Analysis of USA Olympic Swimming team*

USA's Best Men's Performance on 100 (m) Backstroke form 1970 to 2020

Nora Adadurova

September 20, 2025

Research Question

This paper aims to analyze Olympic swimming data and answer questions: for the USA in Men's 100 meters backstroke, by how many seconds on average does the best expected USA time change per Olympic year? Moreover, is the trend improving (negative) or worsening (positive)?

Abstract

This paper analyzes Olympic Men's 100 meters Backstroke performance for the USA team from 1972 to 2020. From swimmer-level per game results, I found the best USA time per Olympiad and fit a simple linear regression of time in seconds on year. The estimated trend is -0.1073 seconds per year with 95% Confidence Interval of $[-0.1316, -0.0829]$ and $R^2 = 0.9063$, i.e., roughly -0.43 seconds per Olympic cycle, indicating significantly faster performances over time. These results show a clear, sustained improvement in peak USA backstroke performance across the observed years.

Introduction

Motivated by my experience in competitive swimming with backstroke as my favorite stroke, I examine where USA performance in the Men's 100 meters Backstroke has improved over time. Using Olympic results (1970-2020), I extract the best USA time per game and fit a simple linear regression of time in seconds on year. The model estimates an average change of b_1

*Project repository available at: https://github.com/peteragao/MATH261A-project-template.

= -0.1073 seconds per year with 95% critical interval of [-0.1316, -0.0829] seconds per year, indicating faster performance over time. This negative trend suggests steady gain in speed for American swimmers in Olympic cycles, and the $R^2 = 0.9063$ quantifies how much of the variation is explained by year alone.

Data

Data come from the SCORE Sports Data Repository (Brendan Karadenes, olympic_swimming.csv, published June 13, 2024; 606 rows × 10 variables originally). For this paper, I restricted the scope of the analysis to Men's Backstroke at 100m. Team codes were standardized to handle historical naming: RUS, ROC, URS, and EUN were unified as RUS; GER, FRG, and GDR were unified as GER. Entries for the Czech/Slovak lineage (TCH, CZE, SVK) were dropped to avoid cross-entity aggregation. No missing values were detected in the retained variables for this subset. Each row is one athlete's official result in the Olympic Men's 100m Backstroke (100 meters only) for a given games. Relevant variables retained for analysis are: year (competition year), stroke (categorical), gender (categorical), athlete (name), results (race time in seconds), and team_std (standardized team code). The raw dataset originally included Location, dist_m, Team, Rank, and time_period; these were used for filtering/cleaning and then dropped from the modeling table. Times (results) are official Olympic results in seconds. Potential limitations include (i) changes in competitive context across eras, (ii) policy/technology/regulations changes that can induce structural shifts, and (iii) historical team-code changes; I mitigated the last via team_std. Because the dataset includes only the 100m distance for backstroke, conclusions are specific to that event. Comparable or validating sources include official Olympic/World Aquatics results archives and reputable third-party compilations. These could enrich covariates in future work.

Methods

The observational unit is a single Olympic Games for the USA in the Men's 100 m Backstroke. From the raw swimmer-level file, I restricted to stroke equal to only "Backstroke" and gender is only equal to "Men", then aggregated to one record per Olympic year by taking the best USA time per games:

$$\text{best\_time}_t = \min\{\text{Results}_i : \text{year}_i = t,\ \text{team\_std}_i = \text{USA}\}.$$

The analysis dataset therefore has two columns: year (calendar year of the Games) and best_time (seconds). Team codes were standardized in preprocessing (e.g., ROC/URS/EUN → RUS; FRG/GDR → GER) in a separate cleaning step; only USA rows are used here.

Simple Linear Regression

I fit a simple linear regression (SLR) of time on year:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Model notation: - Y_t — best USA final time (seconds) in Olympic year t. - X_t — calendar year t. - β_0 — intercept (expected time when $X_t = 0$; typically not directly interpretable). - β_1 — average change in seconds per calendar year; $\beta_1 < 0$ indicates improving (faster) times, $\beta_1 > 0$ indicates slowing. - ε_t — mean-zero, homoscedastic, independent error term.

Assumptions and diagnostic checks

SLR inference relies on: - **Linearity:**

$$\mathbb{E}[Y_t \mid X_t] \;=\; \beta_0 + \beta_1 X_t.$$

- **Independence:** Error terms ε_t are independent across Games.

- **Homoscedasticity:** Constant error variance across years (i.e., $\mathrm{Var}(\varepsilon_t) = \sigma^2$).
- **Approximate normality:** Residuals are approximately normal (primarily important for small-sample inference).

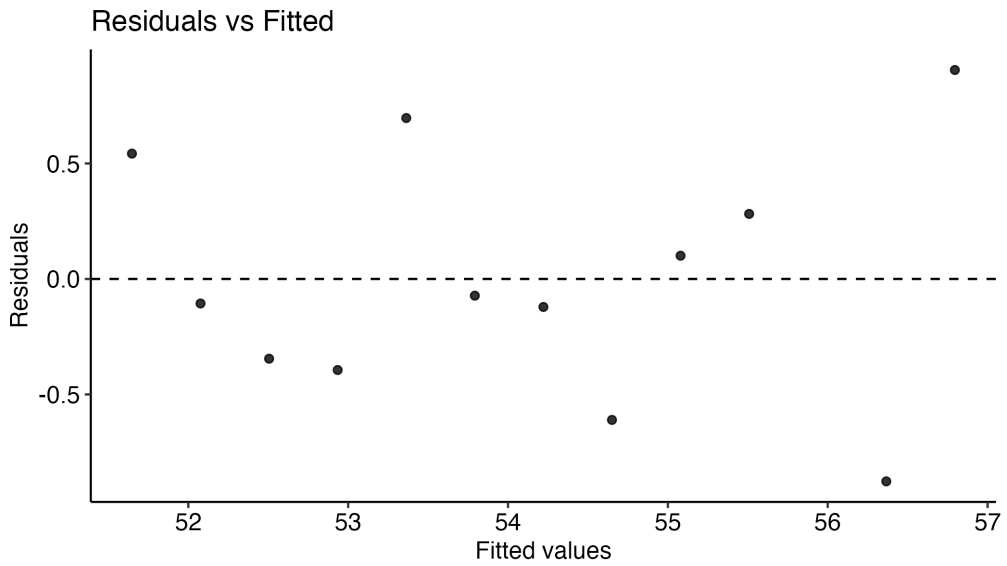

Figure 1: Residuals vs Fitted — USA Men's 100 m Backstroke SLR (best_time ~ year)

When checking, I got a roughly horizontal, band-shaped residuals-vs-fitted plot supports linearity and homoscedasticity. Moreover, a QQ-plot close to the line supports normality.

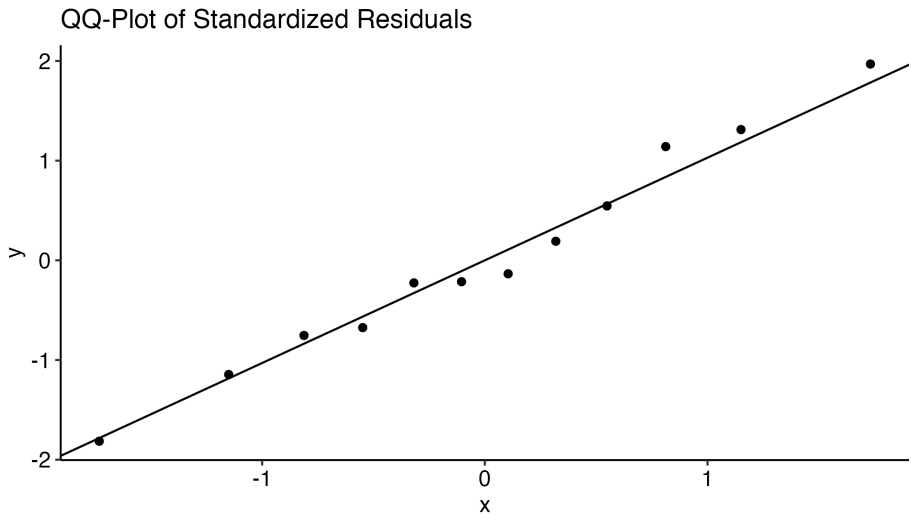

Figure 2: QQ plot comparing standardized residuals to a normal distribution for the model best_time ~ year.

However, independence is plausible because games are spaced four years apart, but mild serial dependence is possible, hence, further investigation will be needed.

Model Selection

The research question targets the average linear trend over time for a single event and team, so SLR is the minimal, directly interpretative model. I chose "best USA time per games" to maintain a consistent performance target across years.

Limitation

- **Event Scope:** conclusions apply to Men's 100 m Backstroke and could not be generalizes to other strokes/distances without additional supporting evidence.
- **Unobserved Covariates:** changes in suit regulations, training, selection, and pool technology could influence improvement performance making a structural break not captured by the current linear trend.
- **Aggregation Choice:** using the minimum (best) USA time focuses on peak performance but is sensitive to single-swimmer outcomes. So, for future analysis is better to focus on the median-based sensitivity checks help address this issue. Moreover, I did not

check is it the same or different swimmers performing in different Olympic games which should be adressed in future analysis.

- **Small N:** the number of Olympic cycles is limited and the number of best performing men per Olympiad is limited to one creating a very small sample size; inference is correspondingly less precise, and larger sample size would improve the accuracy of the analysis.

Software

All analyses were conducted in R with dplyr for wrangling, ggplot2 + ggpubr for graphics, broom for model augmentations, and (optionally) sandwich/lmtest for robust inference.

Results

I modeled the USA Men's 100 m Backstroke best time per Olympiad as a linear function of year:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t,$$

Fitted model. Using OLS on (n=12) Olympic years, the fitted line is

$$\widehat{Y}_t = 268.296 - 0.1073\, X_t$$

with residual SE = 0.560 (s) and $R^2 = 0.906$ (adjusted $R^2 = 0.897$). The slope is negative and statistically significant t = -9.83, p = 1.85×10^{-6}, indicating faster times over time.

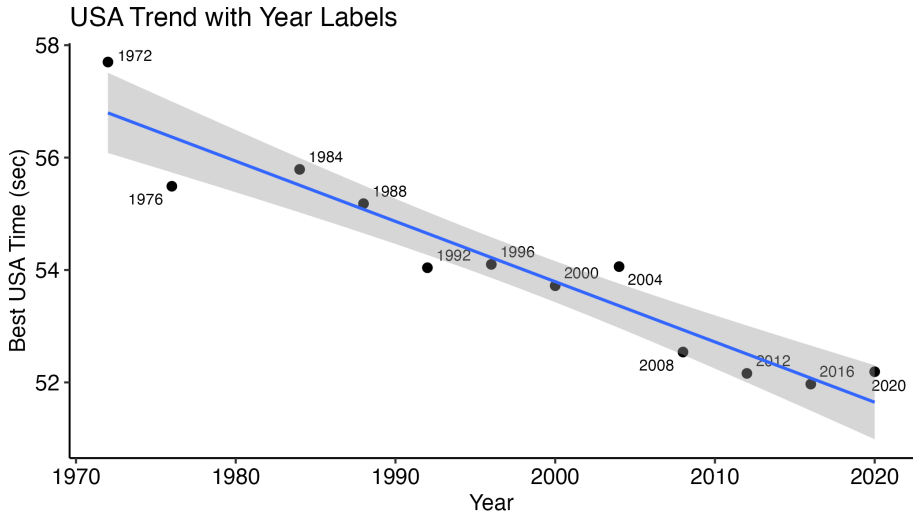

Figure 3: USA Men's 100m Backstroke: Best Time per Olympics

Effect size. The estimated trend is −0.107 seconds per calendar year. Interpreted per Olympic cycle (~4 years), that is about −0.43 seconds per games. Over a decade, the model implies an average improvement of ~1.07 seconds.

Confidence interval. A 95% CI for the slope is [−0.132, −0.083] seconds per year (computed from the reported SE = 0.0109 and (df=10)).

Coefficient table.

Term	Estimate	Std. Error	t value	Pr(>|t|)	95% CI
(Intercept)	268.296	21.785	12.315	2.29e-07	—
year	-0.1073	0.0109	-9.833	1.85e-06	[-0.1316, -0.0829]

Implications for the research question.
- *By how many seconds per year does the best USA time change?* It changes for approximately 0.107 seconds faster per year.
- *Is the trend overall positive or negative?* The trend overall is negative. So, the time is improving over the years, large in magnitude, and highly significant. Hence, this suggests a sustained long-run improvement in peak USA performance in the Men's 100 meters Backstroke across the observed Olympic years.

References

- Gao, P. *MATH261A Project Template (GitHub).* Retrieved September 24, 2025, from https://github.com/peteragao/MATH261A-project-template/tree/main
- SCORE Sports Data Repository. (2024, June 13). *Olympic Swimming.* Retrieved September 24, 2025, from https://data.scorenetwork.org/swimming/olympic_swimming.html
- International Olympic Committee. *Olympic Results Database.* Retrieved September 24, 2025, from https://www.olympics.com/en/olympic-games/olympic-results
- Olympedia. *Countries (National Olympic Committees).* Retrieved September 24, 2025, from https://www.olympedia.org/countries