Consignes projet
4M071 – Modélisation Statistique

Maud Thomas

1 Projet

Ce projet constitue l'évaluation de ce cours. L'objectif est d'illustrer sur un jeu de données réelles les différents modèles et méthodes vus au cours du semestre.

1.1 Choix du jeu de données

A priori, vous êtes libre de travailler sur le jeu de données de votre choix (une liste de sites est disponible ci-dessous). Si vous êtes à court d'idées vous pouvez choisir un jeu de données dans la banque proposée sur Moodle. Votre choix doit **obligatoirement** être validé par Mme Thomas ou votre chargé de TP. Vous devez avoir fait validé votre jeu de données au plus tard le **30 janvier**.

Votre jeu de données doit contenir (les observations correspondent aux lignes et les variables aux colonnes)

- Au moins 100 observations
- Au moins 1 variable d'intérêt **quantitative continue** : c'est la variable que vous aimeriez expliquer en fonction des autres
- Au moins 3 ou 4 variables explicatives (les variables permettant d'expliquer la variable d'intérêt). Il faut à la fois des variables **quantitatives continues, discrètes et qualitatives**
- **Optionnel** : 1 variable d'intérêt discrète et/ou 1 variable d'intérêt qualitative avec deux modalités
- **Attention !** Pour les étutiants incrits en présentiel, deux binômes ne peuvent pas avoir le même jeu de données.

Où trouver un jeu de données ?

1. Les packages `R` :
   - `datasets` : à installer directement sur R
   - `DAAG` : à installer directement sur R
   - `insuranceData` : à installer directement sur R
   - `CASdatasets` : http://dutangc.perso.math.cnrs.fr/RRepository/
2. Les sites :
   - Kaggle https://www.kaggle.com
   - Google dataset https://datasetsearch.research.google.com
   - Data.gouv : https://www.data.gouv.fr/fr/
   - Page professionnelle de F. Husson : https://husson.github.io/data.html
3. En dernier recours, la banque proposée sur Moodle.

1.2 Consignes du projet

1. Vous devez réaliser ce projet en **binôme** (dans le même groupe de TP). *Pour les étudiants en FOAD, vous pouvez le faire en binôme ou seul.*
2. Votre projet doit contenir :
   — Une problématique, un objectif
   — Une analyse descriprive de vos données. **Attention !** Cette partie doit contenir uniquement des informations qui sont utiles pour le reste du projet
   — Régression linéaire multiple
   — **1 modèle au choix** parmi
      — ANOVA
      — ANCOVA
      — 1 GLM : logistique, Poisson, Gamma...

      Attention ! Chaque modèle doit être justifié et comprendre une étape de validation ! Il n'est pas demandé que les modèles soient bien ajustés aux données, mais que l'étape de validation soit menée correctement (pour les GLM la partie validation est assez succinte évidemment).
   — Vous devez rendre `Rmarkdown` (ou `R`, l'utilisation de `Rmarkdown` n'est pas obligatoire mais fortement recommandé) ainsi que votre rapport sur Moodle **avant le 3 mai**.
   — Vous devez aussi rendre votre jeu de données avec votre code R (même s'il a été déposé avant pour validation). Dans votre code, dans l'importation des données, il ne doit pas y avoir de chemin, juste le nom de votre jeu de données qui doit donc se trouver dans le même dossier que votre code. Votre code doit s'exécuter sans action de ma part.
   — Vous devrez également rendre un petit document texte donnant le descriptif de votre jeu de données notamment le nom des variables et la source.
   — **Attention !** Si vous utilisez les outils que nous n'avons pas vu en cours ou en TP, vous devez expliquer cet outil dans votre projet et vous devez être en mesure de répondre aux réponses à son sujet.

1.3 Consignes du rapport

1. Votre projet fera l'objet d'un rapport commum aux deux membres du binôme.
2. Le rapport doit contenir la présentation, l'explication et l'interprétation des différents éléments et modèles demandés ci-dessus. Vous appuyerez vos propos avec des graphiques et des tables judicieusement choisis. Vous devez présenter votre jeu de données, énoncer clairement la problématique et faire un choix judicieux des outils et des modèles que vous utilisez.
3. **Attention !** Tous les graphiques doivent être faits avec le package `ggplot2`.
4. Ce rapport devra être rendu au format **uniquement html ou pdf** (`Rmarkdown` permet notamment de générer un document html ou pdf à partir du code). **Tout autre format ne sera pas accepté**
5. **Attention !** Le code ne doit pas apparaître sur le rapport.
6. La note du rapport est commune aux deux membres du binôme. Celle-ci tiendra compte de la qualité de la rédation, de la clarté, de la justification du choix des modèles et de leur validation, du choix et de l'interprétation des graphiques et des tables.

1.4 Consignes de la soutenance

1. Votre projet donnera lieu à une soutenance qui aura lieu **lors de la semaine des examens** de 10 min par groupe (soit 5 min par membre du binôme).
2. Vous devez déposer votre soutenance sur Moodle **avant le 3 mai** au **format pdf**.
3. Votre exposé ne doit pas s'adresser uniquement au jury mais doit aussi convenir à une audience d'étudiants de niveau M1. Votre présentation doit donc être concise et pédagogique.
4. La note de la soutenance peut être différente pour les deux membres du binôme. Celle-ci tiendra compte de la qualité pédagogique et de la clarté de la présentation et de la pertinence des réponses aux questions posées.

1.5 Consignes du Contrôle continu

Ce contrôle continu correspond aux premiers résultats de votre projet. Ce premier travail vous permettra d'avoir un retour sur les premiers résultats de votre projet et sur ce qui est attendu dans votre rapport final.

1. Vous devez déposer votre code `Rmarkdown` (ou `R`) sur Moodle et le mini-rapport correspondant **avant le 22 mars** au **format pdf**.
2. Dans ce travail, vous devez
   — Présenter une petite problématique
   — Proposer une analyse descriptive de votre jeu de données
   — Effectuer une régression linéaire multiple avec son étape de validation.

2 Calcul de la note finale

Le non-respect d'une consigne sera pénalisé par le retrait de deux points sur la note finale.
— R = note du rapport
— CC = note du contrôle continu
— S = note de la soutenance
— F = note finale

$$\mathrm{F} = \max\left(0.65\mathrm{R} + 0.35\mathrm{S}, 0.5\mathrm{R} + 0.25\mathrm{S} + 0.25\mathrm{CC}\right)$$