Projet modèle linéaire sous R (M1–MIDO, 2024-2025)

Table des matières

Contexte, base de données et objectifs

Contexte

Un système de partage de vélos, un programme de vélos en libre-service, un programme de vélos publics ou un programme de vélos publics en libre-service est un service de transport partagé dans lequel des vélos sont mis à la disposition de particuliers pour un usage partagé à faible coût. Grâce à ces systèmes, l'utilisateur peut facilement louer un vélo à un endroit donné et le rendre à un autre endroit. En juillet 2020, Google Maps a commencé à inclure les systèmes de vélos en libre-service dans ses recommandations d'itinéraires. En 2022, environ 3000 villes dans le monde offriront des systèmes de partage de vélos, par exemple Dubaï, New York, Paris, Mexico, Montréal et Barcelone. Aujourd'hui, ces systèmes suscitent un grand intérêt en raison du rôle important qu'ils jouent dans la circulation, de l'environnement et de la santé.

Base de données

Le processus de location de vélos en libre-service est étroitement lié aux conditions environnementales et saisonnières. Par exemple, les conditions météorologiques, les précipitations, le jour de la semaine, la saison, l'heure de la journée, etc. peuvent affecter les comportements de location. Les données correspondent au journal historique pour une année du système de partage de vélos d'une certaine ville. Les données sont agrégées sur une base journalière. Une observation est une période de la journée dans laquelle on a compté le nombre de vélos loués. Le nombre d'observations est $n = 1817$.

Objectif du projet

Le but du projet est de déterminer le meilleur modèle qui prédit/explique le nombre de location de vélos (`velos`), en fonction de diverses caractéristiques que vous trouverez dans le codebook çi-dessous.

Vous décrirez la question d'intérêt, les étapes que vous avez suivi dans l'analyse des données, vos conclusions sur la question d'intérêt et les limites éventuelles de votre étude. Vous décrirez précisément les méthodologies utilisées et vous appuierez votre rapport sur des tableaux et des graphiques. Votre rapport peut contenir ce qui suit:

1. Une discussion de la question d'intérêt. Pourquoi cette question vaut-elle la peine d'être étudiée? Description des variables prédictives et hypothèses sur les effets de ces différentes variables.
2. Partie Analyse: distribution des variables (tableaux, graphs, …), exploration des données, discussion des étapes de modélisation, sélection du model de régression final. Indicateur de validité du modèle etc …
3. Résultats et inférence sur la question d'intérêt: quelles sont les facteurs qui prédisent/expliquent significativement ($p < 0.05$) de façon indépendante le location de vélo.
4. Limite de l'étude et conclusion: amélioration de l'étude possible? biais? etc…

Attentes pour le projet (idéal)

Pensez à (ne pas faire attention à l'ordre des points proposés) :

- Effectuer le data-management nécessaire pour créer les variables pour l'analyse, en particulier les variables de types `factor` (catégorielles)
- Examiner les variables avant de les inclure dans un modèle de régression: tableaux et/ou graphiques présentant la distribution de chaque variable
- Transformations de variables (à expliquer ou explicative) si nécessaire.
- Corrélation (association) entre les variables quantitatives/qualitatives (Tableaux et/ou graphiques). Problème de multicolinéarité éventuel. Supprimer les variables redondantes éventuelles.
- Association entre le nombre de vélos loués et chaque variable prédictive pris une à une (Tableaux et/ou graphiques)
- Peut être centrer (et/ou changer l'échelle de) certaines variables continues pour une meilleure interprétation des réultats
- Méthode pour sélectionner le modèle de régression multiple final
- Écrire et interpréter l'équation du modèle de régression linéaire multiple final
- Interpréter les coefficients de régression estimés
- Vérification des hypothèses du modèle, analyse des résidus…
- Rechercher les valeurs aberrantes et les observations influentes…
- Réajuster le modèle en retirant les valeurs atypiques pour améliorer l'adéquation, si nécessaire.
- Réajuster le modèle avec des interactions éventuelles entre certaines variables: tester et visualiser les interactions entre les prédicteurs, si vous en proposez.

Important :

La capacité prédictive du modèle sera notée par les correcteurs.
Diviser aléatoirement les données (70%/30% ou 80%/20%) en une base d'apprentissage et une base de test. Entraîner le modèle sur l'apprentissage et évaluer l'erreur du modèle sur la base de test. La fonction *loss* attendue pour évaluer cette erreur est le *Mean Square Error (MSE)*.

Variable dans la base	Valeurs numériques	Label
obs		**Numéro de l'observation**
saison		**Saison**
	1	Hiver
	2	Printemps
	3	Eté
	4	Automne
mois		**Mois**
	1	Janvier
	2	Février
	3	Mars
	4	Avril
	5	Mai
	6	Juin
	7	Juillet
	8	Août
	9	Septembre
	10	Octobre
	11	Novembre
	12	Décembre
jour_mois		**Jour du mois**
jour_semaine		**Jour de la semaine**
	1	Dimanche
	2	Lundi
	3	Mardi
	4	Mercredi
	5	Jeudi
	6	Vendredi
	7	Samedi
horaire		**Heure dans la journée**
	1	[24h-07h[
	2	[07h-11h[
	3	[11h-15h[
	4	[15h-19h[
	5	[19h-24h[
jour_travail		**Jour de travail (Oui/Non)**
	1	Non
	2	Oui
vacances		**Vacances (Oui/Non)**
	1	Non
	2	Oui
meteo		**Météo**
	1	Claire
	2	Nuageux/Brumeux
	3	Pluie/Neige
velos		**Nombre de locations de vélos**
humidite		**Pourcentage d'humidité**
vent		**Vitesse du vent (Km/h)**
temperature1		**Température moyenne mesuré (°celcius)**
temperature2		**Température moyenne ressentie (°celcius)**