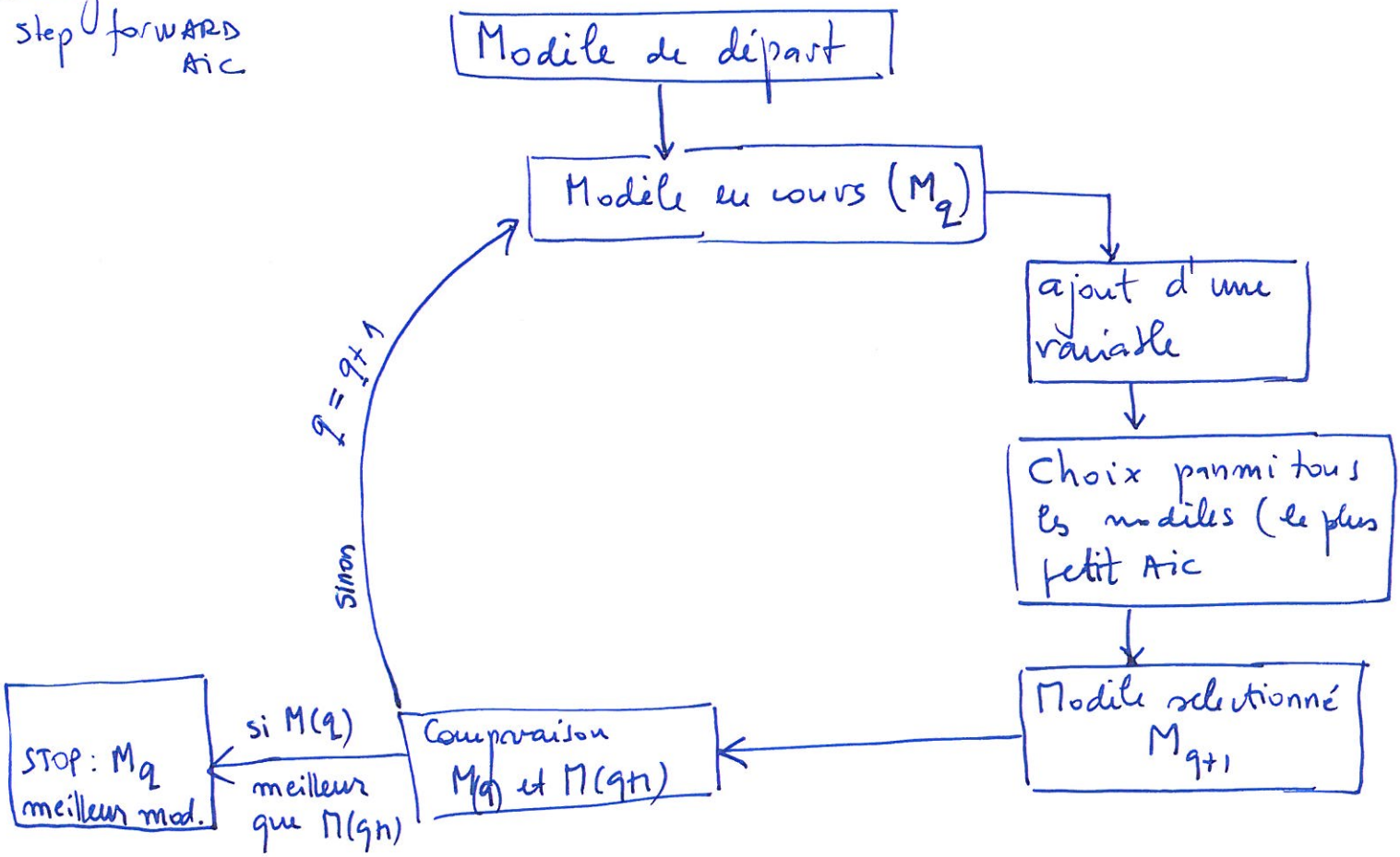
- On part avec un modèle avec la seule "variable", la constante "1".
- On ajoute à chaque pas la variable X_j qui optimise un critère de choix (R^2_a, C_p, AIC, ...)

→ On s'arrête lorsque : - toutes les variables ont été intégrées
ou
- détérioration du critère lors de l'ajout

(2) Méthode pas-à-pas descendante (backward)

- On part avec le modèle complet (p variables)
- à chaque pas on enlève la variable X_j qui optimise le critère pour les modèles sans la variable X_j ;
- arrêt lorsque toutes les variables sont retirées ou que le critère ne s'améliore pas par rapport au pas précédent.

Exemple
step forward Aic



③ Méthode progressive (step-wise).

Même principe que la sélection ascendante mais on peut éliminer des variables déjà introduites (celles qui ne sont pas significatives).

Fonctions utiles en R et SAS

en R : package "leaps"

fonction : regsubsets et summary

: package "MASS"

fonction stepAic

en SAS : proc reg

model Y = ... /selection = forward

/selection = rsquare cp adjrsq

bic best = 1 ...