Introduction à R

Programmation et Logiciel

Cours 2

Salim Lardjane - Université de Bretagne-Sud

Modèles Statistiques en R

- La part systématique d'un modèle est spécifiée par une formule de la forme

  réponse~principale*co-variable+autres

- Le terme de gauche désigne la réponse (variable expliquée ou indépendante). Le terme de droite regroupe les variables explicatives (dépendantes, prédicteurs).
- Le * spécifies une interaction et les effets principaux correspondants
- Par défaut, les facteurs sont codés sous la forme de variables indicatrices pour toutes les catégories sauf la première.
- Les termes de la formule peuvent être des variables, des expressions ou des objets plus complexes

Exemples

- depress~rural*agegp+partner+parity+income

  Le risque de dépression post-natale varie-t-il entre régions urbaines et rurales, séparément pour chaque groupe d'âge, en prenant en considération l'existence d'un conjoint, le nombre précédent de grossesses et le revenu?

- asthma~pm25+temp+I(temp^2)+month

  Comment le nombre d'admissions en hôpital pour crise d'asthme varie-t-il en fonction de la pollution par particules fines, en prenant en considération la température, le carré de celle-ci et le mois ?

Exemples

- log(pm25)~temp+stag+month+lag(temp,1)

  Prévoir le logarithme de la pollution par particules fines à partir de la température, la stagnation de l'air, le mois et la température du jour précédent

- Surv(ttoMI,MI)~LDL+age+sex+hibp+diabetes

  Le niveau de cholestérol permet-il de prévoir le temps écoulé jusqu'à un infarctus du myocarde en prenant en considération l'âge, le sexe, la présence d'hypertension et la présence d'un diabète?

GLM

- L'estimation de modèles linéaires généralisés (régression linéaire, régression logistique, régression de Poisson) peut être effectuée à l'aide de la fonction glm(). Celle-ci requiert
- Une formule de modèle
- Un tableau de données contenant les variables [optionnel]
- Une famille de modèle :

  binomial() régression logistique

  gaussian() régression linéaire

  poisson() régression de Poisson

  et d'autres moins fréquemment utilisées

```
glm(asthma~pm25+temp+I(temp^2)+month,
    data=pmdat,family=poisson())
```

Objets Modèle

- Les logiciels typiques de statistique ajustent un modèle et renvoient les résultats en sortie. Sous R, un objet est créé à l'issue de l'ajustement. Celui-ci regroupe toute l'information requise à propos du modèle ajusté. Les coefficients, diagnostics et autres aspects du modèles peuvent être récupérés à l'aide des méthodes définies pour cet objet.

Autres modèles de régression

- Outre les GLM, R permet d'estimer un grand nombre de modèles de régression
- lm() régression linéaire simple
- coxph() modèle de Cox (package survival)
- clogit() régression logistique conditionnelle (package survival)
- gee() Equations estimantes généralisées (packages gee et geepack)
- lme() modèle linéaire mixte (package nlme)
- polr() modèle de hasards proportionnels (package MASS)

Data mining

- Le Data mining consiste, entre autres, à
- Regrouper les données pour effectuer des typologies ou des segmentations de marché
- Utiliser des modèles de régression adaptative pour la prédiction
- Il est typiquement réalisé sur de grands ensembles de données

Data mining

- R dispose de plusieurs fonctions de classification et de régression adaptative, mais elles ne sont pas adaptées en général à des tableaux de données en Go.
- R peut être utile pour la modélisation détaillée d'un échantillon aléatoire ou systématique issu d'un entrepôt de données, ou pour la modélisation prédictive de tableaux de données restreints. Il existe des packages de connexion à des bases de données permettant de faciliter ce type d'utilisation. On peut également passer par SAS par exemple.

Data mining

- Le package cluster propose diverses fonctions pour la classification.
- La plupart requièrent des capacités de stockage en O(n^2), excepté clara qui en requiert O(n).

Data mining

- D'autres techniques de data mining disponibles sous R sont :
- rpart : L'algorithme CART pour la construction et l'élagage d'arbres de régression et de classification.
- MARS : Le package polymars implémente la méthode MARS pour les variables catégorielles ou dichotomiques.
- k-NN : classification par k plus proches voisins. Elle est implémentée dans le package class.
- Réseaux de Neurones : implémentés dans le package nnet, entre autres.
- SVM : implémentées dans le package e1071.

Data mining

- La plupart des fonctions de data mining de R conservent la totalité du tableau de données en mémoire vive et ne peuvent donc être utilisées que pour des tableaux de quelques dizaines voire quelques centaines de milliers d'observations. Pour de plus grand tableaux, on doit avoir recours à l'échantillonnage ou au partitionnement.
- R n'est pas particulièrement bien adapté au data mining vu comme l'extraction automatisée d'informations dans de grandes bases de données. R est plus adapté à la modélisation interactive basée sur une bonne connaissance du domaine d'application.