Modèles Linéaires

Contrôle Continu - Deuxième Epreuve - 1h30
Licence 3 MIASHS (2023-2024)

Guillaume Metzler

Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Résumé

L'examen sera entièrement effectué sur R et vous devrez déposer votre fichier dans l'espace prévu à cet effet sur Moodle.

La description et la présentation de la démarche sera prise en compte dans l'évaluation de votre copie.

Un problème de régression multiple

De nombreux facteurs peuvent impacter le nombre de spectateurs à un match de Baseball comme la température extérieure, le pourcentage de victoires des deux équipes, le moment où a lieu le match ou encore si les billets sont vendus à prix réduits ce jour là.

Le jeu de données contient les informations suivantes :

- *Team* : nom de l'équipe.
- *Attendance* : nombre de spectateurs
- *Temp* : température maximale du jour
- *Win%* : le pourcentage de succès de l'équipe à cette période de la compétition
- *OpWin%* : le pourcentage de succès de l'équipe adverse à cette période de la compétition
- *Weekend* : prend la valeur 1 si le match a lieu le week-end et 0 sinon
- *Promotion* : prend la valeur 1 si les billets ont été vendus à prix réduits et 0 sinon

Votre objectif est de construire le meilleur modèle de régression capable de prédire les valeurs de la variable *Attendance*. **Au préalable, vous choisirez une seule et unique équipe parmi les 4 disponibles !**.

Vous devrez penser à écarter les variables redondantes et construire votre modèle en vous basant sur un critère statistique comme l'AIC ou le BIC. Enfin, on pensera à analyser les résidus du modèle et on pourra aussi regarder ses performances via le critère d'*Erreur Quadratique Moyenne*.

Un problème de régression logistique

Une étude a été conduite afin de savoir s'il existe un biais de genre dans le cadre de recrutement dans des universités de Sciences. Pour cela, des entretiens ont été mené et plusieurs informations ont été collectées afin de trancher ou non sur ce phénomène là. Les données collectées se trouvent dans le fichier ...

Afin de mener à bien votre étude, on dispose des informations suivantes :

- *offered starting salary* : le salaire à l'embauche
- *competence rating* : compétences du candidat
- *gender of the candidate* : genre du candidat. 0 - Masculin et 1 - Féminin
- *gender of the rater* : genre du recruteur. 0 - Masculin et 1 - Féminin
- *type of school* : nature de l'école. 0 - Ecole publique et 1 - Ecole privée

- *age of rater* : age du recruteur
- *age of candidate* : age du candidate

Votre objectif est déterminer si le modèle est biaisé ou non. Pour cela, on va considérer un modèle de régression logistique où la variable à prédire (Y) est la variable *Genre du candidat*.

Après avoir supprimé les informations redondantes, construire un modèle de régression logistique qui permettra de déterminer si oui ou non le recrutement est biaisé.
On se basera sur la notion d'accuracy pour évaluer le biais de notre modèle.
On pourra aussi séparer notre jeu de données en deux ensembles train/test si on le souhaite.

Biais d'un modèle Notre modèle sera dit biaisé s'il est parfaitement capable de discriminer les candidats du sexe masculin des candidats de sexe féminin.