ENSC Bordeaux INP – 2ème année

Module CO7SFMA1 : Modélisation statistique

Année universitaire 2021-2022

Projet sur la régression linéaire multiple

Consignes.

- *Le travail est à réaliser en trinôme.*
- *La date du remise du projet est fixée au* ***lundi 13 décembre 2021 à 14h*** *au plus tard.*
- *Le format du rapport est un fichier* ***PDF*** *de* ***maximum 8 pages*** *(annexes comprises).*
- *Le rapport doit être envoyé à l'adresse : jerome.saracco@ensc.fr*

Les attentes.

- *Le travail devra comprendre les codes R, des sorties numériques et graphiques issues de R, mais il faudra aussi et surtout fournir des commentaires pertinents et des conclusions compréhensibles par une personne qui n'est pas nécessairement spécialiste en statistique.*
- *Il est fortement recommandé de visualiser les données avant de faire de la modélisation statistique afin de repérer d'éventuelles observations hors normes ou aberrantes.*

Travail demandé : Étude du jeu de données « station » (voir feuille de TP 2)

La direction marketing d'un distributeur d'essence souhaite établir un modèle expliquant les ventes de ses stations-services situées dans les grands centres urbains. Le tableau de données précise, pour n=45 stations de ce type, les informations suivantes :

- les ventes de la station exprimées en milliers de litres (variable ventes),
- le nombre de pompes de la station (variable nbpompes),
- le nombre de concurrents dans la zone desservie par la station (variable nbconc),
- le trafic quotidien exprimé en milliers de voitures (variable trafic).

Les données sont contenues dans le fichier station.txt. L'objectif de l'étude est donc d'essayer de modéliser les ventes d'une station en fonction des autres variables disponibles.
La personne de l'étude doit mettre en œuvre des modélisations via des régressions linéaires (simples ou multiples) afin d'expliquer au mieux le nombre de ventes.

Vous pourrez suivre la démarche suivante :

1. Faire tout d'abord une étude descriptive des données (statistique descriptive, graphiques). Par exemple, vous pouvez opter pour une analyse multidimensionnelle des données en faisant une ACP (Analyse en Composantes Principales) et commenter les résultats obtenus (en gardant bien en mémoire la problématique initiale de régression linéaire multiple).

2. Expliquer pourquoi le modèle à trois variables ne peut être retenu.
   Quelle variable faut-il éliminer de ce modèle pour obtenir un modèle à deux variables qui semble mieux convenir ? Justifier votre choix.
   Ne pas oublier d'étudier les résidus...
   Remarque : en étudiant les résidus, vous allez vous apercevoir qu'une station pose un problème et qu'il sera alors nécessaire de supprimer cette station pour poursuivre votre modélisation.

3. a) Faire une analyse des indices de qualité du modèle à deux variables obtenu à la question précédente et en tirer une conclusion quant à sa validité statistique.
   b) Faire une interprétation du signe des coefficients de régression de ce modèle. Que pensez-vous de la validité économique du modèle ? Permet-il de mieux comprendre la réalité ?
   c) Préciser ce que l'on prédirait comme nombre de ventes étant donné un nouveau couple de valeurs des variables explicatives retenues (valeurs à choisir de manière raisonnable sur le support d'observation des deux variables sélectionnées dans le modèle final).

4. Justifier pourquoi le modèle à deux variables obtenu est préférable aux trois différents modèles de régression linéaire simple (i.e. à une seule variable).