SAE DOMAINE D'APPLICATION

BUT SD 2023-2024 – 2ÈME ANNÉE

-

Analyse approfondie de la relation âge-performance

-

Ndoumbe SECK
Merlin SIMSEN BARATAULT

-

IUT de Paris - Rives de Seine
Université Paris Cité

SOMMAIRE

INTRODUCTION

Présentation du sujet

La relation entre l'âge et la performance physique est un domaine d'étude crucial pour la compréhension des mécanismes du vieillissement et de leurs impacts sur les capacités humaines. Ce projet vise à approfondir cette relation en analysant les données réelles provenant de compétitions d'athlétisme, tout en essayant d'utiliser des modèles mathématiques pour mieux appréhender cette relation.

-

Objectifs du projet

L'objectif principal de cette analyse est de comprendre comment les performances physiques évoluent en fonction de l'âge des individus, en prenant en compte diverses variables telles que la vitesse de course ou d'autres mesures de performance. À travers cette étude, nous cherchons à déterminer s'il existe des modèles adéquats pour décrire cette relation et à identifier les facteurs qui influent sur la performance à différentes étapes de la vie.

Nous poserons les bases de notre analyse en présentant les jeux de données disponibles après le nettoyage de la base de données, en définissant les objectifs de l'étude et en exposant la méthodologie utilisée pour atteindre ces objectifs. Nous aborderons également les enjeux théoriques liés à la relation entre l'âge et la performance, ainsi que les implications pratiques de nos résultats. Enfin, nous énoncerons les différentes questions auxquelles nous tenterons de répondre au cours de ce projet, en mettant en évidence l'importance de comprendre cette relation.

MÉTHODOLOGIE

Description des jeux de données

Le jeu de données "Résultats Sportifs" constitue une compilation des performances dans diverses épreuves d'athlétisme, organisées en deux ensembles distincts :

- "resultats_men.csv" pour les compétitions masculines
- "resultats_women.csv" pour celles féminines.

Ces ensembles de données fournissent un aperçu détaillé des résultats obtenus par les athlètes dans différentes disciplines.

Variables incluses :

Rank (Classement) : Position occupée dans le classement annuel

Mark (Temps réalisé) : Temps enregistré par l'athlète, reflétant sa performance dans l'épreuve

Competitor (Concurrent) : Nom de l'athlète ayant participé à la compétition

DOB (Date de naissance) : Date de naissance de l'athlète

Nat (Nationalité) : Nationalité de l'athlète

Pos (Position) : Classement de l'athlète lors de l'épreuve

Venue (Lieu de l'épreuve) : Endroit où l'épreuve s'est déroulée

Date (Date de l'épreuve) : Date à laquelle l'épreuve a eu lieu

Results.Score (Score IAAF) : Score attribué par l'IAAF (Association Internationale des Fédérations d'Athlétisme)

Annee (Année du classement/épreuve) : Année à laquelle le classement a été établi ou à laquelle l'épreuve s'est déroulée

Dis (Discipline) : Discipline sportive de l'épreuve

Le fichier "resultats_hommes.csv" nous fournit 120 454 entrées pour 17 variables.
Et le fichier "resultats_femmes.csv" contient 20 846 entrées pour 17 variables.

Ces données offrent une base solide pour nous permettre d'analyser et comprendre les performances des athlètes dans le domaine de l'athlétisme, nous permettant ainsi d'explorer de manière approfondie la relation entre l'âge et la performance sportive, ainsi que selon le sexe.

Explication des étapes de traitement des données

Avant de pouvoir utiliser les données pour nos analyses, nous avons dû passer par une série d'opérations de nettoyage et de préparation afin d'assurer la qualité et la cohérence des données avant l'analyse.

Tout d'abord, les données des résultats sportifs des femmes sont soumises à un processus de nettoyage rigoureux. Les variables inutiles ou ayant un pourcentage de valeurs manquantes trop élevé sont supprimées du jeu de données. Concernant les variables comme la date de naissance ou la date de l'épreuve, les valeurs manquantes sont supprimées, et les informations partielles sont conservées si au moins l'année est disponible. Les dates de naissance et d'épreuves sont également formatées dans un format commun afin d'uniformiser les données. Les distances des épreuves sont normalisées (toutes mises en mètres) afin de faciliter les comparaisons entre les différentes épreuves et différentes performances. Les temps de performance sont convertis en secondes pour toutes les épreuves, afin d'uniformiser les observations et donc faciliter l'utilisation future.

À la suite de ce nettoyage complet, nous calculons des variables supplémentaires afin d'enrichir le jeu de données pour des analyses ultérieures, comme l'âge au moment de l'épreuve, le groupe d'âge, et la vitesse moyenne lors de l'épreuve (en m/s).

De manière similaire, les données des hommes sont soumises au même processus de nettoyage pour garantir la qualité et la cohérence entre les différents jeux de données. Les mêmes étapes sont suivies pour normaliser les distances, formater les dates et convertir les temps en secondes.

Enfin, une fois que toutes les données ont été nettoyées et préparées, les meilleures performances sont extraites pour chaque âge entier afin de créer des jeux de données contenant uniquement les meilleures performances par âge, pour chaque épreuve.

En résumé, le processus de traitement des données que nous avons établi dans ce projet permet de garantir la qualité, la cohérence et la pertinence des données tout en supprimant le moins d'informations possible, fournissant ainsi une base solide pour nos analyses.

-

Présentation des modèles à ajuster

Dans ce rapport nous utiliserons trois modèles d'ajustements dans l'analyse de donnée. Chacun de ces modèles offre une approche différente pour la modélisation.

Le premier modèle d'ajustement suit une forme classique de régression linéaire, suivant l'équation $P(t) = at + b$ Cette forme linéaire simple est souvent utilisée pour modéliser des relations linéaires entre deux variables.

Le deuxième modèle améliore le premier en ajoutant des termes quadratiques, ce qui permet de modéliser des relations non linéaires entre deux variables. Son équation $P(t) = at^2 + bc + c$ nous donne la possibilité de modéliser des courbes en forme de paraboles, contrairement au modèle linéaire.

Enfin le troisième modèle, présente une approche plus complexe en combinant deux fonctions exponentielles. L'équation de Moore est la suivante $P(t) = a(1e^{-bt}) + c(1e^{dt})$ Entre contrepartie de sa complexité, ce modèle nous permet de modéliser la relation entre deux variables de façon plus flexible que les deux modèles précédents.

-

Détails sur les techniques d'ajustement

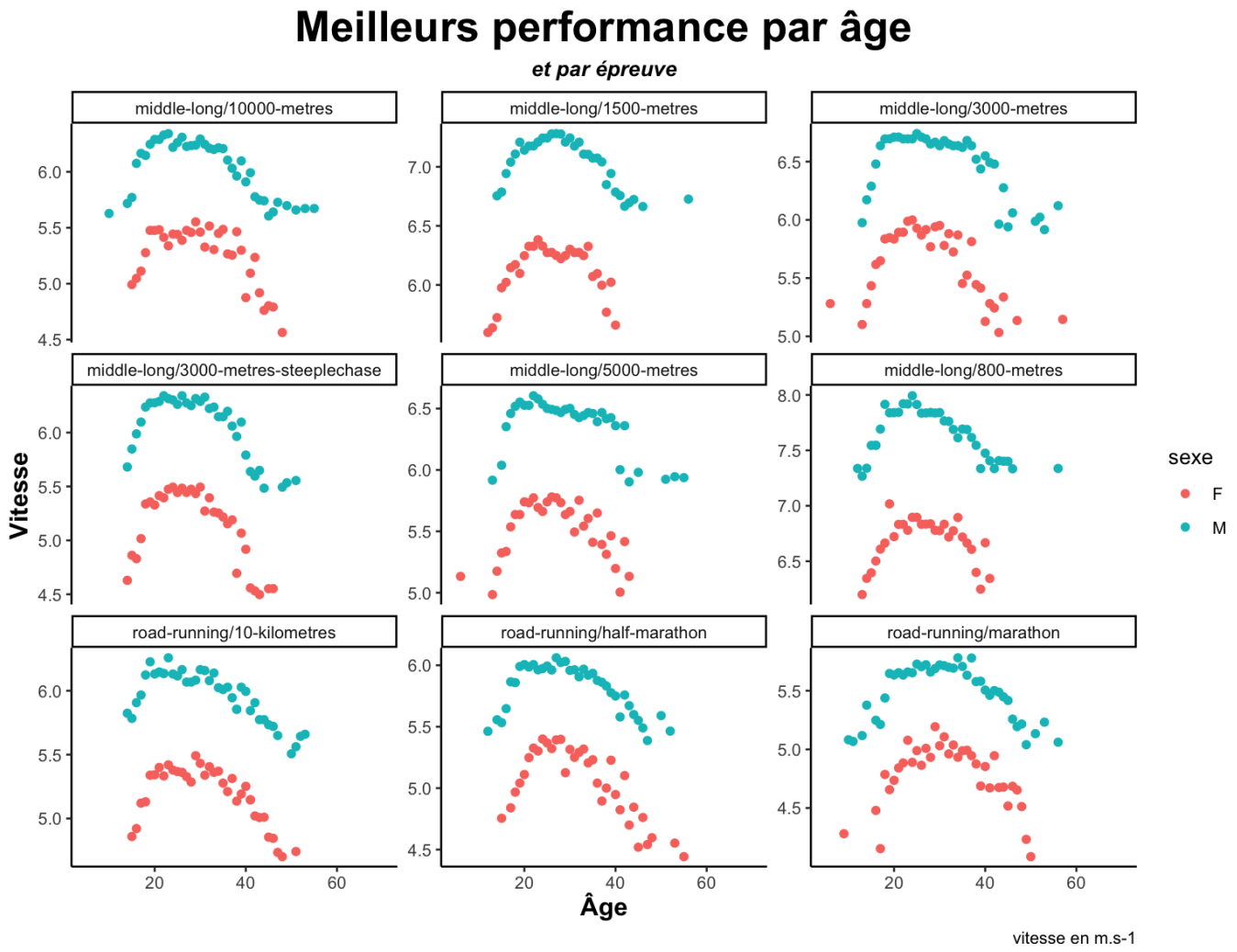
Pour l'ajustement des modèles sur nos données, compte tenu du nombre important de modélisation à faire (trois par épreuve, et par sexe), nous avons décidé d'écrire une fonction sur R. Les trois fonctions (une par modèle) suivent la même structure, seul le modèle et l'ajustement diffèrent. Les fonctions nous permettent de choisir différents arguments comme : le jeu de donnée, les deux variables sur lesquels on veut appliquer le modèle de régression, la liste des épreuves que l'on veut étudier, ainsi que plusieurs éléments esthétiques pour les graphiques, et la possibilité de sortir ou non un dataframe contenant des informations sur l'ajustement.

En sortie la fonction nous donne un graphique par épreuve, avec le nuage de points des données ainsi que la courbe de régression. Le dataframe optionnel contient une ligne par épreuve, avec les informations suivantes pour chaque : le nom de l'épreuve, les coefficients du modèle (2, 3 ou 4 selon le modèle), le R2 ajusté, l'AICc, le BIC, le maximum et l'objectif estimé par le modèle, le RSS, la moyenne et l'écart-type des résidus, ainsi que les résultats du test de normalité de ShapiroWilk.

RÉSULTATS

Présentation de la relation âge-performance

Dans cette partie, nous cherchons à voir si l'âge et la performance ont une relation dépendante ou pas pour chaque épreuve et par sexe. La performance est représentée en vitesse(m/s).



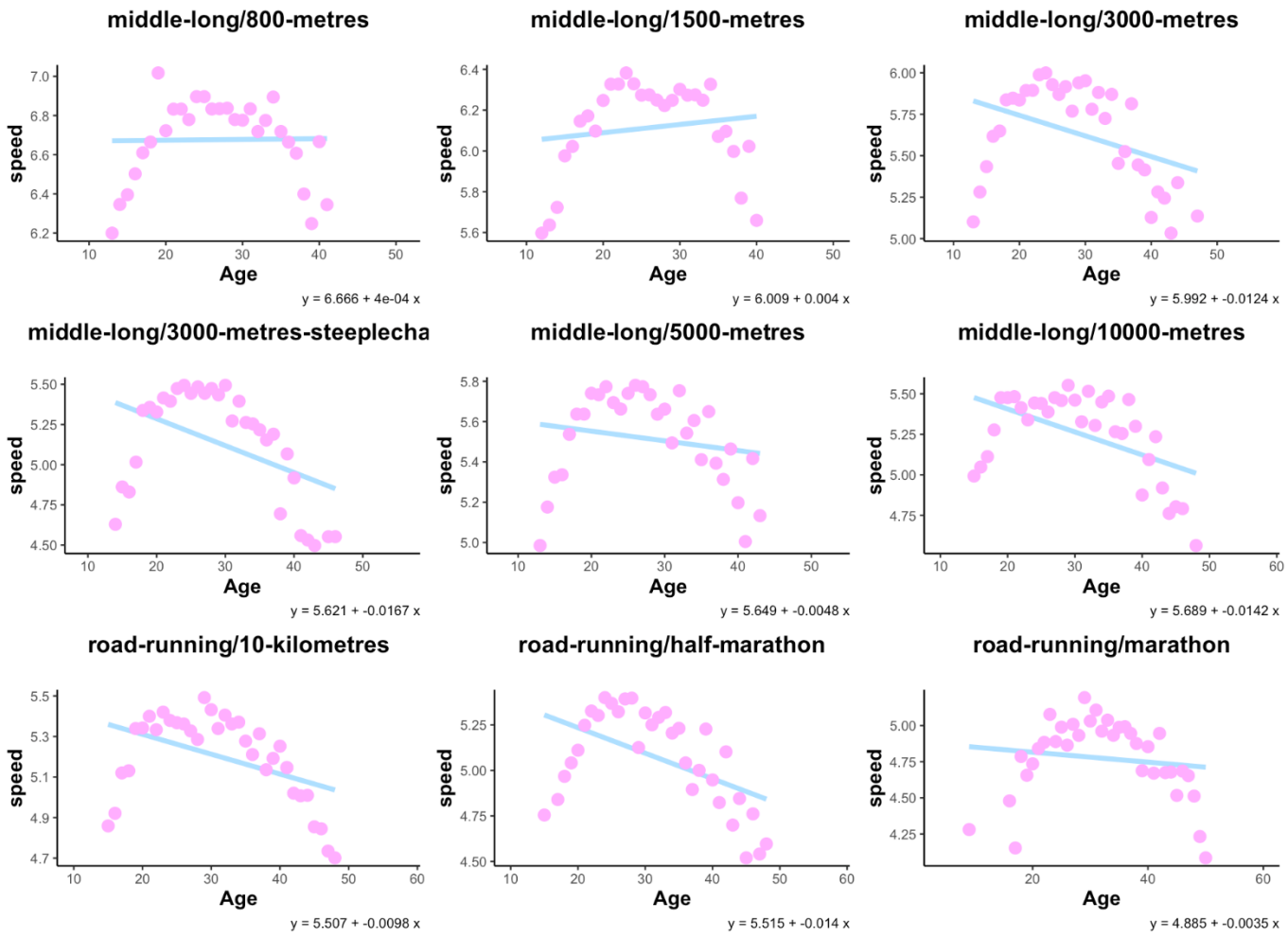
Ce graphique représente les meilleures performances des athlètes, par âge et par épreuves. La performance correspond ici à la vitesse en mètre par seconde. Les données sont séparées par sexe avec les femmes en rouge et les hommes en bleu.

On observe une forme de parabole dans les données, avec des performances qui augmentent jusqu'à atteindre un pic entre 25 et 35 ans dépendant des épreuves, puis déclinent progressivement. Cette tendance est observée tant chez les femmes que chez les hommes et peu importe l'épreuve, bien que la vitesse maximale ne sont pas les mêmes en fonction du sexe.

Résultats de l'ajustement des modèles

Nous allons maintenant pouvoir appliquer nos différents modèles sur nos deux jeux de données, nous avons choisi de réaliser nos analyses sur l'ensemble des épreuves sportives afin de voir si celle-ci à un facteur d'influence significatif sur les résultats. Chaque combinaison de jeu de donnée / modèle sera présenté sous forme de graphique comme celui-ci-dessous :

Modèle linéaire - Femmes

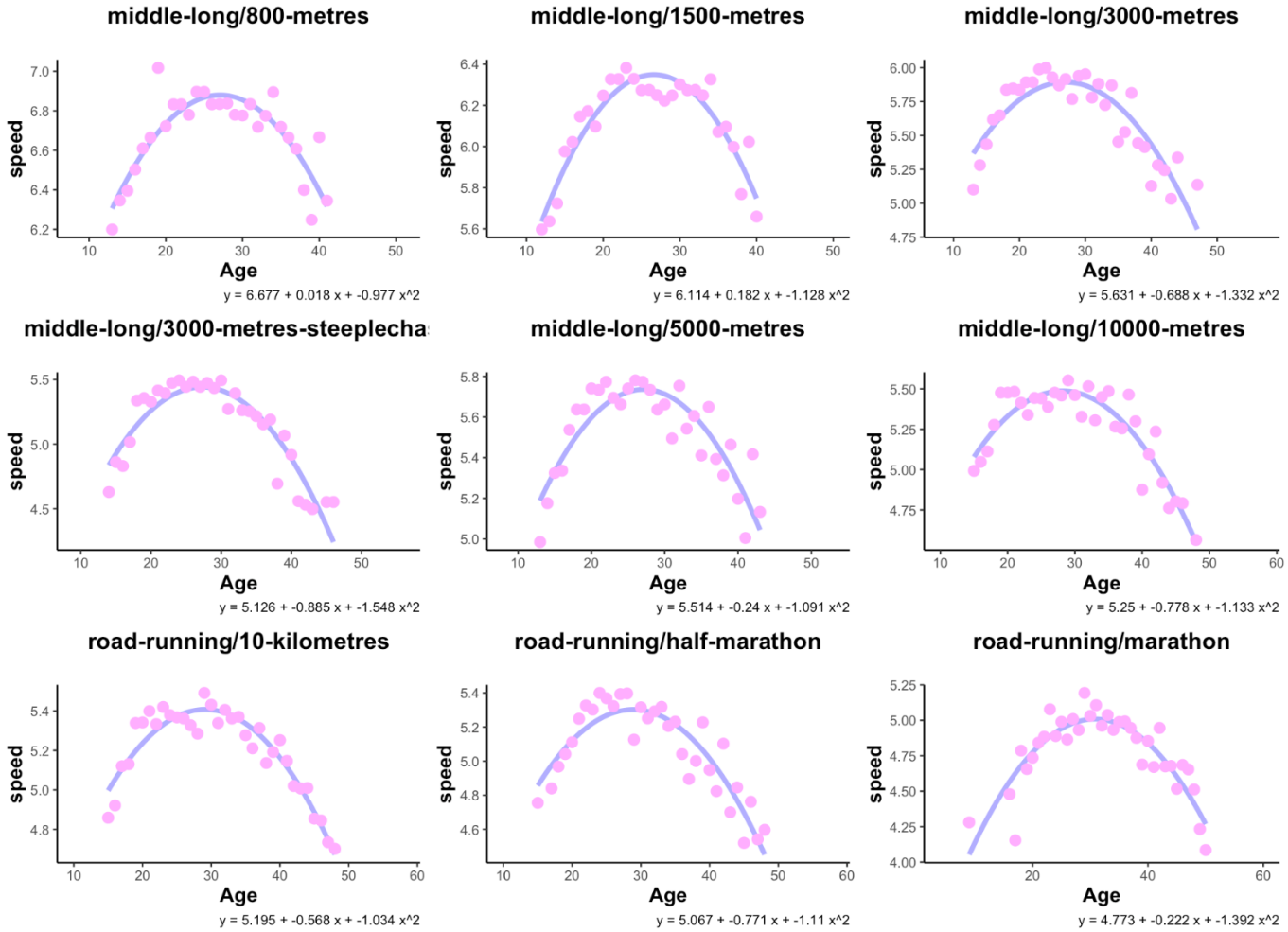


En appliquant le modèle de régression linéaire sur notre jeu de donnée des performances sportives féminines, on s'aperçoit très vite des limites de ce modèle, la droite tracée par ce modèle ne permet pas d'approcher correctement les données. Cela s'explique simplement du fait que la relation entre l'âge et la vitesse n'est pas une relation linéaire.

Nous n'allons donc pas appliquer ce modèle sur le reste des données car nous obtiendrons un résultat similaire.

Nous allons maintenant appliquer le modèle quadratique, qui devrait en principe être mieux ajuster à nos données car il est plus adapté pour modéliser des relations non-linéaire, ce qui est le cas de notre relation âge/performance comme nous avons pu le voir précédemment.

Modèle quadratique – Femmes



Cet ensemble de graphique nous montre l'ajustement du modèle quadratique sur les données âge/performance des femmes en fonction des différentes épreuves.

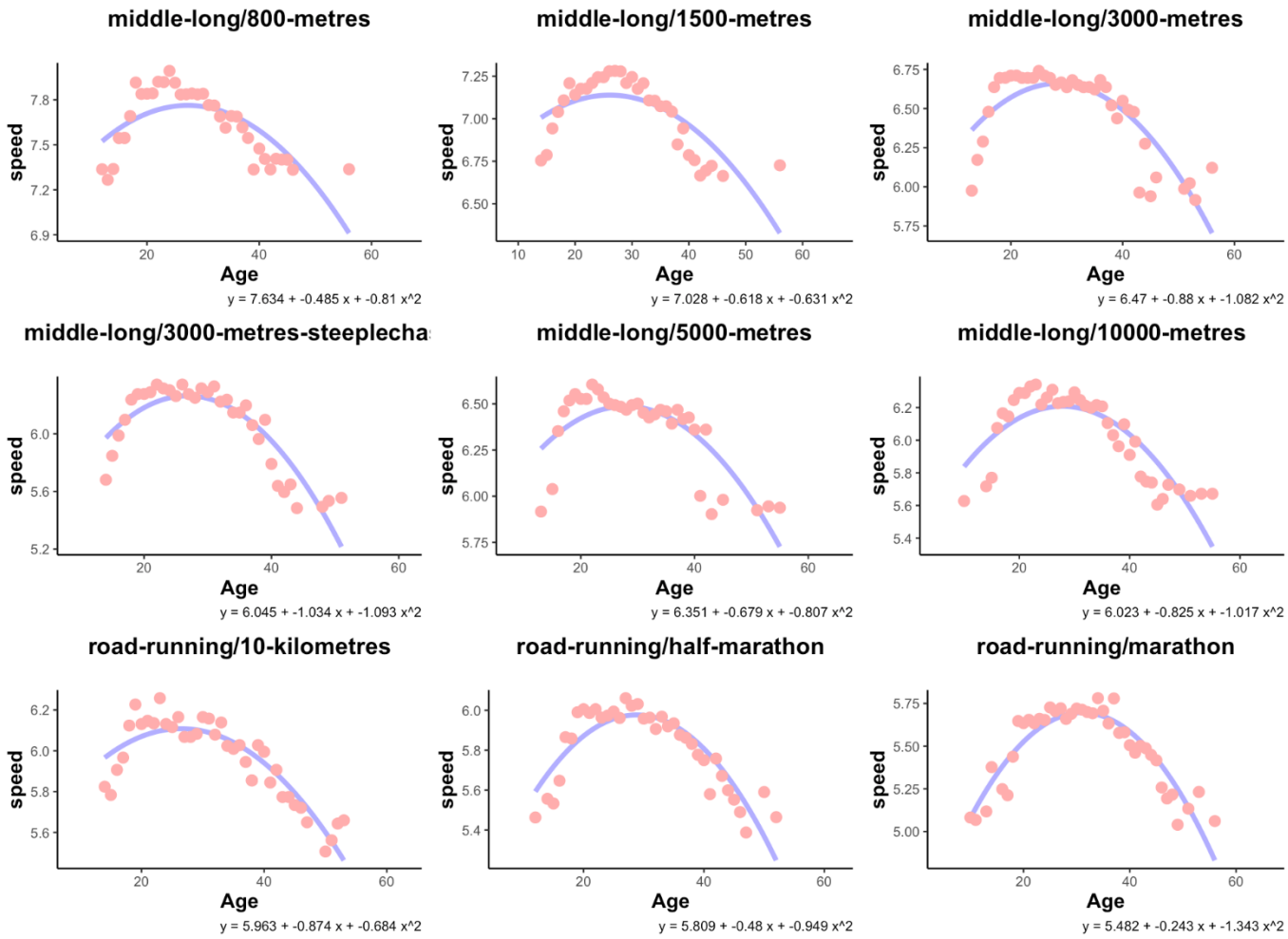
Les graphiques montrent une tendance en forme de parabole, croissante, puis décroissante à partir de l'âge de performance pic, généralement situé aux alentours de 28 ans. On voit que l'épreuve n'est pas un facteur déterminant de l'âge au pic de performance car toutes les épreuves ont approximativement le même.

L'épreuve est cependant un facteur influant sur la performance, on observe une vitesse moyenne de 5.5m/s, mais certaines épreuves comme le 800m, le 1500m, le 3000m et le 5000m ont des vitesses maximales plus élevées, allant jusqu'à environ 7m/s pour le 800m.

Ce modèle est bien plus adapté pour ajuster nos données en formes de parabole, que le modèle précédent qui été linéaire.

Nous allons maintenant appliquer ce même modèle quadratique sur les données des hommes.

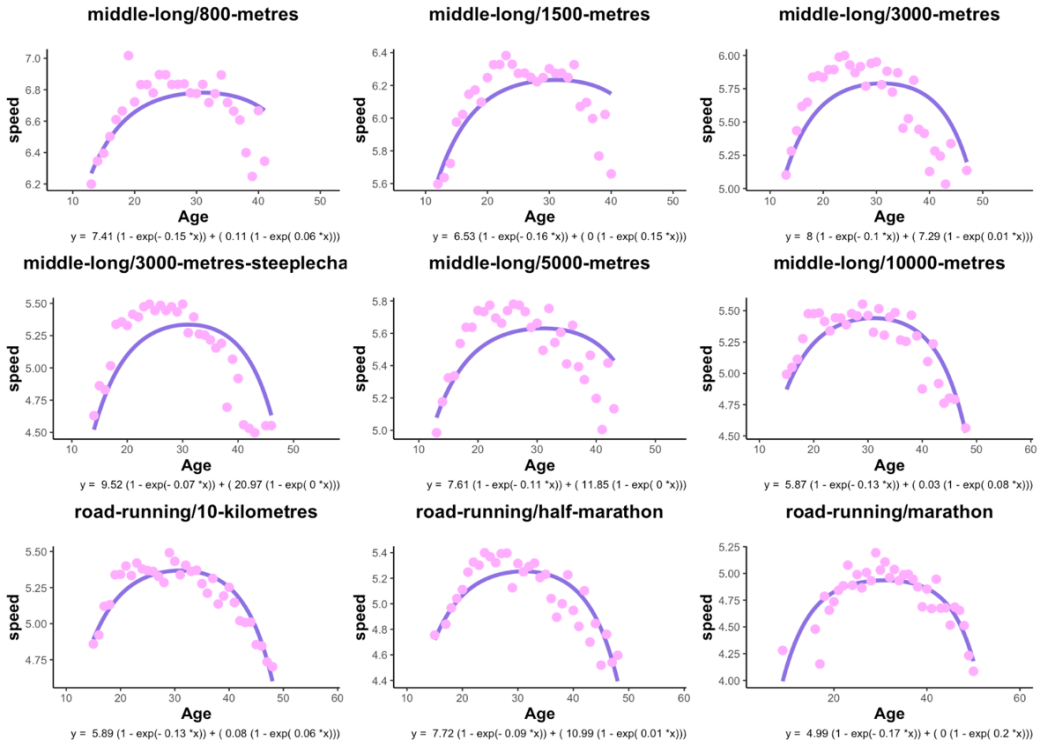
Modèle quadratique – Hommes



On voit que ce modèle fonctionne aussi bien sur les données des hommes. On observe la même tendance que pour les femmes, hormis que l'âge de performance maximale semble être atteint plus tôt que chez les femmes. On observe aussi que la vitesse moyenne est plus élevée, avec des performances maximales allant de 5.75m/s à 8m/s.

Nous allons maintenant utiliser le modèle de Moore. Ce modèle avec 4 paramètres est comme le modèle quadratique, bien plus adapté à la modélisation de nos données que le modèle linéaire simple.

Modèle de Moore – Femmes

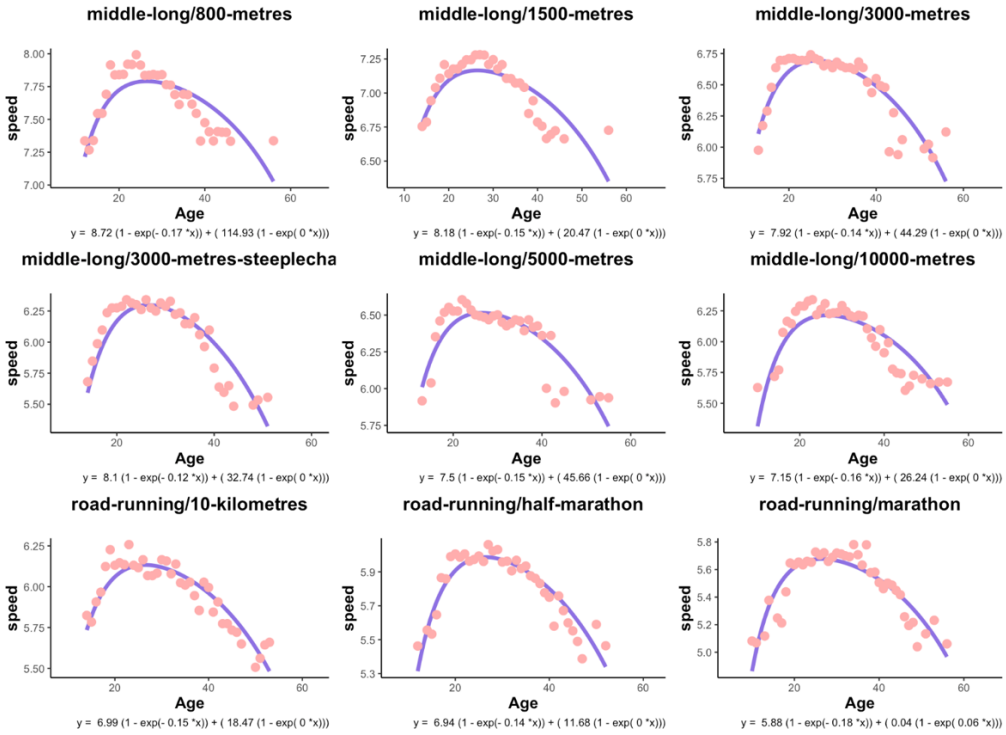


Cet ensemble de graphique nous montre l'ajustement du modèle de moore sur les données âge/performance des femmes en fonction des différentes épreuves.

Les graphiques montrent une tendance en forme de parabole, croissante, puis décroissante à partir de l'âge de performance pic, généralement situé aux alentours de 40 ans. On voit que l'épreuve n'est pas un facteur déterminant de l'âge au pic de performance car toutes les épreuves ont approximativement le même.

L'épreuve est cependant un facteur influant sur la performance, on observe une vitesse moyenne de 5.5m/s, mais certaines épreuves comme le 800m, le 1500m, le 3000m et le 5000m ont des vitesses maximales plus élevées, allant jusqu'à environ 7m/s pour le 800m.

Modèle de Moore – Hommes



On voit ici que le modèle de Moore ajuste plutôt bien nos données, avec une allure différente que le modèle quadratique, mais à l'œil nu il est difficile de définir lequel est le mieux ajusté à nos données.

Nous allons donc passer par une méthode plus précise que l'œil pour choisir entre les deux modèles.

ANALYSES

Comparaison des modèles

Afin de comparer les trois modèles, linéaire, quadratique et Moore, nous allons utiliser différents indicateurs comme le coefficient de détermination ajusté, l'AICc et le BIC.

Afin de travailler de manière efficace, nous avons créé une fonction par modèle. Les 3 fonctions sont identiques, seul le modèle diffère. Cet outil nous permet d'indiquer le jeu de données sur lequel on veut travailler, ainsi que certains paramètres comme par exemple une variable stratifiante (dans notre cas « épreuve »), et la fonction va en sortie nous fournir les graphiques avec le modèle ajusté, ainsi que la table de données avec pour chaque épreuve, les coefficients, ainsi que les indicateurs.

Voici la table de sortie de l'ajustement quadratique sur les données des femmes.

Epreuve	a	b	c	R2_ajuste	AICc	BIC	residus.sum	residus.sd	ShapiroWilk_normality.W	ShapiroWilk_normality.p.value
middle-long/800-metres	6.68	0.02	-0.98	0.72	-39.71	-34.24	-6.3e-14	0.11	0.88	0.00
middle-long/1500-metres	6.11	0.18	-1.13	0.86	-53.92	-48.45	2.1e-14	0.08	0.97	0.53
middle-long/3000-metres	5.63	-0.69	-1.33	0.75	-25.87	-19.88	-2.8e-14	0.15	0.98	0.87
middle-long/3000-metres-steeplechase	5.13	-0.88	-1.55	0.84	-29.78	-23.91	9.1e-14	0.14	0.98	0.73
middle-long/5000-metres	5.51	-0.24	-1.09	0.74	-38.31	-32.57	-2.4e-14	0.12	0.97	0.43
middle-long/10000-metres	5.25	-0.78	-1.13	0.83	-46.64	-40.65	-6.1e-14	0.11	0.98	0.64
road-running/10-kilometres	5.20	-0.57	-1.03	0.88	-78.95	-72.73	-2.8e-14	0.07	0.97	0.53
road-running/half-marathon	5.07	-0.77	-1.11	0.79	-40.72	-34.74	-3.9e-14	0.12	0.97	0.43
road-running/marathon	4.77	-0.22	-1.39	0.75	-36.44	-30.11	5.3e-14	0.13	0.92	0.01

Les valeurs de R^2 ajusté sont élevées, particulièrement pour les épreuves de 1500 mètres et 10 kilomètres. Avec une moyenne de 0.80 cela nous indique que le modèle quadratique explique bien la variance des performances.

Les valeurs AICc les plus basses sont observées pour les épreuves de 10 kilomètres (-78.9516) et de 1500 mètres (-53.9235), indiquant que ces modèles ont un bon ajustement avec une pénalisation pour la complexité du modèle.

Les valeurs BIC suivent une tendance similaire, confirmant que les modèles pour les épreuves de 10 kilomètres (-72.7302) et de 1500 mètres (-48.4543) sont parmi les meilleurs en termes de compromis entre ajustement et complexité.

Les p-values du test de Shapiro-Wilk montrent que pour certaines épreuves (par exemple, 800 mètres et marathon), les résidus ne suivent pas une distribution normale, ce qui suggère des possibles anomalies dans les données, ou des effets non capturés par le modèle quadratique.

Voici les résultats de l'ajustement quadratique sur les données des hommes.

Epreuve	a	b	c	R2_ajuste	AICc	BIC	residus.sum	residus.sd	ShapiroWilk_normality.W	ShapiroWilk_normality.p.value
middle-long/800-metres	7.63	-0.49	-0.81	0.48	-25.49	-19.05	-8.3e-14	0.16	0.98	0.79
middle-long/1500-metres	7.03	-0.62	-0.63	0.51	-27.79	-21.80	-9.8e-15	0.14	0.93	0.05
middle-long/3000-metres	6.47	-0.88	-1.08	0.66	-24.69	-18.14	-1.8e-14	0.16	0.87	0.00
middle-long/3000-metres-steeplechase	6.04	-1.03	-1.09	0.77	-32.03	-25.93	3.9e-14	0.14	0.92	0.02
middle-long/5000-metres	6.35	-0.68	-0.81	0.61	-30.69	-24.58	-9.3e-14	0.14	0.83	0.00
middle-long/10000-metres	6.02	-0.82	-1.02	0.68	-36.40	-29.74	6.5e-14	0.14	0.95	0.08
road-running/10-kilometres	5.96	-0.87	-0.68	0.79	-75.59	-68.74	-5.1e-14	0.09	0.98	0.77
road-running/half-marathon	5.81	-0.48	-0.95	0.75	-62.11	-55.56	-1.0e-13	0.10	0.99	0.88
road-running/marathon	5.48	-0.24	-1.34	0.82	-66.93	-60.08	6.9e-14	0.10	0.98	0.80

Les valeurs de R^2 ajusté sont correct, mais avec une moyenne de 0.67, plus basse que sur les femmes. On voit que pour l'épreuve 800m le R^2 ajusté est particulièrement bas (0.48) ce qui peut indiquer que le modèle quadratique n'est pas particulièrement bon pour ajuster les données de cette épreuve, qui à peut-être des particularités que les autres épreuves n'ont pas.

Concernant l'AICc et le BIC, les valeurs restent très bonnes, mais légèrement plus grande que pour les femmes. On peut donc conclure que le modèle quadratique ajuste mieux les données des femmes.

-

Nous allons maintenant regarder les mêmes critères, mais pour le modèle de Moore.

Voici la table du modèle de Moore pour les données des femmes :

Epreuve	a	b	c	d	R2_ajuste	AICc	BIC	estimated_maximum	estimated_objective	residus.mean	residus.sum	ShapiroWilk_normality.W	ShapiroWilk_normality.p.value
middle-long/800-metres	7.41	0.15	0.11	0.06	0.78	-46.52	-42.41	25.27	6.87	-1.7e-16	0.10	0.87	0.00
middle-long/1500-metres	6.53	0.16	0.00	0.15	0.89	-61.60	-57.50	26.29	6.33	-5.6e-17	0.08	0.97	0.49
middle-long/3000-metres	8.00	0.10	7.29	0.01	0.86	-45.40	-40.91	24.74	5.94	5.6e-17	0.11	0.97	0.51
middle-long/3000-metres-steeplechase	9.52	0.07	20.97	0.00	0.90	-45.60	-41.20	25.76	5.47	2.8e-16	0.11	0.97	0.54
middle-long/5000-metres	7.61	0.11	11.85	0.00	0.83	-52.88	-48.58	24.82	5.75	1.2e-16	0.10	0.98	0.81
middle-long/10000-metres	5.87	0.13	0.03	0.08	0.84	-50.65	-46.17	27.06	5.49	1.2e-16	0.10	0.98	0.69
road-running/10-kilometres	5.89	0.13	0.08	0.06	0.92	-91.87	-87.21	27.58	5.41	5.9e-17	0.06	0.99	0.95
road-running/half-marathon	7.72	0.09	10.99	0.01	0.85	-51.76	-47.27	27.07	5.33	-8.3e-17	0.10	0.97	0.58
road-running/marathon	4.99	0.17	0.00	0.20	0.67	-27.60	-22.85	31.04	4.95	5.7e-16	0.15	0.91	0.01

Les valeurs de R^2 ajusté varient de 0.67 à 0.92, indiquant que le modèle de Moore explique bien la variance des données, surtout pour les épreuves de 10 kilomètres (0.92) et de 3000 mètres steeplechase (0.90). Cela montre que ce modèle est particulièrement efficace pour ces épreuves. Le test de Shapiro-Wilk indique que pour certaines épreuves, les résidus ne suivent pas une distribution normale, en particulier pour le 800 mètres (p-value = 0.00) et le marathon (p-value = 0.01). Cela suggère que le modèle peut encore être amélioré pour ces épreuves, malgré son bon ajustement global.

Afin de déterminer lequel du modèle quadratique et du modèle de Moore ajuste le mieux nos données, nous regardons les mêmes indicateurs que précédemment.

Le modèle de Moore montre des R^2 ajustés généralement plus élevés par rapport au modèle quadratique, indiquant une meilleure capacité à expliquer la variance des données. Pour les hommes le R^2 ajustés du modèle de Moore dépasse systématiquement le modèle quadratique.

Les valeurs AICc et BIC sont plus basses pour le modèle de Moore, en particulier pour les épreuves de 10 kilomètres et de 1500 mètres, ce qui nous montre une meilleure adéquation du modèle de Moore malgré sa complexité. En revanche pour les hommes, les valeurs AICc et BIC confirment que le modèle de Moore est supérieur au modèle quadratique, avec les différences les plus marquées dans les épreuves de 10 kilomètres et de half-marathon.

Et pour finir nos comparaisons, les tests de Shapiro-Wilk montrent que les résidus du modèle de Moore sont plus proches d'une distribution normale comparé au modèle quadratique, suggérant un meilleur ajustement. Même si pour certaines épreuves des hommes comme le 1500 mètres et le marathon, des améliorations pourraient encore être nécessaires afin d’obtenir un meilleur ajustement.

Globalement, le modèle de Moore surpasse le modèle quadratique en termes de R^2 ajusté, AICc, et de BIC, pour les hommes et les femmes dans la plupart des épreuves d'athlétisme. Cela suggère que le modèle de Moore est plus efficace pour capturer la relation entre l'âge et la performance sportive ici représenté par la vitesse. Les meilleures valeurs des critères d'information (AICc et BIC) pour le modèle de Moore indiquent qu'il offre un meilleur compromis entre précision et complexité, malgré le fait qu'il nécessite plus de paramètres à ajuster.

Cependant, des améliorations peuvent encore être envisagées, surtout pour assurer la normalité des résidus dans certaines épreuves spécifiques. Dans l'ensemble, le modèle de Moore semble plus adapté pour l’analyse de la relation des performances sportives en fonction de l'âge.

Pic de performance

Les valeurs maximales estimées et les objectifs estimés fournissent des informations cruciales sur les performances optimales attendues pour différentes épreuves d'athlétisme en fonction de l'âge. Nous allons analyser ces valeurs pour les hommes et les femmes, en utilisant les résultats obtenus avec le modèle de Moore.

Les valeurs maximales estimées représentent l'âge auquel les athlètes atteignent leur performance optimale pour chaque épreuve.

Courtes distances :
Les athlètes féminines atteignent leur pic de performance pour les courtes distances au milieu de la vingtaine (25.27 ans). Les hommes atteignent leur pic de performance pour les courtes distances à un âge plus jeune que les femmes (23.35 ans), ce qui peut être dû à des différences physiologiques, notamment une plus grande capacité musculaire et une récupération plus rapide.

Moyennes distances :
Les performances maximales pour les moyennes distances pour les femmes montrent un pic légèrement plus tardif que pour les courtes distances 26.29 ans (1500 mètres) et 24.74 ans (3000 mètres), probablement en raison de la nécessité d'une plus grande endurance. Les pics de performance pour les moyennes distances chez les hommes sont légèrement plus tardifs que pour les courtes distances 24.85 ans (1500 mètres) et 24.87 ans (3000 mètres), mais reste plus tôt que chez les femmes pour les mêmes épreuves.

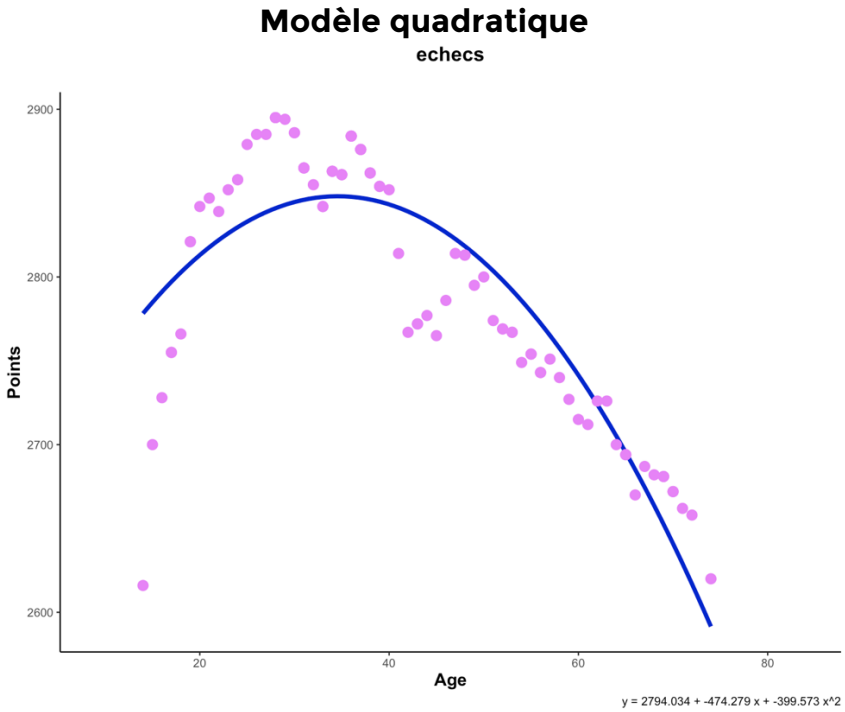
Longues distances :
Pour les longues distances, les femmes atteignent leur pic de performance plus tard 27.21 ans (10 kilomètres), 27.07 ans (half-marathon), 31.04 ans (marathon), avec le marathon atteignant un sommet encore plus tardif. Les hommes atteignent leur pic de performance pour les longues distances plus tôt que les femmes 24.75 ans (10 kilomètres), 24.84 ans (half-marathon), 26.45 ans (marathon), et avec une variation moins prononcée entre les différentes distances. Cela peut être lié à une capacité biologique/naturelle plus élevée à maintenir des performances élevées sur de plus longues périodes.

La comparaison des valeurs maximales estimées et des objectifs estimés entre les hommes et les femmes révèle des différences significatives dans les âges de performance maximale et les performances optimales pour diverses épreuves d'athlétisme. Les hommes tendent à atteindre leur pic de performance à un âge plus jeune que les femmes pour la plupart des distances, et leurs objectifs estimés sont généralement plus élevés, reflétant des différences physiologiques et de conditionnement.

Nouveau jeu de données

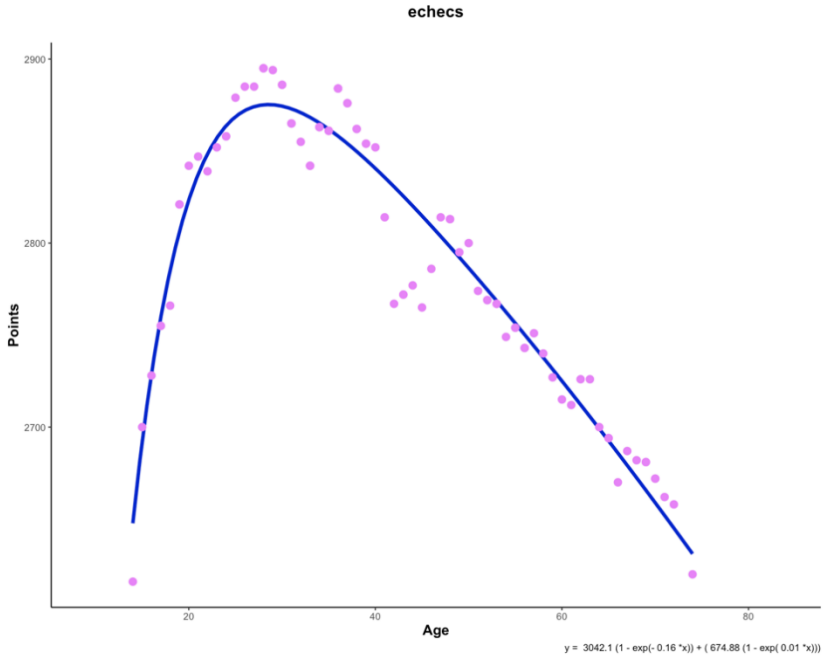
Nous allons maintenant introduire un nouveau jeu de données afin de voir comment se comportent nos modèles sur d'autres données. Ce nouveau jeu de données contient des performances aux échecs, mesurées en scores. Comme pour les données d'athlétisme, nous allons effectuer une analyse de la relation performance/âge.

Nous allons uniquement travailler avec le modèle quadratique et le modèle de Moore, car comme nous l'avons vu précédemment, le modèle linéaire n'est pas du tout adapté à ce type de relation. Les graphiques ne seront pas directement commentés, mais plutôt analysés et comparés dans l'analyse.



Ce graphique illustre l'ajustement d'un modèle quadratique appliqué aux scores de performance aux échecs en fonction de l'âge des joueurs. Les graphiques révèlent une tendance parabolique : les scores augmentent jusqu'à atteindre un sommet vers l'âge de 28 ans, avec un score de 28 500 points, puis diminuent progressivement. Cela indique que l'âge est un facteur crucial influençant les performances, avec un pic de performance généralement situé autour de 28 ans. Au-delà de cet âge, les scores tendent à diminuer à mesure que l'âge avance.

Modèle de Moore



Ce graphique illustre l'ajustement du modèle de Moore aux scores de performance aux échecs en fonction de l'âge des joueurs. Les résultats montrent une tendance en forme de parabole : les scores augmentent jusqu'à atteindre un pic vers l'âge de 28 ans, avec un score de 28700 points, puis diminuent progressivement. Cela souligne que l'âge joue un rôle déterminant dans les performances, le sommet étant généralement atteint autour de 28 ans. Au-delà de cet âge, les scores diminuent progressivement à mesure que l'âge avance.

CONCLUSION

Principales découvertes

- Tendance parabolique : Les performances athlétiques et échiquéennes montrent une tendance parabolique par rapport à l'âge, avec un pic de performance généralement autour de 28 ans.

- Modèle de Moore : Ce modèle s'est révélé le plus adéquat pour modéliser la relation âge-performance, surpassant les modèles linéaire et quadratique en termes de précision et de fiabilité.

- Différences de sexe et de discipline : Bien que le pic de performance varie légèrement entre les sexes et les disciplines, l'âge optimal reste relativement constant.

L'étude confirme l'importance de l'âge dans la performance sportive et échiquéenne, avec des implications pratiques pour l'entraînement et la gestion des carrières. Le modèle de Moore s'avère être un outil puissant pour l'analyse de telles relations non linéaires, offrant des perspectives précieuses pour la planification stratégique dans le domaine du sport et au-delà.

Ces découvertes peuvent aider les entraîneurs, les athlètes et les joueurs d'échecs à optimiser leurs stratégies de formation et de compétition en fonction des âges de pic de performance. De plus, ces insights peuvent être appliqués à d'autres domaines nécessitant une gestion optimale de la performance en fonction de l'âge.

ANNEXE

Code R

Lors de ce projet, après avoir nettoyé les différents jeux de données fournis, nous nous sommes très vite rendu compte du temps que prendrait l'ajustement des trois modèles sur les 19 sous-ensembles de données. Afin de réaliser ces 57 ajustements, nous avons décidé de développer une fonction qui nous permettrait de faire l'ajustement sur tous les sous-ensembles bien plus rapidement.

Cette fonction prend en entrée :
- Le modèle (formule à suivre pour l'ajustement)
- Une base de données
- La variable stratifiante (pour créer les sous-ensembles de la base)
- Un vecteur contenant les modalités de cette variable stratifiante (ce qui nous permet de choisir un ou plusieurs sous-ensembles spécifiques)

Il y a aussi des arguments binaires qui nous permettent de choisir ce que l'on veut en sortie. On peut choisir de générer les graphiques des ajustements, mais aussi d'obtenir un dataframe dans l'environnement R qui contiendra pour chaque sous-ensemble d'ajustement les différents coefficients et les indicateurs statistiques tels que le R^2 ajusté, le BIC, l'AICc, etc.

De plus, il y a des arguments esthétiques que l'on peut renseigner afin de modifier les graphiques générés.

Le développement de cette fonction nous a pris un certain temps car nous n'avions jamais fait ce type de fonction auparavant, donc nous avons dû apprendre comment faire. Mais finalement, cette fonction nous a fait gagner énormément de temps. En plus d'avoir appris de nouvelles compétences en programmation, cela nous a été grandement utile pour ce projet et nous a permis de prendre notre temps pour le reste du projet.

Sur la page suivante, vous trouverez le code R de cette fonction, si cela vous intéresse.

```r
graph_moore <- function(dataset, epreuve,
                        x.axis, x.label, y.axis,
                        printTable=TRUE, valueTable = F,
                        ncol=1, point.size=3.6, point.col="#E683FG", point.shape = 16, line.size=1.5, line.col="blue", table_name="result") {
  if (!is.vector(epreuve)) {
    epreuve <- as.vector(epreuve)
  }
  plots <- list()
  results <- data.frame(Epreuve = character(),
                        a = numeric(), b = numeric(), c = numeric(), d = numeric(),
                        R2_ajuste = numeric(), AICc = numeric(), BIC = numeric(),
                        estimated_maximum = numeric(), estimated_objective = numeric(),
                        residus.mean = numeric(), residus.sum = numeric(),
                        ShapiroWilk_normality.W = numeric(), ShapiroWilk_normality.p.value = numeric()
                        )

  for (ep in epreuve) {
    data <- dataset %>%
      filter(epreuve == ep)
      if (printTable == TRUE) {
      print(data)
    }

    coef <- MMC(data[[x.axis]], data[[y.axis]], methode = modele_moore, nbpara = 4, precision=100, borne = 0, initial = c(max(data[[y.axis]]), 0.1554491, 0.40300051, 0.025860582))
    model <- lm(data[[y.axis]] ~ modele_moore(data[[x.axis]], coef$parametre))

    residus <- residuals(model)
    residus.sum <- sum(residus)
    residus.sd <- sd(residus)
    ShapiroWilk_normality.W <- shapiro.test(residus)$statistic
    ShapiroWilk_normality.p.value <- shapiro.test(residus)$p.value
    R2.adjusted <- adjR2(model)
    AICc <- AIC(model)
    BIC <- BIC(model)

    optimize_result <- optimize(function(x) modele_moore(x, coef$parametre), interval = c(0, 100), maximum = TRUE)
    estimated_maximum <- optimize_result$maximum
    estimated_objective <- optimize_result$objective

    results[nrow(results) + 1, ] <- list(ep, coef$parametre[1], coef$parametre[2], coef$parametre[3], coef$parametre[4], R2.adjusted, AICc, BIC, estimated_maximum, estimated_objective, residus.sum,
residus.sd, ShapiroWilk_normality.W, ShapiroWilk_normality.p.value)

    plot <- ggplot(data) +
      aes_string(x = x.axis, y = y.axis) +
      geom_smooth(method = "lm", formula = y ~ modele_moore(x, coef$parametre), se = FALSE, color = line.col, size = line.size) +
      geom_point(shape = point.shape, size = point.size, colour = point.col) +
      labs(x = x.label, y = y.axis, title = ep, subtitle = "",
        caption = paste("y = ", round(coef$parametre[1], 2),"(1 - exp(-", round(coef$parametre[2], 2),"*x)) + (", round(coef$parametre[3], 2), "(1 - exp(", round(coef$parametre[4], 2), "*x)))")) +
      theme_classic() +
      theme(
        plot.title = element_text(size = 15L, face = "bold", hjust = 0.5),
        plot.subtitle = element_text(face = "bold.italic", hjust = 0.5, size = 10),
        axis.title.y = element_text(size = 13L, face = "bold"),
        axis.title.x = element_text(size = 13L, face = "bold")
      ) +
      xlim(min(data[[x.axis]])-5, max(data[[x.axis]])+10)
    plots[[length(plots) + 1]] <- plot
  }
  grid.arrange(grobs = plots, ncol = ncol)
  if (valueTable == T) {
    assign(table_name, results, envir = .GlobalEnv)
  }
}
```

nota bene : la fonction MMC() que l'on appelle dans notre fonction nous permet d'estimer les différents paramètre du modèle sur plusieurs itération, et de ressortir les paramètres du meilleur ajustement trouvé.