Travaux dirigés - Régression linéaire multiple

Julien Chiquet et Guillem Rigaill

15 octobre 2015

Objectifs de la séance

- maîtriser la régression linéaire multiple
- interpréter les sorties de `R` associées

Remarques

- Les TD doivent être faits en binômes
- Vous préparerez un rapport (fichier `R markdown`) que vous enverrez à julien.chiquet@gmail.com.
- Pour toute question d'ordre technique, ne pas hésiter à solliciter les encadrants.
- Pour les questions de cours, référez-vous ans la mesure du possible... au cours !

Les rapports sont à rendre (au sens de "mail envoyé") dès la fin de la séance. Ils seront notés et compteront dans la note finale: soignez vos commentaires. Ils doivent montrer que vous comprenez ce que vous faites et que vous commencer à apprivoiser le modèle linéaire et son intégration dans `R`.

Vote aux présidentielles US de 2000 en Georgie

Le jeu de données `gavote` décrit le vote présidentiel aux États-Unis en 2000, dans l'état de Géorgie. Chacun des 159 cantons est décrit par les variables suivantes:

- `atlanta`, indicateur de l'appartenance ou non à Atlanta
- `ballots`, nombre de bulletins
- `bush`, nombre de votes pour Bush
- `econ`, le statut économique du canton (`middle`, `poor`, `rich`).
- `equip`, le système physique de vote
  - LEVER: machine à levier
  - OS-CC: Scan optique comptage centralisé ("central count"),
  - OS-PC: Scan optique comptage local ("precinct count")
  - PAPER: vote par bulletin papier
  - PUNCH: vote par poinçon
- `gore`, nombre de votes pour Gore
- `other`, nombre de votes pour les candidats autres que Bush et Gore
- `perAA`, le pourcentage d'afro-américains
- `rural`, indicateur de la ruralité du canton (`urban`, `rural`)
- `votes`, nombre de votes validés

1. Préliminaires

- Charger le jeu de données `gavote` à la page [http://julien.cremeriefamily.info/reglin/gavote.txt]

- Créer la variable `undercount` (proportion de bulletins de vote considérés comme nuls) et l'ajouter à la table. Nous allons nous intéresser à la prédiction de cette variable par les autres. Elle porte donc le statut de *variable réponse*, ou *à expliquer*. Supprimer les variables `votes` et `ballots` du tableau (expliquer pourquoi).
- Créer les variables `pergore`, `perbush` et `perother` (pourcentage de votants pour Gore, Bush et pour d'autres candidats). Les ajouter à la table. Supprimer les variables `gore`, `bush` et `other` du tableau.

2. Analyse descriptive

Faire une analyse descriptive des données. Cette analyse vous aidera dans vos choix de modèles et pour l'interprétation des résultats.

Vous disposez d'un large choix parmi les outils de `R`: résumés numériques, histogramme, boxplot, barplot, fonction de répartition, graphes pair à pair, matrice de corrélation, clustering hiérarchique, etc. soyez imaginatif ! L'idée n'est pas d'intégrer au rapport le plus de graphes possible mais uniquement ceux pour lesquels vous avez quelque chose à dire.

3. Quelques modèles linéaires simples

Votes parmi les Afro-Américains

- Tracer le diagramme de dispersion croisant `pergore` avec `perAA`. Ajuster un modèle de régression simple et tracer la droite de régression entre `pergore` et `perAA`.
- Régresser `undercount` sur `perAA` puis sur `pergore`, et enfin sur `pergore` *et* `perAA`. Comparer ces modèles et interpréter les résultats.

Votes selon le niveau de vie et ANOVA à un facteur

Pour étudier l'effet du facteur `econ` sur la variable `undercount`, on propose le modèle d'ANOVA 1 suivant :

$$Y_i = \mu + \mathbf{1}_{\{i \in k\}} \mu_k + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

où Y_i décrit la variable `undercount` associée au $i$ème individu, μ l'intercept et μ_k un terme additif associé au groupe k (c'est-à-dire à la modalité k de la variable `econ`).

- Observer les modalités de la variables catégorielle `econ`. Recoder cette variable en classant ces modalités dans un ordre facilement interprétable.
- Représenter la distribution de la variable `econ`.
- Montrer que le modèle ci-dessus peut s'écrire comme un modèle linéaire de la forme

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- Étudier l'effet du facteur `econ`.

Attention à l'interprétation des coefficients de régression: le modèle d'ANOVA étant surparamétré, des contraintes sur les valeurs des μ_k sont nécessaires. Sous `R`, on impose que le premier niveau du facteur soit nul. Il est donc utilisé comme référence pour les valeurs des coefficients associés aux autres niveaux.

Votes selon...

Illustrer un comportement pertinent de votre choix dans le jeu de donnée en ajustant un modèle linéaire simple différent de ceux considérés jusqu'à présent.

4. Construction de modèles linéaires multiples

Construction d'un modèle sur base de tests

- Ajuster un modèle de régression linéaire qui explique `undercount` en fonction des variables quantitatives `perbush`, `pergore`, `perother` et `perAA`. Comment expliquer vous le comportement obtenu ? Proposer un modèle alternatif en supprimant une variable (justifier). Cette variable sera dorénavant ôter du tableau de données `gavote`.
- Ajuster un modèle qui explique `undercount` en fonction de toutes les variables explicatives (restantes). Construire un modèle `model.test` qui intègre uniquement les variables significatives.
- Étudier les résidus de `model.test`.

Construction par régression stepwise

La régression "Stepwise" adopte une stratégie qui construit un modèle de proche en proche en partant du modèle nul et en ajoutant/supprimant une nouvelle variable explicative sur la base d'un critère (le R^2 par exemple). Nous verrons en cours comment motiver un critère de choix de modèle visant un compromis entre l'ajustement aux données et le nombre de paramètres. Dans cet optique, l'AIC et le BIC sont les critères les plus classiquement utilisés.

- En utilisant la procédure `step`, proposer un modèle à l'aide de la procédure forward/backward utilisant les critères AIC puis BIC. On appelera `model.AIC` et `model.BIC` les modèles respectivement obtenus.
- Étudier les résidus de ces modèles.

Recherche exhaustive

Lorsque le nombre de prédicteurs reste raisonable (< 30), il est possible de tester tous les modèles et de choisir le meilleur au sens d'un critère.

- En utilisant la procédure `regsubsets` du package `leaps`, déterminer le meilleur modèle au sens du R^2 ajusté. Commentez.
- Étudier les résidus de `model.R2`.

5. Étude et raffinement du modèle final

- Choisissez un modèle entre `model.test`, `model.AIC`, `model.BIC` et `model.R2`. Faites un diagnostic complet (résidus, test des hypothèses associées au modèle linéaire, distance de Cook, leviers). On pourra utiliser les fonctions `plot.lm`, `cook.distance`, `rstudent`, ...
- Essayez d'améliorer l'ajustement en excluant les points aberrants, en transformant la réponse ou certains prédicteurs. On pourra aussi fusionner les niveaux de certains facteurs dont les effets sont similaires.
- Interpréter le modèle final que vous retenez.