Analysing trends in performances of elite swimmers

Samuel Arrowsmith

1 Introduction

In competitive swimming, there are four strokes (freestyle, backstroke, breaststroke, butterfly) and the individual medley, whereby a swimmer races all four in one event. Events vary by distance, with all strokes having a 200m distance. Races can also be swum in pools of varying length, however 50m (long course) is the most common format used internationally, such as in the Olympics [1]. For many athletes, some events will suit them better due to a variety of factors, like training frequency, strength, flexibility and morphology, as different body types have advantages in different strokes, such as butterfly swimmers having broad shoulders [2]. By analysing large databases of performances, it is possible to identify trends based on various factors, such as age, sex, distance, stroke, etc. Patterns in elite swimmers and their progress can be used to model target times, allowing swimmers and coaches to plan training blocks and season goals more effectively - i.e. aiming to drop a certain percentage of time in a year based on typical trends seen in the best athletes. Some research questions which can be answered using the datasets include: how does the percentage change in personal best times (PBs) differ between event and gender as elite athletes go through puberty? Are there common race tactics for longer events such as 200m and above which elite athletes employ - furthermore, can these split times be adjusted for developing athletes? To answer these, we can use Event Rankings [3], a comprehensive database of races licensed by British Swimming, with records since 1997 and over 24 million swims in the database. Data can be scraped using BeautifulSoup4 [4] - a Python library which processes HTML as a tree of DOM elements. There are two main approaches for getting data: querying for top n swims based on age, sex, event, year, etc. or using a swimmer's ID number to access their times. All aspects of the site are publicly available so there are no issues with data protection; moreover, age is given based on birthyear so specific birthdays aren't available. Getting data in an appropriate format to use for machine learning and modelling means testing that the data provided in the queries is formatted in seconds and isn't missing key details. Some uncommon issues include members' IDs being hidden and split times not being available, which interrupt scraping of large amounts of data and allow erroneous data to impact models. To address this, a modular approach which is easy to test and verify is needed, thus I integrated mapping, filtering and function composition to concisely ensure that data being scraped and processed was in the right format at each step [5]. By having pipelines consisting of functions like this, I was able to get reliable data with no anomalies and few outliers.

2 Linear Regression on PB Progressions over Puberty

Swimmers with records further back in the database (i.e. 90s and 00s) don't always have comprehensive records of their swims available; to ensure that swimmers we get data from have an extensive history of swims through puberty, I queried the top 40 open-age athletes from 2022 as most of these will have swims from about age 12 in the 2010s, as well as reflecting the latest training programs and stroke techniques being used. Once IDs for the top 40 athletes have been gathered, further queries can be done to find the table of their times for the event, which can be filtered for PBs. For fair comparisons between gender and event, we need to find the percentage changes from PB to PB; as birthdays aren't accessible and most meets are done based on an athlete's age at the end of the year, I used the year of birth to calculate how far into their year of being each age they were when they performed a swim. For simplicity, I just compared 200m free and medley, making a plot for each, with male and female swimmers on both, age on the x-axis and PB change on the y-axis. Clear gaps between clusters on the x-axis intuitively corresponds to short course seasons in latter months of each year when swimmers won't tend to race long course. A polynomial regression model would allow us to compare these continuous variables (age: independent variable; PB change: dependent variable), although it should be noted that many factors impact progression which aren't accounted for by the model, like training hours, diets and schoolwork. To determine the degree of polynomial, I used the Bayes Information Criterion [6] and in each case, the best-fitting model which minimised residue was linear - this makes sense as by considering changes as a percentage instead of in raw time, it would correlate as tapering off as a swimmer ages and stops seeing strength and fitness gains from puberty. An advantage is we don't need to normalize the data as each component of the model is differentiated separately [7]. Despite this, the models still have a lot of outlying points at the earlier stage (12-14 years old) for both genders as there is a lot of development in a swimmer's life at this stage, not just in terms of puberty but also with increased schoolwork, changes in training, etc. Another interesting factor in both plots is the large gap around the 17-18 age group, which likely correlates to swimmers studying for A-levels and adjusting to university life, with progress resuming after as shown by the denser cluster of points at around 19. Both models start with 2% change in PB at age 12 for both genders which tapers towards 0.6-0.7% around age 21, however, interestingly, women see bigger percentage changes up to around 18 in freestyle than men according to the model, with this dropping below the mens rate after this age, whereas in medley, men consistently have a higher percentage change in PB compared to women up to 21. An issue with regression models is their sensitivity to outliers, although removing points with <10% PB change still led to the BIC suggesting a linear model.

3 DBSCAN Clustering of 200m splits to identify tactics

Clustering athletes based on race splits can allow us to infer tactics. I decided to get top 40 all-time swimmers' ID numbers (for men and women) in an event and find all their swims within 2% of their best time where split times are accessible. Then, we can plot the first 100m as a percentage of the swim (e.g. if their splits by 100m are (1:00.00-1:08.00), then proportionally, their splits are 46.875%, 53.125%) and also plot how far off their 200m PB they were as a percentage, capped at 2%. Clusters higher up are closer to 2% off PB but still account for good swims whereas those lower down are when the swimmer was on PB. Left to right accounts for a faster front end vs backend. For simplicity, I chose to do 200m medley and a scatter plot to ascertain which clustering method to use. Men and women were dispersed evenly across the plot with points covering large parts of it, thus an unsupervised approach is required to eliminate bias when identifying clusters. K-nearest-neighbour is not inherently for clustering and focuses on classification, as well as being supervised; K-means is unsupervised but requires the number of clusters and the starting means to be predetermined, which we can't do [8]. Both are skewed by outliers and due to needing some label or predetermined info on the clusters beforehand, are unsuitable compared to DBSCAN which is unsupervised, robust to noise, and calculates clusters by itself using densities of points [9]. I applied DBSCAN to the combined set of male and female points as they overlapped a lot, implying similar tactics. A drawback is the need to scale the data due to the use of distance metrics like points falling within a certain radius - I did this with MinMax, applied DBSCAN (eps=0.06, min_samples=5) and then plotted the clusters with the original scale, as well as mean averages of the clusters. Due to DBSCAN yielding clusters of various shapes and densities, mean points can be misleading; however, in our case, clusters were fairly uniform in shape. The main thing to note with the clusters was that they mostly ranged between 46-48% for the first 100m when excluding outliers, with the mean points of the fastest cluster (closest to PB) being just above 47%.

4 Conclusion

In summary, there is a lot of data which can be collected from Event Rankings to perform analysis on, based on stroke, age and gender. This report should hopefully have demonstrated how several techniques can be combined and applied to data to develop an insight into how different swimmers perform and how it can be used to influence race plans. To demonstrate how they can be used to model a season goal, suppose we have a 16-year-old boy whose 200m medley PB is 2:10.00 = 130 secs. The regression model implies an improvement of 1.5%: 130 * 0.015 = 1.95; 130 - 1.95 = 128.05 (2:08.05). The clustering model shows that when the best performers are on PB pace, their first 100m makes up 47.1% of the swim: 0.471 * 128.05 = 60.31 (1:00.31). Hence according to our models, our swimmer should aim to get a PB of 2:08.05 next season by splitting the first 100m in 1:00.31. Future developments could include modeling top performers in specific age groups and comparing split percentages with elite performers.

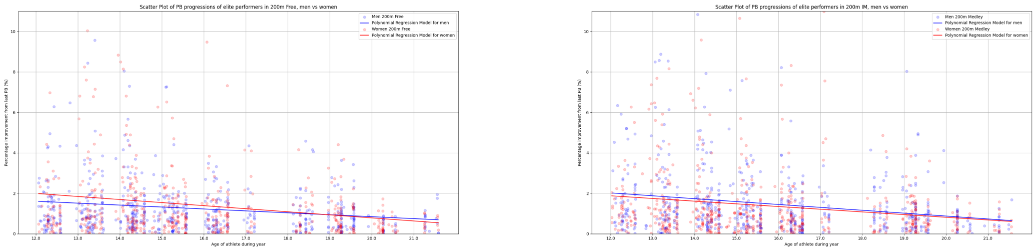
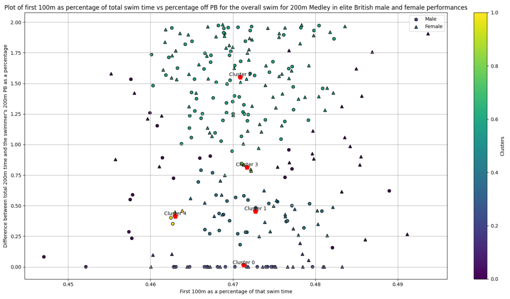


Figure 1: Above are the plots of pb progression through puberty for the top 40 athletes in 200m free and medley from 2022 for both men and women.

This is a scatter plot of 1st 100m of a 200m medley as a percentage of the total time on the x-axis against percentage off pb for the 200m on the y-axis. Data was collected from the top 40 men and women of all-time in the UK, with swims being selected if splits were available and if the swims were within 2% of the athletes' PBs. DBSCAN was applied and mean points found for the clusters.



References

[1] Olympics. Olympics swimming rules: Know events and format, 2024. Available at: https://olympics.com/en/news/olympics-swimming-rules.

[2] Bri Groves. A look at swimmer muscles by stroke, 2024. Available at: https://www.swimmingworldmagazine.com/news/a-look-at-swimmer-muscles-by-stroke/.

[3] AquaticsGB SPORTSYSTEMS. Swimming results, 2024. Available at: https://www.swimmingresults.org.

[4] Leonard Richardson. Beautiful soup documentation, 2007. Available at: https://beautiful-soup-4.readthedocs.io/en/latest/.

[5] D Karasek. Unleashing the power: The advantages of functional programming in the digital age, 2023. Available at: [https://scalac.io/blog/unleashing-the-power-the-advantages-of-functional-programming-in-the-digital-age/].

[6] H. S. Bhat and N Kumar. On the derivation of the bayesian information criterion, 2010. Archived from the original (PDF) on 28 March 2012; available at: https://arxiv.org/abs/1001.0774.

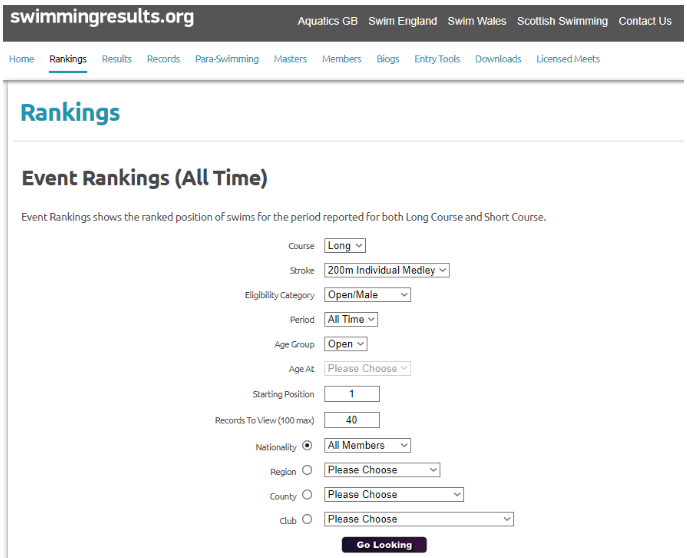
[7] Gianluca Malato. Which models require normalized data?, June 2022. Available at: https://www.yourdatateacher.com/2022/06/13/which-models-require-normalized-data/.

[8] GeeksforGeeks. How do k-means clustering methods differ from k-nearest neighbor methods?, November 2024. Available at: https://www.geeksforgeeks.org/how-do-k-means-clustering-methods-differ-from-k-nearest-neighbor-methods/.

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996. Available at: https://cdn.aaai.org/KDD/1996/KDD96-037.pdf.

A Appendix

This screenshot shows the inputs for a query for top swims.



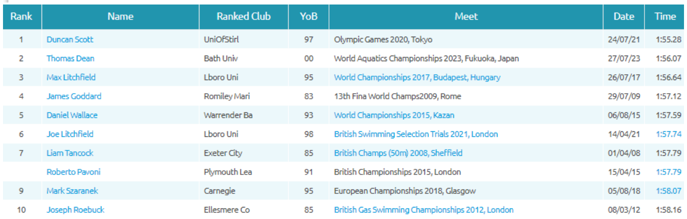
Rankings

Event Rankings (All Time)

Event Rankings shows the ranked position of swims for the period reported for both Long Course and Short Course.

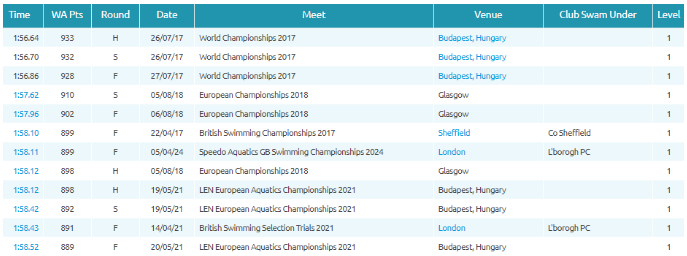
Course: Long
Stroke: 200m Individual Medley
Eligibility Category: Open/Male
Period: All Time
Age Group: Open
Age At: Please Choose
Starting Position: 1
Records To View (100 max): 40
Nationality ● All Members
Region ○ Please Choose
County ○ Please Choose
Club ○ Please Choose

This screenshot shows the table returned from a query. Note how swimmers' names are hyperlinks. These contain their ID numbers which can be used to search biogs.



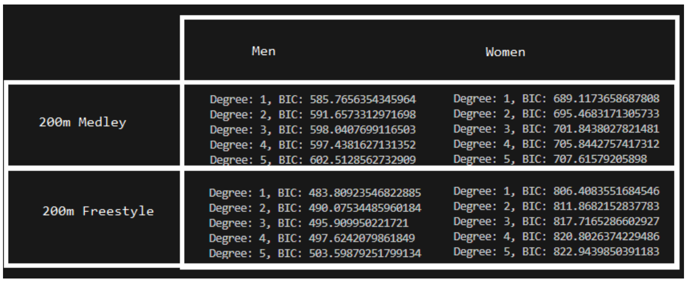
Rank	Name	Ranked Club	YoB	Meet	Date	Time
1	Duncan Scott	UniOfStirl	97	Olympic Games 2020, Tokyo	24/07/21	1:55.28
2	Thomas Dean	Bath Univ	00	World Aquatics Championships 2023, Fukuoka, Japan	27/07/23	1:56.07
3	Max Litchfield	Lboro Uni	95	World Championships 2017, Budapest, Hungary	26/07/17	1:56.64
4	James Goddard	Romiley Mari	83	13th Fina World Champs2009, Rome	29/07/09	1:57.12
5	Daniel Wallace	Warrender Ba	93	World Championships 2015, Kazan	06/08/15	1:57.59
6	Joe Litchfield	Lboro Uni	98	British Swimming Selection Trials 2021, London	14/04/21	1:57.74
7	Liam Tancock	Exeter City	85	British Champs (50m) 2008, Sheffield	01/04/08	1:57.79
	Roberto Pavoni	Plymouth Lea	91	British Championships 2015, London	15/04/15	1:57.79
9	Mark Szaranek	Carnegie	95	European Championships 2018, Glasgow	05/08/18	1:58.07
10	Joseph Roebuck	Ellesmere Co	85	British Gas Swimming Championships 2012, London	08/03/12	1:58.16

This screenshot shows the table of performances when you input a person's ID number in biogs and select the relevant event. Notice we can sort the rows by time if we convert it to seconds or by date the swim was performed.



Time	WA Pts	Round	Date	Meet	Venue	Club Swam Under	Level
1:56.64	933	H	26/07/17	World Championships 2017	Budapest, Hungary		1
1:56.70	932	S	26/07/17	World Championships 2017	Budapest, Hungary		1
1:56.86	928	F	27/07/17	World Championships 2017	Budapest, Hungary		1
1:57.62	910	S	05/08/18	European Championships 2018	Glasgow		1
1:57.96	902	F	06/08/18	European Championships 2018	Glasgow		1
1:58.10	899	F	22/04/17	British Swimming Championships 2017	Sheffield	Co Sheffield	1
1:58.11	899	F	05/04/24	Speedo Aquatics GB Swimming Championships 2024	London	L'borogh PC	1
1:58.12	898	H	05/08/18	European Championships 2018	Glasgow		1
1:58.12	898	H	19/05/21	LEN European Aquatics Championships 2021	Budapest, Hungary		1
1:58.42	892	S	19/05/21	LEN European Aquatics Championships 2021	Budapest, Hungary		1
1:58.43	891	F	14/04/21	British Swimming Selection Trials 2021	London	L'borogh PC	1
1:58.52	889	F	20/05/21	LEN European Aquatics Championships 2021	Budapest, Hungary		1

This table shows the outputs for the BIC calculations for the regression model - clearly linear regression minimizes the residue. It was calculated using the following formula: $BIC = -2\ln\hat{L} + k\ln n$ where $\hat{L}$ is the maximum likelihood estimate of the likelihood function, k is the number of parameters in the model, and n is the number of data points.



	Men	Women
200m Medley	Degree: 1, BIC: 585.7656354345964 Degree: 2, BIC: 591.6573312971698 Degree: 3, BIC: 598.0407699116503 Degree: 4, BIC: 597.4381627131352 Degree: 5, BIC: 602.5128562732909	Degree: 1, BIC: 689.1173658687808 Degree: 2, BIC: 695.4683171305733 Degree: 3, BIC: 701.8438027821481 Degree: 4, BIC: 705.8442757417312 Degree: 5, BIC: 707.61579205898
200m Freestyle	Degree: 1, BIC: 483.80923546822885 Degree: 2, BIC: 490.07534485960184 Degree: 3, BIC: 495.909950221721 Degree: 4, BIC: 497.6242079861849 Degree: 5, BIC: 503.59879251799134	Degree: 1, BIC: 806.4083551684546 Degree: 2, BIC: 811.8682152837783 Degree: 3, BIC: 817.7165286602927 Degree: 4, BIC: 820.8026374229486 Degree: 5, BIC: 822.9439850391183

For simplicity, I have uploaded the scraped data used to GitHub - filenames should make it clear which models they were used to make. Something to note is that the linear regression models used data in the form of days which was then converted to age in years for formatting in the plot, but the models are specifically built using days for accurate measuring and precision: Click me for the data

Generative AI Statement

AI-supported/AI-integrated use is permitted in this assessment. I acknowledge the following uses of GenAI tools in this assessment:

- (NO) I have used GenAI tools for developing ideas.
- (NO) I have used GenAI tools to assist with research or gathering information.
- (YES) I have used GenAI tools to help me understand key theories and concepts.
- (NO) I have used GenAI tools to identify trends and themes as part of my data
- analysis.
- (NO) I have used GenAI tools to suggest a plan or structure for my assessment.
- (NO) I have used GenAI tools to give me feedback on a draft.
- (NO) I have used GenAI tool to generate image, figures or diagrams.
- (NO) I have used GenAI tools to proofread and correct grammar or spelling errors.
- (NO) I have used GenAI tools to generate citations or references.
- (YES) Other: I have used GenAI tools to summarise documentation for specific libraries such as BeautifulSoup, MatPlotLib and SKLearn so that I can debug and restructure Python code
- I have not used any GenAI tools in preparing this assessment.

I declare that I have referenced use of GenAI outputs within my assessment in line with the University referencing guidelines.