Proposition TD/TP relation âge-performance

Ce sujet à pour objectif d'étudier la relation entre l'âge et la performance physique 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.

1 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: temps réalisé
- 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 l'âge et la performance, qu'il faut calculer. On pensera à convertir ces deux variables textes en valeurs numériques en utilisant les fonctions R suivantes: str_split(), gsub() et as.Date() dans R.
On exprimera l'âge en année (discrétisation annuelle) en prenant la meilleure performance pour chacun de ces âges.
On exprimera la vitesse en $m.s^{-1}$.
Ensuite, tracer les courbes de performance en fonction de l'âge pour 3 épreuves différentes que vous choisirez sur le même graphique, en utilisant l'âge pour l'axe des X et la vitesse (en $m.s^{-1}$) pour l'axe des Y. Que remarquez-vous?

2 Question 2

Notons $P(t)$ la performance à un temps (ou âge) t. Ajuster ces 3 équations à ces 3 jeux de données:

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

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

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

Ou l'équation 3 est l'équation de Moore. Pour cette dernière équation, a, b, c, d sont 4 constantes positives et il faudra utiliser la régression non-linéaire (voir la fonction MMC() mise à disposition).

3 Question 3

On cherche à savoir si ces équations décrivent bien nos données. Pour cela, on calculera les quantités suivantes pour chaque équation:

- étude des résidus (normalité) et l'homoscédasticité.
- 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?

4 Questions Bonus 1

Comparer les 2 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 de fit) et pourquoi?