Choix de modèle

Critères de choix

Procédures de sélection

stat4all

Introduction

Remarque

Dans tous les modèles que nous allons considérer, la constante sera par défaut automatiquement présente. Ainsi, si le modèle comporte $p-1$ variables explicatives, il y aura p paramètres dans le modèle.

Critères

Il existe différentes façon de choisir k variables explicatives parmi P disponibles. Nous allons en détailler certaines.

R^2

Le R^2 est le critère le plus simple pour commencer :

- Il a déjà été introduit dans le cours portant sur la régression linéaire.
- C'est une mesure qui permet d'évaluer le degré d'adéquation du modèle.

Attention

Le problème de critère est qu'il croit de façon monotone avec le nombre de variables dans le modèle.

stat4all

R^2 ajusté

La formule du R^2 ajusté, noté R^2_{adj}, est :

$$R^2_{adj} = 1 - \frac{SC_{res}}{SC_{tot}} \frac{n-1}{n-P}$$

Propriétés

- R^2_{adj} peut prendre des valeurs négatives,
- $R^2_{adj} < r^2$ pour $P > 1$,
- R^2_{adj} n'augmente pas forcément avec le nombre de variables.

stat4all

C_p de Mallow

Définition

Le C_p de Mallow est définit de la façon suivante :

$$C_p = \frac{SC_{res}}{\hat{\sigma^2}} - (n - 2p)$$

Il faut donc disposer d'une estimation de σ^2. Nous avons à notre disposition :

$$s^2 = \frac{SC_{res}}{n - p}$$

stat4all

C_p de Mallow

- Un rapide calcul donne alors : $C_p = p$. Ce qui n'a plus aucun intérêt. Pour contrer cela, nous prendrons pour estimation de la variance celle obtenue dans le modèle complet. Ainsi, seul $C_P = P$.
- Nous prendrons alors le modèle tel que C_p est le plus proche possible de p.

stat4all

Critère AIC

- Le critère d'information AIC s'applique aux modèles estimés par une méthode du maximum de vraisemblance : les analyses de variance, les régressions linéaires multiples, les régressions logistiques et de Poisson peuvent rentrer dans ce cadre
- Le critère AIC représente donc un compromis entre le biais diminuant avec le nombre de paramètres libres, et la parcimonie , volonté de décrire les données avec le plus petit nombre de paramètres possibles.
- Dans le cadre du modèle linéaire gaussien, ce critère s'écrit :

$$AIC = 2k^* + n\left[ln\left(\frac{2\pi \times SC_{res}}{n}\right) + 1\right]$$

stat4all

Critère AIC

Remarques

- La rigueur voudrait que tous les modèles comparés dérivent tous d'un même « complet » inclus dans la liste des modèles comparés.
- Il est nécessaire de vérifier que les conditions d'utilisation du modèle complet et de celui sélectionné sont remplies.
- Le meilleur modèle est celui possédant l'AIC le plus faible.

stat4all

Critère AIC corrigé

Lorsque le nombre de paramètres libres est grand par rapport au nombre d'observations, c'est-à-dire si :

$$n/k^* < 40,$$

il est recommandé d'utiliser l'AIC corrigé définit par :

$$AIC_c = AIC + \frac{2k^*(k^*+1}{n-k^*-1}.$$

stat4all

Critère BIC

- Il se définit par :

$$BIC = -2\log(L) + k^* \log(n).$$

- Il est plus parcimonieux que le critère AIC puisqu'il pénalise plus le nombre de variables présentent de le modèle. Ripley en 2003, souligne que l'AIC a été introduit pour retenir des variables pertinentes lors de prévisions, et que le critère BIC vise la sélection de variables statistiquement significatives dans le modèle.

Choix de modèle

Critères de choix

Procédures de sélection

stat4all

Introduction aux procédures de sélection

- Maintenant que nous disposons de critères pour comparer deux modèles, reste à trouver une procédure donnant les modèles à comparer.
- Si je dispose de P variables explicatives, le nombre de modèles possibles est 2^P. Ce nombre augmente rapidement :
  - $2^{10} = 1024$
  - $2^{20} = 1048576$
  - $2^{30} = 1073741824$
  - ...

stat4all

Recherche exhaustive

Une façon simple de procéder est, lorsque c'est possible, de faire une recherche exhaustive.

Recherche exhaustive

- Pour tous les modèles possibles, calculer le critère.
- Choisir le modèle qui optimise le critère.

Hélas, ceci n'est pas toujours possible, bien qu'il s'agisse de la meilleure manière de procéder.

stat4all

Procédures

Nous avons vu en cours les procédures suivantes :

- Les procédures pas à pas, ou stepwise :
  - forward (partir du modèle sans variable, et les ajouter fur à et à mesure),
  - backward (partir du modèle avec toutes les variables, et les retirer fur et à mesure),
  - backward-forward (dans les deux directions)
- La procédure stagewise