Modèle Linéaire Gaussien
Chapitre 5 – Autres méthodes de sélection de variables

Achille Thin

09 Janvier 2024

Executive Master Statistique et Big Data

Objectifs du Chapitre

Sélection de variables :: Choisir parmi un ensemble de variables qualitatives et quantitatives, ainsi que leurs potentielles interactions, quelles sont les variables pertinentes pour modéliser une variable cible continue.

Possibilités:

1. Différents tests statistiques vus lors des parties précédentes quant aux effets des variables explicatives sur la variable cible. Ne retenir que les variables avec un effet significatif.
2. Utilisation de critère de diagnostic du modèle, comme le R^2 ajusté.
3. Utilisation de critères de sélection de variables (Aikakke Information Criterion, Bayesian Information Criterion), et selection ascendante, descendante ou mixte.
4. Utilisation de la performance de prédiction sur un ensemble test, ou en cross validation.

Tests Statistiques et Effets des Variables Explicatives

- **Objectif :** Identifier les variables explicatives ayant un impact significatif sur la variable cible.
- **Méthodologie :** Utiliser des tests statistiques (t-tests, ANOVA) pour évaluer l'impact de chaque variable.
- **Significativité :** Une variable est considérée significative si la p-valeur du test est inférieure au seuil de significativité (souvent 0,05).

Exemple de code R :

Test de significativité pour une régression linéaire

```
modèle <- lm(Y ~ X1 + X2 + X3, data=mes_données)
résumé <- summary(modèle)
résumé$coefficients
```

Ce code réalise une régression linéaire et affiche les p-valeurs pour chaque coefficient, permettant de juger leur significativité.

Critère de Diagnostic du Modèle : R^2 Ajusté

- R^2 **Ajusté :** Mesure la proportion de la variance de la variable cible expliquée par le modèle, ajustée pour le nombre de prédicteurs.
- **Interprétation :** Un R^2 ajusté élevé indique un modèle plus explicatif. Attention à ne pas confondre avec la causalité.
- **Utilisation :** Comparer des modèles avec un nombre différent de prédicteurs.

Exemple de code R :

Calcul du R^2 ajusté

```
modèle <- lm(Y ~ X1 + X2 + X3, data=mes_données)
résumé <- summary(modèle)
résumé$adjusted.r.squared
```

Ce code calcule le R^2 ajusté pour un modèle de régression linéaire.

Critère AIC : Fondements Théoriques

- **Origine :** L'AIC, développé par Hirotugu Akaike, est basé sur les concepts d'information et d'entropie en théorie de l'information.
- **Principe :** L'AIC évalue la qualité d'un modèle en considérant à la fois la complexité du modèle et la qualité de l'ajustement.
- **Formule :** $\text{AIC} = 2p - 2\ln(L_p)$ où p est le nombre de paramètres à estimer et L_p la vraisemblance du modèle ajusté avec p paramètres.
- **Interprétation :** Plus l'AIC est faible, meilleur est le modèle. Un modèle est considéré comme significativement meilleur si son AIC est inférieur d'au moins 2 à celui d'un autre modèle.

Utilisation en R : *Les fonctions R telles que* `AIC()` *ou* `step()` *permettent de calculer l'AIC et de sélectionner le modèle optimal. Pouvez-vous expliquer comment fonctionne la commande* `step()` *?*

Critère BIC : Fondements Théoriques et Application

- **Origine :** Le BIC, ou critère d'information bayésien, a été développé par Gideon Schwarz, s'inspirant des principes bayésiens.
- **Formule :** $\text{BIC} = \ln(n)k - 2\ln(L)$ où n est la taille de l'échantillon, k le nombre de paramètres et L la vraisemblance maximale.
- **Comparaison avec l'AIC :** Le BIC pénalise plus fortement les modèles avec un grand nombre de paramètres, en particulier pour les grands échantillons.
- **Choix du modèle :** Un modèle avec un BIC inférieur est préféré. Le BIC est souvent utilisé pour la sélection de modèles en présence de grands échantillons.

Utilisation en R : *Utiliser* `BIC()` *ou* `step()` *avec un critère basé sur le BIC pour sélectionner le modèle optimal. Pouvez-vous expliquer comment fonctionne la commande* `step()` *?*

Évaluation de la Performance : Ensemble Test et Métriques

- **Ensemble Test :** Utiliser un sous-ensemble de données non inclus dans l'entraînement du modèle pour évaluer sa performance. Se servir de cette performance "de généralisation" pour déterminer le nombre de paramètres optimal du modèle.
- **Métriques de Performance :**
  - **Erreur Quadratique Moyenne (MSE) :** Moyenne des carrés des écarts entre les valeurs prédites et réelles.
  - **Coefficient de Détermination (R^2) :** Proportion de la variance de la variable cible expliquée par le modèle.
- **Importance :** Ces métriques aident à évaluer l'efficacité du modèle dans des conditions réelles, en évitant le surajustement.

Exemple en R : *Utilisation de fonctions telles que* `predict()` *pour évaluer les performances sur un ensemble test.*

Validation Croisée : Théorie et Application

- **Principe :** Diviser les données en plusieurs sous-ensembles et utiliser chaque sous-ensemble comme un ensemble test, tandis que les autres servent à l'entraînement.
- **Types :**
  - **Validation Croisée K-fold :** Diviser les données en K sous-ensembles. Chaque sous-ensemble est utilisé une fois comme test.
  - **Validation Croisée Leave-One-Out (LOO) :** Chaque observation est utilisée une fois comme ensemble test (K = nombre total d'observations).
- **Avantages :** Fournit une estimation robuste et moins biaisée de la performance du modèle.
- **Application :** Idéale pour les petits échantillons ou lorsque chaque observation est importante.

Exemple en R : *Utilisation du package* `caret` *pour implémenter une validation croisée.*

Comprendre la Validation Croisée

- **Objectif :** Évaluer la capacité d'un modèle à généraliser sur de nouvelles données.
- **Méthodologie :** Répéter le processus d'entraînement et de test sur différents sous-ensembles pour obtenir une estimation fiable de la performance du modèle.
- **Importance :** Permet d'éviter le surajustement (overfitting) et de tester la robustesse du modèle sur différentes partitions des données.
- **Choix du nombre de plis (K) :** Dépend de la taille de l'échantillon et des objectifs spécifiques. Un K plus élevé implique une évaluation plus précise mais augmente le coût computationnel.

Note : *La validation croisée est une technique standard en apprentissage automatique pour l'évaluation et la sélection de modèles, surtout lorsque les données disponibles sont limitées.*