Choix du modèle

Frédéric Bertrand et Myriam Maumy[1]

[1]IRMA, Université Louis Pasteur
Strasbourg, France

Master 1ère Année 11-02-2008

- Il existe plusieurs critères pour sélectionner $(p-1)$ variables explicatives parmi k variables explicatives disponibles.

- Le critère du R^2 se révèle le plus simple à définir.

- **Un inconvénient majeur du R^2 :** Il augmente de façon monotone avec l'introduction de nouvelles variables même si celles-ci sont peu corrélées avec la variable expliquée.
- **Pour parer à cet inconvénient,** il est conseillé de se tourner vers l'utilisation des alternatives suivantes :
  - le R^2 ajusté
  - le C_p de Mallows qui est un autre critère relatif au biais.
  - les critères AIC et $AICc$.
  - le critère BIC.

Ces six critères vont être maintenant présentés.

Le coefficient de détermination multiple R^2 :

- Il a déjà été introduit dans le cours portant sur le modèle linéaire.
- C'est une mesure qui permet d'évaluer le degré d'adéquation du modèle.
- Lors de l'introduction du test F partiel : accroissement de R^2 au fur et à mesure de l'introduction de variables dans le modèle. Il atteint son maximum lorsque toutes les variables disponibles au départ sont incluses.

- **Pour comparer deux modèles ayant le même nombre de variables explicatives :** comparer les R^2 obtenus et choisir le modèle pour lequel R^2 le plus grand.

- **Pour comparer un modèle avec** $(p-1)$ **variables avec un modèle** $(p-1+r)$ **variables :** utiliser le test du F partiel.

  Que nous dit ce test ?
  Ce test dit si l'introduction des variables supplémentaires augmente suffisamment le R^2 ou non.

Le coefficient de détermination multiple ajusté R^2_{aj} :

- Introduire un R^2 qui concerne la population et non plus l'échantillon défini par :

$$R^2_{pop} = 1 - \frac{\sigma^2}{\sigma^2(Y)}.$$

- Estimer R^2_{pop} par :

$$R^2_{aj} = 1 - \frac{s^2}{s^2(Y)} = 1 - \frac{SC_{res}}{SC_{tot}} \frac{n-1}{n-p}.$$

Propriété sur R^2_{aj} :

- $R^2_{aj} < R^2$ dès que $p \geq 2$.
- R^2_{aj} peut prendre des valeurs négatives.

Intérêts de R^2_{aj} :

- R^2_{aj} n'augmente pas forcément lors de l'introduction de variables supplémentaires dans le modèle.
- possibilité de comparer deux modèles n'ayant pas le même nombre de variables à l'aide du R^2_{aj} et choisir le modèle pour lequel R^2_{aj} est le plus grand.

Le critère du C_p de Mallows

Le critère du C_p de Mallows est défini par :

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

Mais il y a un problème ! : Lequel ?
On ne peut plus estimer σ^2 par

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

Pourquoi ?

Car C_p vaudrait toujours p et alors il ne serait plus intéressant.

Que fait-on dans la pratique ?

- On estime σ^2 par le s^2 du modèle qui fait intervenir toutes les k variables explicatives du modèle à disposition.

  Pour ce modèle on a : $C_p = p$. Et pour les autres ? C_p prendra d'autres valeurs que p.

- On choisit parmi les modèles le modèle où C_p de Mallows est le plus proche de p.

Le critère d'information d'Akaike (AIC)

Le critère 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 est défini par :

$$AIC = -2\log \widetilde{L} + 2k$$

où $\widetilde{L}$ est la vraisemblance maximisée et k le nombre de paramètres dans le modèle.
Avec ce critère, la déviance du modèle $-2\log(\widetilde{L})$ est pénalisée par 2 fois le nombre de paramètres.

L'***AIC*** représente donc un **compromis** entre le **biais**, diminuant avec le nombre de paramètres, et la **parcimonie**, volonté de décrire les données avec le plus petit nombre de paramètres possibles.

- 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**.

Le critère d'information d'Akaike corrigé (AIC_c)

Lorsque le **nombre de paramètres k est grand** par rapport au nombre d'observations n, c'est-à-dire si $N/k < 40$, il est recommandé d'utiliser **l'*AIC* corrigé**.
Le critère d'information d'Akaike corrigé, AIC_c, est défini par :

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

Hurvich, C. M. and Tsai, C.-L., 1995. Model selection for extended quasi-likelihood models in small samples. *Biometrics* **51** : 1077-1084.

Le critère *BIC*

Le Bayesian information criterion BIC est défini par :

$$BIC = -2\log(\widetilde{L}) + k\log(n).$$

Il est plus parcimonieux que le critère AIC puisqu'il pénalise plus le nombre de variables présentent de le modèle.

Ripley, 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 significative dans le modèle.

Plusieurs types de procédures de sélection de variables :

- la recherche exhaustive
- les méthodes de type pas à pas.

L'efficacité de ces méthodes n'est pas démentie mais il est impossible de se fier aux résultats fournis par un programme informatique.

Quant à décider ou supprimer une variable dans un modèle, il faut conserver :

- une part d'intuition
- une part de déduction
- une part de synthèse.

Surtout ne pas oublier l'objectif recherché !

Exemples :

- Pour rendre l'équation utile à des fins de prévision, il est souhaitable que le modèle contienne un maximum de variables explicatives afin d'en augmenter le pouvoir explicatif.
- En raison du coût élevé relatif à l'obtention d'informations pour un grand nombre de variables explicatives, l'analyste souhaite avoir un modèle avec un minimum de variables explicatives.

Pour obtenir un compromis entre ces deux extrêmes, le statisticien dispose d'**une gamme de méthodes** :

- la recherche exhaustive
- la méthode descendante
- la méthode ascendante
- la régression stepwise
- deux variantes des quatre méthodes précédentes
- la régression stagewise.

Toutes ces procédures ne mènent pas forcément à la même solution quand elles sont appliquées au même problème.

Lorsque le nombre k de variables explicatives à disposition n'est pas trop élevé, il est envisageable de considérer tous les modèles possibles.

- On a
$$C_k^r = \frac{k!}{r!(k-r)!}$$
modèles différents faisant intervenir r variables explicatives.
- Cela fait
$$\sum_{r=0}^{k} C_k^r = 2^k$$
modèles possibles à considérer.

On choisit ensuite le modèle pour lequel, par exemple, le R^2_{aj} est le maximum.

Les méthodes de type pas à pas consistent à considérer d'abord un modèle faisant intervenir un certain nombre de variables explicatives. Puis on procède par élimination ou ajout successif de variables.

- On parle de la **méthode descendante** lorsqu'on **élimine des variables** (elle sera développée dans le paragraphe 4)
- On parle de la **méthode ascendante** lorsqu'on **ajoute des variables** (elle sera développée dans le paragraphe 5).
- La **méthode stepwise** est une **combinaison de ces deux méthodes** (elle sera développée dans le paragraphe 6).

La recherche exhaustive

est une méthode fastidieuse et difficile à utiliser sans un ordinateur rapide.

Pourquoi ?

Il faut calculer toutes les régressions possibles impliquant un sous-ensemble des k variables explicatives à disposition, soit un total de 2^k régressions.

Que fait-on ensuite ?

- Ces équations sont réparties selon le nombre r de variables explicatives qu'elles contiennent.
- Chaque ensemble d'équations est ordonné selon le critère choisi, souvent le R^2 ou l'AIC.
- Les meilleures équations de régression issues de ce classement sont ensuite sélectionnées pour un examen plus détaillé.

Méthode descendante

(ou élimination en arrière) est une simplification de la méthode de la recherche exhaustive.

En quoi est-elle une simplification ?

Cette méthode examine non pas toutes les régressions possibles mais uniquement une régression pour chaque nombre r de variables explicatives.

En pratique comment fait-on ?

- Calculer la régression pour le modèle incluant toutes les k variables explicatives à diposition.
- Effectuer un test de Student pour chacune des variables explicatives. **Deux cas se présentent :**
    - Les variables sont trouvées significatives. Ce modèle est alors choisi. On stoppe là notre analyse.
    - Éliminer la variable la moins significative du modèle.
- Recommencer le processus avec une variable en moins.

Le modèle final est donc un modèle au sein duquel toutes les variables sont significatives.

Conclusions :

- La méthode descendante est très satisfaisante pour l'utilisateur préférant avoir toutes les variables possibles afin de ne rien ignorer.
- C'est une procédure plus économique en terme de temps et d'interprétation
- Mais il y a un inconvénient majeur. Il n'est plus possible de réintroduire une variable une fois qu'elle a été supprimée !

Méthode ascendante (ou sélection en avant) :

- C'est également une simplification de la méthode de la recherche exhaustive.
- Cette méthode procéde dans le sens inverse de la méthode descendante.
- Cette méthode examine un modèle avec une seule variable explicative puis introduction une à une d'autres variables explicatives.

En pratique comment fait-on ?

- Effectuer les k régressions possibles avec une seule variable explicative. Pour chacune d'elles, **effectuer le test de Student**. Retenir le modèle pour lequel la variable explicative est la plus significative.

- Effectuer les $(k-1)$ régresions possibles avec deux variables explicatives. Pour chacune d'elles, **effectuer le test de Student pour la nouvelle variable**. Retenir le modèle pour lequel la variable est la plus significative. Si aucune variable est retenue, alors on stoppe le processus.

Sinon

- Réitérer le processus en effectuant les $(k-2)$ régressions possibles avec trois variables explicatives. Pour chacune d'elles, **effectuer le test de Student pour la nouvelle variable.** Retenir le modèle pour lequel la variable est la plus significative. Si aucune variable est retenue, alors on stoppe le processus.

Sinon

- Réitèrer le processus en effectuant les $(k-3)$ régressions possibles avec quatre variables explicatives...

Le processus se termine lorsqu'on ne peut plus introduire des variable significatives dans le modèle.

Parmi les méthodes présentées, **la méthode ascendante est la plus économique**.

Les avantages de la méthode ascendante :

- éviter de travailler avec plus de variables que nécessaire,
- améliorer l'équation à chaque étape.

L'inconvénient majeur de la méthode ascendante :

- une variable introduite dans le modèle ne peut plus être éliminée.

Le modèle final peut alors contenir des variables non significatives.

Ce problème est alors résolu par la procédure stepwise.

Procédure stepwise :

amélioration de la méthode descendante.

- **Pourquoi ?** À chaque étape, on réexamine toutes les variables introduites précédemment dans le modèle.

  En effet, une variable considérée comme la plus significative à une étape de l'algorithme peut à une étape ultérieure devenir non significative. **Pourquoi ce phénomène ?**

  - En raison de ces corrélations avec d'autres variables introduites après coup dans le modèle.

Comment remédie-t-on à cela ?

La procédure stepwise propose après l'introduction d'une nouvelle variable dans le modèle :

- **réexaminer les tests de Student** pour chaque variable explicative anciennement admise dans le modèle,
- après réexamen, si des variables ne sont plus significatives, alors **retirer du modèle la moins significative d'entre elles.**

Le processus continue jusqu'à ce que plus aucune variable ne puisse être introduite ni retirée du modèle.

La procédure stepwise

semble être la meilleure procédure de sélection de variables.

Mais

- **la procédure stepwise** peut facilement abuser l'utilisateur qui a tendance à se focaliser exclusivement sur le résultat de la sélection automatique proposé par l'outil informatique.
- En effet, il faut se méfier de certaines situations : celles où apparait **un phénomène de multicolinéarité** que nous étudierons dans un prochain chapitre.

Un exemple :

X_1 et X_2 expliquant significativement Y, sont fortement corrélées entre elles.

L'introduction de X_1 dans le modèle masquera le pouvoir explicatif de X_2.

En effet, l'introduction de X_1 en premier va accroitre le R^2 alors que l'introduction ultérieure de X_2 provoquera un faible effet.

Et réciproquement.

Premières remarques :

Les méthodes présentées jusqu'ici :

- ne sélectionnent pas nécessairement le meilleur modèle absolu,
- mais donnent généralement un modèle acceptable.

Des procédures alternatives et combinatoires sont apparues. On présentera très brièvement seulement **deux méthodes** dans ce cours.

- **La première méthode :** effectuer d'abord la procédure stepwise.

  Supposons que le modèle sélectionné contient r variables explicatives.

  Effectuer alors C_k^r régressions possibles utilisant r variables.

  Choisir comme modèle final celui pour lequel R^2 est maximal.

D'un point de vue pratique, **les améliorations apportées par cette procédure sont faibles**...et nécessitent beaucoup de calculs.

- **La deuxième méthode :** utiliser deux seuils de signification différents lors de l'utilisation de la méthode stepwise :
  - un certain seuil (par exemple $\alpha = 0,05$) lors de la procédure d'élimination de variables,
  - un seuil plus petit (par exemple $\alpha = 0,01$) lors de la procédure d'introduction de variables.

On favorise l'introduction des variables les plus significatives, tout en acceptant de les maintenir dans le modèle si elles deviennent par la suite un peu moins significatives.

C'est intéressant d'utiliser cette méthode.

Pourquoi ?

Cette méthode propose un modèle alternatif au modèle donné par **la méthode stepwise** tout en maintenant son pouvoir explicatif.

La procédure de régresion stagewise

diffère des procédures présentées ci-dessus.

Pourquoi ?

La procédure de régresion stagewise n'aboutit pas toujours à une équation obtenue par la méthode des moindres carrés.

Cette méthode se déroule de la façon suivante :

- effectuer la régression avec la variable la plus corrélée avec Y.
- Calculer les résidus obtenus avec cette régression.
- Considérer ensuite ces résidus comme une nouvelle variable dépendante que l'on veut expliquer à l'aide des variables explicatives restantes.

- Sélectionner ainsi celle d'entre elles qui est la plus corrélée avec les résidus.
- Effectuer une nouvelle régression avec les résidus de la première régression dans le rôle de la variable expliquée.
- Également calculer les résidus obtenus avec cette nouvelle régression.
- Réitérer le processus avce des variables explicatives restantes, jusquà ce que plus aucune d'entre elles ne soit corrélée significativement avec les résidus.

Quelques conclusions

- D'un point de vue théorique, **la recherche exhaustive est la meilleure**.
- **Pour les autres méthodes :** Des résultats semblables sans nécessiter autant de calculs. Ces derniers dépendent du choix des seuils de signification utilisés lors des diverses procédures.
- **Dans la pratique, la procédure stepwise et la procédure descendante** sont les plus utilisées.
- En cas de doute et si les conditions le permettent, toutes les régressions doivent être examinées.
- **Les autres méthodes** peuvent trouver leur utilisation dans des applications plus spécifiques.