ISYE6416 - Computational Statistics - Spring 2017
"Recovering the Real World Records in Men's Swimming"
Final Report

Zhenyu Gao, Yixing Li and Zhengxin Wang

April 16, 2017

Contents

List of Figures

List of Tables

1 Problem Statement

Swimming, a competitive sport, emerged in the 1830s in England and was an important part of the Olympics history since the first modern Olympic Games. Thanks to the development of technologies and swimming techniques, time records of all swimming events has been improved in the past 30 years. For example, in the 100m Freestyle, the world record went from more than 49 seconds in the 1990s to today's less than 47 seconds. However, the introduction of Non-textile Full-body Swimsuit (as <Full-body Swim Suit> in the rest of the report), a high-tech swimsuit that can help reducing the skin friction under the water, seemed to have disturbed the improvement curve of human athletes. As indicated in Figure 1, the Full-body Swimsuit basically covers the whole body except for the arms and neck, yet the standard swimsuit only covers the lower part of the body above knees.

Figure 1: Full-body swimsuit (left) and standard swim suit (right)

During the years that the Full-body Swimsuit was used in Men's swimming, results in the world-class competitions had been improved to some extent. The suit was widely used since the year 2007 and was banned in 2010.

Results of two top swimming events were affected by the Full-body Swimsuit: the 2009

World Championships and the 2008 Olympics. Many World Records were created during these two years, and nobody was able to surpass the winners' times in those two events since the ban of the Full-body Swimsuit in 2010. However, World Records created in the Full-body Swimsuit Era still count. For example, the current World Record of 100m Freestyle (46.91 s) was also created in 2009 in the Full-body Swimsuit Era and had since become a huge mountain for recent athletes to climb over. As a reference, even with further improvements in swimming techniques in recent years, fastest time every year ranged from 47.04 s to 47.65 s. **The objective of this project is to explore the effect brought by the Full-body Swimsuit and recover the real world records in several events created in the Full-body Swims Suit era. We want to know the real best records of human athletes without the Full-body Swim Suit**

2 Data Source

All the data are open and can be found on Wikipedia and official websites. Our study main focused on results from the two major competitions in the past 10 years. For each competition, results of all eight swimmers in the final races were used. These results were treated equally, based on the assumption that athletes had equivalent desire for the championship, whether it is the Olympics or World Championships. Data for other swimming events were not considered, as the range of participants are limited or events themselves are not appealing enough.

3 Methodology

One good aspect of this topic is that we have the chance to use various methods taught in the class to solve the real world problem. Three methods in the Computational Statistics class were used here: **Spline**, **Logistic Regression & Newton's Method** and **EM Algorithm**. Also, useful methods in other Statistics class were used, including **ARIMA** (in Time Series) and **ANOVA** (in Design of Experiments). Methodology flowchart of this project is shown in Figure 2, and the set-ups of all these methods will be introduced in details in their respective parts.

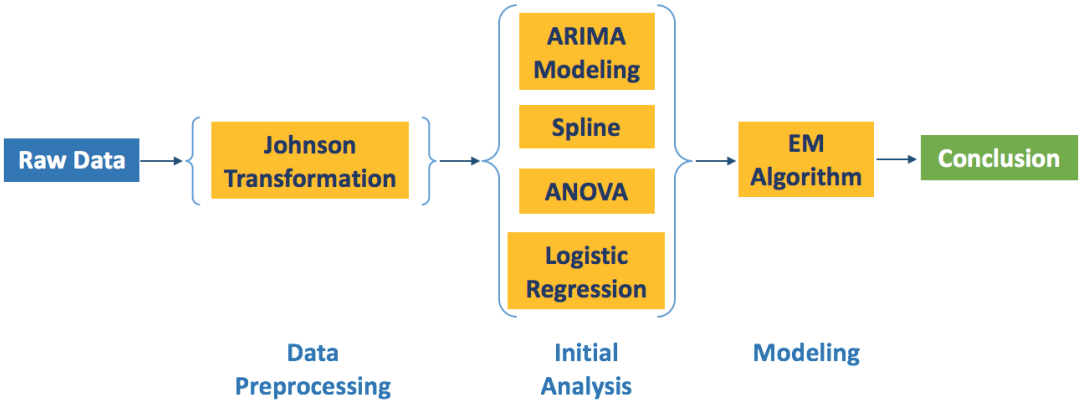

Figure 2: Methodology Flowchart

4 Data Preprocessing

Since later in the EM algorithm, data are required to be normally distributed, **Johnson Transformation** was used to ensure that all groups of data are distributed normally. In this step, data transformation function of Minitab was used. Among all six groups of data (Three events, with or without Full-Body Swim Suit), five of them already follow the normal distribution such that no Johnson Transformation was performed. The sixth group showed slight deviation from normal distribution. Considering the fact that all other groups of data are normally distributed, we deemed that the slight deviation in the sixth group was negligible and the data are good to proceed.

5 Initial Analysis

5.1 ANOVA

A one-way ANOVA was conducted to compare the effect of the Full-Body Swim Suit on time in <With Full-Body Swim Suit> and <Without Full-Body Swim Suit> conditions. With the small p values shown below, we can claim that there was a significant effect of the Full-Body Swim Suit on time at the $p < 0.05$ level in every stroke. In other words, wearing a Full-Body Swim Suit indeed significantly improved the results. Results of the ANOVA are shown in Table 1.

Table 1: ANOVA Table

	Source	SS	df	MS	F	p
100m Freestyle	Swimsuit	2.65	1	2.65	22.81	0
	Error	3.486	30	0.116		
	Total	6.135	31			
50m Freestyle	Swimsuit	1.596	1	1.596	43.3	0
	Error	1.106	30	0.037		
	Total	2.701	31			
4*100m Relay	Swimsuit	48.472	1	48.472	19.37	0
	Error	75.066	30	2.502		
	Total	123.538	31			

5.2 ARIMA Modeling

For ARIMA method, a new dataset was employed. This dataset contains best time records in final races for Olympic games and World Championships. For 100m Freestyle, it follows ARIMA (0, 1, 0). Based on current ARIMA model, we used Kalman Filter to handle missing values (here, records in year 2008 and 2009, the Full-Body Swim Suit Era were treated as missing data). We took state space form of ARIMA model from the output returned by ARIMA and passed it to Kalman. Based on results before 2008 and after 2009, figures are plotted below. For year 2008 and 2009, real results with Full-Body Swim Suit are given in red, while estimated results given by ARIMA and Kalman Filter are given in blue, as shown in Figure 3. Estimated values are much larger than the real data, showing the huge impact of Full-Body Swim Suit on the swimming results.

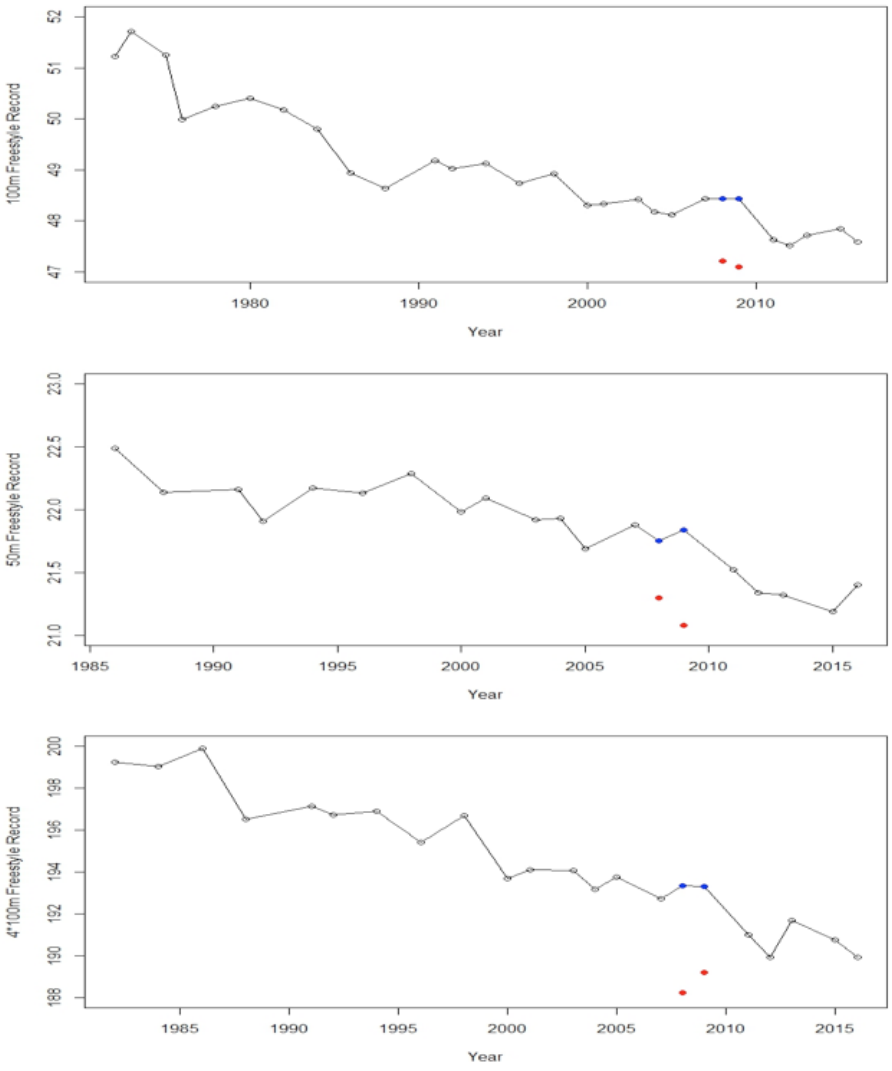

Figure 3: ARIMA Results for the Three Events

5.3 Logistic Regression

Logistic Regression was also used to identify if athletes swim faster with the Full-Body Swim Suit. Since the dependent variable needs to be binary in logistic regression, a term "World's Top-level Results" was used on the finishing times and the standards for the term are shown in Table 2:

Table 2: Standards for Top-level Results

	50m Freestyle	**100m Freestyle**	**4*100m Relay**
Top Level (1)	$<$ 21.5s	$<$ 48s	$<$ 192s
Non Top Level (0)	$\geq$ 21.5s	$\geq$ 48s	$\geq$ 192s

After the standards were set, **Logistic Regression** was used in all three events and **Newton's Method** was used to optimize the log-likelihood function to fit the model. Plots for the values of the log likelihood function versus iterations are in Figure 4:

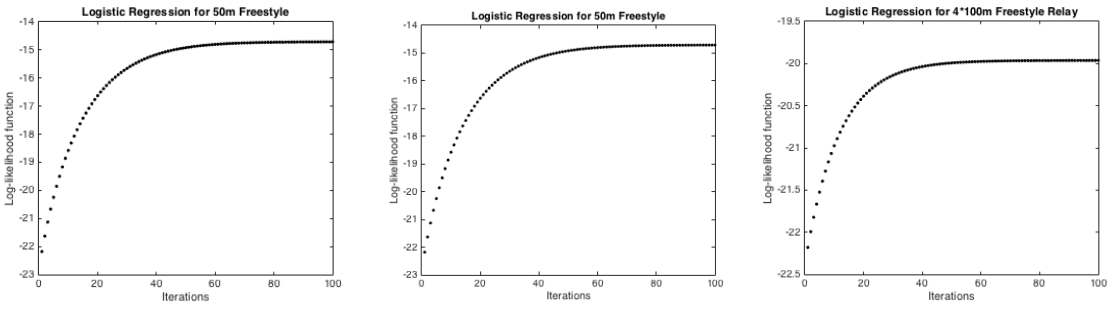

Figure 4: Logistic Regression: values of the log likelihood function vs. iterations

After the modeling for all the data we have, we went to find the chances of reaching the Top-level Results with and without Full-Body Swim Suit (FB). As can be seen in Table 3, when swimmers wear the Full-Body Swim Suit (FB), they got much higher chances to swim the Top-level Results, which again confirms that the Full-Body Swim Suit can improve swimmers' speed.

Table 3: Chances of Reaching Top-level Results

	100m Freestyle		**50m Freestyle**		**4*100m Relay**	
Swim Suit	FB	Normal	FB	Normal	FB	Normal
Top Level Chance (%)	43.81	6.70	68.56	12.88	56.19	25.25

5.4 Spline

Knowing the fact that swimmers' techniques had been improved year after year, we used Splines to model the improvement curves for all the three events in order to compare results from different years at the same level. In this study, best yearly records for the three events were believed to best represent improvements of human beings, and were used to build models. Since our data went across a time span from 2008 to 2012, representative results for those four years were obtained, using 2008 as a base. The differences in Table 4 were used to correct results from the year 2009, 2011 and 2012. After the correction, improvement effects of human swimmers during the time span can no longer affect the following analyses.

Table 4: Spline Predictions of Improvements

	50m Freestyle	**Delta**	**100m Freestyle**	**Delta**	**4*100m Relay**	**Delta**
2008	21.671 s		48.032 s		192.493 s	
2009	21.634 s	-0.037 s	47.966 s	-0.066 s	192.214 s	-0.279 s
2011	21.560 s	-0.111 s	47.834 s	-0.198 s	191.657 s	-0.796 s
2012	21.523 s	-0.148 s	47.768 s	-0.264 s	191.377 s	-1.116 s

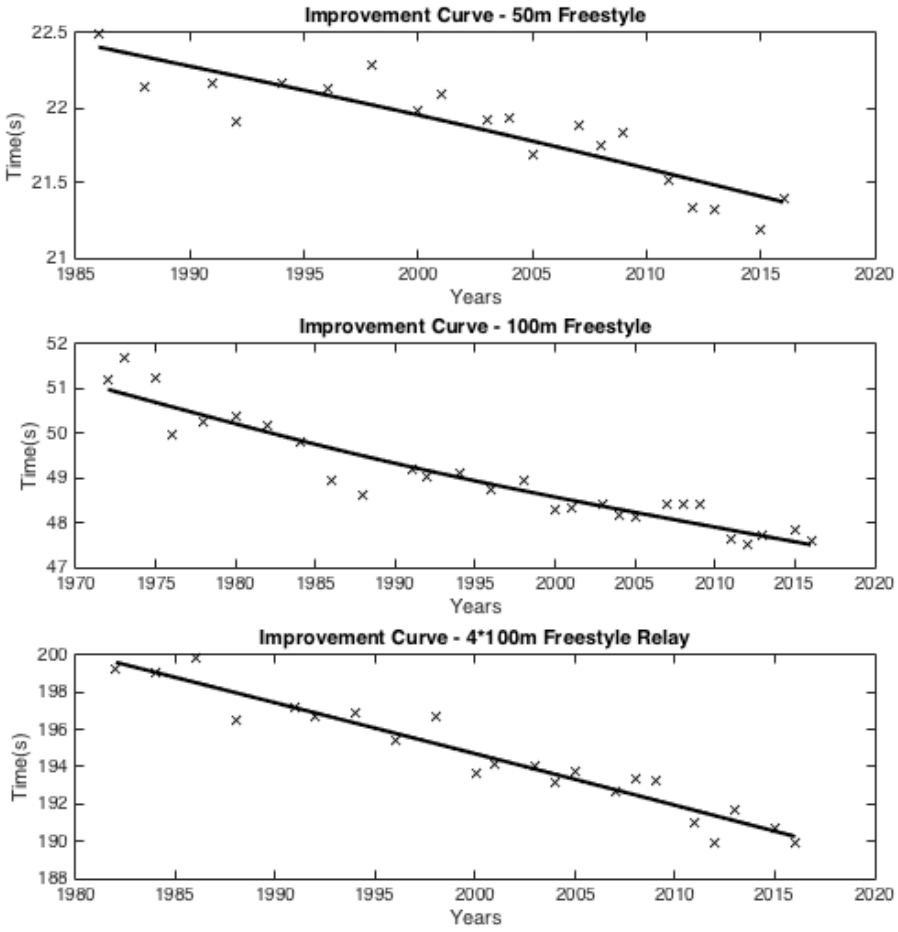

Figure 5: Spline Regression for the Best Yearly Records for All Three Events

6 EM Algorithm

In order to circumvent influence of the Full-Body Swim Suit (FB) and study the "true" world records in 50m Freestyle, 100m Freestyle and 4*100m Freestyle Relay, **EM Algorithm** was finally used. And the set-up is described below:

True times for all athletes in the finals can be treated as $T \sim \mathcal{N}(u_t, \sigma_t^2)$
Swimsuit bias can be treated as $D \sim \mathcal{N}(v_d, \tau_d^2)$
times with Full-Body Swim Suit will be $R|T + D \sim \mathcal{N}(T + D, \sigma^2)$

Final form for the E-step is:

$$u_{td} = \begin{bmatrix} u_{td,1} \\ u_{td,2} \end{bmatrix} = \begin{bmatrix} u_t + \dfrac{\sigma_t^2}{\sigma^2 + \sigma_t^2 + \tau_d^2}(x^{td} - u_t - v_d) \\ v_d + \dfrac{\tau_d^2}{\sigma^2 + \sigma_t^2 + \tau_d^2}(x^{td} - u_t - v_d) \end{bmatrix} \tag{1}$$

$$\Sigma_{td} = \begin{bmatrix} \Sigma_{td,11} & \Sigma_{td,12} \\ \Sigma_{td,21} & \Sigma_{td,22} \end{bmatrix} = \frac{1}{\sigma^2 + \sigma_t^2 + \tau_d^2} \begin{bmatrix} \sigma_t^2(\tau_d^2 + \sigma^2) & -\sigma_t^2\tau_d^2 \\ -\sigma_t^2\tau_d^2 & \tau_d^2(\sigma_t^2 + \sigma^2) \end{bmatrix} \tag{2}$$

Final form for the M-step is:

$$u_t = \frac{1}{D}\sum_{d=1}^{D} u_{td,1} \tag{3}$$

$$v_d = \frac{1}{T}\sum_{t=1}^{T} u_{td,2} \tag{4}$$

$$\sigma_t^2 = \frac{1}{D}\sum_{d=1}^{D}(\Sigma_{td,11} + u_{td,1}^2 - 2u_{td,1}u_t + u_t^2) \tag{5}$$

$$\tau_r^2 = \frac{1}{T}\sum_{t=1}^{T}(\Sigma_{td,22} + u_{td,2}^2 - 2u_{td,2}v_d + v_d^2) \tag{6}$$

And here, we primarily care about the Swimsuit bias, which is the $D \sim \mathcal{N}(v_d, \tau_d^2)$. Finally, the EM algorithm gives the result as shown in Table 5.

Table 5: EM Algorithm Results

	50m Freestyle		**100m Freestyle**		**4*100m Relay**	
	With FB	**W/o FB**	**With FB**	**W/o FB**	**With FB**	**W/o FB**
v_d	0.2809	0.7191	0.2253	0.7746	-0.6449	1.6314
τ_d	0.0425	0.0427	0.0690	0.0656	0.0177	0.1076
v_d **Diff**	0.4382		0.5493		2.2763	

7 Final Results

After we have the final results from EM Algorithm, we can recover the true world records of the three events without the influence of Full-Body Swim Suit. The current world records were all created in the Full-Body Swim Suit Era, and shown in Table 6.

Table 6: Current World Records

	50m Freestyle	**100m Freestyle**	**4*100m Relay**
Time (s or min)	20.91	46.91	3:08.24
Year	2009	2009	2008

And the following formula had been used to do the correction:

Real World Record = Current World Record + FB Swim Suit Bias + Year Correction

Finally, we have the recovered World Records as follows (round up to 2 decimal places):

Real World Record - 50m Freestyle = 20.91s + 0.4382s - 0.037s = **21.31s**

Real World Record - 100m Freestyle = 46.91s + 0.5493s - 0.066s = **47.39s**

Real World Record - 4*100m Relay = 3:08.24 + 2.2763s - 0 = **3:10.52**

Any swimmer or team who swam faster than the records above, and maintained the fastest since then, should own the World Record.

8 Work Breakdown

All the three team members contributed equally in this project as of ideas and workloads.