ENSC Bordeaux INP – 2A

Année universitaire 2020-2021

Module CO7SFMA1 – Modélisation statistique

Contrôle continu sur la régression linéaire

Consignes.
Le travail est à réaliser par groupe de 3-4 élèves/
La date du remise du CC est fixée au vendredi 27 novembre 2020 à 14h au plus tard.
Le format du rapport est un fichier PDF de maximum 12 pages (annexes comprises).

Le travail devra comprendre les codes R, des sorties numériques et graphiques issues de R, des commentaires pertinents et des conclusions compréhensibles par une personne qui n'est pas spécialiste en statistique.

Remarque. Il est fortement recommandé de visualiser les données avant de faire de la modélisation statistique.

Travail demandé :

Traiter les deux exercices de la section 5 de la feuille de TP 2 « Régression linéaire multiple avec le logiciel R ».

Plus précisément, il y a deux jeux de données à considérer :

- **Le jeu de données « ozone ».**
  Vous pouvez suivre la démarche suivante :
  1. Faire une ACP de ces données en ne prenant en compte que les variables quantitatives et commenter les résultats obtenus.
  2. D'un point de vue descriptif, quel est l'impact de la variable qualitative ***pluie*** (resp. ***vent***) sur la variable ***maxO3*** ?
  3. Faire une ACP mixte de ces données en prenant en compte l'intégralité des variables (quantitatives et qualitatives), et commenter les résultats obtenus.
  4. A partir de ces données et des variables quantitatives disponibles (variables de température ***T9, T12, T15***, variables de nébulosité ***Ne9, Ne12, Ne15***, variables de vent ***Vx9, Vx12, Vx15***, et mesure du maximum de la concentration en ozone de la veille ***maxO3v***), proposer le meilleur modèle de régression linéaire multiple pour expliquer le maximum journalier de la concentration en ozone (variable ***maxO3***). Justifier votre démarche. Penser à bien regarder les résidus des modèles estimés.

- **Le jeu de données « station ».**
  Vous pouvez suivre la démarche suivante :
  1. Faire tout d'abord une étude descriptive des données. Par exemple, vous pouvez commencer par faire une ACP de ces données et commenter les résultats obtenus
  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...
  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 des coefficients 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.
  4. Justifier pourquoi le modèle à deux variables obtenu est préférable aux différents modèles de régression linéaire simple (i.e. à une seule variable).