Devoir 4

Ce travail est à réaliser en équipe (minimum deux, maximum quatre personnes).

Vous devez remettre les documents suivants:

- votre rapport au format PDF
- votre code **R** ou un fichier Rmarkdown

Utilisez la convention de nomenclature `d4_matricule.extension`, où `matricule` est le matricule de l'étudiant(e) qui soumet le rapport et `extension` est un de `pdf`, `R` ou `Rmd`.

Question 1

Le modèle de Bradley & Terry (1952) décrit la probabilité que le résultat de l' «équipe i» soit supérieur à celui de l'«équipe» j,

$$\Pr(Y_i > Y_j) = \frac{\exp(\beta_i)}{\exp(\beta_i) + \exp(\beta_j)}, \quad i, j \in \{1, \dots, K\},$$

en assumant que les doublons (égalité) ne surviennent pas.

Ce modèle simple peut servir pour prédire le classement d'équipes sportives: si on écrit le modèle en terme de cote, on obtient pour l'équipe i à domicile et l'équipe j en visite

$$\ln \left\{ \frac{\Pr(\text{victoire équipe } i \text{ (domicile)})}{\Pr(\text{victoire équipe } j \text{ (visiteur)})} \right\} = \beta_i - \beta_j.$$

Un paramètre du modèle n'est pas identifiable: une des catégories $\text{ref} \in \{1, \dots, K\}$ sert de référence et le coefficient correspondant est nul, soit $\beta_{\text{ref}} = 0$.

Le modèle décrit ci-dessus peut être ajusté à l'aide d'une régression logistique avec un ensemble de $K-1$ variable explicatives[1] où pour le match i et l'équipe $k = 2, \ldots, K$, on a

$$X_{ik} = \begin{cases} 1, & k = i, \\ -1, & k = j, \\ 0, & \text{sinon.} \end{cases}$$

Le modèle Bradley–Terry de base n'a pas d'ordonnée à l'origine. Si on l'ajoute, l'équation du modèle pour une partie devient

$$\ln\left\{\frac{\Pr(\text{victoire équipe } i \text{ (domicile)})}{\Pr(\text{victoire équipe } j \text{ (visiteur)})}\right\} = \alpha + \beta_i - \beta_j,$$

où β_i représente la force de l'équipe à domicile, β_j la force de l'équipe en visite et l'ordonnée à l'origine α capture l'effet du jeu à domicile.

La base de données `lnh` du paquet `hecmulti` contient les résultats de chaque partie par équipe, tandis que `lnh_BT` fournit les mêmes données, mais dans un format propice pour l'ajustement du modèle de Bradley–Terry.

Ajustez le modèle de Bradley–Terry aux données `lnh_BT` (utilisez la formule `vainqueur ~ .` pour ajuster le modèle avec toutes les équipes). La catégorie de référence est `Anaheim Ducks`, qui n'apparaît pas dans les sorties.

1. Interprétez le coefficient pour l'ordonnée à l'origine α en terme de pourcentage d'augmentation ou de diminution de la cote par rapport à la référence jouer à l'extérieur.
2. Calculez un intervalle de confiance de niveau 95% pour l'ordonnée à l'origine et déterminez si jouer à domicile impacte significativement le score.
3. Fournissez un tableau avec le classement des cinq premières équipes qui ont la plus grande chance de succès selon le modèle.[2]
4. Pour chaque match, utilisez le modèle logistique pour prédire l'équipe gagnante. Construisez une matrice de confusion (1 pour une victoire de l'équipe à domicile, 0 sinon) et rapportez cette dernière.
5. Calculez le taux de bonne classification, la sensibilité et la spécificité avec un point de coupure de 0.5 (assignation à l'événement ou à la classe la plus probable).
6. Produisez un graphique de la fonction d'efficacité du récepteur et rapportez l'aire sous la courbe. Commentez sur la qualité prédictive globale du modèle.
7. Retournez le lift pour le 40e percentile des succès.

[1] Il n'y a pas de variable explicative X_{ref} pour la catégorie de référence, autrement les données seraient colinéaires.

[2] Attention à la catégorie de référence.

Question 2

Dans une étude sur le comportement d'investisseurs et leur perception du comportement aléatoire ou épistémique de la bourse, Walters et al. (2022) étudie la présentation de résultats financiers et l'impact du format (par exemple, par le biais de différents graphiques) sur leur perception du risque. Le but de l'étude 4A de l'article est d'établir l'effet modérateur de l'incertitude (épistémique ou aléatoire) sur le risque et les décisions d'investissement.

Les données `compinvest` fournissent les observations de l'étude.

Ajustez un modèle logistique avec effets principaux (`risque`, `aleatoire`, `epistemique`) et une interaction entre risque et aléatoire, ainsi que risque et épistémique comme dans l'article.

1. Rapportez un tableau des coefficients du modèle et comparez avec les résultats du Tableau 3 de la prépublication.[3]
2. Interpétez l'effet d'une augmentation de l'échelle de risque de forte aversion (`risque = 1`) à en quête de risque (`risque = 4`) en terme de pourcentage d'augmentation ou diminution de la cote pour

    - un score aléatoire (`aleatoire`) de 1 point supérieur à la moyenne des scores de l'échantillon et
    - un score pour `epistemique` égal à la moyenne de l'échantillon.[4]

3. Vérifiez l'hypothèse de modération des auteurs en déterminant si l'effet des termes d'interaction est significativement non-nul à niveau 5%.

    - Rapportez les valeurs des statistiques de test et les valeur-p associées. Commentez sur les résultats obtenus.

4. Produisez un graphique de la probabilité prédite d'investir (axe des ordonnées) en fonction de l'échelle pour la perception de l'aléatoire (`aleatoire`, axe des abcisses) pour une personne avec un score `epistemique` égal à la moyenne.

    - Tracez deux courbes: une pour une personne avec un score de risque de 1 (forte aversion au risque), l'autre avec un score de 4 (en quête de risque).[5]
    - En vous basant sur le graphique, commentez sur les différences estimées en terme de décision d'investissement, selon le profil de risque.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345. https://doi.org/10.2307/2333402

[3] Notez que les auteurs de l'article ont centré préalablement les variables explicatives (c'est-à-dire, en soustrayant la moyenne de chacune des colonnes).

[4] *Astuce*: vous pouvez calculer le rapport de cote en remplaçant les coefficients dans l'équation, ou prédire les probabilités pour les deux profils de personnes présentés et calculer le rapport de cote.

[5] Astuce: créez une base de données (`data.frame`) avec une séquence adéquate de valeurs de `aleatoire` pour les valeurs données de `risque` et de `epistemique`.

Walters, D. J., Ulkumen, G., Tannenbaum, D., Erner, C., & Fox, C. R. (2022). *Investor behavior under epistemic versus aleatory uncertainty.* https://doi.org/10.2139/ssrn.3695316