TD2 : Sélection de modèles/variables en régression

27/02/2025

Le but de ce deuxième TD est de présenter les différents algorithmes de sélection de modèles/variables en régression (linéaire). Ces techniques seront mises en oeuvre à l'aide de R, pour la base de données `Carseats` du package `ISLR`.

Taille d'un modèle et précision

Lorsque la taille du modèle de RLM et petite (ou nombre petit de variables explicatives) :

(i) variance faible, biais très élevé;

(ii) erreur théorique de prévision élevée;

(iii) erreur empirique (d'ajustement) élevée.

Lorsque la taille du modèle et grande (ou nombre élevé de variables explicatives) :

(i) variance très élevée, biais faible;

(ii) erreur théorique de prévision élevée;

(iii) erreur empirique (d'ajustement) très faible (problème de sur-ajustement).

D'où la nécessité de développer des procédures de sélection de modèles (de variables).

Le cadre

On considère un modèle de RLM

$$Y = w_0 + w_1\, X_1 + \ldots + w_p\, X_p + \varepsilon.$$

On dispose d'un échantillon $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ du vecteur $(\mathbf{X}, Y) =: (X_1, \ldots, X_p, Y) \in \mathbb{R}^p \times \mathbb{R}$.

L'objectif est d'obtenir le sous-ensemble de variables explicatives qui conduit au "meilleur'' modèle RLM au sens d'un critère donné.

Avec p variables explicatives candidates, $X_1, \ldots, X_p$, on peut construire $2^p - 1$ modèles de régression linéaires différents (les modèles à une variable, à deux variables, ..., à p variables).

Exemples de critères de sélection de modèles

(1) L'Akaike Information Criterion (AIC), d'un modèle de RLM, constitué de k variables explicatives, est défini par

$$AIC = -2\,\mathcal{L}_n(\widehat{\mathbf{w}}, \widetilde{\sigma}^2) + 2\,(k+2),$$

où $\mathcal{L}_n(\widehat{\mathbf{w}}, \widetilde{\sigma}^2)$ est la log-vraisemblance du modèle, définie ci-dessus, sous les hypothèses d'homoscédasticité et de normalité des erreurs;

(2) Le Bayesian Information Criterion (BIC) :

$$BIC = -2\,\mathcal{L}_n(\widehat{\mathbf{w}}, \widetilde{\sigma}^2) + \log(n)\,(k+2);$$

(3) R^2-ajusté :

$$R_a^2 := 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right);$$

(4) Le C_p de Mallow d'un modèle de régression utilisant k variables explicatives ($1 \leq k \leq p$) est donné par :

$$C_p := \frac{1}{n}\left(\|\mathbf{Y} - \widehat{\mathbf{Y}}_0\mathbf{1}\|^2 + 2(1+k)\,\widehat{\sigma}^2\right),$$

où $\widehat{\mathbf{Y}}_0$ est le vecteur des valeurs ajustées selon le modèle utilisant les k variables explicatives;

(5) Le critère F de Fisher : Notons $\widehat{\mathbf{Y}}$ le vecteur des valeurs ajustées selon le modèle de RLM complet à p variables explicatives, et $\widehat{\mathbf{Y}}_0$ le vecteur des valeurs ajustées selon le modèle réduit à $p-q$ variables explicatives ($1 \leq q < p$). Le critère F du modèle réduit est défini par

$$F := \frac{\|\widehat{\mathbf{Y}}_0 - \widehat{\mathbf{Y}}\|^2/q}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2/(n-p-1)}.$$

Algorithme de recherche exhaustive

(1) Construire les $2^p - 1$ modèles;

(2) Choisir celui qui optimise un critère donné.

Excercice 1

(1) Donner les modèles optimaux selon les critères R^2-ajusté, BIC et C_p, par recherche exhaustive : utiliser la fonction `regsubsets()` du package `leaps`;

(2) Reprendre la question précédente en utilisant cette fois-ci la fonction `glmulti()` du package `glmulti`.

Algorithme de recherche pas-à-pas

L'approche exhaustive permet de comparer tous les modèles; l'inconvénient est que le temps de calcul devient très important si le nombre de variables est grand;

Lorsque le nombre de variables est grand, on privilégie souvent les méthodes pas-à-pas qui consistent à construire les modèles de façon récursive, en ajoutant/supprimant une variable explicative à chaque étape.

Méthode ascendante (forward selection, version 1)

(1) Construire $\mathcal{M}_0$ le modèle trivial (avec uniquement l'intercept);

(2) Pour $k = 0, \ldots, p-1$:

(i) Construire les $p-k$ modèles consistant à ajouter une variable dans $\mathcal{M}_k$;

(ii) Choisir, parmi ces $p-k$ modèles, le modèle $\mathcal{M}_{k+1}$ qui optimise un critère donné;

(3) Choisir, parmi $\mathcal{M}_1, \ldots, \mathcal{M}_p$, le meilleur modèle au sens du critère considéré.

Méthode descendante (backward elimination, version 1)

(1) Construire $\mathcal{M}_p$ le modèle complet (avec les p variables);

(2) Pour $k = p, \ldots, 2$:

(i) Construire les k modèles consistant à supprimer une variable dans $\mathcal{M}_k$;

(ii) Choisir, parmi ces k modèles, le modèle $\mathcal{M}_{k-1}$ qui optimise un critère donné;

(3) Choisir, parmi $\mathcal{M}_1, \ldots, \mathcal{M}_p$, le meilleur modèle au sens du critère considéré.

Exercice 2

Appliquer les deux algorithmes précédents pour sélectionner les modèles optimaux selon les critères BIC, C_p et R^2-ajusté : Utiliser la fonction `regsubsets()`.

Méthode ascendante (forward selection, version 2)

(1) Modèle sans variables;

(2) Insertion de la variable qui diminue le plus le critère;

(3) Insertion de la deuxième variable qui diminue le plus le critère, ... arrêt quand on ne diminue plus le critère.

Méthode descendante (backward elimination, version 2)

(1) Modèle complet;

(2) Enlever la variable qui diminue le plus le critère;

(3) Enlever la deuxième variable qui diminue le plus le critère, ... arrêt quand on ne diminue plus le critère.

Exercice 3

Appliquer les deux algorithmes précédents pour sélectionner les modèles optimaux selon les critères AIC, BIC et Fisher : Utiliser la fonction `step()`.

Méthode ascendante bidirectionnelle (bidirectional selection)

(1) Ascendante avec remise en cause à chaque étape des variables déjà inclues ;
(2) Permet d'exclure des variables qui redeviennent plus significatives compte tenu de celle qui vient d'être intégrée.

Méthode descendante bidirectionnelle (bidirectional elimination)

(1) Descendante avec remise en cause à chaque étape des variables déjà exclues ;

(2) Permet de réintégrer des variables qui redeviennent significatives compte tenu de celle qui vient d'être exclue.

Exercice 4

Appliquer les deux algorithmes précédents pour sélectionner les modèles optimaux selon les critères AIC, BIC et Fisher : Utiliser la fonction `step()`.

Sélection par algorithme génétique

(1) On l'utilise quand le nombre de variables devient de plus en plus grand et qu'une recherche exhaustive est impossible et une recherche pas à pas peut mener à une solution qui n'est pas tout à fait optimale;

(2) Cet algorithme est implémenté dans la fonction `glmulti()` du package `glmulti` en spécifiant l'argument `method = "g"`;

(3) Cette méthode est donc supposée trouver le meilleur modèle sans avoir besoin de calculer le critère à considérer sur tous les modèles possibles (recherche exhaustive).

Exercice 5

Appliquer l'algorithme précédent pour sélectionner les modèles optimaux selon les critères AIC et BIC : Utiliser la fonction `glmulti()` du package `glmulti`.

Exercice 6

Reprendre toutes les questions de l'exercice 1 du TD1 en utilisant cette fois-ci seulement le sous-ensemble de variables explicatives sélectionnées par le critère BIC.