TD MSS2 – TIS4

Régression logistique - Prédiction - Anova

Nicolas Glade – Florent Chuffart

2018-2019

Etude sur les variants d'une graminée

Objectif de cette séance. Au cours cette séance de TD de modélisation statistique, à partir de données sur l'assimilation du CO_2 par une espèce de graminée, nous verrons comment faire une étude d'ANOVA, réaliser un modèle de régression logistique, prédire l'origine des plantes, générer un nouveau jeu de données à partir du modèle. Ce TD est très proche du précédent dans son fonctionnement.

Compte Rendu. Vous disposez de 2 semaines pour rédiger et rendre un rapport au format article par quadrinôme. L'article devra présenter un aspect de ce travail en contextualisant la question traitée, en présentant la méthodologie utilisée et les résultats importants (pas de code, graphiques soignés, ...). Il fera 4 pages maximum et sera rendu au secrétariat TIS en version papier.

Thème du TD. L'espèce *Echinochloa crus-galli* est une graminée très commune rencontrée de partout dans le monde. C'est une mauvaise herbe redoutée dans de nombreuses régions agricoles, qui infeste notamment les rizières italiennes. C'est une plante envahissante en Amérique du Nord. La plante est considérée comme l'une des pires adventices de la planète car elle réduit les rendements des cultures en absorbant jusqu'à 80 % de l'azote disponible dans le sol et sert d'hôte à plusieurs virus mosaïque (c'est donc un réservoir pour ces virus). Les niveaux élevés de nitrates qui s'accumulent en elle peuvent empoisonner le bétail [source Wikipedia]. Les nombreuses études sur son métabolisme et sa variabilité génétique sont donc un enjeu majeur pour l'industrie agroalimentaire et phytosanitaire.

Dans l'étude suivante [Potvin C. et al, Ecology, 71 (1990) 1389], on s'intéresse à l'assimilation du CO2 par ces plantes, assimilation qui conditionne leur croissance. On mesure ainsi, pour des concentrations ambiantes croissantes de CO2 l'assimilation faite par 6 plantes en provenance du Québec (Canada) et 6 autres en provenance du Mississippi (sud des Etats Unis d'Amérique). La moitié des plantes de chaque provenance géographique a été mise au frais (4°C) toute la nuit avant réalisation de l'expérience tandis que l'autre moitié restait à température ambiante (20°C).

Les variables sont *Plant* l'identifiant de chaque plante, *Type* le type désignant la provenance de chaque plante (Québec ou Mississippi), *Treatment* le traitement de température préalable que subissent les plantes avant la mesure (4°C (chilled) ou 20°C (nonchilled)), *conc* la concentration en CO_2 dans l'air ambiant, et *uptake* la quantité surfacique de CO_2 absorbée par les plantes en 1 seconde.

Références bibliographiques.

D.G. Altman (1991), *Practical Statistics for Medical Research*, Table 12.11, Chapman & Hall.

O'Neill et al. (1983), *The effects of chronic hyperinflation, nutritional status, and posture on respiratory muscle strength in cystic fibrosis*, *Am. Rev. Respir. Dis.*, 128:1051--1054.

Réalisation du TD

NB : Une partie du code nécessaire pour réaliser ce TD est fourni : le code pour plusieurs graphiques et le code de calcul des courbes ROC et des matrices de confusion.

1. *Données*

Le jeu de données est natif des installations R. Pour le charger en mémoire, utilisez la commande ***data(CO2)***.

Les noms des variables ne sont pas très explicites. Vous pouvez les renommer en faisant :
names(CO2) <- ***c(****"ID_Plante"****,*** *"Type_Plante"****,*** *"Traitement", "C_CO2", "Assimilation"****)***

Pour travailler plus facilement avec les données, attachez-les : ***attach(CO2)*** ; cela vous donnera directement accès aux variables internes : par exemple, au lieu de taper *CO2$ID_Plante*, vous écrirez directement *ID_Plante*. Pour les détacher, utilisez la commande ***detach(CO2)***

→ Quelles données sont qualitatives ou quantitatives ?
→ Que va-t'on pouvoir faire avec ces données ?

2. Analyse graphique

Utilisez le code fourni (TIS4_TD_CodeGraph1.R) pour tracer le graphique (figure 1).
→ Analysez.

3. Régression linéaire multiple et ANOVA

Réalisez un modèle complet (*CO2.model.complet*) expliquant l'assimilation du CO2 (uptake/Assimilation) par les plantes. Attention, il faut convertir plusieurs variables en valeurs numériques (***as.numeric(****x****)***).

Affichez les résultats de ce modèle (***summary(****CO2.model.complet****)***) et commentez :

→ Qu'est ce que l'AIC d'un modèle ? A quoi correspond le R^2 ? Quelle est la différence entre le R^2 multiple et le R^2 ajusté ?
→ Expliquez ce qu'on cherche à modéliser. De quel type de modèle s'agit-il et pourquoi utilise-t'on ce type de modèle dans ce cas ?
→ Expliquez ce qu'indiquent les p-valeurs associées aux coefficients de régression.

Pour réaliser une anova de type 2, vous pouvez utiliser le package "**car**" fourni ici. Pour cela 3 solutions:

- Décompressez l'archive **car.zip** (bouton droit → « extraire ici »)dans votre répertoire de travail (vous devez alors voir apparaître un nouveau dossier *car*).
- Vous pouvez aussi suivre les instructions (**https://cran.r-project.org/bin/windows/base/rw-FAQ.html#I-don_0027t-have-permission-to-write-to-the-R_002d3_002e4_002e3_005clibrary-directory**) et l'installer à partir de https://cran.r-project.org/web/packages/car/index.html
- Si vous êtes admin (sur votre machine), vous pouvez enfin l'installer à l'aide de la commande **install.packages("car", dependencies=TRUE)**
- Si vous utilisez R-Studio, vous pouvez simplement suivre ces instructions http://web.cs.ucla.edu/~gulzar/rstudio/

Pour vous en servir, chargez la librairie '***car***' à l'aide de la commande : **library(car, lib="./car")**

Faites une anova de type 2 à l'aide de la commande ***Anova(****CO2.model.complet, type="II"****)***. Les anova de type I peuvent être faites avec les commandes ***anova(****CO2.model.complet****)***.

→ A quoi sert de faire une ANOVA ici ? Pourquoi de type II ?

→ Dans cette étude d'ANOVA, expliquez la relation entre les valeurs de la somme des carrés des écarts à la moyenne et les p-valeurs associées.
→ Interprétez.

On génère un nouveau jeu de données prédites en se servant de la commande :
CO2.prd <- ***predict(object=****CO2.model.complet****, newdata=****CO2[,c(1,2,3,4)]****, type=****"response"****, se.fit=****T)***

Utilisez le code fourni (TIS4_TD_CodeGraph2.R) pour tracer le graphique (figure 2).
→ Analysez le code de la figure 2. Que fait-on ?
→ Est-ce que le modèle est efficace ? Vous argumenterez.

Réduisez le modèle en vous servant des p-valeurs associées à l'ANOVA (vous nommerez le nouveau modèle *CO2.model*). Vous pouvez faire de même en utilisant la commande *CO2.model* ***<-step(****CO2.model)*
Calculez l'AIC des 2 modèles à l'aide de la fonction ***AIC(****m)*.

→ Au vu des résultats de l'ANOVA et des tests sur les coefficients sur chacun des modèles (modèle complet et modèle réduit), expliquez pourquoi on peut raisonnablement réduire ainsi ce modèle complet.
→ Au vu des AIC et R^2 respectifs de 2 modèles, expliquez en quoi ce nouveau modèle (sans la variable *Plant / ID_Plante*) est intéressant.

4. Régression logistique

Réalisez un modèle complet (*Plante.model*) permettant de prédire l'origine de la plante (*Type/Type_Plante*) en fonction de l'assimilation et du traitement.

Réduisez ce modèle (*Plante.model.best*) à l'aide de la fonction ***step(****modele)*.
Faites également un modèle réduit (*Plante.model.interact*) de l'origine de la plante en fonction l'interaction *assimilation:traitement*.
→ Comparez ces 2 modèles.

Tracez à l'aide du code fourni (*TIS4_TD_CodeGraph3.R*) la figure 3.
→ Commentez :
→ Que peut-on expliquer à partir de ce graphique ?
→ Est ce que ce modèle vous paraît efficace ?

5. Matrice de confusion - ROC

Calculez les matrices de confusion des prédictions faites pour le type de plante (voir cours - code non fourni).

De même (et en vous servant du code *roc_auc.R*), calculez et tracez les courbes ROC et déterminez l'AUC de ces modèles. Interprétez ! Que peut-on faire de ces informations ?

6. Trouver des groupes sans connaissance *a priori* : classification non supervisée avec k-means

La commande *kmeans(x, centers)* permet de déterminer des groupes sur la base de critères quantitatifs. Essayez cette méthode pour tenter de retrouver les prédictions de la régression logistique (prédire l'origine des plantes à partir de leurs caractéristiques).
→ Qu'apprenez vous ?