Travaux Pratiques Data Science 03
Regression

Gaëtan Constant
gaetan.constant@protonmail.com

Polytech Lyon — Vendredi 12 Octobre 2018

Introduction

Les techniques de régression sont une compétence cruciale dans toute boîte à outils de data scientist ou de statisticien.

Un modèle linéaire explique comment se comporte une variable à réponse continue, en fonction d'un ensemble de covariables ou de variables explicatives. Bien que souvent insuffisants pour expliquer des problèmes complexes, les modèles linéaires présentent des compétences sous-jacentes, telles que la sélection variable et les examens de diagnostic. Par conséquent, une introduction intéressante aux techniques de régression statistique.

Le cadre général de ce TP considère donc les observations d'une variable aléatoire Y dite réponse, exogène, dépendante qui doit être expliquée (modélisée) par les mesures effectuées sur n variables dites explicatives, de contrôle, endogènes, dépendantes ou régresseurs .

Ces variables peuvent être quantitatives ou qualitatives, ce critère déterminant le type de méthode ou de modèle à mettre en œuvre : régression linéaire, analyse de variance et covariance, régression logistique, modèle log-linéaire.

Back to the future (30')

La vie est ce qu'elle est, nous ne disposons pas tout le temps de données propres dès le début des travaux.

Nous avons pour ambition de créer une application qui nous donne la probabilité qu'une équipe gagne un match face à une autre équipe.
Une idée que nous avons est de récupérer toutes les données des matchs des coupes du monde, compétition majeure dans le monde du football et qui nous semble significative pour notre modèle.

Info: La Coupe du monde de football ou Championnat du monde de football ou encore Coupe du monde de la FIFA est une compétition internationale de football qui se déroule ordinairement tous les quatre ans. Cette compétition, créée en 1928 en France, sous l'impulsion de Jules Rimet alors président de la FIFA, est ouverte à toutes les fédérations reconnues par la Fédération internationale de football association (FIFA). Le vainqueur de la Coupe du monde à la fin de la compétition obtient le titre de Champion du monde. La première édition se déroule en 1930 en Uruguay, dont l'équipe nationale sort vainqueur.

Question 1

A partir des 3 fichiers envoyés ("data_2018.json" & matches_19302010.csv & WorldCupMatches2014.csv), créer la table DATA de tous les matchs de coupe du monde.

Les colonnes de cette table sont les suivantes :

- **id_match** : le premier match de l'histoire porte le numéro 1 et le dernier (finale 2018 France Croatie le dernier numéro de la liste)
- **home_team** : nom de l'équipe 1
- **away_team** : nom de l'équipe 2
- **home_result** : buts de l'équipe 1
- **away_result** : buts de l'équipe 2
- **date** : date du match
- **round** : tour du match (finale / demi / poules etc...)
- **city** : ville du match
- **edition** : nom de l'édition de la coupe du monde

Du nettoyage sera à prévoir pour les noms d'équipe et les villes.
Mon code fait 100 lignes. Do it better ;)

Exercice 1 (30')

Notre premier voyage nous emmène à Boston où nous allons essayer de prédire les prix moyens des maisons occupées. On est sur du classique, le voyage commence doucement.

Info: Nous travaillons sur le dataset **BostonHousing** qui contient 506 lignes et 14 colonnes. L'enjeu va être de réaliser une régression.

Le jeu de données contient les variables suivantes :

Variables prédictives :

- **crim** : per capita crime rate by town.
- **zn** : proportion of residential land zoned for lots over 25,000 sq.ft.
- **indus** : proportion of non-retail business acres per town.
- **chas** : Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- **nox** : nitrogen oxides concentration (parts per 10 million).
- **rm** : average number of rooms per dwelling.
- **age** : proportion of owner-occupied units built prior to 1940.
- **dis** : weighted mean of distances to five Boston employment centres.
- **rad** : index of accessibility to radial highways.
- **tax** : full-value property-tax rate per $10,000.
- **ptratio** : pupil-teacher ratio by town.
- **black** : 1000(Bk - 0.63)2 where Bk is the proportion of blacks by town.
- **lstat** : lower status of the population (percent).

Variable à prédire :

- **medv** : median value of owner-occupied homes in $1000s.

Question 2

Charger le jeu de données BostonHousing et décrire les variables présentes.

Info: data("BostonHousing")
str()
summary()

Question 3

Explorer et visualiser la distribution de la variable à expliquer.

Info: ggplot()
stat_density()

Question 4

Explorer les différentes variables explicatives et démontrer la présence de potentielles correlations entre la variable à expliquer et les variables explicatives.

Question 5

Créer deux échantillons test et train (25/75).

Info: Avec le package caret.
createDataPartition()

Question 6

A partir des variables explicatives significatives, essayez différentes combinaisons de modèles linéaires pour créer un modèle performant. (3 combinaisons)

Info: lm()

Question 7

Examiner la performance des modèles créés. SI les modèles ne sont pas satisfaisants, recommencer l'étape précédente de modélisation.

Info: summary(model)$r.squared

Question 8

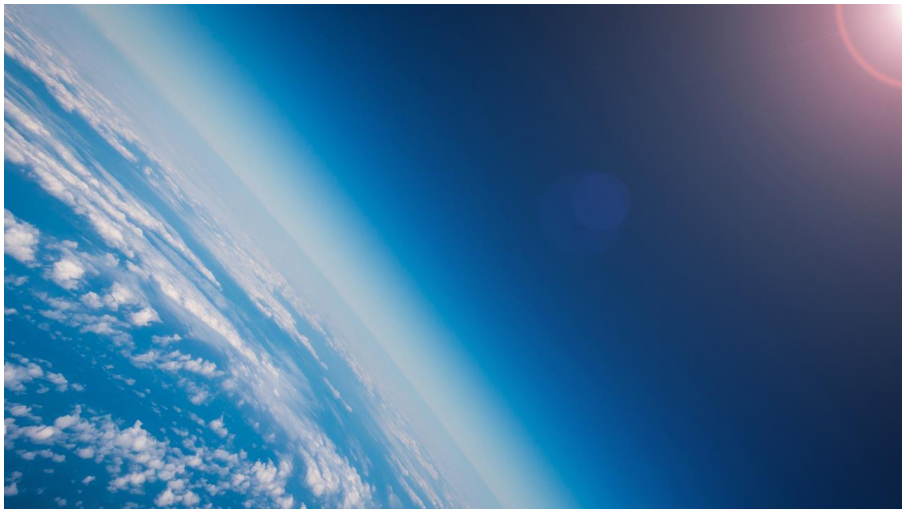
A l'aide des modèles créés, faire la prédiction sur l'échantillon test et tester la performance de prédiction des modèles. Comparer les erreurs de prédiction.

Info: Mettre les valeurs prédites dans un dataset de test et réaliser des calculs d'indicateur type RMSE / MSE, mais également un graphique affichant les différents modèles. predict()

Question 9

Créer d'autres modèles avec différents algorithmes glm. Comparer les performances.

Exercice 2 (30')



Les données ont été extraites et mises en forme par le service concerné de Météo France.

La couche d'ozone ou ozonosphère désigne la partie de la stratosphère contenant une quantité relativement importante d'ozone (concentration de l'ordre de un pour cent mille). Son existence est démontrée en 1913 par les physiciens français Henri Buisson et Charles Fabry grâce à son interféromètre optique.

Cet ozone est produit par l'action des UV, du rayonnement solaire, sur les molécules de dioxygène à haute altitudeN 1. Sydney Chapman propose le mécanisme de formation en 1930. Elle renvoie les rayons solaires et n'en laisse pénétrer que 50 % dans la troposphère.

Info: Nous travaillons sur le dataset **Ozone** qui contient 1041 lignes et 20 colonnes. L'enjeu va être de réaliser une régression.

Le jeu de données contient les variables suivantes :

- **JOUR** Le type de jour ; férié (1) ou pas (0) ;
- **O3obs** : La concentration d'ozone effectivement observée le lendemain à 17h locales correspondant souvent au maximum de pollution observée ;
- **MOCAGE** : Prévision de cette pollution obtenue par un modèle déterministe de mécanique des fluides (équation de Navier et Stockes);
- **TEMPE** : Température prévue par MétéoFrance pour le lendemain 17h ;
- **RMH2O** : Rapport d'humidité ;
- **NO2** : Concentration en dioxyde d'azote ;
- **NO** : Concentration en monoxyde d'azote ;
- **STATION** : Lieu de l'observation : Aix-en-Provence, Rambouillet, Munchhausen, Cadarache et Plan de Cuques ;
- **VentMOD** : Force du vent ;
- **VentANG** : Orientation du vent.

Question 10

La recherche d'une meilleure méthode de prévision suit le protocole suivant.

- Étape descriptive préliminaire uni et multidimensionnelle visant à repérer les incohérences, les variables non significatives ou de distribution exotique, les individus non concernés ou atypiques... et à étudier les structures des données. Ce peut être aussi la longue étape de construction de variables, attributs ou features spécifiques des données.
- Procéder à un tirage aléatoire d'un échantillon test qui ne sera utilisé que lors de la dernière étape de comparaison des méthodes.
- Créer l'échantillon d'apprentissage pour l'estimation des paramètres des modèles.
- Comparaison des qualités de prévision à l'aide de l'échantillon de test qui est resté à l'écart.

Info: Reprendre le protocol de l'exercice précédent est une bonne manière de réaliser cet exercice.