Projet : Analyse approfondie de la relation âge-performance

Ce sujet à pour objectif d'étudier la relation entre l'âge et la performance (physique ou cognitive) à partir de données réelles et de travaux issus de la littérature scientifique. Cette relation est importante pour essayer de mieux comprendre comment les mécanismes liés au vieillissement impactent les performances physiques, cognitives, et plus généralement physiologiques. L'idée générale de ce projet est double: vous préparer à rédiger un document synthétique (comme votre rapport de stage ou d'alternance) d'une part, et vous préparer à des problématiques que vous pourrez rencontrer en entreprise d'autre part:

- lire et comprendre un document vous donnant un objectif,
- se familiariser avec les données:
  - comprendre des données et leur implications,
  - manipuler des données hétérogènes (continues ou discrètes), manquantes, érronnées ou aberrantes,
  - structurer des données pour pouvoir les exploiter statistiquement,
- utiliser le calcul numérique pour répondre à une interrogation (ici: approximer les paramètres d'une équation),
- observer et interpréter des statistiques pour approfondir un sujet et tirer des conclusions les plus précises et rigoureuses possible,
- rédiger un document synthétique (comme votre rapport de stage ou d'alternance) en sélectionnant les informations les plus pertinentes.

Dans ce projet, les objectifs sont clairement explicités, il suffit de bien lire l'énoncé. On vous demande de structurer des données de performance et d'étudier la relation de ces données avec l'âge. Il va vous falloir commencer par comprendre ces données, puis les traiter: certaines seront aberrantes ou manquantes. Il vous faudra manipuler des données temporelles (et les convertir) et choisir une discrétisation, par semestres ou par années par exemple. Vous devrez également apprendre à utiliser une fonction externe (en R) pour estimer les valeurs des paramètres de plusieurs équations non linéaires. Enfin, il vous faudra réfléchir à comment ordonner et mettre en avant vos résultats, notamment au type de tableaux et de graphiques que vous choisirez de présenter dans votre rapport final. Ils doivent tous être lisibles et compréhensibles (avoir une légende qui décrit ce qu'il y a sur l'axe X et sur l'axe Y? par exemple). Vous devrez finalement essayer de tirer quelques conclusions d'ordre plus général sur cette relation âge-performance.

Pour vous aider dans cette tâche, vous aller trouver deux documents sur le Moodle:

- "*Aide fonctions R.txt*" est un fichier donnant quelques exemples d'utilisation des fonctions gsub() et str_split()
- "*fonction_moindre_carrés.R*" qui contient le script de la fonction MMC() dont vous aurez besoin pour la question 3

Barème

Pour résumer ce travail, vous produirez un court document d'une dizaine de page environ par groupe de 2 en répondant aux 5 questions suivantes. Il faudra bien détailler vos données, présenter quelques graphiques et/ou tableaux clairs, et terminer par un paragraphe récapitulant vos découvertes, interprétations, et conclusions sur cette relation âge-performance. Le barème sera le suivant:

- Question 1 : 7 points
- Question 2 : 3 points (1 point pour le linéaire, 1 point pour le polynôme, et 1 point pour l'équation de Moore)
- Question 3 : 3 points
- Question 4 : 3 points
- Question 5 : 4 points
- Question Bonus 1 : 2 points
- Question Bonus 2 : 2 points

Question 1

Les jeux de données pour plusieurs épreuves d'athlétisme sont disponibles dans deux fichiers csv de résultats sportifs:

- 'resultats_men.csv' pour les hommes
- 'resultats_women.csv' pour les femmes

Chacun de ces fichiers comprennent plusieurs variables:

- Rank: la place au ranking de l'année
- Mark: le temps réalisé (la performance donc)
- Competitor: nom de l'athlète
- DOB: date de naissance
- Nat: nationalité
- Pos: position le jour de la course
- Venue: lieu de l'épreuve
- Date: date de l'épreuve
- Results.Score: score IAAF
- Annee: année du ranking / épreuve
- Dis: discipline

On s'interesse en particulier à la relation liant l'âge et la performance. Quand on parle d'âge, on parle de l'âge du compétiteur (de l'athlète) effectuant la performance à un instant donné. Ainsi, chaque valeur de performance est réalisée à un âge donné. Par exemple, un athlète peut courrir le 100m. à 25.1247 ans en 10.12 secondes. L'objectif est donc de structurer ces données et d'obtenir un jeu de données comportant le couple âge et performance réalisée, ainsi que l'épreuve où cette performance est réalisée. En effet, il est évident que la performance obtenue dans le 100m n'est pas la même que dans le marathon.
Dans un premier temps, on pensera à convertir ces deux variables textes (l'âge et la performance donc) en valeurs numériques en utilisant les fonctions R suivantes: str_split(), gsub() et as.Date() dans R (voir fichier d'aide "*Aide_fonctions_R.txt*" sur le Moodle).
On est intéressé pour étudier cette relation âge performance d'un point de vue "des maximums", c'est à dire qu'on souhaite, pour chaque âge, et pour chaque épreuve, observer quelle est la meilleure performance réalisée du point de vue humain. Il faut donc regrouper les performances par catégorie d'âge. Par exemple, on peut choisir de prendre la meilleure performance pour chaque âge par année (en arrondissant par "année d'âge" donc): la meilleure performance pour 18 ans, la meilleure performance pour 19 ans, etc. On peut aussi choisir de prendre la meilleure performance par trimestre: la meilleure performance

entre 18.00 ans et 18.33 ans, puis entre 18.33 ans et 18.66 ans, etc. C'est à vous de décider quelle discrétisation (catégorie, ou répartition) choisir.
Attention aux pièges: on exprimera la vitesse en m.s^{-1} (on rappellera que $v = d/t$ ou d est la distance de course et t le temps réalisé sur cette course).
A vous de choisir le nombre d'épreuves à étudier, plus le nombre d'épreuves est grand, plus vos observations, interprétations et conclusions seront robustes.

Question 2

Dans cette question, nous allons ajuster trois modèles pour chaque épreuve. C'est à dire que nous allons ajuster 3 modèles pour le 100m, 3 modèles pour le marathon, 3 modèles pour le 800m, etc. Notons $P(t)$ la performance à un âge t. Ajustez ces 3 équations à aux jeux de données que vous avez sélectionné:

$$P(t) = at + b, \qquad P(t) \geq 0 \tag{1}$$

$$P(t) = at^2 + bt + c, \qquad P(t) \geq 0 \tag{2}$$

$$P(t) = a(1 - e^{-bt}) + c(1 - e^{dt}), \qquad P(t) \geq 0 \tag{3}$$

Ou l'équation 3 est l'équation de Moore. Pour cette dernière équation, a, b, c, d sont 4 constantes positives et il vous faudra utiliser la régression non-linéaire à l'aide de la fonction MMC() mise à votre disposition. L'entête de cette fonction est le suivant: MMC=function(age, vitesse, methode, nbpara, precision=10, borne=-Inf, initial=NULL). Les paramètres "age" et "vitesse" correspondent à vos données, c'est à dire à vos couples (age, vitesse). La "methode" est l'équation que vous souhaitez utiliser, un exemple est donné ici:

```
une_fonction_f = function(x,p){
p[1]*x^3 + exp(p[2]*x^2)
}
```

ou "p" sont les paramètres (coéfficients) à estimer de votre équation, et "x" la variable. Dans cet exemple on spécifie l'équation d'une droite $ax + b$ ou a et b sont les paramètres à estimer et dénotés p[1] et p[2] dans ce code R. Le paramètre "nbpara" correspond au nombre de paramètres de votre équation (2 dans l'exemple présenté), et "precision" est le nombre de répétitions différentes de l'algorithme. Plus ce nombre est grand, plus vous avez de chances de trouver une bonne estimation des coefficients. Attention cependant à ne pas mettre une trop grande valeur car cela pourrait fortement ralentir voire bloquer vos ordinateurs (trop de répétitions). Le paramètre "borne" est une valeur seuil pour contraindre l'algoritme à estimer les paramètres dans un domaine dont la partie inférieure est bornée. Par défaut, le borne inférieure est égale à $-\infty$ (borne = -Inf), ce qui signifie qu'elle n'est pas bornée, mais si vous souhaitez contraindre l'estimation de vos paramètres par une borne inférieure spécifique, vous pouvez utiliser ce paramètre. Par exemple: borne = rep(0,4) vous donnera une borne inférieure de 0 pour vos 4 paramètres. Enfin, le paramètre "initial" est un paramètre optionnel qui contient les valeurs initiales de vos paramètres ((a, b) dans notre exemple). Si vous ne le spécifiez pas, l'algorithme tirera des valeurs aléatoires comme conditions initiales pour chacun de vos paramètres. Si au contraire, vous avez une bonne idée des valeurs que devraient prendre vos paramètres a, b, etc. alors vous pouvez les spécifier en utilisant "initial".

Un exemple de figure qui pourrait figurer dans votre rapport est présenté (voir Fig. 1). Dans cette proposition de figure, les performances de plusieurs épreuves différentes sont tracées en fonction de l'âge sur le même graphique. L'axe des X représente l'âge et l'axe des Y, la vitesse (en m.s^{-1}).

Question 3

On cherche à mesurer quel modèle s'ajuste le mieux aux données : s'agit-il du modèle polynomial? du modèle de Moore? Nous avons donc besoin d'indicateurs statistiques puor mesurer les ajustmenets de chaque modèle, afin de savoir si ces modèles dévcrivent bien nos données. Pour cela, on calculera les quantités suivantes pour chaque équation (modèle):

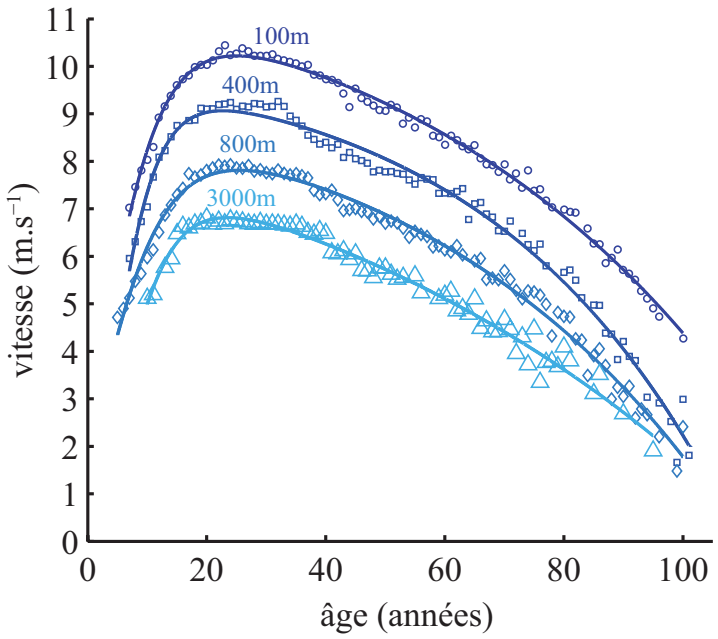

Figure 1: Exemple de réalisation d'une figure avec 4 épreuves différentes contenant les données et un ajustement de modèle (Moore). La relation âge-performance est tracée avec une valeur de performance par année d'âge. C'est à dire que la meilleure performance est choisie pour chaque âge entier (les valeurs d'âge sont arrondies en entiers donc). Les valeurs de vitesses sont exprimées en m.s^{-1}. Les cercles (valeurs du 100m), carrés (valeurs du 400m), losanges (800m) et triangles (3000m) représentent ces meilleures performances par âge entier. Les courbes sont les modèles de Moore ajustés à chacune de ces épreuves.

- étude des résidus (normalité).
- le coefficient de détermination ajusté:

$$R^2 \text{ ajusté} = 1 - \frac{(1-R^2)(n-1)}{n-k-1} \tag{4}$$

  avec n le nombre d'observations (taille de l'échantillon) et k le nombre de paramètres de l'équation.

- L'AICc (corrected Akaike information criterion):

$$\text{AIC} = n \log\left(\frac{\text{RSS}}{n}\right) + 2k + \frac{2k\,(k+1)}{n-k-1} \tag{5}$$

- Le critère d'information bayésien (BIC):

$$\text{BIC} = n \log\left(\frac{\text{RSS}}{n}\right) + k \log\,(n) \tag{6}$$

On rappelle que la somme des résidus s'écrit:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \tag{7}$$

avec y_i la valeur à estimer et $\tilde{y}_i$ la valeur estimée par l'équation.

Que pouvons-nous en déduire en terme d'ajustement des modèles? Y a t'il un meilleur modèle?

Question 4

A ce stade et à partir des résultats obtenus, vous constatez que la relation âge-performance ressemble à un "U" inversé. Cette relation admet donc un "pic", autrement dit, il existe un âge t ou la performance est maximale. Ce pic de performance est simplement le maximum de l'équation dans l'intervalle d'âge étudié, autrement dit $\max(P(t))$. La question est: est-ce que l'âge t ou ce pic intervient est toujours le même, quelque soit l'épreuve observée ? En d'autre mot, est-ce qu'il existe un âge ou la performance est maximale, et ce, quelque soit la distance que nous courons ou le sport que nous pratiquons? Vous pouvez obtenir cette valeur de performance maximale $\max(P(t))$ à partir de la fonction MMC (valeur de "peak" dans les résultats) et pouvez la calculer pour chaque épreuve que vous avez traité. Cependant, nous recherchons l'âge t ou cette valeur de pic $\max(P(t))$ intervient. Pour obtenir cet âge t de performance maximale, il faut donc trouver la valeur de t dans les équations 1, 2 et 3 en fonction de $\max(P(t))$. La fonction R *optimize(fct, intervalle)* permet d'obtenir cette valeur t ainsi que la valeur maximum de performance dans le modèle (c'est à dire $\max(P(t))$, qui est également founi par la fonction MMC(), voir "peak" dans les résultats).

Que pouvez-vous déduire de cette analyse des âges ou la performance est maximale pour chaque épreuves?

Question 5

On ajoute un nouveau jeu de données avec le fichier "resultats_joueurs_echecs.csv". Il s'agit des performances des joeurs d'échecs. Identiquement à précédemment, chaque individu (défini par la colonne "nom") possède une performance (nombre de points "Points") à un âge donné. Il vous faudra répéter l'expérience pour afficher la courbe de performance (Points) en fonction de l'âge. Il faudra d'abord caculer les âges (quand c'est possible) et répéter les étapes précédentes.

Questions Bonus 1

Comparer les équations précédentes à cette nouvelle équation:

$$P(t) = \beta_0 N_\infty \cdot e^{-\frac{\alpha_0}{\alpha_r} e^{-\alpha_r t}} \cdot \left(1 - e^{\beta_r (t - t_d)}\right) \tag{8}$$

où α_0, α_r, β_0, et β_r et t_d sont les constantes à ajuster. Quelle équation vous semble la meilleure (en terme d'ajustement) et pourquoi?

Questions Bonus 2

En premier lieu, utilisez ChatGPT pour répondre à une question de ce TD, celle de votre choix. Produisez ensuite des éléments écrits permettant de démontrer quels sont les avantages et les faiblesses de la réponse de chatGPT par rapport à la réponse que vous avez produite dans ce TD. En d'autres mots: estimer la différence entre la réponse de ChatGPT et la vôtre, en argumentant pourquoi c'est mieux / moins bien. Ces éléments écrits peuvent être un court paragraphe, accompagné de résultats chiffrés ou d'un code informatique.