Journal of the Royal Statistical Society Series C: Applied Statistics, 2023, **00**, 1–16
https://doi.org/10.1093/jrsssc/qlad065

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Original Article

A flexible mixed model for age-dependent performance: application to golf

Rose D. Baker[1] and Ian G. McHale[2]

[1]Salford Business School, University of Salford, Greater Manchester, M5 4WT, UK
[2]University of Liverpool Management School, University of Liverpool, Liverpool, L69 7ZH, UK

Address for correspondence: Ian G. McHale, University of Liverpool Management School, University of Liverpool, Liverpool, L69 7ZH, UK.
Email: ian.mchale@liverpool.ac.uk

Abstract

We present a new mixed linear model for the relationship between age and performance. The model allows for random effects at the nodes of a barycentric interpolation, such that performance evolves with age in a non-prescriptive way. We use the model to investigate the effects of age on performance in golf and find that performance peaks in the 30s and then declines after that. We disaggregate performance into its constituent components and find that driving, which tends to require power and speed, deteriorates consistently from the early 20s, whilst putting, which requires touch and finesse, remains strong until the late 40s. Our model can be used in other settings, and requires only that measures of performance exist.

Keywords: ageing, barycentric rational interpolation, mixed-effects model, optimum age, statistics in sports

1 Introduction

It is natural for humans to wonder about the effects of ageing. Understanding, or knowing, the age at which we are 'at our best' can, for those of us past that age, be somewhat sobering. However, for those performing at elite levels in their field of work, understanding when peak performance occurs can help recruitment (e.g. of athletes to sports teams), manage expectations (e.g. of productivity levels of individuals), and the setting of targets (e.g. publications of scientific papers). As such, in sports and other elite fields, understanding peak age of performance, and measuring the effects of age on level of performance, is more than simply an intriguing problem. Since the first look at this problem in Lehman (1945), researchers have considered age of peak performance in fields from conducting Nobel Prize winning research to 100 m sprinting. Here, we present a new flexible mixed effects model that can be used to estimate the relationship between age and performance in any field and we use it to study ageing in golf.

Much of the literature on age and performance has been descriptive in nature. Various authors in the sports science community have tabulated ages of achievement in, for example, setting records, and winning competitions. Schulz and Curnow (1988) look at several sports. Tabulating the ages of Olympic gold medallists in a number of disciplines, they find that the average age of winners of the 100 m sprint is 22.85 years, and that the age increases monotonically with race distance to a peak of 27.85 years for marathon runners. They perform similar analyses for swimming, baseball and indeed golf. In golf they find that the average age of a golfer who reaches number one in the world rankings is 33.67, though they find that the age is decreasing.

Koensigsberg et al. (2020) focuses on the age of peak performance of golfers and looks at the mean age of golfers reaching the top 100 in the world rankings. The descriptive statistics suggest that this age is now 23.91, but it has decreased significantly from 33.29 years a few decades ago.

Received: November 7, 2022. **Revised:** March 23, 2023. **Accepted:** June 26, 2023

Despite presenting clear complexities such that understanding the relationship between age and performance requires statistical modelling, there is surprisingly little in the scientific literature on the topic. Roring and Charness (2007) present a multi-level model for chess rankings as a function of age and find that age is kinder (a slower rate of decay in performance) for those with initially higher levels of expertise. They calculate a confidence interval for the peak age of performance being between 39 and 49. Taken together with findings from other activities, it seems clear that performance in activities requiring speed and power peaks (much) earlier than in more cerebral activities such as chess.

Unlike Jones (2010), in this paper we are not concerned with the changing age of peak performance. Instead, we build a model to estimate the age of peak performance, and subsequently estimate the magnitude of the effects of age.

To do so, we present a new type of random-effects model. We allow for random effects at the nodes of a barycentric rational interpolation. This allows performance to evolve with age in a non-prescriptive way and means that as well as estimating a mean relationship between age and performance, each golfer's performance is allowed to vary from the mean through random fluctuations in performance at the nodal ages.

Golf is a particularly attractive laboratory in which to test our model as it requires several distinct skill sets to perform at the elite level. Speed and power are required for hitting the ball a long way (driving), whilst touch and finesse are essential for short shots and putting. Examining golf, and disaggregating performance across its various component skills, allows a unique insight into how individuals are affected by age across their personal skill sets.

The paper is structured as follows. Next, in Section 2, we present the dataset we use to model performance as a function of age. In Section 3, we describe the model we use. In Section 4, we present some properties of the model, including how to use it to forecast an individual's expected future performance. Section 5 provides our results and findings before we comment on other fields in which the model can be used in Section 6. We close the paper with some concluding remarks in Section 7.

2 Data

We obtained data for golfers on the Professional Golfers' Association of America (PGA) Tour from 2004 to 2021. The data were scraped from the PGA Tour website which includes round-by-round scores for players at all PGA Tour tournaments (including the tournament date and the course name).

It seems obvious, at first, that one should use a player's score in a round as a measure of the performance of that player. A player who has a lower score (as he has taken fewer shots to complete the round) has performed better than a player who has a higher score (who has taken more shots to complete the round). However, not all golfers play in the same tournaments or on the same courses. Some golf courses are much more difficult than others. Further, the same golf course can play much harder due to a number of factors. For example, the weather may make the course harder, or the grass can be grown in certain areas to make it more difficult, or the pin positions can be placed in places on the greens that are hard to access. As such, using the round scores to measure and monitor performance is fraught with problems that need to be taken account of.

Fortunately, Broadie (2012) presented an intuitive solution to measuring performance in golf whilst taking account of the complexities described above. The basic idea of Broadie's *strokes gained* is, for each shot, to compare the average number of shots expected to be required to finish the hole from the starting position, to the average number of shots expected to be required to finish the hole from the end position. For example, suppose a player is to take his first shot on a hole measuring 450 yards. The average number of shots required to complete a hole, from a tee box 450 yards away is, say, 4.2, based on historic data from the PGA Tour. Now suppose the player takes his tee shot and hits it 350 yards down the middle of the fairway such that he has 100 yards remaining. From this new position, the average number of shots required is 2.8 (say). The player's strokes gained from his tee shot is thus $4.2 - 2.8 - 1 = 0.4$. Intuitively, the player has gained 0.4 of a shot by hitting a good (long and straight) tee shot. Strokes gained are commonly calculated in five categories, accumulated over a round (or a tournament):

Table 1. Descriptive statistics for round scores and strokes-gained metrics

Variable	Mean	Median	St. Dev.
core	70.13	70	1.81
SG_{total}	0.758	0.796	1.356
SG_{driving}	0.032	0.04	0.153
SG_{app}	0.068	0.073	0.214
SG_{atg}	0.027	0.028	0.141
SG_{putt}	0.062	0.063	0.217

Note. St. Dev. = standard deviation.

- strokes gained off-the-tee (driving) (SG_{driving}): the sum of strokes gained for tee shots on par 4 and par 5 holes;
- strokes-gained around-the-green (SG_{atg}): the sum of strokes gained for all shots within 30 yards of the putting surface;
- strokes-gained putting (SG_{putt}): the sum of strokes gained for all shots played from a green (all putts);
- strokes-gained approach (SG_{app}): the sum of strokes gained for all strokes not included in 'off-the-tee', 'putting', or 'around-the-green'. Strokes-gained approach includes tee shots on par 3 holes.
- total strokes gained (SG_{total}): the sum of strokes gained for all shots. Note that strokes gained is additive such that $SG_{\text{total}} = SG_{\text{driving}} + SG_{\text{app}} + SG_{\text{atg}} + SG_{\text{putt}}$.

Each of these strokes-gained metrics attempts to quantify performance in a different area of 'skill' as each category requires different skill sets. For example, off-the-tee requires power and speed to hit the ball long and straight. Putting, on the other hand, requires touch and finesse to guide the ball to its target.

As for the round-by-round scores of players in tournaments, the PGA Tour website now offers round-by-round strokes-gained data from 2004, and we scraped strokes-gained data for golfers in each tournament.

In addition to scoring and performance data on golfers, we obtained date of birth, height, weight, nationality, and year turned professional for all golfers who had played more than five tournaments. In total, we had complete information on 944 golfers, playing 168,296 rounds in 597 tournaments. Table 1 provides descriptive statistics for the round scores, and strokes-gained metrics.

3 A model for estimating the variation in golfers' performance with age

3.1 Why fit a parametric model?

Before describing our model, we can ask why one should fit a parametric model at all rather than calculating descriptive statistics as has been done many times in the literature. For example, one could compute mean scores and standard errors on the mean for (say) yearly age-bands. Such a non-parametric 'age-band' approach is extremely simple, but has three disadvantages in comparison to building a mathematical model. First, there is a secular trend in scores, such that round scores decrease by about two strokes over the 18-year period of the data. There is also a slight drop in average age by about 3.5 years over the period. Only a parametric model can disentangle age-based and secular trend effects. Second, non-parametric results would need smoothing to be interpretable, which occurs automatically in a parametric model. Finally, a parametric model allows the performance of individual golfers to be modelled and forecast far into the future, e.g. to age 60, which is of interest. The non-parametric approach cannot do this and so one must adopt a parametric model.

3.2 Modelling choices

We have several years of performance data for each golfer, y_{it} indexed by golfer i and match number t, occurring at calendar time s_{it}. The variable y_{it} can be one of the five strokes-gained metrics or any combination of them, or total score for a tournament. Of course, the model can be applied outside of golf and y_{it} can then be any measure of performance the user thinks appropriate in the field of study.

Two modelling choices are how to flexibly model individual variations of performance with age, and whether to use fixed or random effects.

To model how a golfer's performance evolves with age, we interpolate performance between a series of 'nodes' at regularly spaced ages. This could be done with spline-interpolation where the nodes correspond to the spline knots. However, we use barycentric interpolation, because it is simpler to describe and implement, and has no discontinuities in derivatives. This means that a function minimiser has a smooth function to minimise. Also, the methodology has been shown to interpolate at least as accurately as a spline (Baker & Jackson, 2014).

Barycentric intepolation was introduced by Berrut (1988) and further developed by Floater and Hormann (2007). The interpolant is the ratio of two polynomials. Baker and Jackson (2014) discuss its use in statistical applications. Baker and McHale (2015a) use barycentric interpolation to model the evolution of the strength of golfers, whilst Baker and McHale (2015b) use it to compare the strengths of football teams from different eras. To date, it has been little used by statisticians, despite its attractive properties.

We use fixed effects to allow performance to respond to a golfer's height (measured in cm from the mean height of all golfers), and a variable representing the date of play to allow for secular (temporal) trend in y_{it}.

Golfer-specific effects are needed to allow each golfer to have a personalised, and arbitrary variation in performance through time. The performance of some golfers, for example, will decline more slowly than others and using a golfer-specific effect will allow for this in the model. Random, rather than fixed, effects are used to model this, for two reasons. First, some of the age-dependence would be absorbed into fixed effects, because in the 18 years data, some golfers are young and some much older. Second, there are nearly a thousand golfers, so having fixed effects for each golfer would require many parameters. An age-based random-effects model allows a smooth autocorrelated deviation for each golfer from the mean age-trajectory. Our chosen model is thus a mixed effects model.

3.3 The model in detail

We model y_{it} as the sum of a barycentric interpolation of performance to the ith golfer's age at the tth match, a random error for the golfer's performance, and a random residual error. A barycentric interpolation coefficient $\mu_j + \gamma_j s_{it} + \epsilon_{ij}$ appears at each of m ages, a_j, that are the nodes of the interpolant, and this coefficient is a normally distributed random variate. The $\gamma_j s_{it}$ term allows performance to change with calendar time s, and the addition of the ϵ_{ij} allows the performance of individual golfers to diverge from the mean age-trajectory. We take age a ranging from a minimum of 17.5 to a maximum of 62.5 with $m = 8$ regularly spaced nodes; the model fit was insensitive to the number of nodes and their placement. Suppressing the golfer suffix i, with the golfer's age at the tth match a_t, the model for the tth score y_t is

$$y_t = h_t + \sum_{j=1}^{m} r_{tj}\zeta_j + \eta + \epsilon_t, \tag{1}$$

where $h_t = \frac{\sum_{j=1}^{m} w_j(\mu_j+\gamma_j s_t)/(a_t-a_j)}{\sum_{j=1}^{m} w_j/(a_t-a_j)} + \boldsymbol{\beta}^T \mathbf{x}(a_t)$ is the deterministic part of the age-dependence, plus a regression on fixed or age-dependent covariates $\mathbf{x}$, and

$$r_{tj} = \frac{w_j/(a_t - a_j)}{\sum_{j=1}^{m} w_j/(a_t - a_j)}. \tag{2}$$

The w_j are the barycentric weights for which Floater and Hormann (2007) provide formulae. Here, we use weights of ‘order zero’ given by $w_j = (-1)^j$.

The expression for h_t is mathematically well defined at the nodal ages, as $a_t \to a_j$, where the barycentric interpolation simply yields the nodal value. Computationally, one can either simply test that $|a_t - a_j| < b$, where $b \simeq 10^{-3}$, and if so use the nodal value, or else use the reformulation in Baker and Jackson (2014) that avoids the problem completely.

The errors are $\epsilon_t \sim N[0, \sigma^2]$, $\eta \sim N[0, \lambda^2]$, and $\zeta_j \sim N[0, \phi^2]$ and all errors are independent. A golfer thus has a random shock ζ_j occurring at each tabulated age (node) and an overall random level of performance η. Each score y_t is also subject to a random error ϵ_t. The parameters are $\boldsymbol{\mu}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and σ, λ, ϕ.

3.4 The likelihood function

Using equation (1), the likelihood for a golfer is

$$\mathcal{L} = \frac{F}{(2\pi)^{(m+1)/2}} \int_{-\infty}^{\infty} \exp(-\psi/2)\, \mathrm{d}\boldsymbol{\zeta}\, \mathrm{d}\eta, \tag{3}$$

where

$$F = (2\pi\sigma^2)^{-n/2} \lambda^{-1} (\phi)^{-m}. \tag{4}$$

With $\Delta_t = y_t - h_t$,

$$\psi = \frac{\sum_{t=1}^{n} \left(\Delta_t - \sum_{j=1}^{m} r_{tj}\zeta_j - \eta \right)^2}{\sigma^2} + \frac{\sum_{j=1}^{m} \zeta_j^2}{\phi^2} + \frac{\eta^2}{\lambda^2}. \tag{5}$$

Writing $\mathbf{v} = (\zeta_1 \cdots \zeta_m, \eta)^T$, we have

$$\psi = A - 2\mathbf{B}^T\mathbf{v} + \mathbf{v}^T\mathbf{M}\mathbf{v}, \tag{6}$$

where $\mathbf{M}$ is symmetric, $A = \sum_{t=1}^{n} \Delta_t^2/\sigma^2$ and

$$B_i = \begin{cases} \sum_{t=1}^{n} \Delta_t r_{ti}/\sigma^2 & \text{if } i \le m \\ \sum_{t=1}^{n} \Delta_t/\sigma^2 & \text{if } i = m+1. \end{cases} \tag{7}$$

Also,

$$M_{ij} = \begin{cases} \sum_{t=1}^{n} r_{ti} r_{tj}/\sigma^2 + \delta_{ij}/\phi^2 & \text{if } i, j \le m \\ \sum_{t=1}^{n} r_{ti}/\sigma^2 & \text{if } i \le m, j = m+1 \\ n/\sigma^2 + 1/\lambda^2 & \text{if } i = j = m+1. \end{cases} \tag{8}$$

3.5 Model fitting

To fit the model to data by likelihood-based methods, the likelihood function must be integrated over the $m + 1$ normal random variates $\boldsymbol{\epsilon}$, η.

The likelihood for a model with just one random effect for the golfer's performance, i.e. $\phi = 0$, can be found analytically, by doing the integration over the random effect corresponding to differences in golfer performance (see Appendix A).

For the full model with $\phi > 0$, we can evaluate the integral, allowing for random effects at each node, by completing the square in the exponent. Then

$$\psi = A - 2\mathbf{B}^T\mathbf{v} + \mathbf{v}^T\mathbf{M}\mathbf{v} = C + (\mathbf{v} - \boldsymbol{\delta})^T\mathbf{M}(\mathbf{v} - \boldsymbol{\delta}), \tag{9}$$

from which we read off $\mathbf{M}\boldsymbol{\delta} = \mathbf{B}$, $C = A - \boldsymbol{\delta}^T\mathbf{B}$. The vector $\boldsymbol{\delta}$ is found by solving the $m+1$ linear equations $\mathbf{M}\boldsymbol{\delta} = \mathbf{B}$. The distribution of $\mathbf{v}$ given $y_1 \cdots y_n$ is multivariate normal with mean $\boldsymbol{\delta}$ and covariance matrix $\mathbf{M}^{-1}$.

Maximum-likelihood estimators are known to underestimate scale parameters, such as σ, ϕ and λ, but because of the large sample size, this bias will be negligible. For small samples, corrections can be made. One of many methods is to use the parametric bootstrap to simulate fresh datasets from the fitted model, and so estimate the bias of the maximum-likelihood estimates.

Note that one can think of this random-effects model in Bayesian terms; the normal distribution of the errors would be the prior pdf, and our likelihood would then become the posterior probability. This approach would then be empirical Bayes, based on maximum posterior probability.

The vector of realised random effects, $\boldsymbol{\delta}$, is found by solving the $m+1$ linear equations. Doing a Cholesky decomposition $\mathbf{M} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is lower-diagonal, the $m+1$ linear equations for $\boldsymbol{\delta}$ can be solved, and the determinant $|\mathbf{M}| = \prod_{i=1}^{m+1} L_{ii}^2$ calculated. There is no need to invert the matrix $\mathbf{M}$. Thus, finally

$$\mathcal{L} = F \exp\left(-A/2 + \boldsymbol{\delta}^T\mathbf{B}/2\right)/|\mathbf{M}|^{1/2}. \tag{10}$$

Likelihood maximisation was done using a function minimiser that does not require derivatives: this requires only computation of the log-likelihood and is quick. On a standard desktop running Windows the model was fitted in less than a minute. Note that this computational scheme means that we are not restricted to a linear model, although we have used one here.

The computation proceeds by direct maximisation of the likelihood function, whereas economists usually use feasible generalised least squares. This is a 2-step iteration in which generalised least squares fit residuals are used to estimate the random-effect parameters, which are then used to re-estimate the covariance matrix across all observations. We have instead regarded this as a general likelihood-maximisation problem, which was simpler to compute.

In order to check the model (and coding), data were simulated from the fitted model and the model refitted to that data. The resulting parameter estimates were very close to the ones simulated (within the estimated standard errors). This gave us confidence in the correctness of the model development and programming.

Finally, we comment on the novelty of this methodology. Mixed models are widely used, and Straube et al. (2015) have used splines to model temporal variation, although this is unusual. Using barycentric interpolation in a mixed model is new. Further, the method of model fitting by directly maximising the log-likelihood function after solving linear equations for the realised random effects is also new.

4 Model properties

Using random variables at the nodes of the barycentric interpolation as random effects has not, as far as we know, been done before. As such, it is useful to explore the model properties. Figure 1 shows four random realisations from the model, which shows that it can indeed model a wide range of individual performance and its change over time. The lower line shows how that hypothetical golfer tends to have below average performance and this 'trait' is carried throughout his career.

4.1 Additivity

The model allows for individual stochastic variation in scores. The model can be fitted to each of the four strokes-gained metrics (to see how age effects putting, and driving separately), or for

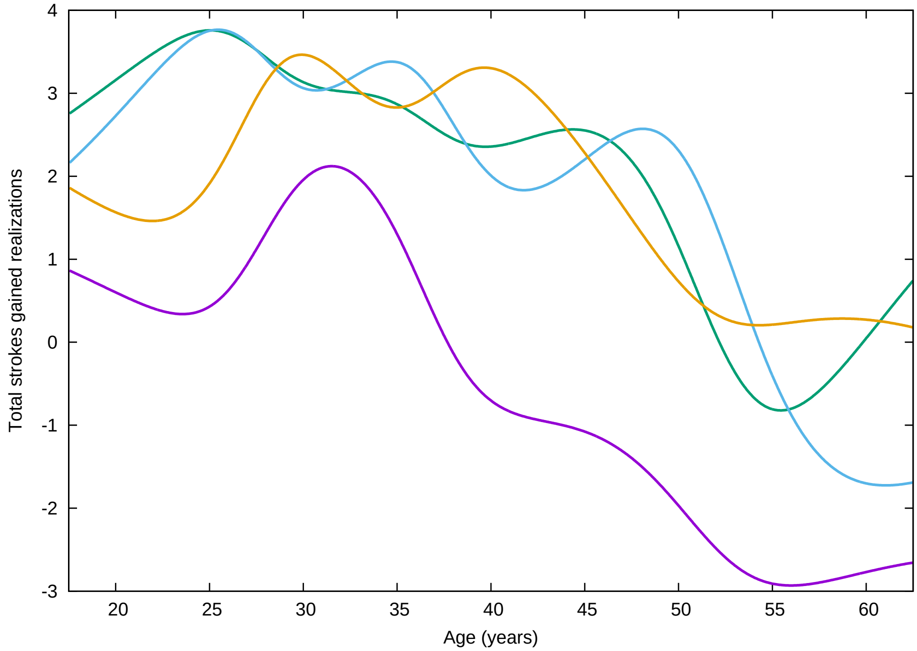

Figure 1. Total strokes gained for four random realisations of the fitted model.

combinations of them, such as long game minus short game scores. The model has an additivity property, in that any linear combination of two uncorrelated scores will obey the same form of model. This follows from the fact that the normal distribution has the same property.

4.2 An identity for the expected value of the random effect η given the data

From the properties of $\mathbf{M}$ and $\mathbf{B}$, it may be shown that $\delta_{m+1} = (\lambda^2/\phi^2)\sum_{i=1}^{m}\delta_i$, where δ_{m+1} is the mean of η.

To show this, we take the equation $\mathbf{M}\boldsymbol{\delta} = \mathbf{B}$. The equation for the ith row, where $i \leq m$, is

$$\sum_{t=1}^{n} r_{ti}r_{tj}\delta_j/\sigma^2 + \delta_i/\phi^2 + \sum_{t=1}^{n} r_{ti}\delta_{m+1}/\sigma^2 = \sum_{t=1}^{n}\Delta_t r_{ti}/\sigma^2, \tag{11}$$

The $m+1$th row gives

$$\sum_{t=1}^{n} r_{tj}\delta_j/\sigma^2 + (n/\sigma^2 + 1/\lambda^2)\delta_{m+1} = \sum_{t=1}^{n}\Delta_t/\sigma^2. \tag{12}$$

Summing equation (11) from 1 to m and recalling that $\sum_{i=1}^{m} r_{ti} = 1$ we have that

$$\sum_{t=1}^{n}\sum_{j=1}^{m} r_{tj}\delta_j/\sigma^2 + \left(\sum_{i=1}^{m}\delta_i\right)/\phi^2 + n\delta_{m+1}/\sigma^2 = \sum_{t=1}^{n}\Delta_t/\sigma^2, \tag{13}$$

and subtracting this from equation (12), we obtain

$$\delta_{m+1}/\lambda^2 = \sum_{i=1}^{m}\delta_i/\phi^2. \tag{14}$$

Intuitively, the sum of the expected values of the random effects at the m nodes could be taken as a measure of lifetime achievement, and the expected value of the random effect η is proportional to this.

4.3 Autocorrelation of scores

From equation (1), for a particular golfer, the covariance of scores between two ages a_s, a_t is

$$\mathrm{Cov}(Y_s, Y_t) = \sum_{j=1}^{m} r_{sj} r_{tj}. \tag{15}$$

The correlation is

$$r(Y_s, Y_t) = \frac{\sum_{j=1}^{m} r_{sj} r_{tj}}{\left\{\left(\sum_{j=1}^{m} r_{sj}^2 + \sigma^2\right)\left(\sum_{j=1}^{m} r_{tj}^2 + \sigma^2\right)\right\}^{1/2}}. \tag{16}$$

The correlation is largest when $a_s = a_t$ and decreases when a_s, a_t are far apart.

Autocorrelation of performances by a single golfer at different ages is a reasonable, and logical property of a model for the evolution of performance with age. A golfer who has a high performance at age a_s is likely to possess superior talent for example, and is therefore more likely to have a high performance at age a_t $(t > s)$, and the model captures this.

4.4 Probability of a golfer beating another

Given two random golfers, one of age a and the other of age b, the predicted 18-hole scores for golfer a and golfer b are $Y(a)$ and $Y(b)$, respectively. These scores are normally distributed, so the probability that the golfer of age a 'wins' (i.e. has a lower score over the 18 holes) is

$$\mathrm{Prob}(Y_a - Y_b > 0) = \Phi\left(\frac{h(a) - h(b)}{\sum_{j=1}^{m} \{r_{aj}^2 + r_{bj}^2\}\phi^2 + 2\lambda^2 + 2\sigma^2}\right), \tag{17}$$

where Φ is the normal distribution function. The probability that a particular golfer of age a performs better than he did at age b is

$$P(a, b) = \Phi\left(\frac{h(a) - h(b)}{\sum_{j=1}^{m} \{r_j(a) - r_j(b)\}^2 \phi^2 + 2\sigma^2}\right). \tag{18}$$

4.5 Forecasting

Given results $y_1 \cdots y_n$, we seek the distribution of $Y(a)$, when the golfer is of age a. Writing $f(\cdot)$ for pdfs, the conditional pdf for $\mathbf{v}$ by Bayes theorem is

$$f(\mathbf{v} \mid y_1 \cdots y_n) = \frac{f(y_1 \cdots y_n \mid \mathbf{v}) f(\mathbf{v})}{\int_{-\infty}^{\infty} f(y_1 \cdots y_n \mid \mathbf{v}) f(\mathbf{v})\, \mathrm{d}\mathbf{v}}. \tag{19}$$

Hence, the pdf of $Y(a)$ is

$$f(y(a) \mid y_1 \cdots y_n) = \frac{\int_{-\infty}^{\infty} f(y(a) \mid \mathbf{v}) f(y_1 \cdots y_n \mid \mathbf{v}) f(\mathbf{v})\, \mathrm{d}\mathbf{v}}{\int_{-\infty}^{\infty} f(y_1 \cdots y_n \mid \mathbf{v}) f(\mathbf{v})\, \mathrm{d}\mathbf{v}}. \tag{20}$$

Thus, $f(y(a) \mid y_1 \cdots y_n) = \mathcal{L}(y(a), y_1 \cdots y_n)/\mathcal{L}(y_1 \cdots y_n)$. Define $\boldsymbol{\xi}^T = (r_{a1} \cdots r_{an}, 1)$. Then for fixed $\mathbf{v}$, $Y(a)$ is normally distributed with mean $h(a) + \boldsymbol{\xi}^T \mathbf{v}$ and variance σ^2. Hence, $Y(a)$ is the sum of two

Table 2. Model parameter estimates with standard errors for (minus) score relative to par in 2012–2013

Parameter	Estimate	Standard error	Trend γ	Standard error
Height	0.0365	0.0112	–	–
Residual error σ	6.7695	0.0035	–	–
Performance random effect sd λ	1.2596	0.0632	–	–
Age variation random effect sd ϕ	0.5006	0.2073	–	–
Score at age 17.5	2.9094	0.4455	0.2464	0.0946
Score at age 25.0	3.3174	0.1328	0.2067	0.0247
Score at age 30.0	3.4546	0.1078	0.1333	0.0189
Score at age 35.0	3.4285	0.1092	0.1345	0.0190
Score at age 40.0	2.9905	0.1230	0.1220	0.0217
Score at age 47.5	2.2309	0.1841	0.0011	0.0338
Score at age 55.0	−0.0413	0.7470	0.0599	0.1393
Score at age 62.5	−0.2697	1.5975	0.3829	0.3036

Note. The mean correction term was 0.0105. The p-value for height is 0.0012. Units are strokes per tournament, except for trend, where the units are strokes per tournament per year.

independent normally distributed random variables, and the unconditional distribution of $Y(a)$ is normal with mean $h(a) + \xi^T \boldsymbol{\delta}$ and variance $\sigma^2 + \xi^T \mathbf{M}^{-1} \xi$.

Forecasting a golfer's future performance level is of interest for several reasons. For example, it allows sponsoring companies to understand the future expectations of a golfer they may be considering sponsoring. Bookmakers may be interested in setting odds on future tournament wins, for example, on the number of Majors won in a career. Or, perhaps fans and the media are simply intrigued to have a prediction on how good a young golfer is expected to become.

4.6 A correction term

The means μ_j give the mean score for the average golfer at the jth age node, when covariates such as height are centred on their means. The secular trend term had origin mid-way through the time span so the μ_j refer to a golfer in that year. The vector $\boldsymbol{\delta}$ gives a correction to the mean for a particular golfer, and the mean over all golfers is zero, as it should be.

However, it is a fact that better golfers play more games than less able golfers. If by golfer performance at some age we mean the average performance of a randomly selected golfer, in computing the mean $\boldsymbol{\delta}$, we must weight the $\boldsymbol{\delta}$ values for each golfer by the number of games played. In other words, in observing a random golfer we have 'length-biased sampling'. Doing this, the mean is no longer zero, and the mean $\boldsymbol{\delta}$ gives a small correction to mean golfer performance. This small correction is mentioned in the table captions, but is not added in the plots.

5 Results

The quality of the model was measured by the Akaike Information Criterion. This showed that the fit could not be improved by adding more nodes or changing their placement. Fit residuals were normally distributed (skewness −0.0054, excess kurtosis 0.355) and had zero autocorrelation.

Table 2 shows the results of the model when the dependent variable is score relative to par and Table 3 shows the results of the model when the dependent variable is total strokes gained. Results for the other strokes-gained metrics are shown in tables in Appendix B.

It is perhaps easier to grasp the results of these models visually, and Figure 2 shows the fitted evolution of score relative to par, with the different lines showing its variation with calendar time. The time-dependent model this figure is based upon shows how younger golfers have improved with calendar time, as can be seen by the blue line being higher than the other two lines. The age-dependence of total score is fairly flat until the 40s. The upturn in performance at an age of around 52 years shown for the blue line is likely to be a consequence of 'survival' bias in

Table 3. Model parameter estimates with standard errors for total strokes gained 2012–2013

Parameter	Estimate	Standard error	Trend γ	Standard error
Height	0.0406	0.0114	–	–
Residual error σ	5.2815	0.0035	–	–
Performance random effect sd λ	1.3919	0.0503	–	–
Age variation random effect sd ϕ	0.7947	0.0798	–	–
Score at age 17.5	2.0988	0.3858	0.1185	0.0826
Score at age 25.0	2.1744	0.1268	0.0657	0.0233
Score at age 30.0	2.4714	0.1062	0.0006	0.0179
Score at age 35.0	2.3448	0.1080	−0.0127	0.0180
Score at age 40.0	1.8906	0.1210	−0.0607	0.0204
Score at age 47.5	0.8002	0.1758	−0.1707	0.0316
Score at age 55.0	−0.6898	0.6320	−0.2017	0.1153
Score at age 62.5	−0.8368	1.3303	0.0078	0.2481

Note. The mean correction term was 0.029. The p-value for height is 0.0004. Units are strokes per tournament, except for trend, where the units are strokes per tournament per year.

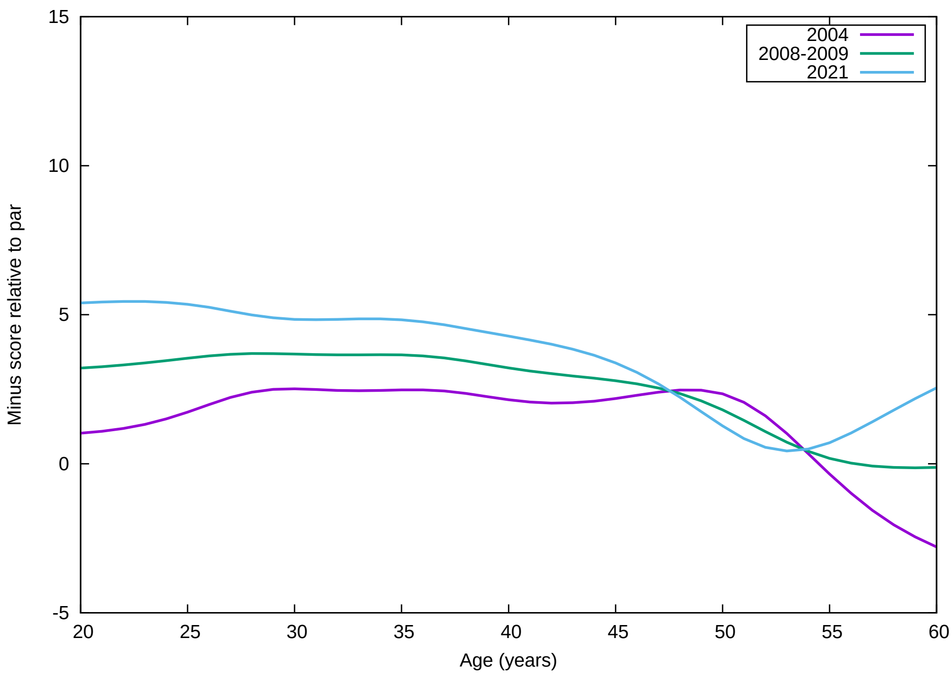

Figure 2. Minus score relative to par and its variation with age and calendar time.

that the golfers still competing at that age are highly likely to be elite golfers, with talent levels in the tails of the distribution.

Figure 3 shows the evolution for total strokes gained, strokes-gained approach, strokes-gained around-the-green, and strokes-gained putting, respectively. The yellow line in Figure 3 shows how overall scoring performance improves until the mid-30s (peaking at 35), which is later than in other sports. A somewhat rapid decay in performance begins thereafter. Our results on ‘off-the-tee’ vs. the other strokes gained categories highlight the different skills required to excel in each area of

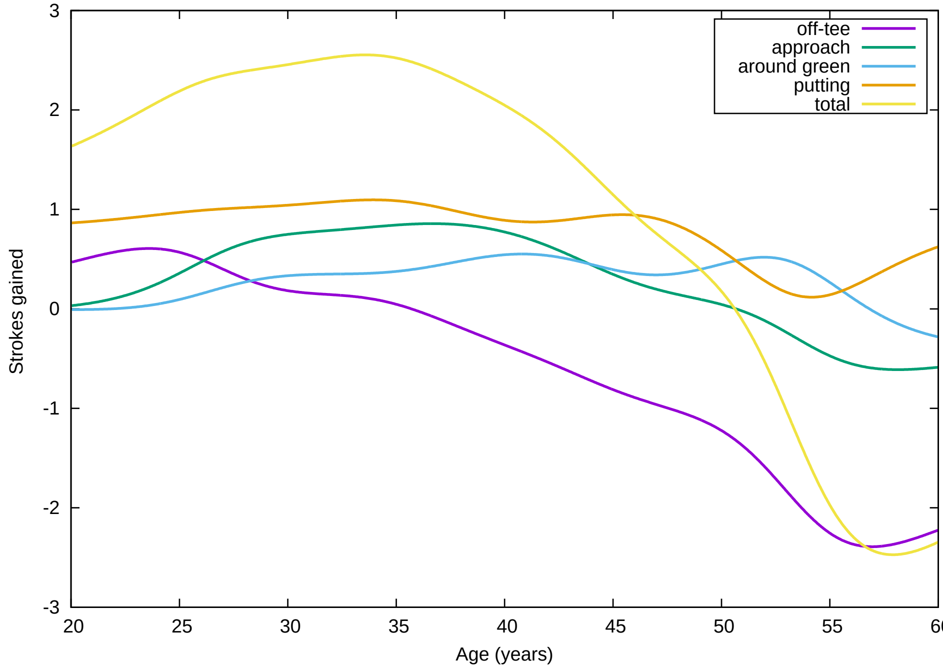

Figure 3. Synopsis of the age-dependence of the four golfing skills.

golf. Like sprinting, driving performance peaks early and effectively declines from 20 years old. Putting and around-the-green performance appear to be almost constant with age, at least until the late forties. Like total strokes-gained, approach play appears to peak in the mid- to late-30s. These results highlight, to some extent, the different physical (and to some extent mental) attributes required to excel in golf.

It is worth highlighting the magnitude of the age effect. Considering total strokes gained (the yellow line in Figure 3), we see that for a golfer at his peak age of around 35, he is nearly one full shot per round better than himself at age 20, and more than two full shots per round better than himself at age 50.

As was evident in Figure 2, the slight upturns in performance for strokes-gained putting, off-the-tee and total, are likely to be a consequence of 'survival bias' such that only the very best golfers survive on the PGA Tour to compete at that age.

The coefficient on height in Tables 2 and 3 is positive and statistically significant such that taller golfers have a slight advantage in terms of scoring and total strokes gained. The standard deviation of height is 6.08 cm, and moving from one standard deviation below mean height to one standard deviation above gives a mean drop of 0.45 strokes. Going from the bottom 5% to the top 5% height gives an advantage of 0.73 strokes. We also experimented with including a golfer's weight and/or BMI in the model but it appears not to have an effect on performance. Height had no effect on putting strokes gained (see Table B4 in Appendix B), which is as one might expect.

Figure 4 shows the long-term forecast of expected performance for Tiger Woods to age 60. Visually, one can see how the model fit tracks through the middle of his total strokes gained. The forecast for Tiger was, as of 2021, that he would continue to attain a positive strokes gained until his 60s. However, the picture may have changed considerably, as Tiger Woods had a car accident and broke both legs. In mid-2022, Tiger struggled to even complete a round of golf, let alone reach the level of performance the world had come to expect from him. Our model cannot of course account for such random events. Note that the error bars shown here represent plus and minus the standard error on the performance, not a prediction interval. The results for any match are subject to a further variation of about five strokes. Similarly, Figure 5 shows the long-term forecast for Scottie Scheffler. Scheffler is still relatively young in golfing terms at only 25, yet he

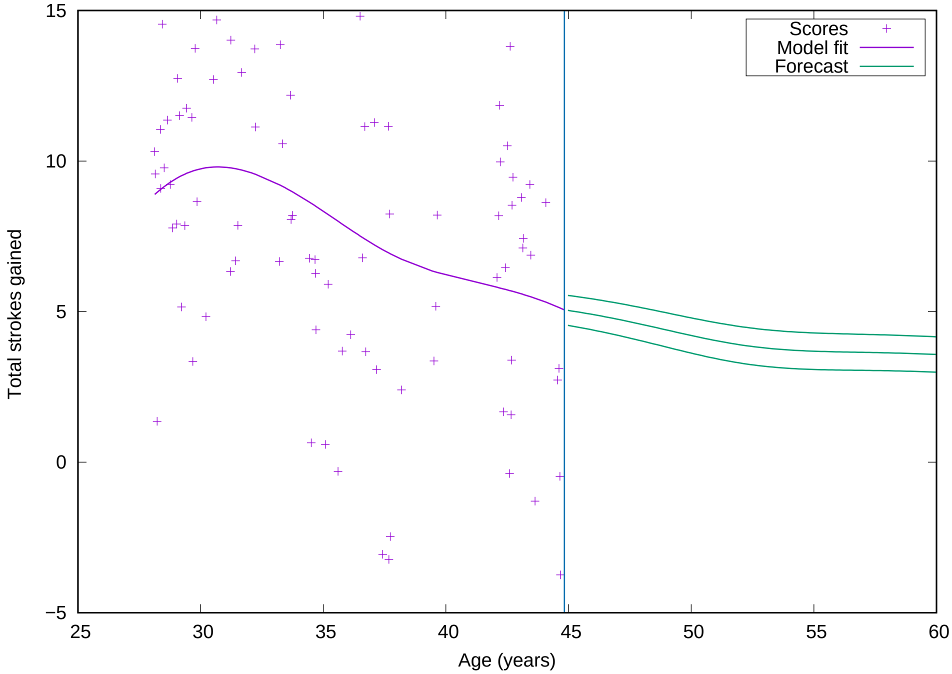

Figure 4. Fitted curve of total strokes gained for Tiger Woods, showing forecast performance to age 60 with standard error.

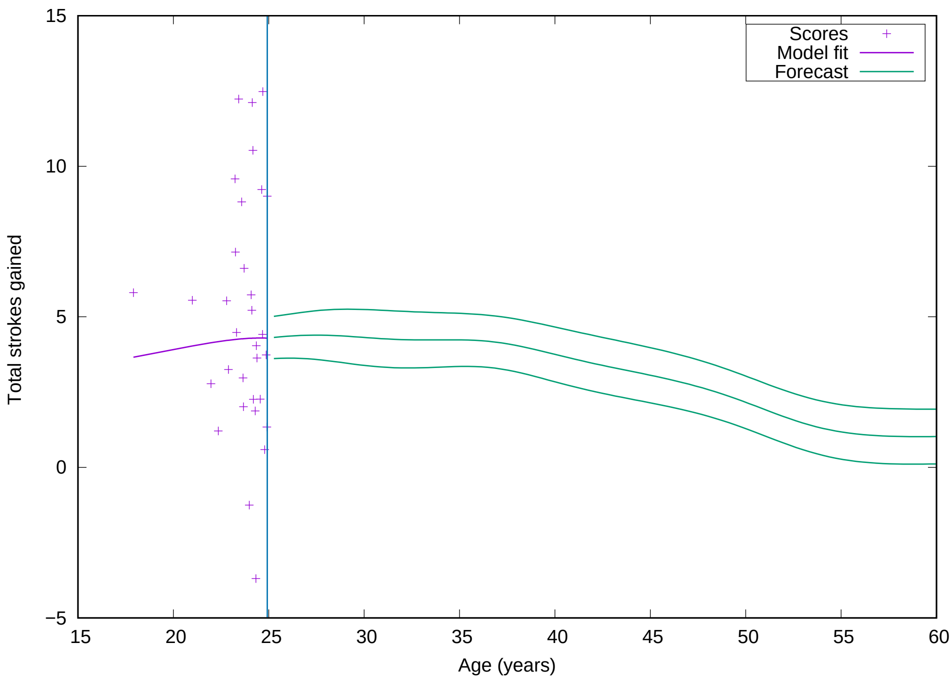

Figure 5. Fitted curve of total strokes gained for Scottie Scheffler, showing forecast performance to age 60 with standard error.

has already won many tournaments and reached the top of the Official World Golf Rankings. Whilst at his peak Tiger was gaining nearly 10 shots per round over the average (see Figure 4), Scheffler is currently 'only' gaining around five shots, which is expected to begin to decrease around his early to mid-30s.

6 Other application areas

Performance in all studied fields improves from childhood before gradually declining. What differs across type of activity is the age of peak performance, and the rate of decline. The persistent finding is that activities relying on speed and power result in younger peak age of performance (early to mid-20 s), whilst activities relying on skill, experience, or intelligence have a peak age of performance in the early 30 s, or for invention, the late 30 s.

Beyond the context of sport, Jones (2010) considers the 'age of great invention' by modelling the age at which an individual conducts Nobel Prize winning research or is responsible for technological innovation. He finds that, within an activity, the age of peak performance has increased over the course of the last century. Having been in the early 30s at the start of the 20th century, it is now in the late 30s. He speculates that the reason for increasing age of great invention is that it now requires longer to make it through 'training' before great invention is allowed to take place.

Most recently, Choi et al. (2022) look at the performance of entrepreneurs as a function of age and experience and, as in other fields, find an inverse U-shaped relationship with a firm's performance. Similarly, Zhang et al. (2022) find that ageing of the population impedes corporate innovation.

For further references of the descriptive work in this area, we point the reader towards Hertzog (2020) who provides a review of the literature.

The model described here can be used in all these areas.

7 Conclusions

The paper presents a new type of mixed model which allows for random effects in a barycentric rational interpolation. We use the model to understand the effects of age on the performance of professional golfers. The model allows for smooth variation in performance, but cannot allow for sudden changes in performance, e.g. as a result of injury.

Golf provides an interesting laboratory for such analysis as it requires different physical and mental attributes for each different part of the game. As has been shown in previous studies, we find that activities requiring power and speed (here this is driving, in other studies it has been 100 m sprinting, for example), peak performance occurs in the early 20s, and declines thereafter. We have the additional insight that the other skills in golf, namely putting and around-the-green peak much later, and, to some extent, are fairly insensitive to age, until the late 40s at least.

Apart from the intrinsic interest of these findings, the model enables long-term forecasts of performance for a golfer. This could be of use to several parties. For example, agents wondering which young golfers to sign up, and sponsoring companies thinking about which golfer to sign, and the size of the contract to be offered, would be interested in the forecast performance of individual golfers.

A use of the model we have yet to explore, but which looks promising, is to use the model to help golfers improve their total performance by identifying strengths and weaknesses in the different areas of the game. However, to do this, we would need to know the relative effort required to achieve improvements in the four skills (putting, driving, approach play, and around-the-green).

Our model could certainly be used to help understand the effects of age in other sports and other fields. In football, for example, players are given contracts worth millions, even tens of millions of pounds. It would be valuable for clubs to know when a player is likely to peak and how his or her ability to play at a particular level will evolve. Outside of sport, it would be interesting to see how age effects performance in other areas such as business. Sales commissions, for example, could be modelled as a function of age.

Conflict of interest: None declared.

Funding

No funding was received for this research.

Data availability

All data were obtained from the PGA Tour website (https://www.pgatour.com/). Code is available upon request from the authors.

Appendix A: Likelihood for simple model

We need to estimate the n parameters μ_i, with $\phi_i = 0$, as well as λ and σ. However, we can do the integral over η_i analytically. The likelihood for a golfer is

$$\mathcal{L}_i = \int_{-\infty}^{\infty} \frac{\exp\left(-\sum_{i=1}^{n_i} \exp\{-(y_{it} - h_{it} - \eta_i)^2/2\sigma^2 - \eta_i^2/2\lambda^2\}\, \mathrm{d}\epsilon_i\right.}{(2\pi\sigma^2)^{n_i/2}(2\pi\lambda^2)^{1/2}}.$$

Completing the square in the exponent allows this integral to be evaluated, as

$$\mathcal{L}_i = \frac{\exp\left\{-\sum_{t=1}^{n_i}(y_{it} - h_{it})^2/2\sigma^2 + \left(\sum_{t=1}^{n_i}(y_{it} - h_{it})\right)^2/(2\sigma^2(n_i + \sigma^2/\lambda^2)\right\}}{(2\pi\sigma^2)^{n_i/2}(\sigma^2/(n_i\lambda^2 + \sigma^2))^{1/2}}.$$

Appendix B: Parameter estimates for the different strokes-gained metrics

Table B1. Model parameter estimates with standard errors for strokes-gained around-the-green in 2012–2013

Parameter	Estimate	Standard error	Trend γ	Standard error
Height	0.0650	0.0179	–	–
Residual error σ	5.9475	0.0035	–	–
Performance random effect sd λ	2.4346	0.0385	–	–
Age variation random effect sd ϕ	1.3740	0.0541	–	–
Score at age 17.5	0.4506	0.4921	0.0239	0.1037
Score at age 25.0	−0.0106	0.1796	−0.0205	0.0324
Score at age 30.0	−0.1883	0.1556	−0.0096	0.0252
Score at age 35.0	−0.4983	0.1593	−0.0455	0.0252
Score at age 40.0	−0.9614	0.1762	−0.0555	0.0283
Score at age 47.5	−1.7924	0.2496	−0.1214	0.0437
Score at age 55.0	−2.8796	0.7982	−0.0318	0.1416
Score at age 62.5	−4.6404	1.6552	0.1315	0.2970

Note. The mean correction term was 0.0028. The p-value for height is 0.0209. Units are strokes per tournament, except for trend, where the units are strokes per tournament per year.

Table B2. Model parameter estimates with standard errors for strokes-gained off-the-tee in 2012–2013

Parameter	Estimate	Standard error	Trend γ	Standard error
Height	0.0339	0.0073	–	–
Residual error σ	2.1513	0.0036	–	–
Performance random effect sd λ	0.9585	0.0426	–	–
Age variation random effect sd ϕ	0.7487	0.0419	–	–
Score at age 17.5	1.0581	0.1984	−0.0585	0.0410
Score at age 25.0	0.6803	0.0757	−0.0056	0.0136
Score at age 30.0	0.3247	0.0662	−0.0190	0.0107
Score at age 35.0	0.1081	0.0679	−0.0196	0.0107
Score at age 40.0	−0.2991	0.0755	−0.0428	0.0120
Score at age 47.5	−0.7933	0.1054	−0.0607	0.0182
Score at age 55.0	−1.7888	0.3159	−0.0459	0.0553
Score at age 62.5	−2.4217	0.6505	0.0560	0.1136

Note. The mean correction term was 0.0208. The p-value for height is 0.0003. Units are strokes per tournament, except for trend, where the units are strokes per tournament per year.

Table B3. Model parameter estimates with standard errors for strokes-gained approach in 2012–2013

Parameter	Estimate	Standard error	Trend γ	Standard error
Height	0.0200	0.0077	–	–
Residual error σ	3.3121	0.0035	–	–
Performance random effect sd λ	0.9748	0.0461	–	–
Age variation random effect sd ϕ	0.5485	0.0712	–	–
Score at age 17.5	0.1914	0.2475	0.1300	0.0530
Score at age 25.0	0.3325	0.0828	0.0240	0.0153
Score at age 30.0	0.7452	0.0700	0.0117	0.0118
Score at age 35.0	0.7292	0.0713	−0.0118	0.0118
Score at age 40.0	0.6970	0.0796	−0.0224	0.0134
Score at age 47.5	0.1968	0.1153	−0.0866	0.0207
Score at age 55.0	−0.0402	0.4053	−0.0711	0.0736
Score at age 62.5	−0.3457	0.8507	0.0220	0.1574

Note. The mean correction term was 0.015. The p-value for height is 0.01. Units are strokes per tournament, except for trend, where the units are strokes per tournament per year.

Table B4. Model parameter estimates with standard errors for strokes-gained putting 2012–2013

Parameter	Estimate	Standard error	Trend γ	Standard error
Height	−0.0019	0.0067	–	–
Residual error σ	3.3914	0.0035	–	–
Performance random effect sd λ	0.7834	0.0494	–	–
Age variation random effect sd ϕ	0.4661	0.0833	–	–
Score at age 17.5	0.6327	0.2396	0.0096	0.0518
Score at age 25.0	0.9410	0.0747	0.0397	0.0142
Score at age 30.0	0.9761	0.0622	−0.0072	0.0109
Score at age 35.0	1.0296	0.0632	0.0090	0.0110
Score at age 40.0	0.8251	0.0706	0.0004	0.0125
Score at age 47.5	0.8055	0.1035	−0.0141	0.0194
Score at age 55.0	0.4852	0.3953	−0.0991	0.0727
Score at age 62.5	1.3121	0.8393	−0.0706	0.1574

Note. The mean correction term was 0.0022. The p-value for height is 0.77 (n/s). Units are strokes per tournament, except for trend, where the units are strokes per tournament per year.

References

Baker R. D., & Jackson D. (2014). Statistical application of barycentric rational interpolants: An alternative to splines. *Computational Statistics*, *29*(5), 1065–1081. https://doi.org/10.1007/s00180-014-0480-7

Baker R. D., & McHale I. G. (2015a). Deterministic evolution of strength in multiple comparisons models: Who is the greatest golfer? *Scandinavian Journal of Statistics*, *42*(1), 180–196. https://doi.org/10.1111/sjos.12101

Baker R. D., & McHale I. G. (2015b). Time varying ratings in association football: The all-time greatest team is···. *Journal of the Royal Statistical Society, Series A*, *178*(2), 481–492. https://doi.org/10.1111/rssa.12060

Berrut J. P. (1988). Rational functions for guaranteed and experimentally well-conditioned global interpolation. *Computers and Mathematics with Applications*, *15*(1), 1–16. https://doi.org/10.1016/0898-1221(88)90067-3

Broadie, M. (2012). Assessing golfer performance on the PGA tour. *Interfaces*, *42*(2), 146–165. https://doi.org/10.2307/41472743

Choi Y., Eum W., Kang T., & Lee J.-D. (2022). Impacts of entrepreneurial experience on firm performance in IPO firms. *Managerial and Decision Economics*, *43*(8), 1–13. https://doi.org/10.1002/mde.3647

Floater M. S., & Hormann K. (2007). Barycentric rational interpolation with no poles and high rates of approximation. *Numerische Mathematik*, *107*(2), 315–331. https://doi.org/10.1007/s00211-007-0093-y

Hertzog, C. (2020). Aging and peak human performance: A glance back. *Journals of Gerontology, Series B: Psychological Science*, *75*(8), 1621–1624. https://doi.org/10.1093/geronb/gbaa057

Jones, B. F. (2010). Age and great invention. *The Review of Economics and Statistics*, *92*(1), 1–14. https://doi.org/10.3386/w11359

Koensigsberg A., Pilgrim J., & Baker J. (2020). Generational differences in the ranking pathways of top 100 ranked golfers. *Journal of Sports Science*, *18*(38), 2047–2053. https://doi.org/10.1080/02640414.2020.1769814

Lehman H. C. (1945). Man's most creative years quality versus quantity of output. *Scientific Monthly*, *61*, 127–137. http://www.jstor.org/stable/18302

Roring, R. W., & Charness, N. (2007). A multilevel model analysis of expertise in chess across the life span. *Psychology and Aging*, *22*(2), 291–299. https://doi.org/10.1037/0882-7974.22.2.291

Schulz R., & Curnow C. (1988). Peak performance and age among superathletes: Track and field, swimming, baseball, tennis and golf. *Journal of Gerontology: Psychological Sciences*, *43*(5), 113–120. https://doi.org/10.1093/geronj/43.5.P113

Straube J., Gorse A. D., Bevan E. H., & Le Cao K. A. A. (2015). Linear mixed model spline framework for analysing time course 'omics' data. *PloS One*, *10*(8), e0134540. https://doi.org/10.1371/journal.pone.0134540

Zhang B., Zhou R., Yang L., & Zhang X. (2022). Population aging and corporate innovation: Evidence from China. *Asia-Pacific Journal of Accounting and Economics*, *30*(4), 986–1007. https://doi.org/10.1080/16081625.2022.2047741