TD1 : Rappels sur le modèle linéaire

M2 MMA - Université de Paris

2020-2021

Exercice 1 : Petits exemples

Exercice 1.1 : jeu de données `iris`

1. Charger et explorer les données `iris`.
2. Apprendre un modèle linéaire gaussien expliquant la longueur des pétales (`Sepal.Length`) en fonction des autres variables. Analyser les résultats. Commenter notamment les coefficients correspondant à la variable `Species`.
3. Déterminer à l'aide d'un modèle logistique quelle variable caractérise l'espèce `versicolor`.

Exercice 1.2 : jeu de données `airquality`

1. Charger et explorer les données `airquality`.
2. Indiquer les variables qui ont une influence linéaire sur le taux d'Ozone. Les signes des estimateurs correspondants sont-ils en accord avec ce que l'on pourrait penser?
3. Comment peut-on interpréter le signe affecté au coefficient correspondant à la variable `month`? Pour mieux analyser cette variable, tracer les boxplots représentant la taux d'ozone en fonction du mois. Réduire les données en ne gardant que celles du mois de mai à juillet. Quel signe attendre pour l'influence du mois? Relancer l'apprentissage d'un modèle linéaire puis commenter.

Exercice 2 : Problème de colinéarité

Exercice 2.1 : jeu de données `Prostate`

On travaille sur le jeu de données `Prostate` du package `lasso2` qui contient des informations ($p = 9$) concernant $n = 97$ patients atteints par un cancer de la prostate. : On commence par charger les données à l'aide des commandes suivantes :

```
library(lasso2)
data("Prostate")
```

1. Mettre en place un modèle linéaire gaussien expliquant l'importance du cancer (variable `lcavol`) en fonction des autres variables. Analyser les résultats.
2. Utiliser la fonction `cor` puis la fonction `corrplot` du package du même nom pour représenter la corrélation entre les variables. Que peut-on remarquer?

3. En pratique, une grande collinéarité des variables fait *planter* le modèle linéaire. La collinéarité est mesurée par le VIF (*Variance Inflation Factor*), qui mesure qualitativement la dépendance linéaire d'une variable X^j par rapport aux autres variables :

$$\mathrm{VIF}_j = \frac{1}{1 - \mathrm{R}_j^2}.$$

   Utiliser la fonction `vif()` du package `car` pour mesurer la collinéarité dans ce jeu de données.

On considère souvent qu'un VIF≥ 10 est un signe de colinéarité importante.

Exercice 2.2 : passage à la grande dimension

1. Pour cette question, on considère deux entiers p et n et une covariance $0 \leq \rho \leq 1$. Construire un jeu de données contenant $2p+1$ variables $(X^1, ..., X^{2p}, Y)$ mesurées sur n échantillons, et telles que :

— $Var(X^i) = 1$, pour tout $1 \leq i \leq p$,

—

$$cov(X^i, X^j) = \begin{cases} 0 & \text{si } i \leq p \text{ et } j \geq p+1, \\ \rho & \text{sinon.} \end{cases}$$

— $Y = X^1 + X^{p+1} + \varepsilon$, ε suivant une loi normale centrée et d'écart-type 0.5.

Pour cela, créer une fonction qui prend en argument p, n et ρ et génère les données correspondantes.

2. Pour $\rho = 0.1$, générer des données et apprendre un modèle linéaire gaussien pour différentes valeurs de p et n selon les trois scenarii suivants $n >> p$, $n > 2p$ et $n \leq 2p$.
3. Même question pour $\rho = 0.9$. Commenter les résultats.

Exercice 3 : Sélection de variables

Pour cet exercice, on travaille avec le jeu de données `chenilles`, qui contient des informations concernant des chenilles processionnaires. Celles-ci se développent de préférence sur des pins et peuvent causer des dégâts considérables. On souhaite étudier l'influence de certaines caractéristiques de peuplements forestiers sur leurs développements à partir d'un échantillon de 32 parcelles forestières de 10 hectares. Chaque parcelle est échantillonnée en placettes de 5 ares pour lesquelles on dispose des mesures suivantes :

- Y : le **nombre de nids** de chenilles processionnaires par arbre,
- X^1 : l'**altitude** en mètres,
- X^2 : la **pente** en degrés,
- X^3 : le **nombre de pins** dans la placette,
- X^4 : la **hauteur** en mètres de l'arbre échantillonné au centre de la placette,
- X^5 : le **diamètre** de cet arbre,
- X^6 : la note de **densité** de peuplement,
- X^7 : l'**orientation** de la placette, allant de 1 (sud) à 2 (autre),
- X^8 : la **hauteur** en mètres des arbres dominants,
- X^9 : le **nombre** de strates de végétation,
- X^{10} : le **mélange** du peuplement, allant de 1 (non mélangé) à 2 (mélangé).

1. Charger les données à l'aide de la commande suivante :

```
chen <- read.table("chenilles.txt",header=TRUE)
attach(chen)
```

2. Construire un modèle linéaire gaussien expliquant le nombre de nids en fonction des autres variables et identifier les variables les plus importantes.

3. A l'aide de la fonction `anova()`, qui permet d'effectuer des tests de Fisher de comparaison de modèles, comparer le modèle complet au modèle comprenant toutes les variables sauf `Densite` et `HautMax`. Indiquer combien de tests sont alors nécessaires pour identifier le meilleur modèle.

4. En pratique, lorsque p est grand, on utilise des méthodes de sélection pas à pas pour choisir le meilleur modèle. L'objectif de ces méthodes consiste à introduire ou à supprimer les variables du modèle les unes après les autres :

- sélection **forward** : on part du modèle constant. A chaque étape de l'algorithme, la variable la plus significative (au sens d'un critère à définir) est ajoutée au modèle. En pratique, cette méthode est mal fondée théoriquement et donc déconseillée.
- sélection **backward** : il s'agit de la version symétrique du forward, qui part donc du modèle complet, auquel on enlève itérativement des variables.
- sélection **stepwise** : cet algorithme introduit une étape d'élimination de variable après chaque étape de sélection, afin de retirer du modèle d'éventuelles variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites.

Effectuer une sélection descendante des variables explicatives avec le test de Student. Quel est le modèle final retenu?

5. Effectuer une sélection descendante des variables explicatives avec le critère AIC à l'aide de la fonction `step()`. Quel est le modèle final retenu?