

# Using Linear Mixed Effects Models to Model Performance in Male High Jumpers

Jack Andrew

A thesis submitted to the School of Mathematical and Statistical Sciences, National  
University of Ireland, Galway for the degree of Masters in Biostatistics



Supervisor: Prof. John Newell

Date: October 2021

## Table of Contents

Table of Contents .....	i
Figures.....	iii
Tables.....	iv
Declaration .....	v
Acknowledgements .....	vi
Abstract.....	vii
1 Chapter 1: Introduction and Background .....	1
1.1 Introduction.....	1
1.2 Project Work packages .....	2
2 Chapter 2: Acquire .....	4
2.1 Introduction.....	4
2.2 Data Overview .....	4
2.3 Data Collection.....	4
2.3.1 Data Scraping Process: Step One (Athlete ID) .....	6
2.3.2 Data Processing: Step Two (Performance Data) .....	7
2.4 Data Cleaning and Storage .....	8
2.5 Chapter Summary.....	8
3 Chapter 3: Visualise .....	9
3.1 Introduction.....	9
3.2 Data Hierarchy.....	9
3.3 Response and Explanatory variables.....	11
3.3.1 Overview .....	11
3.3.2 Summarising the Response Variable.....	12
3.3.3 Summarising the Explanatory Variables .....	13
3.4 EDA Summary .....	18
3.4.1 Interactive EDA App .....	18
3.5 Chapter Summary.....	20
4 Modelling the Relationship between Age and High Jump Performance .....	21
4.1 Introduction.....	21
4.1.1 Linear Mixed Models .....	21
4.1.2 Smoothing Methods .....	24
4.1.3 Model Fitting Workflow .....	35

4.1.4	Model Comparisons .....	36
4.1.5	Model Selection and Performance .....	37
4.1.6	Model Fitting Process.....	38
4.1.7	Model Comparison Summary.....	48
4.1.8	Random Effect Plots.....	48
4.1.9	Plots of the Fitted Data .....	49
4.2	Modelling Summary.....	50
5	Chapter 5: Using the Model to Predict future Performance.....	52
5.1	Introduction.....	52
5.2	Prediction for Athletes Currently in the Model.....	52
5.3	Sequential Predictions for a New Athlete .....	53
6	Conclusion and Further Works .....	56
6.1	Project Limitations.....	57
6.2	Practical Applications .....	58
6.3	Future Research .....	59
6.4	Conclusion .....	60
7	References.....	61

## Figures

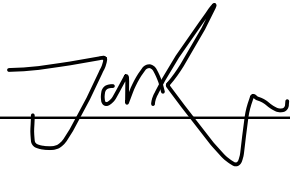
Figure 1.1: An Olympic High Jump Athlete.....	2
Figure 2.1: Chrome HTML inspector on the ranking page with highlighted cell within webpage and HTML code.....	6
Figure 2.2: Red - URL; Blue - Athlete ID; Green - Athlete Results .....	7
Figure 2.3: Chrome HTML inspector on the profile page with highlighted cell within webpage and HTML code.....	8
Figure 3.1: Density plot of High Jump Performance.....	13
Figure 3.2: Distribution of Age.....	13
Figure 3.3: Boxplot of Performance by Age Group.....	14
Figure 3.4: Non-linear relationship between age and performance (left) and repeated measurements per athlete (right). .....	15
Figure 3.5: Varying performances by year (left) and month (right).....	16
Figure 3.6: Varying performance by country (left) and venue (right). .....	17
Figure 3.7: Variation in performance by Data Source. ....	17
Figure 3.8: Performance distributions by age bin of athletes with a lifetime PB of over 2.00m. ....	18
Figure 3.9: Percentage improvement between age bin of athletes with a PB over 1.00m..	19
Figure 3.10: Example of an athlete overview and bench marked against percentile ranks for age bin. ....	19
Figure 3.11: Chance and conditional chance of performance improvement given current performance level. ....	20
Figure 4.1: Polynomial fits of Age vs Performance. ....	26
Figure 4.2: Step function fit of Age vs Performance. ....	27
Figure 4.3: Performance progressions for a sample of athletes. ....	29
Figure 4.4: Individual step functions fitted to a sample of athletes. ....	30
Figure 4.5: Linear regression and polynomial models fitted to a sample of athletes.....	30
Figure 4.6: Polynomials fitted to a sample of athletes.....	31
Figure 4.7: B-splines of 3 and 5 degrees of freedom fitted to a sample of athletes.....	31
Figure 4.8: B-splines of 3 and 10 degrees of freedom fitted to a sample of athletes. ....	32
Figure 4.9: B-splines with 3 through to 6 degrees of freedom.....	33
Figure 4.10: Example of overfitting with B-spline. ....	33
Figure 4.11: Natural splines fitted to a sample of athletes. ....	34
Figure 4.12: P-splines with varying lambda penalties fitted to sample of athletes. ....	34
Figure 4.13: Diagnostic plots for model 1 and model 2. ....	39
Figure 4.14: Diagnostic plots for model 3.....	40
Figure 4.15: Diagnostic plots for models 4, 5 and 6. ....	42
Figure 4.16: Diagnostic plots for model 7.....	43
Figure 4.17: Diagnostic plots for model 8.....	45
Figure 4.18: Diagnostic plots for model 9.....	46
Figure 4.19: Plot of random effects across b-spline terms. ....	49
Figure 4.20: Venue, year, and country random effects plot.....	49
Figure 4.21: Smoothed data vs fitted data.....	50
Figure 5.1: Model with prediction interval for a selection of athletes. ....	53
Figure 5.2: Change in model as new data is added. ....	55

## Tables

Table 3.1: Hierarchical structure 1 .....	9
Table 3.2: Example of hierarchical structure 1. ....	10
Table 3.3: Hierarchical structure 2. ....	10
Table 3.4: Example of hierarchical structure 2 .....	10
Table 3.5: Hierarchical structure 3 .....	11
Table 3.6: Example of hierarchical structure 3 .....	11
Table 3.7: Variable description and information.....	12
Table 3.8: Performance summary. ....	13
Table 3.9: Age summary. ....	13
Table 3.10: Observations per athlete. ....	15
Table 4.1: Summary of 2nd degree polynomial fit. ....	25
Table 4.2: Summary of 3rd degree polynomial fit.....	25
Table 4.3: Polynomial fit model metrics. ....	26
Table 4.4: Model 1 and model 2 model metrics. ....	39
Table 4.5: Unexplained variance accounted for in Model 3.....	40
Table 4.6: Model 3 metrics. ....	41
Table 4.7: Model 3 fixed effects summary.....	41
Table 4.8: Model 4 to Model 6 metrics.....	42
Table 4.9: Model 7 unexplained variance accounted for. ....	44
Table 4.10: Variance explained by components of the random effects in Model 8.....	45
Table 4.11: Model 9 and model 9 variations metrics.....	46
Table 4.12: Variance explained by components of the random effects in Model 9.....	47
Table 4.13: Model 1 to Model 9 metrics.....	48
Table 5.1: Point estimate and prediction intervals for an athlete.....	53
Table 5.2: Initial estimate from final model for 18-year-old athlete. ....	54
Table 5.3: Point estimates and prediction intervals for sequential data.....	54

## Declaration

I declare that the present thesis is a record of my own work and has been completed by myself with full acknowledgement to the individual and institute in which assisted in the research. I have not obtained a degree in NUI Galway, or elsewhere, on the basis of the work described in the thesis.

A handwritten signature in black ink, appearing to read 'Jack', is written over a horizontal line.

Jack Andrew

## Acknowledgements

The main aims of the project were to build an athlete performance database and then build an appropriate statistical model for the modelling of an athlete's high jump performance over time.

First, I would like to thank the tutors of the statistical modules that I was enrolled in prior to commencing this project. The MSc project allowed me to apply specific learnings from these modules.

I have received a great deal of support and assistance throughout the MSc project of which I am extremely grateful for. I would first like to thank my supervisor, Professor John Newell, whose expertise was invaluable in formulating the research question and methodology. Your insightful feedback pushed me to sharpen my statistical thinking and brought my statistical knowledge and application to a higher level.

I would also like to acknowledge my friend and line manager Dr. Kenneth McMillan for his help in creating the opportunity to pursue this MSc and for his continual guidance and support throughout my studies and final project.

Finally, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me.

Jack Andrew

7<sup>th</sup> October 2021

## Abstract

Identifying Youth athletes who have the potential to excel in Senior Athletics Championships is of great importance for Athletic coaches, Federations and National Government Organisations worldwide. Several online athlete performance databases exist where Youth to Senior athlete performance records are recorded, albeit in different formats to each other.

The main aims of this project were to, i) create a database of athlete performance records from these multiple data sources, and ii) utilise this harvested data to build an appropriate statistical model to model an athlete's performance in the high jump over time.

Data were harvested from online sources using web scraping techniques. An interactive application was built to explore and visualise these data. Several approaches to modelling the data were considered, with Linear Mixed Models being selected as an appropriate model of choice. Several model metrics were compared to select an appropriate model.

The results from this project suggest that Linear Mixed Models, incorporating B-splines at both the random and fixed effect level, are appropriate models for modelling the relationship between performance in the high jump over time.

The models presented would benefit from having additional covariates such as temperature, humidity weather and competition category as including these are likely to improve the model fit further.



# 1 Chapter 1: Introduction and Background

## 1.1 Introduction.

Athletic coaches, Federations and National Government Organisations (NGO's) are always searching for ways to model potential future performance of their Youth athletes. Experienced coaches draw on their experience to estimate a reasonable performance prediction. However, Federations and NGOs that are usually held accountable by boards of business-orientated directors need to be more scientific and robust in their methods for quantifying talent and predicting future performance.

Budget-constrained organisations need to be able to identify athletes who have the best possible chance of making future Senior championships. This identification ensures the efficient allocation of financial support to athletes with the highest potential for Senior athletics success.

The main aims of this project were to:

- i) compile a database of athlete performance records,
- ii) create an intuitive web-based application that allows athletics coaches to explore, visualise and summarise the athlete performance data in a useful way, and
- iii) build an appropriate statistical model to model an athlete's performance in the high jump over time.

The methodology presented here is applicable to all gender identities and all athletic sports. As the target population of interest in this project is an all-boys sports academy, data only relating to boys were analysed and modelled. The high jump event was chosen as the main sport of interest. However, summary statistics and visualisations of the performance distributions for the other athletics events that data was collected on are included in the web-based application that was created.

A variety of statistical modelling techniques were progressively evaluated in terms of their ability to correctly model high jump performance over time, with the most suitable model identified.



*Figure 1.1: An Olympic High Jump Athlete.*

## 1.2 Project Work packages

The project consisted of four work packages: acquire, visualise, model, and communicate. A brief outline of each work package is provided below.

**Acquire:** This section describes the web scraping process used to collect the data and provides the reader with a contextual overview of the dataset. A description of how the dataset was cleaned and stored is also provided.

**Visualise:** The first part of the Visualise work package involved carrying out an extensive Exploratory Data Analysis (EDA) to provide a visual overview of the key covariates and the structure of the datasets. There was an emphasis on subjective use of the data in this stage given the extensive domain knowledge available. An R Shiny application was created that allows the user to explore graphical and numerical summaries of certain covariates. This application also gives empirical estimates of performance improvement probabilities. The Visualise work package section concludes with a summary of the EDA findings which informs the decision-making processes needed for the ‘model’ work package.

**Model:** The model work package focuses primarily on ‘objective use’ of the data. A brief overview of potential statistical modelling methodologies is outlined before a more in-depth overview of the method of choice, Linear Mixed Models (LMM), is presented. Models were fitted with increasing levels of complexity and each time their summaries and diagnostic

plots were reviewed. To conclude, an illustrative example of how the model would work with unseen data is provided.

Communicate: The project report concludes with discussions on the main findings of the project and the overall conclusions. Limitations and practical applications of the project are also acknowledged, and recommendations for future work in the field are provided.

## 2 Chapter 2: Acquire

### 2.1 Introduction

In this chapter the datasets that were used for modelling high jump performance are introduced. A description of the variables contained in each dataset are presented. Details on how the data were scraped, cleaned, tidied, and stored are also presented.

### 2.2 Data Overview

Longitudinal data were collected from three on-line databases on athletes who had competed in the high jump event. In total, 248,151 results were collected for 26,045 athletes in 145 countries across 5 Age Groups (Under 13 to Senior level).

The primary area of interest of this project was to model the change in performance over time where time is measured using athlete age at the event in question.

### 2.3 Data Collection

Two of the three datasets are contained on publicly available resources and therefore no informed consent was needed. The third database belonged to an established sports youth academy. The two publicly available resources were the UKA Power of Ten (PoT) ranking database (<https://www.thepowerof10.info/>) and the World Athletics (WA) results database (<https://www.worldathletics.org/>).

The PoT database provides all historical results for every British athlete since 2006. Athletes who made the all-time lists prior to 2006 have an incomplete performance history but have records of their results at key athletic competitions.

The WA database contains worldwide results for athletes who have attained a designated standard at an international level. Therefore, the WA dataset is not as extensive as the PoT dataset in covering the full range of an athlete's performance history, but it contains a greater number of athletic performance data records.

The third database was the Aspire Academy (AA) database which contains all competition results of student athletes who have attended the academy since 2014. The database

contains performance information from both current and former student-athletes, as well as results from higher-level athletes that have since graduated from AA. The AA dataset was obtained via a simple no-code query using the reporting module of the database it is housed in.

The PoT and WA datasets were scraped from their respective websites using the `rvest` and `curl` packages belonging to the R programming language. The `dplyr` package was also used to clean and gather the dataset into a data frame.

Websites are built using HyperText Markup Language (HTML) which is the code that defines the meaning and structure of a webpage and its content. To access data held within websites, a HTML file can be read using the `read_xml` function in R via the uniform resource locator (URL). Using additional functions belonging to the `rvest` and `rcurl` R packages, it is possible to manipulate this HTML file to locate and extract information of interest.

The data in the PoT website were not written into the HTML code on each single page but were stored in a database hosted elsewhere. The HTML is used to create a template for displaying the data and a language called PHP (Hypertext Pre-processor) is used to retrieve the data from the database and create a dynamic page depending on certain query parameters usually contained in the pages URL. When a specific URL is called during a web-scrape the data are already rendered. When writing web-scraping code in such instances PHP code is not required as the data are available directly in the HTML code. The script then simply accesses the data using the R functions that manipulate the HTML.

Using these principles, an R code script was written to scrape a 'template' page. This script was then programmatically looped to extract data from targeted pages.

A simplified outline of the two-step process of how data were scraped for the PoT site is now presented. The first step involved retrieving a list of 'names' of all the athletes' data to scrape, while the second step involved scraping all the athletes' profile pages.

### 2.3.1 Data Scraping Process: Step One (Athlete ID)

A specific web-scraping script programmed in R first accesses a URL to extract all the athlete names for a specific gender, age group, year, and athletic event. An example URL of a webpage that contains data on performance measures on all age-groups for men competing in the 100m in the year 2020 is given below and in the accompanying screen shot in Figure 2.1.

*<https://www.thepowerof10.info/rankings/rankinglist.aspx?event=100&agegroup=ALL&sex=M&year=2020>*

The web-scraping R script reads in the HTML using specialised functions. Additional functions are then utilised to access various nodes within the page. Google Chrome inspector was used to determine the underlying structure of a webpage and relevant nodes and locate specific parts of nodes where relevant data are contained. The rvest package in R has functions that can extract the data from the node locations of interest and write the data into a data frame. Figure 2.1 shows how Google Chrome inspector was used to highlight the nodes corresponding to the athlete profile link (the athlete ID).



Figure 2.1: Chrome HTML inspector on the ranking page with highlighted cell within webpage and HTML code.

Step one was completed once the code looped through every ranking page for every year, event and age group and extracted the athlete's ID. Duplicates were then removed (i.e. athletes who are ranked over multiple years and events) and this final list was passed into a second R script which was written to extract an athlete's performance results.

### 2.3.2 Data Processing: Step Two (Performance Data)

The goal of Step Two was to extract results data for each of the athletes identified in Step One. Figure 2.2 below gives an example of an athlete's profile data.



Figure 2.2: Red - URL; Blue - Athlete ID; Green - Athlete Results

The information that was required to be harvested is displayed in the green box in Figure 2.2. An R script was written to extract these results (at the athlete level) by accessing the table within the HTML code using its specific page ID tag. The script looped through each row, extracted, and stored the results to a data frame labelled with the relevant athletes' ID.

Due to the vast amount of information stored on the PoT and WA databases, and the subsequent number of loops needed to be performed by the R scripts to harvest the relevant data, the runtime of the R scripts took several hours. To speed up this process, the R scripts were run on a distributed computing network, thereby harnessing parallel processing power.

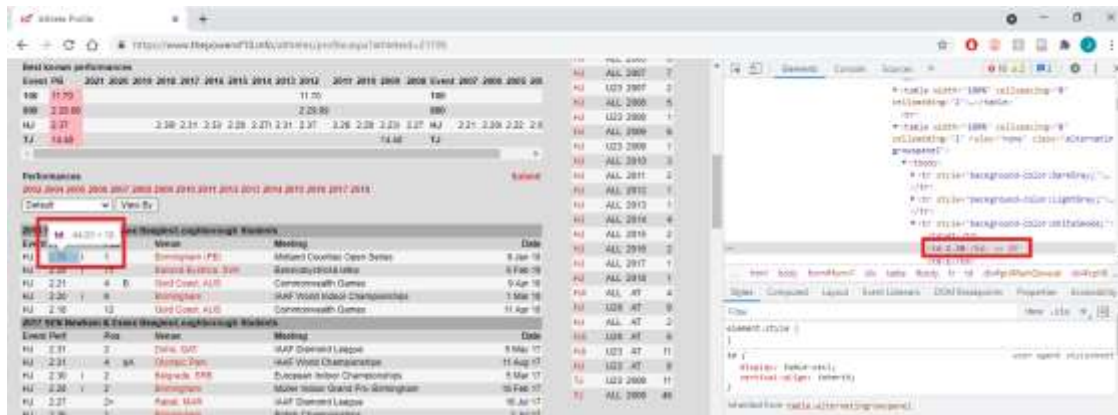


Figure 2.3: Chrome HTML inspector on the profile page with highlighted cell within webpage and HTML code.

## 2.4 Data Cleaning and Storage

The uncleaned data were initially stored as three comma separated value (csv) files - one csv file for each database. All three datasets were cleaned using appropriate R packages. All UK athletes, along with a select number of Qatari Athletes, were removed from the WA dataset to remove any potential duplication within the three datasets. Field names were standardised, and rows that contained missing performance values were removed. The athletes' date of birth was not recorded in the PoT dataset therefore an approximate age at performance was calculated using a simple linear regression model (SLR) where the factor age group was used to predict the age at performance.

After the data cleaning was completed the three datasets were combined, final checks performed, and the data was stored as an R data file (.rds).

## 2.5 Chapter Summary

During the **Acquire** stage a vast amount of data were scraped from online resources, which were then cleaned and stored. Bespoke R functions were written to identify the correct page locations, the relevant nodes within each page and finally to harvest the relevant data. Extensive data cleaning was then carried out to clean and filter the data. Sensible estimates for missing age values were generated by identifying which age bands were missing for an athlete. The next step of the project was to generate suitable numerical and graphical summaries that allowed athletics coaches to explore the data in ways that are most relevant to them.



## 3 Chapter 3: Visualise

### 3.1 Introduction

Given the large number of records of the combined dataset, it could be argued that the dataset is in a sense the population of interest. This would allow the direct computation of population parameters. However, the WA database only contains results of certain performance standards at high-level athletic competitions, meaning that an athlete's performance history may be incomplete. Conversely, the PoT database could be considered the population of all British athletes' performances since 2006 as it contains every result from every competition in the UK for athletes aged 11 years old and upwards. The datasets are constantly expanding (addition of new results) so any analysis can be considered a sample of the population up to that time point.

The first stage in any statistical analysis is to explore the data in detail, both numerically and graphically and communicate any findings. All visualisation and statistical analyses were performed using R version 4.0.3 (1) and all code was written, compiled and communicated using R Markdown under the principles of reproducible research.

### 3.2 Data Hierarchy

The dataset had some clear hierarchical structures which are outlined below. The hierarchical nature dictated the modelling approach which is discussed later in this thesis.

Structure 1 (Table 3.1) shows a hierarchical data structure where each athlete is nested within a single country.

*Table 3.1: Hierarchical structure 1*

Country
Athletes

An example of Structure 1 from the dataset is shown in Table 3.2.

Table 3.2: Example of hierarchical structure 1.

USA	United Kingdom	Russia
Jamie Nieto	Chris Baker	Ilya Ivanyuk
Jesse Williams	Martyn Bernard	Andrey Silnov
...	...	...

Structure 2 (Table 3.3) shows a cross-classified data structure. Each athlete can participate in multiple competitions throughout a season.

Table 3.3: Hierarchical structure 2.

Year
Competition
Athletes

An example of Structure 2 from the dataset is show in Table 3.4.

Table 3.4: Example of hierarchical structure 2

2017	2017	2018
World Championships (London)	Müller Grand Prix (Birmingham)	World Indoor Championships (Birmingham)
Robbie Grabarz	Robbie Grabarz	Robbie Grabarz
Mutaz Barshim	Mutaz Barshim	Mutaz Barshim
Mateusz Przybylko	Mateusz Przybylko	Mateusz Przybylko
...	...	...

Structure 3 (Table 3.5) shows another cross-classified data structure. Each athlete can participate in the same competition but in different seasons (years).

Table 3.5: Hierarchical structure 3

Competition
Year
Athletes

An example of structure 3 from the dataset is show in Table 3.6.

Table 3.6: Example of hierarchical structure 3

Doha Diamond League	Doha Diamond League	Doha Diamond League
2014	2017	2018
Donald Thomas	Donald Thomas	Donald Thomas
Mutaz Barshim	Mutaz Barshim	Mutaz Barshim
Andrii Protsenko	Andrii Protsenko	Andrii Protsenko
...	...	...

### 3.3 Response and Explanatory variables

#### 3.3.1 Overview

Table 3.7 displays the response and explanatory variables of the dataset. There is only one response variable in this project (Numeric\_Performance) and it is a continuous variable by nature. The units of the response variable will depend on the sport under consideration (e.g., seconds for the 100m, minutes for the marathon, and metres for the high jump). All but one of the explanatory variables are categorical, with age at performance (age\_perform) being the exception (continuous in nature).

Table 3.7: Variable description and information

Variable	Role	Type	Description
ID	Explanatory	Categorical	Individual athlete identification number.
Country	Explanatory	Categorical	Athlete nationality.
Year	Explanatory	Categorical	Year of athlete performance.
Date of Event	Explanatory	Categorical	Date of athlete performance.
Event Group	Explanatory	Categorical	Event Group. Jumps.
Event	Explanatory	Categorical	Event. High Jump.
Venue	Explanatory	Categorical	Competition Name
Numeric Performance	Response	Continuous	Athlete performance.
age_perform	Explanatory	Continuous	Athlete age at performance (years).
Age_Group	Explanatory	Categorical	Athlete age group.
whole_age_perform	Explanatory	Continuous	Athlete age rounded down to nearest whole number.
age_bin	Explanatory	Categorical	Age at perform binned into 1-year increments. Used for EDA.

### 3.3.2 Summarising the Response Variable

Table 3.8 shows a summary of the performance response variable. The median is greater than the mean suggesting that the variable has a left skew. This is confirmed in Figure 3.1 which shows a clear left skew in the data. The left skew indicates that there were a greater number of performances at a higher standard than compared to lower standards. The uneven ‘wiggle’ through the centre of the distribution is likely attributable to the greater number of results at 5cm increments (i.e., 1.50m, 1.55m, 1.60m etc.), that are common progressions at lower-level high jump competitions.

Table 3.8: Performance summary.

Event	Minimum	Mean	Median	Maximum
High Jump	0.60	1.75	1.80	2.45

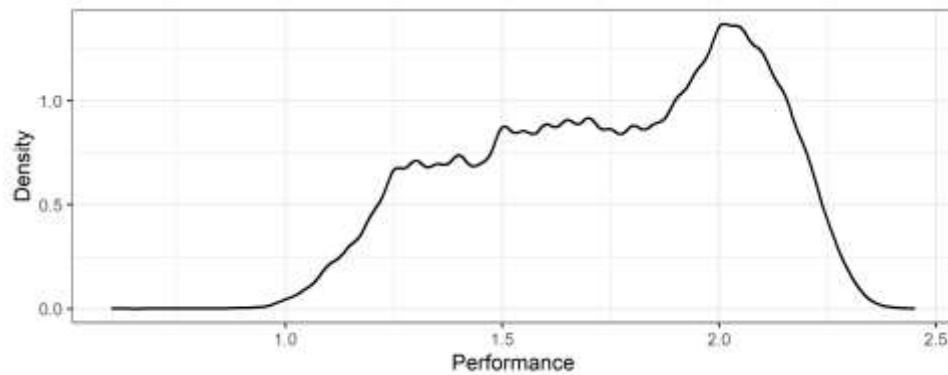


Figure 3.1: Density plot of High Jump Performance.

### 3.3.3 Summarising the Explanatory Variables

Table 3.9 and Figure 3.2 show an overview of the age explanatory variable.

Table 3.9: Age summary.

Event	Minimum	Mean	Median	Maximum
High Jump	9.00	18.38	17.30	48.02

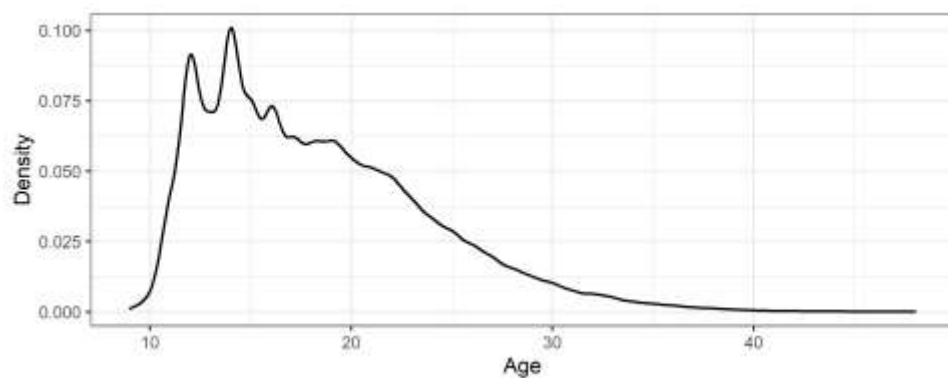


Figure 3.2: Distribution of Age.

The median is less than the mean suggesting that age is right skewed which is confirmed in the density plot (Figure 3.2). This suggests there are more results (performances) at a younger age than at older ages. There are two spikes evident which is likely due to the linear model that was used to estimate age in the PoT dataset giving similar 'age-bin' values for certain athletes with missing age. Overall, there are no implausible outliers in the age covariate.

The performance of elite high jumpers peaks at age 26.0 ( $\pm 2.9$ ) years (2) whilst the performance of super elite high jumpers peaks at an earlier age of 23.2 ( $\pm 2.7$ ) years (3). Data from high jumpers over the age of 30 years were removed and the remaining EDA was carried out on this age-filtered dataset.

Figure 3.3 shows the clear difference in performance across the age groups.

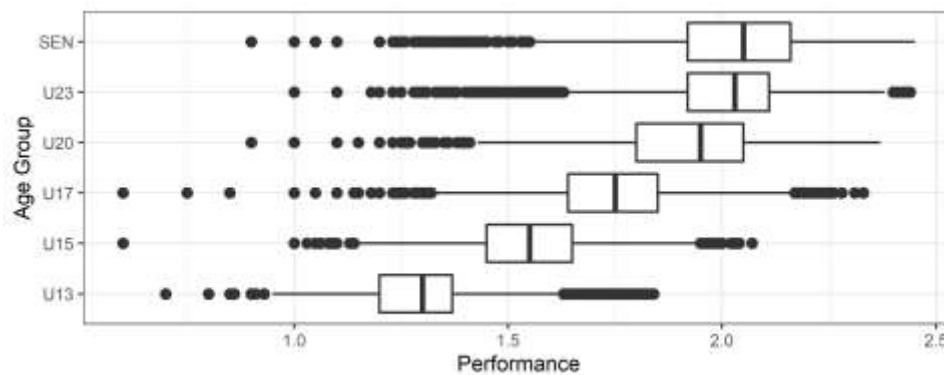


Figure 3.3: Boxplot of Performance by Age Group.

Figure 3.4 (left) on the next page clearly demonstrates a non-linear relationship between age and High Jump performance. **Therefore, fitting a model that allows for a non-linear relationship (i.e., functional form) would be more appropriate than fitting a model assuming a linear relationship.**

Athletes compete at many competitions over their career, therefore there are repeated measures in the dataset (Figure 3.4, right).

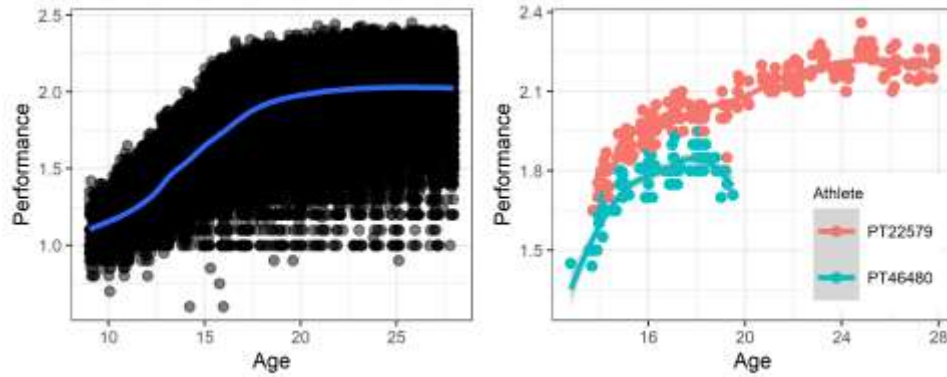


Figure 3.4: Non-linear relationship between age and performance (left) and repeated measurements per athlete (right).

It was found that 7,895 athletes had only one observation, with 838 athletes having 50 observations or more. A full breakdown of the number of records by athlete is shown in Table 3.10.

Table 3.10: Observations per athlete.

Observations	Athletes
1	7,895
2-9	11,474
10-49	5,474
50-99	726
>99	112

Figure 3.5 (left) shows the large variation in performance by year. There is a considerable drop in performance from 2004 onward as this is when the PoT database was introduced. The PoT contains results from Youths (U11) through to Senior (20 years and above), whereas the WA dataset only contains results for elite athletes of 16 years of age and above. Therefore, the performance range of the PoT dataset is much greater than that of the WA dataset. Reported (and presumably actual) performance was considerably impacted by COVID-19 in 2020 and 2021.

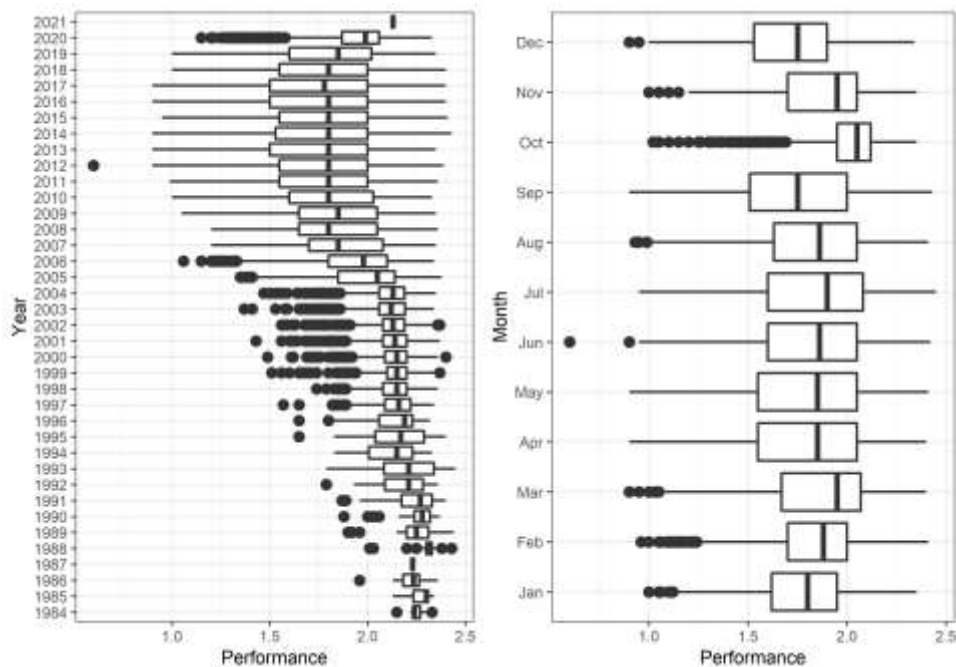


Figure 3.5: Varying performances by year (left) and month (right).

The change in performance by month (Figure 3.5, right) suggests that the trend is consistent with that of a typical athletics calendar competition year that runs from October to the following September. There are fewer competitions in October and November and more during the (northern hemisphere) summer months. The performance distribution in October is also slightly shifted higher which is likely due to the 2019 World Athletics Championships which were held in this month.

Based on these initial findings, only results from 2005 to 2019 were used. This ensures that the data was current and eliminates athletes who may or may not have benefited from the use of performance enhancing drugs.

There were 1125 venues and 122 countries in the dataset. For illustration purposes, a sample of 20 of these two covariates were selected and the performance results for each displayed in Figure 3.6. It can be seen that there are large variations in performance between venues and between countries.



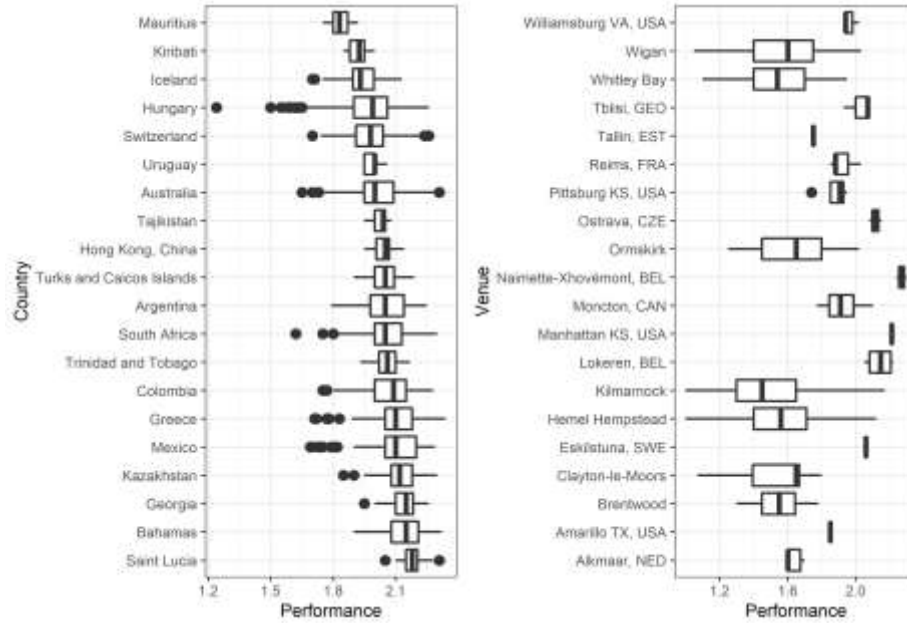


Figure 3.6: Varying performance by country (left) and venue (right).

There is clear difference between the datasets with the PoT and IAAF datasets having a symmetric distribution (albeit with different distribution centres) and the AA dataset having a highly left skewed distribution. The AA dataset is considerably smaller in size ( $n = 286$ ) than both the PoT ( $n = 146,042$ ) and IAAF ( $n = 74,832$ ) datasets used in the analyses.

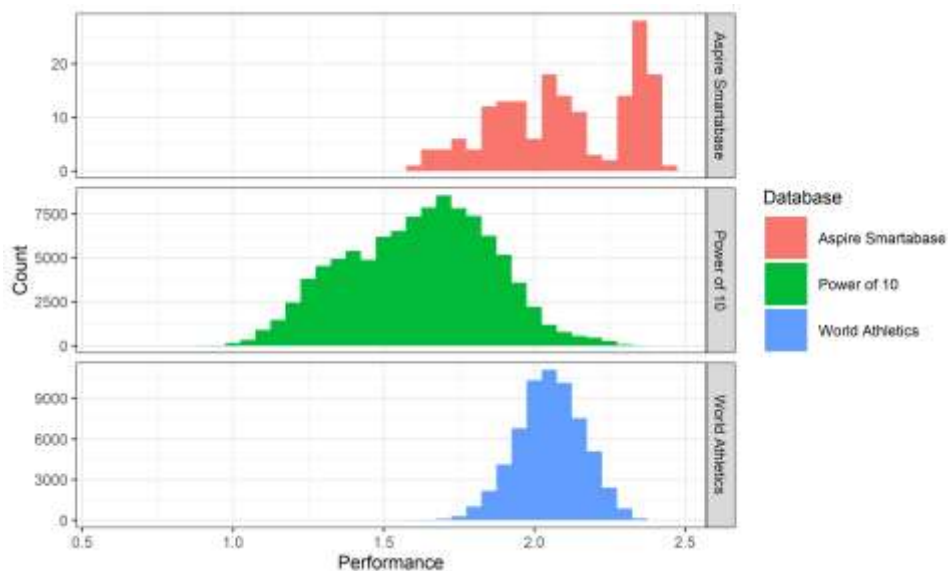


Figure 3.7: Variation in performance by Data Source.

## 3.4 EDA Summary

The most important findings of the EDA were:

- Repeated measurements – some athletes have multiple observations (performances) over time.
- A non-linear relationship is evident between age and performance.
- Hierarchical structures are present - Athletes within countries, Athletes within Competitions, Competitions within Seasons.
- The High Jump Performance response variable is left skewed.
- The explanatory variable age is right-skewed.

### 3.4.1 Interactive EDA App

An interactive dashboard was created using R shiny to make the results of the EDA more translatable to athletic coaches. Some screenshots of the dashboard are presented below along with an explanation of the functionality and insight they provide.

The dashboard displays plots of the distribution of performance, absolute progression, and progression percentage by age.

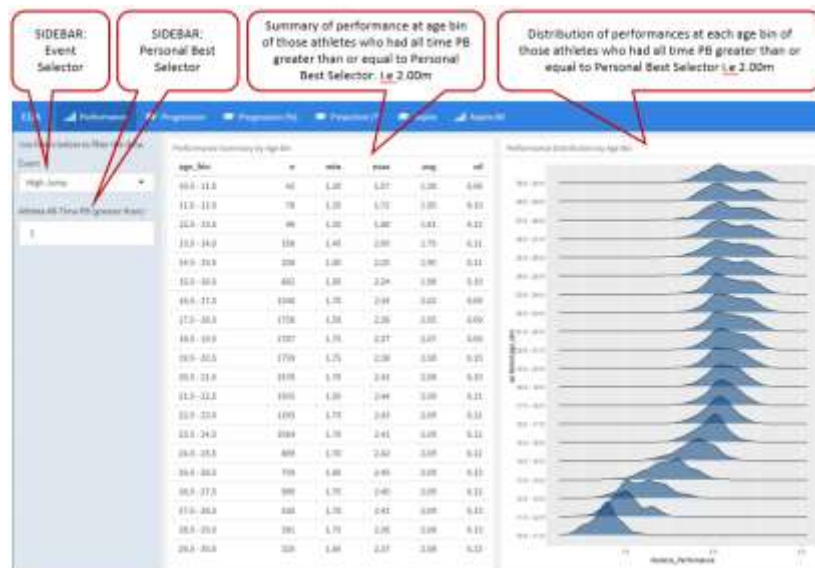


Figure 3.8: Performance distributions by age bin of athletes with a lifetime PB of over 2.00m.



Figure 3.9: Percentage improvement between age bin of athletes with a PB over 1.00m.

Percentile ranks across age categories were calculated and used to benchmark athlete performances over time and displayed alongside their actual performance progression.



Figure 3.10: Example of an athlete overview and bench marked against percentile ranks for age bin.

Plots were created to empirically estimate the probability of an athlete progressing by a certain amount conditional on their current performance and age.

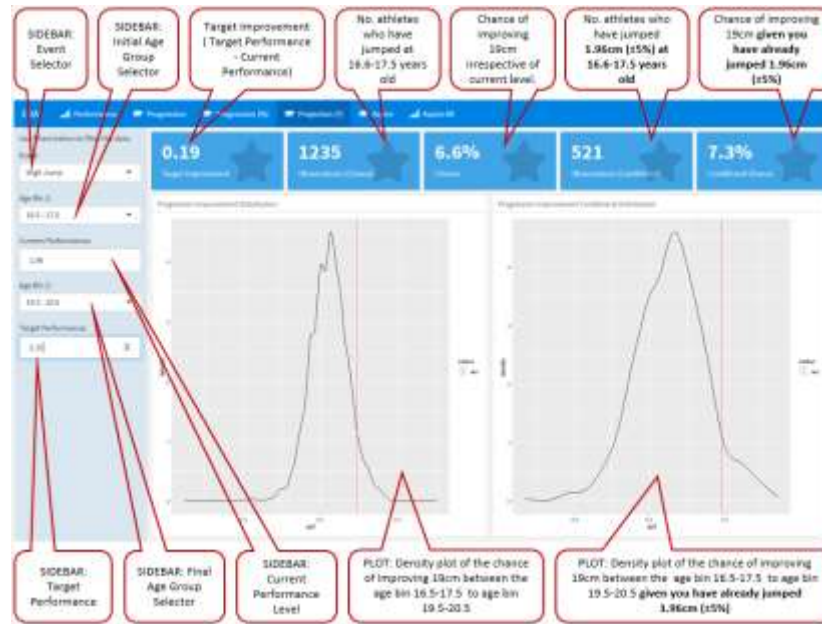


Figure 3.11: Chance and conditional chance of performance improvement given current performance level.

One key finding of the EDA was that the average improvement of athletes with higher personal best was positive for longer (more years), whereas athletes with lower personal bests their average improvement became negative earlier. This is also consistent with the findings of Boccia et al (4).

The EDA shiny application is a very useful tool for an athletics coach to explore the datasets across several athletic events.

### 3.5 Chapter Summary

An extensive EDA was carried out to gain a greater understanding of the data and its structure. An R shiny application was then created that benchmarked athletes and provided data driven empirical estimates for the probability of an athlete improving given their current performance.

To gain greater insight into the performance trajectory of athletes as they age, the next step of the project was to explore appropriate statistical modelling techniques that can model the relationship between High Jump performance and age.

## 4 Modelling the Relationship between Age and High Jump Performance

### 4.1 Introduction

One of the main aims of this project was to identify a statistical model to gain insight into the change in high jump over time for athletes in the population of interest.

The data were observational, hierarchical in nature (i.e., repeated measures for each athlete over time) and exhibited a non-linear relationship between the response (performance) and covariate (age). Although this posed a challenge from a modelling perspective, all these technicalities can be accommodated nicely using the LMM framework.

#### 4.1.1 Linear Mixed Models

The EDA highlighted that the relationship between performance and age is not necessarily linear over time. In addition, the data are hierarchical in nature as there are repeated measurements at the athlete and clustering at the country level. This type of scenario can arise when data collection is undertaken in a hierarchical manner, when several observations are taken on the same observational unit over time, or when observational units are in some other way related, violating the assumptions of independence. The violations of independence rules out linear regression and opens the door to a LMM approach. This method offers a flexible framework by which to model the sources of variation and correlation that arise from grouped data (5) and also allows for the inclusion of smoothing methods (such as splines) to model non-linear functional forms.

LMM's are an extension of a general linear model (GLM). The GLM has this basic form:

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Where  $\mathbf{X}$  represents the design matrix containing the predictors in the model,  $\boldsymbol{\beta}$  represents a vector of regression coefficient and  $\boldsymbol{\varepsilon}$  represents the (stochastic) error in the model with:

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$$

Alternatively, GLMs can be written as follows,

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon_i$$

where  $y_i$  represents the response for the  $i^{th}$  individual,  $\beta_0$  represents the intercept,  $\beta_1$  to  $\beta_p$  represents the regression coefficient for predictors  $X_1$  to  $X_p$ , and finally  $\varepsilon_i$  the  $i^{th}$  population error, where

$$\varepsilon_i \sim N(0, \sigma^2).$$

LMM's contain two parts - a fixed effects part and a random effects part. The general form of a LMM (also called the Laird and Ware model formulation) (6) is:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

$$\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$$

$$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i)$$

where

- $\mathbf{y}_i$  is the  $n_i \times 1$  response vector for observations in the  $i^{th}$  group
- $\mathbf{X}_i$  is the  $n_i \times p$  model matrix for the fixed effects for observations in group  $i$ .
- $\boldsymbol{\beta}$  is the  $p \times 1$  vector of fixed-effect coefficients.
- $\mathbf{Z}_i$  is the  $n_i \times q$  model matrix for the random effects for observations in group  $i$ .
- $\mathbf{b}_i$  is the  $q \times 1$  vector of random-effect coefficients for group  $i$ .
- $\boldsymbol{\varepsilon}_i$  is the  $n_i \times 1$  vector of errors for observations in group  $i$ .
- $\boldsymbol{\Psi}$  is the  $q \times q$  covariance matrix for the random effects.
- $\sigma^2 \boldsymbol{\Lambda}_i$  is the  $n_i \times n_i$  covariance matrix for the errors in group  $i$ .

The fixed effect part of the model contains the covariates of direct interest which, in this thesis, is the variable age. However, the dataset is rich and contained other fields such as venue, country, and year. Whilst the estimated effect of these covariates is not of primary interest, the LMM approach allows for the effect of these covariates on the response to be

accounted for by focusing purely on the variance they explain rather than on estimating their individual level effects.

In a general linear model these covariates would have to be added in as fixed effects which use up valuable degrees of freedom. The random effects part of the LMM gives the opportunity to include additional covariates to help explain the variance structure whilst using less degrees of freedom than estimating the effects of each covariate as a fixed effect. This translates to less degrees of freedom being 'spent' in estimating unnecessary effects leaving the model with more power to test the covariate(s) of interest (age).

Random effects can be included in a LMM as either random intercepts or random slopes. Random intercepts account for variation where individuals are sampled repeatedly such as athlete's performance over time and random slopes account for variation in group responses or 'within individual' variation such as athletes from different countries improving at different rates. An adjustment can be made to the model depending on whether the random effects are correlated (e.g., an intercept and a slope) or are independent (e.g., no relationship between an intercept and slope).

The EDA demonstrated that athletes had repeated measurements over time and that **athletes from different countries improve at different rates** suggesting both random intercepts (i.e., athlete within country) and random slopes (i.e., performance as age increases over time) are required.

To capture these effects in a linear regression model an interaction term would have to be included. For example, given the number of athletes in the dataset, the number of parameters used by this model would be excessively high especially considering that the direct effect of athlete/country is not of interest but rather how much of the variability in performance can be attributed to this.

If there is any between country variation or an age by country interaction then these terms cannot be ignored because systematic variation would end up in the residuals causing a potentially biased inference. To estimate model degrees of freedom more efficiently, a mixed effects model with a random intercept and a random slope can be applied.

Shielzeth and Forstmeier 2009 (7) suggest to always include both random slopes and intercepts when possible. Incorporating random intercepts and slopes collectively reduces the incidence of Type 1 and Type 2 errors and reduces the chance of overconfident estimates (unrealistically low standard error). However fitting random slopes requires relatively large sample sizes for model convergence, especially if the data set contains many groups with only a few observations (8). The dataset used in this thesis is large and therefore it is unlikely that convergence will be a problem.

Given the above, LMM's are an attractive solution to model both the hierarchical structure present in the data and the serial correlation in each athlete's performance over time by incorporating a random effect in order to model each athlete's serial correlation in a flexible way (6).

The issue of non-linearity in performance over time still needs to be addressed and can be done by incorporating flexible functional forms to model non-linear relationships between the response and covariates. As it is possible to include smoothing functions in the linear mixed model framework, whilst still maintaining linearity in the model's parameters, this makes this approach particularly suitable to the problem at hand.

#### **4.1.2 Smoothing Methods**

There are some simple extensions to linear models to accommodate non-linear functional forms such as polynomial regression and step functions. These have limitations, which will be discussed later, and the focus then changes to more sophisticated and powerful smoothing methods, namely splines.

Given the size of the data it is computationally (and time) expensive to use all the data to explore what an appropriate functional form might be. For this reason, a (random) sample of 50 athletes with over 90 observations was used to investigate various smoothing methods using solely age as the predictor of performance.



#### 4.1.2.1 Polynomial Regression

Polynomial regression extends the linear model by adding extra predictors, obtained by raising each of the original predictors to a power (9). For example, in the high jump data using age as the covariate, a cubic regression could be fit using a polynomial of degree 3 for age i.e.,  $age$ ,  $age^2$  and  $age^3$  as covariates. This approach provides a simple way to incorporate a non-linear functional form for the covariate on the response.

Table 4.1: Summary of 2nd degree polynomial fit.

Term	Estimate	SE	Statistic	p-value
(Intercept)	1.891	0.002	1087.949	0.000
$age$	12.307	0.130	94.425	0.000
$age^2$	-5.713	0.130	-43.829	0.000

Table 4.2: Summary of 3rd degree polynomial fit.

Term	Estimate	SE	Statistic	p-value
(Intercept)	1.891	0.002	1123.970	0.000
$age$	12.307	0.126	97.551	0.000
$age^2$	-5.713	0.126	-45.281	0.000
$age^3$	2.457	0.126	19.478	0.000

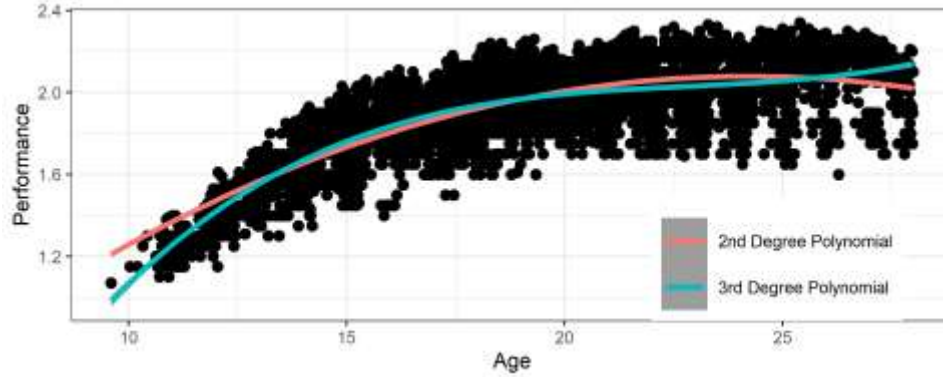


Figure 4.1: Polynomial fits of Age vs Performance.

Summaries for linear models with polynomial fits (for age) of degree 2 and 3 are provided in Table 4.3. In both models all the terms are significant, however neither model seems to fit the data that well with r-squared (adjusted) values of 0.658 and 0.68 respectively.

Table 4.3: Polynomial fit model metrics.

Model	$R^2$	Adjusted $R^2$	p-value	AIC
$performance \sim age + age^2$	0.658	0.658	0.000	-6954.024
$performance \sim age + age^2 + age^3$	0.680	0.680	0.000	-7319.405

#### 4.1.2.2 Step Functions

Step functions cut the range of a covariate into  $k$  distinct regions to produce a categorical variable. This has the effect of fitting a piecewise constant function (9) in each level. Using polynomial functions of the covariate as predictors in a linear model imposes a global structure on the non-linear functional form for the covariate. The step functions can be used to avoid imposing such a global structure. Here the age range is broken into bins and fit a different constant in each bin. This amounts to converting a continuous variable into an ordered categorical variable. An example of step functions using age as a predictor of performance is fitted and plotted in Figure 4.2.

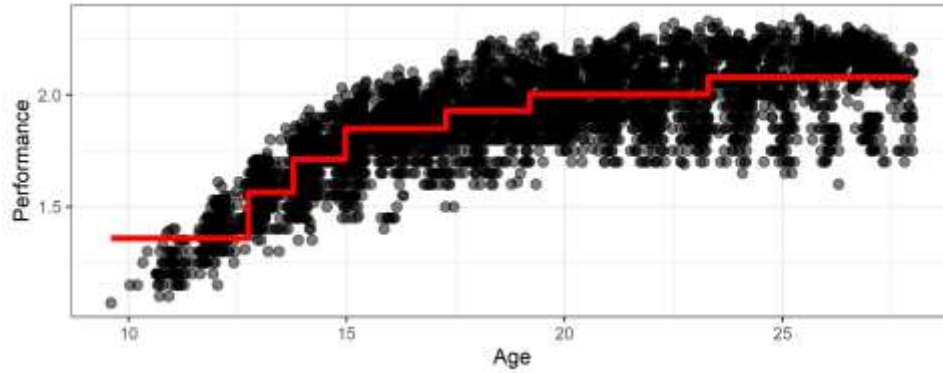


Figure 4.2: Step function fit of Age vs Performance.

Figure 4.2 above gives a visual representation of why a step function is not a suitable method to model the non-linearity in performance over time. A step function approach would suggest that performance is constant for a set period before levelling up (or down) to a new constant level. This is clearly not the case and regular set jumps like this are simply not realistic.

Whilst step functions are clearly not going to be useful in modelling high jump performance, the idea of breaking the range of a covariate down into regions provides a useful introduction to smoothing methods which employ a variation of such techniques.

#### 4.1.2.3 Advanced Smoothing Methods

Both polynomial regression and step function are special cases of what are known as ‘Basis’ functions. Basis functions extend ‘rigid’ step functions to accommodate piecewise polynomials. By incorporating various constraints on continuity (i.e., continuity of the first and second derivatives) the approach can be extended further to what is known as spline regression.

One example of spline regression is called B-splines which utilise the idea of a basis function.

The basis functions (10) approach is to essentially utilise a family of functions or transformations (that are fixed and known) that can be applied to a variable  $X$ .

$$X: b_1(X), b_2(X), \dots, b_K(X)$$

Instead of fitting a linear model in  $X$ , the following model can be fit:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \dots + \beta_K b_K(x_i) + \varepsilon_i.$$

More simply, B-splines allow us to ‘tie’ together (via constraints) multiple curves of lower degree at knots spread out throughout the data to produce a reasonable fit to the data. In regions of the data where the data change more rapidly more knots are placed and where it is more stability fewer knots as placed.

B-splines give more stable estimates than polynomial regression. To fit data that are changing rapidly over a certain region, a polynomial must be of a high degree which causes erratic fits at the boundaries, mainly because the function is applied globally to the data. In comparison, a B-spline simply uses more knots where the data changes rapidly and keeps the degree small (usually degree = 3). The constraints in place at the knots enforce the B-spline to be smooth and continuous and the resulting curves are interpreted as perfectly smooth to the naked eye (11).

Natural splines are an extension of B-splines and further attempt to ameliorate the problem of erratic boundary fits by adding the additional constraints that the function is linear beyond the boundaries of the data.

P-splines, as proposed by Eilers and Marx (1996)(12), have attempted to bridge the gap between regression splines and smoothing splines. They combine a B-spline basis, with a discrete penalty on the basis coefficients, where the same combination of penalty and basis order is allowed (13). P-splines can perform better in particular cases where it is advantageous to mix different orders of basis and penalty and perform almost as well as conventional splines in many standard applications.

B-, natural and P-splines are all forms of regression splines. Regression splines (usually) place knots at equidistant/equi-quantile points whereas smoothing splines use the data points themselves as potential knots and then explicitly penalise roughness.

To investigate which smoothing method is the most suitable, in terms of representing what is known based on domain knowledge of high jump performance and what appears ‘best’

based on the data provided, a comparison of several approaches is now presented and a discussion on the merits (and lack of) of each is given.

As a starting point, a plot of a subset of the random sample of 50 athletes (12 to be precise to make plotting clearer) of age against performance faceted by athlete (ID) is given below in Figure 4.3. It is evident that individual athletes' performances can change at different rates. As discussed previously, rate of performance change can be captured by a LMM through including a random slope at the athlete level.

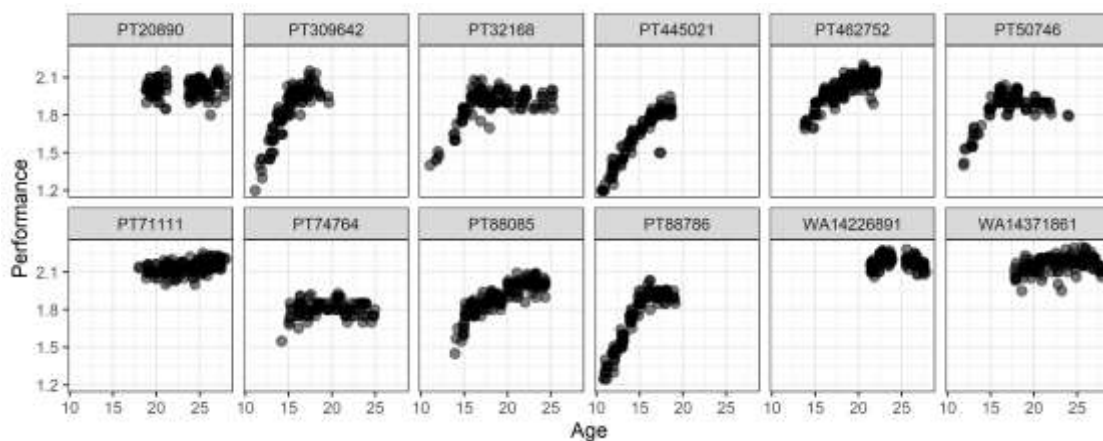


Figure 4.3: Performance progressions for a sample of athletes.

#### 4.1.2.4 Smoother Comparisons

For illustration purposes the R package ggplot was used to visualize the different smoothing methods faceted by athlete (ID). Visual inspection was used to compare smoothing methods whereas more robust statistical comparisons were made in the modelling section.

Figure 4.4 overpage shows piecewise constant functions fitted for a sample of 12 athletes. As discussed previously, performance is best considered as a smooth function of time rather than a big 'step' every so often as assumed when using piecewise constants. These plots provide convincing evidence **that piecewise constant functions are not a suitable candidate to capture the functional form correctly.**

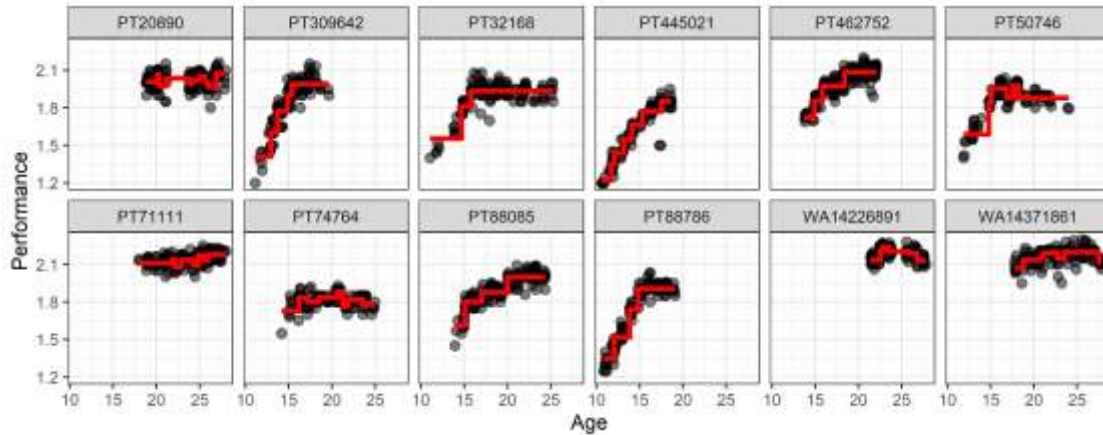


Figure 4.4: Individual step functions fitted to a sample of athletes.

A visual comparison of the fits from a simple linear regression compared to a polynomial regression are displayed below.

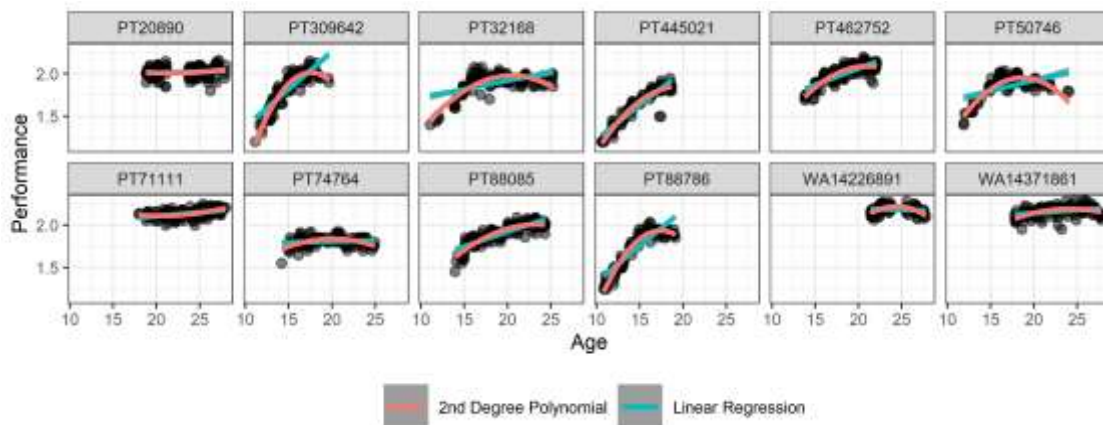


Figure 4.5: Linear regression and polynomial models fitted to a sample of athletes.

A second-degree polynomial produced a better fit than a linear model. This was to be expected given the non-linear nature of the relationships between age and performance.

A third-degree polynomial produced slightly better fits than a second-degree polynomial. A notable example is seen in Figure 4.6 overleaf by athlete PT50746 (top right) where the second-degree polynomial produced an erratic fit at the boundary.

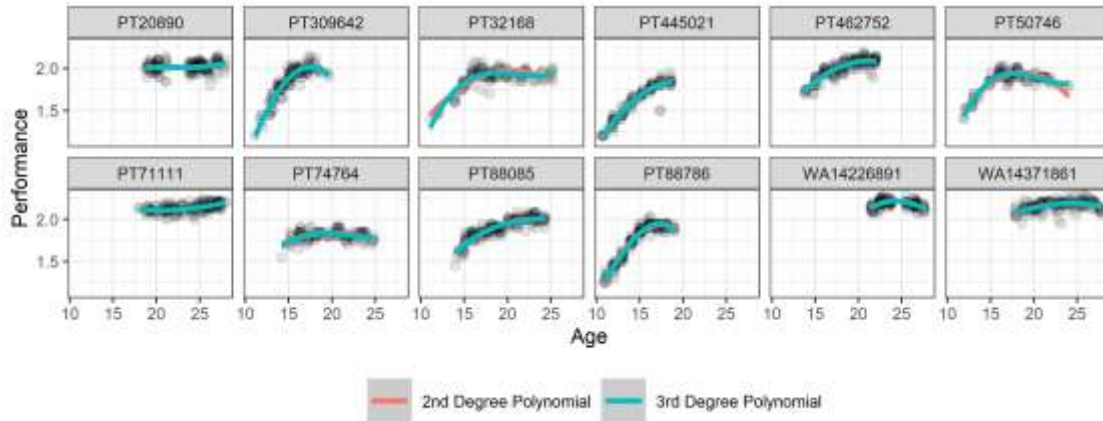


Figure 4.6: Polynomials fitted to a sample of athletes.

Figure 4.7 below shows B-splines with 3 and 5 degrees of freedom fitted over a sample of athletes. Little difference was found between the two approaches in the sample under consideration, suggesting that 3 degrees of freedom may provide enough flexibility.

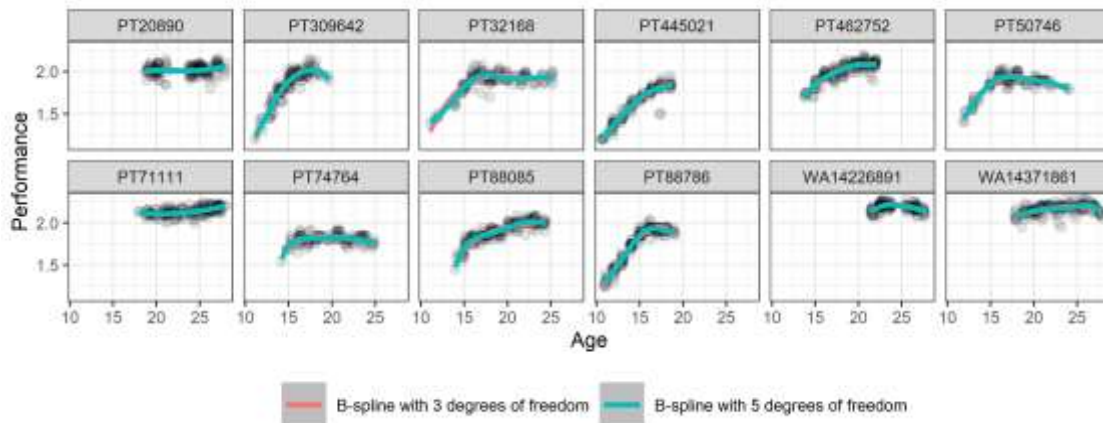


Figure 4.7: B-splines of 3 and 5 degrees of freedom fitted to a sample of athletes.

Figure 4.8 overleaf shows B-splines with 3 and 10 degrees of freedom fitted over the same sample of athletes. There is clear over-fitting in the B-spline when using 10 degrees of freedom across almost all the athletes in the sample as the oscillations in performance evident over time is likely to be spurious.

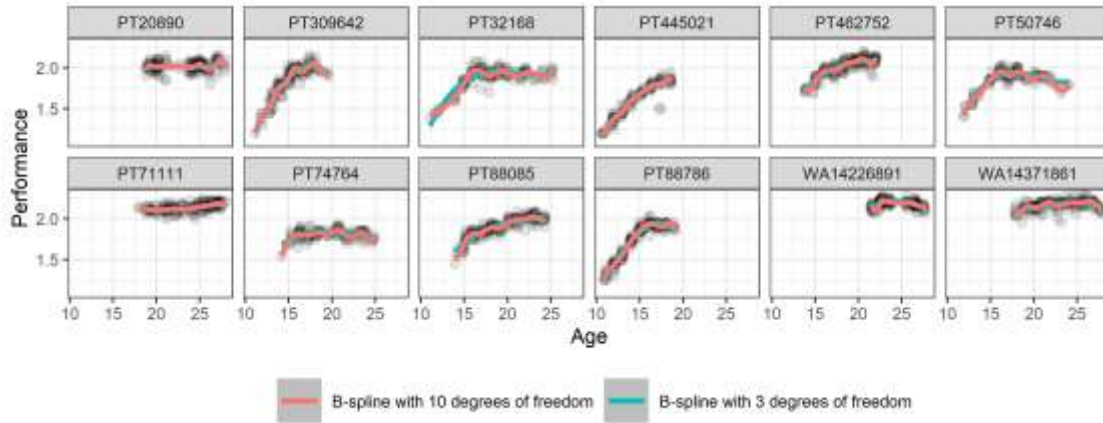


Figure 4.8: B-splines of 3 and 10 degrees of freedom fitted to a sample of athletes.

B-splines are a flexible method when using a large number of knots and it makes sense to place more knots in areas where the data are changing the most. From my domain knowledge as an experienced athlete it is possible to approximate where an athlete's performance will vary the most. However, given the nature of how the smoothing method (i.e. at both the fixed and random effect level) is to be implemented it made more sense to specify the degrees of freedom and let the software place the knots as placement varied from athlete to athlete. **The choice of final model incorporated both the location of the knots (based on the data and assumed model) and the domain expertise at hand.**

Keele (2008) (14) suggests 3 knots should be used if there are less than 30 data points and 5 knots if there are more than 100 data points. This corresponds to 6 and 8 degrees of freedom for a B-spline. The EDA showed that there are not more than 30 data points for every athlete suggesting that 6 degrees of freedom would be appropriate.



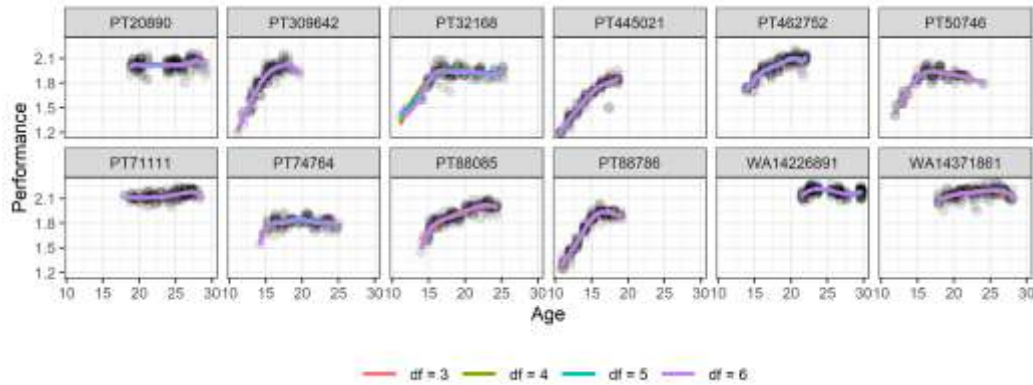


Figure 4.9: B-splines with 3 through to 6 degrees of freedom.

Figure 4.9 graphically illustrates B-splines (of degree 3) with 3 through to 6 degrees of freedom which corresponded to 0 to 3 knots. The B-spline with 3 degrees of freedom produced the most realistic fit from a sport science perspective and the additional degrees of freedom/knots started to cause over fitting. The athlete in Figure 4.10 below has less than 30 data points and is an example of how a B-spline with only 3 degrees of freedom produced the most stable realistic fit.

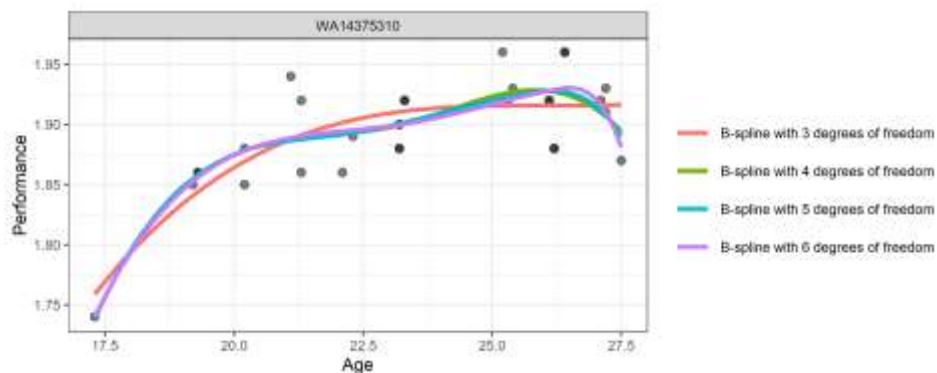


Figure 4.10: Example of overfitting with B-spline.

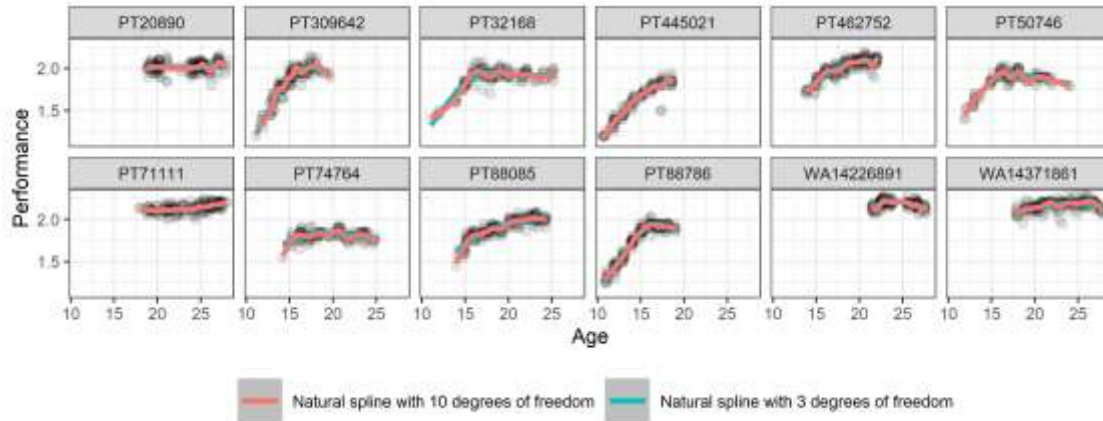


Figure 4.11: Natural splines fitted to a sample of athletes.

The natural splines shown above in Figure 4.11 generally show a similar trend to the B-splines shown previously in Figure 4.8. Natural splines with 10 degrees of freedom look to be over fitting whilst those with 3 degrees of freedom look to fit the data quite well. As expected, there are some differences at the boundaries, where the linearity constraints of the natural splines are enforcing more reasonable results at the boundaries. This is clear when comparing the B-spline and natural spline with 10 degrees of freedom for athlete PT50746.

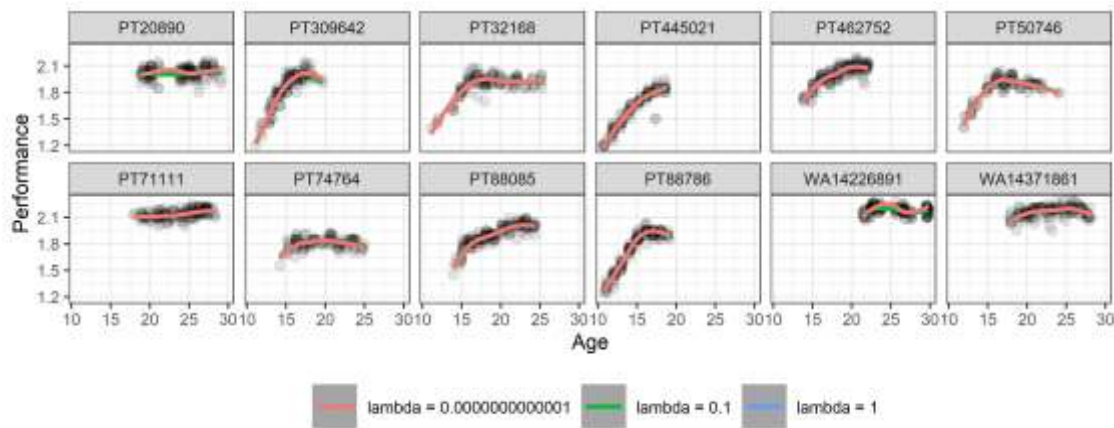


Figure 4.12: P-splines with varying lambda penalties fitted to sample of athletes.

Figure 4.12 shows how different lambda penalties influence the fit of a P-spline on the data. The smaller the penalty the 'wigglier' the fit. There was very little difference in the fit of the P-spline between  $\lambda = 0.000000000001$  and  $\lambda = 0.1$ .

Based on the visualisations presented in this section, all four smoothing methods (i.e., polynomials, B-splines, natural splines, and p-splines) seem to be able to produce reasonable fits to the sample data when smoothing parameters were adjusted accordingly. However, given that the polynomial, B-spline, and natural splines are the easiest to implement from a coding perspective (using the lmer package in R) **a P-spline approach was not considered**, and attention was focused on a formal comparison of the other three smoothing methods.

#### 4.1.3 Model Fitting Workflow

When comparing models of increasing complexity it has been recommended by Gelman (15) that simple linear models are fitted initially and then adapted to generate more complex models (according to some goodness-of-fit criterion) with this process repeated until no further improvements were evident.

Given the aims of this research, age will be the fixed effect component and the remaining covariates used to explain the underlying variance structures.

The following 9 models (written symbolically) were fitted sequentially to model the relationship between the (fixed) covariate Age and the response variable of interest, namely Performance:

1. Simple Linear Regression:

Performance ~ Age

2. Multiple Linear Regression:

Performance ~ Age + ID + Country + Year + Event + Venue

3. LMM with one random intercept:

Performance ~ Age + (1|ID)

4. LMM with random intercept and natural spline smoothing:

Performance ~ ns(Age, df = 3) + (1|ID)

5. LMM with random intercept and B-spline smoothing:

Performance ~ bs(Age, df = 3) + (1|ID)

6. LMM with random intercept and polynomial smoothing:  

$$\text{Performance} \sim \text{poly}(\text{Age}, \text{df} = 3) + (1|\text{ID})$$
7. LMM with B-spline smoothing and hierarchical structure on Country:Athlete as a random intercept:  

$$\text{Performance} \sim \text{bs}(\text{Age}, \text{df} = 3) + (1|\text{ID:Country})$$
8. LMM with hierarchical structure and additional B-spline at random effect level:  

$$\text{Performance} \sim \text{bs}(\text{Age}, \text{df} = 3) + (\text{bs}(\text{Age}, \text{df} = 3)|\text{ID:Country})$$
9. LMM with additional random effects:  

$$\text{Performance} \sim \text{bs}(\text{Age}, \text{df} = 3) + (1|\text{Venue}) + (1|\text{Year}) + (1|\text{Country}) + (\text{bs}(\text{Age}, \text{df} = 3)|\text{ID:Country})$$

All LMM's were fitted using the `lmer` function of the `lme4` R package (16). The `lme4` package default output for LMM's provides t-values but no p-values. The primary motivation for this omission is that in LMM's it is not at all obvious what the appropriate denominator degrees of freedom to use are, except perhaps for some simple designs and nicely balanced data (17). In this project, significance of LMM's were calculated using the `lmerTest` package (18). Prediction intervals from the LMM's were obtained using the '`predictInterval`' function from the `merTools` package (19).  $R^2$  values for LMMs were estimated using the `MuMin` package (20).

#### 4.1.4 Model Comparisons

The Likelihood ratio test (LRT) can compare two different models to determine if one is a better fit to the data than the other. LRTs are commonly used to decide if a particular parameter should be included in a mixed model. LRTs are most commonly used to decide if a particular random effect (say, a random intercept) should be retained in the model by evaluating whether that effect improves the fit of the model, with all other model parameters held constant (17).

When used for evaluating the significance of fixed effects, LRTs have one potential disadvantage: using LRTs to compare two models that differ in their fixed effects structure may not always be appropriate (5). When mixed-effects models are fitted using restricted maximum likelihood (REML), there is a term in the REML criterion that changes when the fixed-effects structure changes, making a comparison of models differing in their fixed effects structure meaningless. Thus, if LRTs are to be used to evaluate significance, models must be fitted using maximum likelihood (ML), and this approach was therefore used in this project.

Before the model fitting process began it was clear what the fixed effect structure would be (age) and that the bulk of the process would be in defining a structure to correctly model the variance. This makes LRTs an attractive option for comparing the models in this project.

The percentage of unexplained variance explained by the random effects was also calculated and compared.

#### 4.1.5 Model Selection and Performance

The Root Mean Squared Error (RMSE) and the Akaike Information Criterion (AIC) (21) were both used to decide between alternative models, by choosing the model with the smaller values. RMSE can be interpreted as the standard deviation of the unexplained variance and has the useful property of being in the same units as the response variable (which for the high jump is metres). Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response and can be regarded as an important criterion for fit if the main purpose of the model is prediction, despite this not being the main aim of this project.

Model performance was evaluated by visually comparing the predicted versus actual performances. Pearson's correlation coefficient ( $R^2$  - predicted v actual), chosen for its robustness and intuitive interpretation (22), was reported alongside AIC and RMSE.

When comparing alternative models, the residuals from the fit were analysed to check for any departures from the model's underlying assumptions. **Validating the assumptions is arguably more important in this thesis than predictive accuracy given the main aim is to develop a model to gain insight into how performance is related to increasing age.**

#### 4.1.6 Model Fitting Process

In this section the iterative process of modelling is outlined. Simple regression models were fitted initially before moving on to more complex LMMs. Model summaries, diagnostic plot and comments are provided for all models. In each case the model is written in compact symbolic form, following the convention used in R for model fitting. In each case the error term is assumed present but not represented.

##### 4.1.6.1 Linear Regression Models

Simple linear regression (SLR) and multiple linear regression models (MLR) were fitted for models 1 and 2 respectively. The specification of the models is outlined below.

Model 1: SLR with age as single covariate.

Performance ~ Age

Model 2: MLR with 6 covariates.

Performance ~ Age + ID + Country + Year + Event + Venue

It is clear from the diagnostic plots (Figure 4.13) that model 1 is not an adequate representation of the relationship between age and performance as the residuals have a slight left skew and there is a clear quadratic pattern in the residuals versus the fitted values. There are some deviations in the upper tail of the QQ-plot and notable deviations in the in the line of fitted versus actual.

There is a noticeable improvement in the diagnostic plots when considering Model 2 (Figure 4.13) as the residuals resemble a normal distribution and there are no obvious patterns in the residuals versus the fitted values. There is strong deviation in the lower tail of the QQ-plot however and whilst the overall trend of the fitted versus actual has improved there is a large spread either side of the reference line. The AIC and RMSE dropped from an AIC of -5302.68 and a RMSE of 0.1509 in model 1 to an AIC of -9074.445 and a RMSE of 0.1001 in model 2.

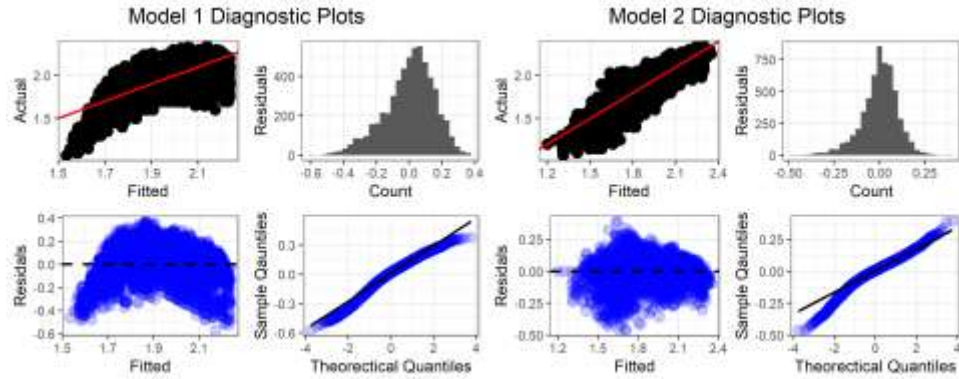


Figure 4.13: Diagnostic plots for model 1 and model 2.

Table 4.4: Model 1 and model 2 model metrics.

Model Terms	$R^2$	AIC	RMSE
~ Age	0.54	-5302.7	0.15
~ Age + ID + Country + Year + Event + Venue	0.80	-9074.4	0.10

It is of no surprise that Models 1 and Model 2 are inadequate given the findings of the EDA which are further validated given the clear violations of several of the underlying assumptions underpinning linear regression models.

#### 4.1.6.2 Linear Mixed Models

Models 3 to 9 are LMMs of increasing complexity chosen to address the results of the EDA and to accommodate additional covariates that are potentially useful variance components.

The key findings of the EDA are repeated here for convenience:

- flexible functional form needed for age
- repeated measures component needed to model serial correlation
- random effects needed to account for hierarchical structure

#### 4.1.6.2.1 Random Intercept Model

##### Model 3: LMM with Random Intercept

The first LMM considered contained a single random effect for athlete (ID) to explain the variance among athletes:

$$\text{Performance} \sim \text{Age} + (1|\text{ID})$$

This model does accommodate the repeated measures within athlete over time and provides an estimate of the correlation of the response within athlete through the intra cluster correlation coefficient which is 0.502. This can be interpreted as amount of unexplained variance accounted for (i.e., the random intercept on Athlete (ID) accounts for 50.2% of the unexplained variance, see Table 4.5).

Table 4.5: Unexplained variance accounted for in Model 3.

Term	var1	Variance	%
ID	(Intercept)	0.0120	50.2%
Residual		0.0119	49.8%

This model produces similar diagnostic plots (Figure 4.14) and AIC and RMSE values (Table 4.6) to the multiple linear regression model (Model 2).

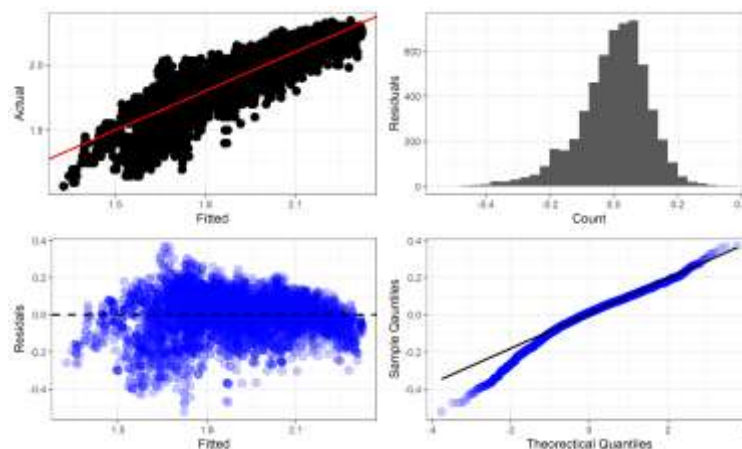


Figure 4.14: Diagnostic plots for model 3.



Table 4.6: Model 3 metrics.

Model Terms	$R^2$	AIC	RMSE
~ Age + (1 ID)	0.70	-8706.4	0.11

Table 4.7 shows that the fixed effect terms of Model 3 are all significant ( $p < 0.05$ ). There was also evidence of a correlation between slope and intercept: high intercepts are correlated with low slopes which can be explained by the ‘ceiling’ on high jump performance.

Table 4.7: Model 3 fixed effects summary.

Term	Estimate	SE	df	t-value	p-value
(Intercept)	1.3396	0.0187	102.6468	71.5117	0.0000
age_perform	0.0291	0.0006	5602.3346	52.5916	0.0000

At this early iterative modelling stage, it can be seen that LMMs are a better choice of model than MLR whilst accounting for an athlete’s repeated performances within the data.

#### 4.1.6.2.2 Adding Fixed Effect Smoothing

In Models 4, 5 and 6 issues of non-linearity were addressed by applying various smoothing techniques at the fixed effects level. The smoothing methods compared were polynomials, B-splines, and natural splines all with three degrees of freedom due to reasons discussed earlier in the smoothing chapter.

#### Model 4: LMM with natural smoothing spline on age and a random intercept for ID

Performance ~ ns(Age, df = 3) + (1|ID)

#### Model 5: LMM with B-spline smoothing on age and a random intercept for ID

$$\text{Performance} \sim \text{bs}(\text{Age}, \text{df} = 3) + (1|\text{ID})$$

#### Model 6: LMM with polynomial smoothing on age and a random intercept for ID

$$\text{Performance} \sim \text{poly}(\text{Age}, \text{df} = 3) + (1|\text{ID})$$

All of these three models produced very similar diagnostic plots with residuals that suggest that the error distribution is acceptable. There were some deviation in the lower tail of the QQ-plots (Figure 4.15) which are not unexpected given the large sample size.

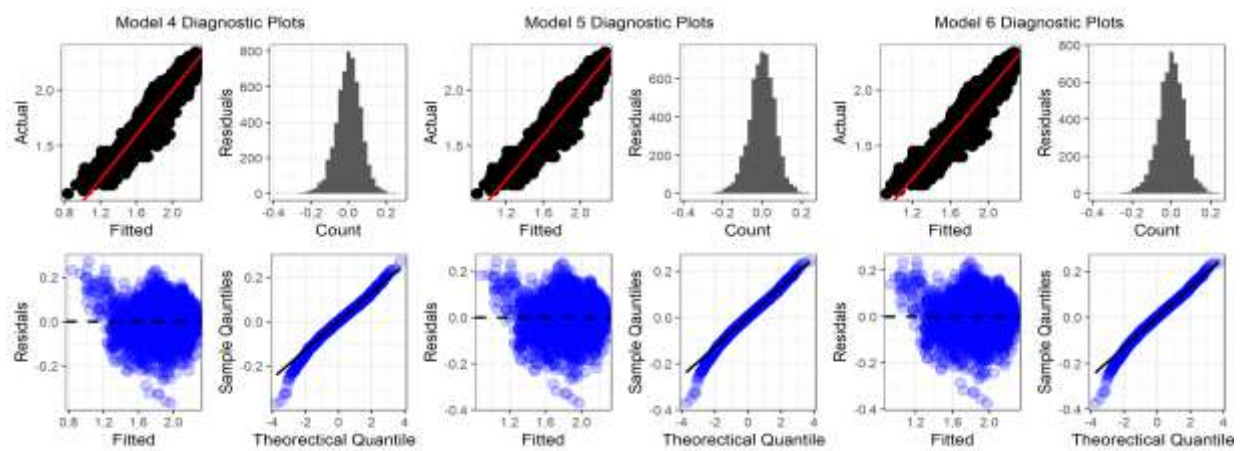


Figure 4.15: Diagnostic plots for models 4, 5 and 6.

All three models had the same RMSE (to 3 decimal places) and similar AIC scores with the polynomial coming out slightly ahead of the B-spline and natural spline (Table 4.8).

Table 4.8: Model 4 to Model 6 metrics.

Model	$R^2$	AIC	RMSE
$\sim \text{ns}(\text{Age}, \text{df} = 3) + (1 \text{ID})$	0.89	-13938.55	0.07
$\sim \text{bs}(\text{Age}, \text{df} = 3) + (1 \text{ID})$	0.89	-13930.61	0.07
$\sim \text{poly}(\text{Age}, \text{df} = 3) + (1 \text{ID})$	0.89	-13956.99	0.07

Given that the  $R^2$  and RMSE values are virtually the same and the AIC scores are similar, the B-spline smoothing method was deemed the most suitable, in terms of flexibility and empirical performance for all of the subsequent models considered.

#### 4.1.6.2.3 Modelling the Hierarchical Structure

In Model 7 the random intercept for Athlete was nested within the athlete's country of origin to capture any correlation within athletes from the same country. This could be a potential surrogate for genetic similarity, between country investment in high jump or environment conditions at the country level.

##### Model 7: LMM with B-spline and hierarchical random effect

$$\text{Performance} \sim \text{bs}(\text{Age}, \text{df} = 3) + (1 | \text{ID:Country})$$

The diagnostic plots (Figure 4.16) do not vary much from those in the previous four models other than some slight deviation in both tails of the QQ-plot.

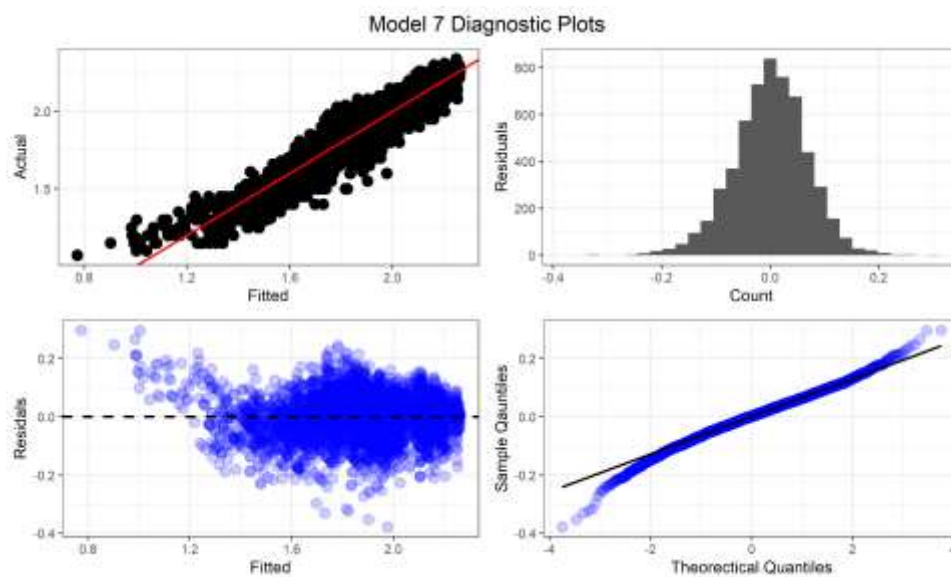


Figure 4.16: Diagnostic plots for model 7.

The nested hierarchical random effect of athlete within country (ID:Country) accounted for 73.2% of the unexplained variance (Table 4.9). This was 23.0% higher than when athlete (ID) alone was used as a random effect.

Table 4.9: Model 7 unexplained variance accounted for.

Term	var1	Variance	%
ID:Country	(Intercept)	0.0131	73.2%
Residual		0.0048	26.8%

$R^2$ , RMSE and AIC remained unchanged but the large improvement in variance structure was a big improvement to the model.

#### 4.1.6.2.4 Adding Smoothing at the Random Effect Level

Model 8 **introduced a non-linear function at the athlete level using B-splines** to attempt to capture variance explained by athletes improving in a non-linear manner and at different rates. Modelling the random effect as a smooth function was considered by Rice and Wu in 2001 (23) .

Model 8: LMM with B-spline on age at both fixed and random levels plus random intercepts for athletes within country.

$$\text{Performance} \sim \text{bs}(\text{Age}, \text{df} = 3) + (\text{bs}(\text{Age}, \text{df} = 3)|\text{ID:Country})$$

The diagnostic plots remained largely unchanged from those in Model 7 except there is now only deviations evident in the lower tail of the QQ-plot. This is likely because there are no results below 1.07m in the model dataset as such results are recorded at this threshold.

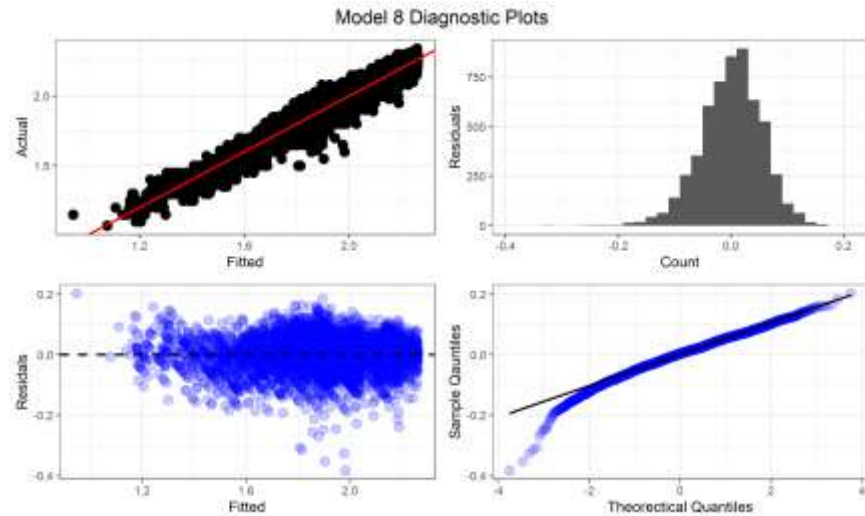


Figure 4.17: Diagnostic plots for model 8.

Using a B-spline with 3 degrees of freedom to model the random slopes within the random effect of athlete within country (`bs(age, df = 3) | ID:Country`) **increased the amount of explained variance considerably from 73.2% to 99.63%**. A full breakdown of the variance explained by each random effect component is displayed below in Table 4.10.

Table 4.10: Variance explained by components of the random effects in Model 8.

Term	var1	var2	Variance	%
ID:Country	(Intercept)		0.1315	15.2%
ID:Country	bs(age_perform, df = 3)1		0.5168	59.7%
ID:Country	bs(age_perform, df = 3)2		0.1170	13.5%
ID:Country	bs(age_perform, df = 3)3		0.1364	15.8%
ID:Country	(Intercept)	bs(age_perform, df = 3)1	-0.2434	-28.1%
ID:Country	(Intercept)	bs(age_perform, df = 3)2	-0.0594	-6.9%
ID:Country	(Intercept)	bs(age_perform, df = 3)3	-0.1213	-14%
ID:Country	bs(age_perform, df = 3)1	bs(age_perform, df = 3)2	0.0681	7.9%
ID:Country	bs(age_perform, df = 3)1	bs(age_perform, df = 3)3	0.2339	27%
ID:Country	bs(age_perform, df = 3)2	bs(age_perform, df = 3)3	0.0824	9.5%
Residual			0.0032	0.4%

#### 4.1.6.2.5 Further Modelling of the Random Effects

In Model 9 additional random effects were added to account for the remaining unexplained variance.

Model 9: LMM with B-spline on age at both fixed and random levels plus random intercepts for athletes within country, Venue, Year and Country.

$$\text{Performance} \sim \text{bs}(\text{Age}, \text{df} = 3) + (\text{bs}(\text{Age}, \text{df} = 3)|\text{ID}:\text{Country}) + (1|\text{Year}) + (1|\text{Country}) + (1|\text{Venue})$$

The diagnostic plots remained very similar to model 8.

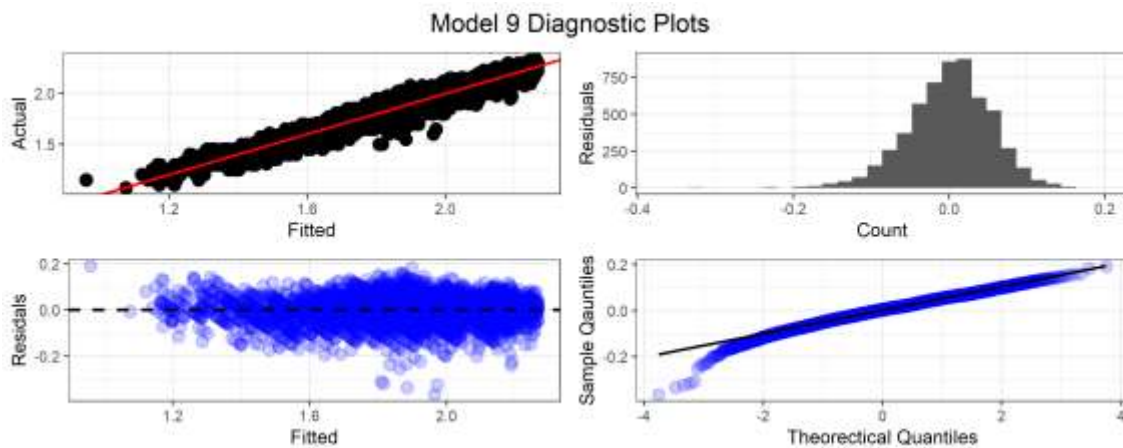


Figure 4.18: Diagnostic plots for model 9.

Table 4.11 compares the model metrics for the Model 9 and its derivatives. The largest change in AIC and RMSE came after the inclusion of the random effect of Venue.

Table 4.11: Model 9 and model 9 variations metrics.

Model	AIC	RMSE
Full Model	-15792.1643	0.0541
Year Random Effect Removed	-15785.1750	0.0542
Country Random Effect Removed	-15779.2834	0.0540
Venue Random Effect Removed	-15723.8953	0.0555

Table 4.12 shows the percentage of explained variance didn't improve but given the findings in the EDA and the nature of the high jump event, it is clear there will be some variance explained by incorporating random effects on year, country, and venue.

Venue could be considered a surrogate for environmental conditions at the time of the event and year will help capture any seasonal variations. The random effect of country will help capture between country variation at a national level.

As the number of additional degrees of freedom needed to include these additional random effects is minimal, these covariates were retained in the final model. Computational time was not adversely affected.

Table 4.12: Variance explained by components of the random effects in Model 9.

Term	var1	var2	Variance	%
Venue	(Intercept)		0.0001	0.02%
ID:Country	(Intercept)		0.0873	10.9%
ID:Country	bs(age_perform, df = 3)1		0.4842	60.45%
ID:Country	bs(age_perform, df = 3)2		0.0955	11.93%
ID:Country	bs(age_perform, df = 3)3		0.1193	14.9%
ID:Country	(Intercept)	bs(age_perform, df = 3)1	-0.1941	-24.23%
ID:Country	(Intercept)	bs(age_perform, df = 3)2	-0.0404	-5.05%
ID:Country	(Intercept)	bs(age_perform, df = 3)3	-0.0912	-11.39%
ID:Country	bs(age_perform, df = 3)1	bs(age_perform, df = 3)2	0.0492	6.15%
ID:Country	bs(age_perform, df = 3)1	bs(age_perform, df = 3)3	0.2086	26.04%
ID:Country	bs(age_perform, df = 3)2	bs(age_perform, df = 3)3	0.0671	8.38%
Year	(Intercept)		0.0000	0%
Country	(Intercept)		0.0122	1.53%
Residual			0.0031	0.38%

#### 4.1.7 Model Comparison Summary

$R^2$ , AIC, RMSE and Unexplained Variance Accounted For are shown in Table 4.13 for all models fitted. The table outlines the change in each metric at each stage in the model fitting process. The largest change in all model metrics came after the inclusion of a non-linear smoothing term at the fixed effect level.

Table 4.13: Model 1 to Model 9 metrics.

Model Terms	$R^2$	AIC	RMSE	Unexplained Variance Accounted For
~ Age	0.54	-5302.70	0.15	
~ Age + ID + Country + Year + Event + Venue + Database	0.80	-9074.40	0.10	
~ Age + (1 ID)	0.70	-8706.40	0.11	50.2%
~ ns(Age, df = 3) + (1 ID)	0.88	-13938.50	0.07	
~ bs(Age, df = 3) + (1 ID)	0.88	-13930.60	0.07	
~ poly(Age, df = 3) + (1 ID)	0.88	-13957.00	0.07	
~ bs(Age, df = 3) + (1  ID:Country)	0.88	-13735.90	0.07	73.2%
~ bs(age_perform, df = 3) + (bs(age_perform, df = 3)   ID:Country)	0.93	-15704.20	0.06	99.6%
~ bs(Age, df = 3) + (bs(Age, df = 3)   ID:Country) + (1 Year) + (1 Country) + (1 Venue)	0.94	-15792.20	0.05	99.6%

#### 4.1.8 Random Effect Plots

The random effect plots (Figure 4.19 and Figure 4.20) show the variability in the levels of the random effects. The red and blue colour coding show positive (blue) and negative (red) effects. For example, there looks like there is little variability between venue (as expected) (Figure 4.20 - left) but considerable variability between athletes and within country (Figure 4.19 and Figure 4.20 - right). Figure 4.20 shows the differences at each B-spline degree of



freedom and shows the important sources of variability that needs to be (and are) captured in the model.

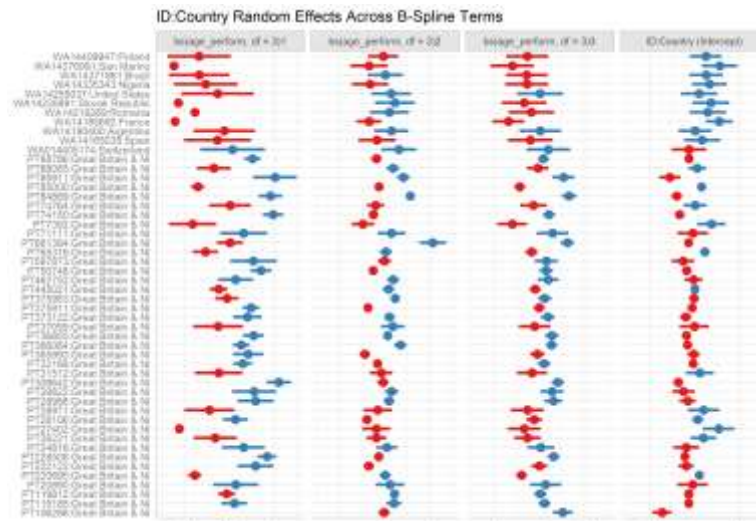


Figure 4.19: Plot of random effects across b-spline terms.

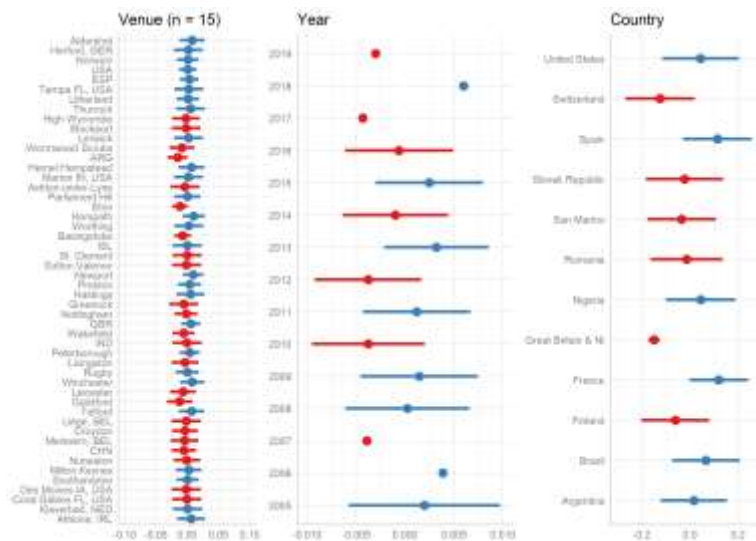


Figure 4.20: Venue, year, and country random effects plot.

#### 4.1.9 Plots of the Fitted Data

Figure 4.21 compares performance for three randomly selected athletes with a smoother fitted through the fitted data from the final model. It provides a convincing visual argument that the model is a reasonable fit to the data.

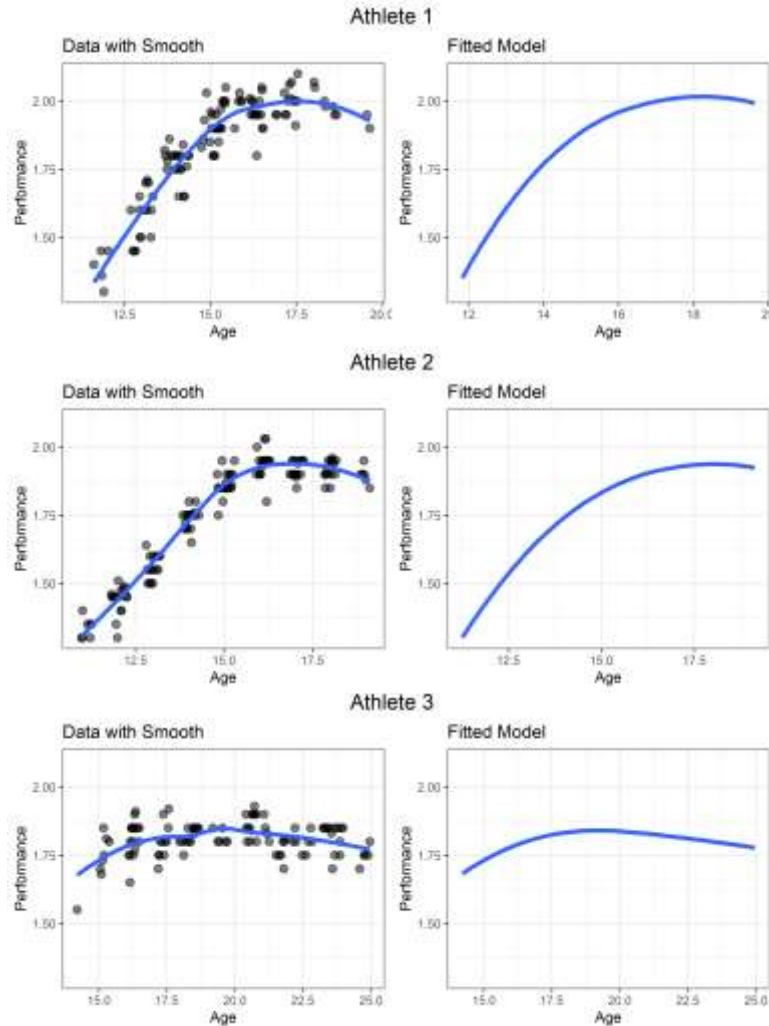


Figure 4.21: Smoothed data vs fitted data.

## 4.2 Modelling Summary

- A set of models, increasing in complexity from a simple linear regression model to a model incorporating B-splines in the athlete level random effect with additional flexible functional forms for covariates and for the inclusion of additional random effects, were considered.
- The final model was based on model adequacy (in terms of plausible underlying assumptions), performance and domain knowledge.

- Visual inspection of residual plots from the final model did not reveal any obvious deviations from homoscedasticity or normality, apart from the QQ plots having a “long tail” appearance.
- All the smoothing methods used had similar RMSE scores and B-splines were chosen due to their flexible nature.
- The inclusion of a hierarchical term did not improve RMSE,  $R^2$  or AIC.
- Venue, Year and Country were included in the model as it was decided that these are useful (based on domain knowledge) in terms of accounting for unexplained variability in the response.
- RMSE of all models was of an acceptable magnitude in practical terms.

Although the main focus of the project was on model building, a discussion on the potential predictive ability of the final model is of interest and will be the focus of the next chapter.

## 5 Chapter 5: Using the Model to Predict future Performance

### 5.1 Introduction

All statistical models can be used for prediction, but this is not often their primary purpose. If the goal had been to generate a single prediction with no regard to uncertainty in the (point) estimate, then a machine learning approach such as a regression tree would have been an approach to consider. There are concerns in using such an approach however as a machine learning approach is not suitable if a measure of uncertainty is needed at the level of the predicted value and not simply for the algorithm as a whole. Given the hierarchical structure of the data, an approach that is capable of modelling both the variation and correlation in the error structure is required, which makes a machine learning approach less attractive.

The suitability of a machine learning algorithm is typically assessed using a classic ‘train/test’ regime where a random sample of athletes were used to ‘train’ the model and the final model was tested on ‘unseen’ athletes comprising the test set. This approach is suitable regardless of how the proposed model is built and is one that can (and should be) used to formally assess the predictive ability of the proposed LMM. What follows however is not an exhaustive formal comparison of the predictive ability of the final model. An informal investigation is presented, based on a simple case study, where an investigation is made on how the predicted values change as more data becomes available on an athlete. This is achieved by iteratively including more data to the model to assess how an athlete’s future predicted performance changes over time as the model adapts and learns from the athletes’ previous performances.

### 5.2 Prediction for Athletes Currently in the Model

In Figure 5.1 overpage the raw data and estimated performance trajectory are presented for three randomly chosen athletes.

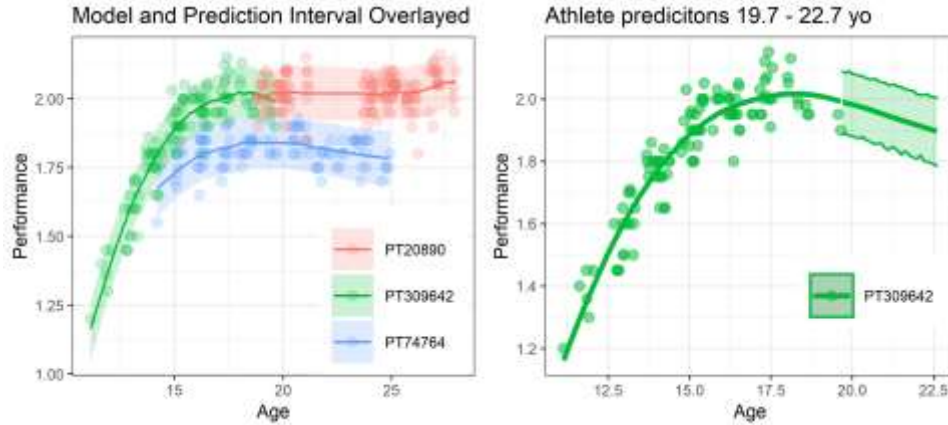


Figure 5.1: Model with prediction interval for a selection of athletes.

The predicted performance for an athlete (PT309642) for 5 different ages are given in Table 5.1 with accompanying upper and lower prediction intervals, as the uncertainty is required at the level of the athlete. In athletic terms this interval is very wide (19-20cm) and therefore may not be of much practical use to a coach.

Table 5.1: Point estimate and prediction intervals for an athlete.

Age	Point Estimate	Upper Bound	Lower Bound	Range
19.7	1.99	2.09	1.88	0.20
20.2	1.98	2.07	1.87	0.20
20.7	1.96	2.07	1.86	0.21
21.2	1.94	2.04	1.85	0.20
21.7	1.92	2.03	1.83	0.20

### 5.3 Sequential Predictions for a New Athlete

The next step undertaken was to investigate how the model adapted as new data for a new (unseen) athlete was added. The aim here was to make a prediction of an athlete's

performance at 18 years of age, initially using the model alone and then by incorporating his performance measures sequentially up to 18 years of age, whilst observing how the model adapts and changes.

The initial prediction of performance was made using only the most ‘basic’ information available i.e., an 18-year-old British athlete in 2018. The model provided the following prediction based on this initial information:

*Table 5.2: Initial estimate from final model for 18-year-old athlete.*

Model	Age	Point Estimate	Upper Bound	Lower Bound
Initial Model	18	1.91	2.00	1.81

The initial model predicted a performance of 1.91m at 18 years with a prediction interval of 1.81m to 2.00m. This is a very reasonable estimate given the sample mean for all athletes aged 18 is 1.92m.

Performance data for the same athlete at age 13 through to age 18 was then added year by year and the model calculated each time and then used to create a point estimate and prediction intervals for that model. The point estimates and prediction intervals for the models with additional data in are shown below in Table 5.3.

*Table 5.3: Point estimates and prediction intervals for sequential data.*

Model	Age	Point Estimate	Lower Bound	Upper Bound
Model - No Data	18	1.905	1.812	1.997
Model w/ 13yo Data	18	1.947	1.814	2.082
Model w/ 13-14yo Data	18	2.015	1.904	2.130
Model w/ 13-15yo Data	18	1.931	1.832	2.033
Model w/ 13-16yo Data	18	1.928	1.822	2.020
Model w/ 13-17yo Data	18	1.926	1.831	2.025
Actual Data	17.9	1.90		

As data were added to the model, the point estimate varied, and the prediction intervals became narrower (Table 5.3).

Figure 5.2 shows how the athletes performance trajectory changes over time as more data were added to the model.

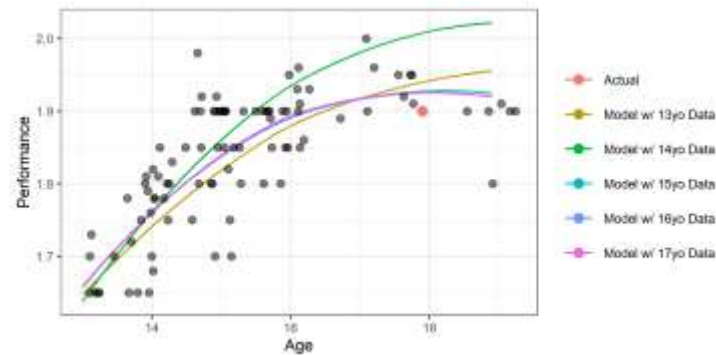


Figure 5.2: Change in model as new data is added.

Up to 14 years of age the athlete was tracking for a high level of performance. However, after results at 15 years of age were added the performance trajectory dropped. The trajectory then stabilised when results at age 16 and age 17 were added, resulting in the last three models having very similar trajectories (Figure 5.2)

The final point estimate was within 2.6cm of the actual performance although the prediction interval was still quite wide (1.83m to 2.03m).

As an illustrative example, the model was tested on a single athlete to demonstrate the fact that the model performs better as more data are available. This exercise suggests that predictions made well into the future may be unreliable and that the model needs to be updated regularly.

Even though the model was tested for a single randomly selected athlete, the final prediction was very close to the one where the actual data was used (1.90m actual to 1.926m predicted). Clearly a more exhaustive analysis is needed to assess the predictive ability of the model, as highlighted in the next chapter where the conclusions and further work are discussed.

## 6 Conclusion and Further Works

The main aims of this project were to, i) create a database of athlete performance records and ii) build an appropriate statistical model to model an athlete's performance in the High Jump over time.

The initial aim was achieved by writing and deploying R scripts to web-scrape publicly available databases of relevance, and then creating a user-friendly R Shiny application to allow coaches to explore the data in an informative manner.

The second aim proved more difficult to achieve due to the repeated measures, hierarchical structure of the data, and the non-linear relationship between age and high jump performance. The final model used a novel approach to allow flexibility at the level of the athlete in an efficient manner while accounting for correlation in the athlete's performance over time, the athlete's country of origin and the competition venue in question.

The results from this project suggest that a LMM using B-splines at both the random and fixed effect level was an appropriate model for estimating the future performance in the high jump. The selected model also accommodates the serial correlation and hierarchical nature of the data.

A simpler model using polynomials to model the functional form for Age produced a slightly better fit (based on AIC). Polynomials however can produce undesirable results at the boundaries whilst B-splines with their flexible nature generally do not have this issue and were therefore chosen as the smoothing method.

The introduction of the B-spline at the random effect level within the nested random effect helped account for a significant proportion of the unexplained variance in performance. This approach accommodated difference in athletes within a country and allowed not only for different athletes within different countries to have a different intercept, but also allowed for athletes within different countries to improve at different rates.



The diagnostic plots generated for the final model suggested that the assumptions underlying the model were plausible. Plots of the fitted values looked sensible, both in shape and value, from what is known about this sport.

## 6.1 Project Limitations

The limitations of this project are:

**Modelling Methods used:** EDA highlighted various structural qualities in the data which lead naturally to the use of a LMM to model the data. A machine learning approach was ruled out as the primary interest of this project was related to modelling and not prediction. However, it could be argued that a Machine Learning approach would have provided a nice comparison to the statistical modelling approach used in terms of variable importance. The suitability of other modelling approaches such as generalised additive mixed models (GAMMs) could also have been explored.

**Limited Covariates and factors:** There were only a small number of explanatory variables available for modelling the response variable. Additional information such as World Athletics competition category, temperature, humidity, an athlete's number of competitions that season/week/etc, travel between said competitions would have been useful. In theory, these explanatory variables could have been collected by linking the venue location and date to other data sources (e.g., competition environmental conditions from linked weather reports). Competitions performed in heavy rain or in humid conditions would likely show lower performances as would competitions with 'less at stake' i.e., local league meetings vs Olympic Games. All the analyses and inferences in this to project related to a target population of male athletes only. The approach presented here is clearly applicable to all gender identities.

**Using the Model to Predict Future Performance:** The suitability of this model for prediction was not considered in detail as this was not the aim of this project. A simple example of the predictive ability for a single athlete (selected at random) demonstrated how the model could be used (and updated) in practice. Clearly a more exhaustive (out of bag) evaluation is needed using either a 'train and test set' approach or bootstrapping.

Care must be taken when making decisions based on a prediction of future performance, regardless of the modelling strategy employed, as it may be considered ‘fool’s gold’ unless a reliable measure of uncertainty for the predicted value is provided. One of the concerns with using an algorithmic approach is that the level of uncertainty is typically reported for the model rather than for individual predictions. The prediction intervals calculated by the model are quite wide and therefore would not be immediately useful to an athletics coach. This is largely down to the variability in performance over time and the lack of additional explanatory variables.

## 6.2 Practical Applications

The practical applications of the results arising from this project are:

**Goal Setting:** The model could be used by coaches and athletes to set realistic goals for the upcoming season and/or competition cycle. Such a model would allow the coach to generate individualised and attainable targets to help develop an athlete whilst trying to prevent injury.

**Funding and Team Selection:** By extending the response variable to incorporate other events the model could be used to help allocate financial funding to appropriate athletes and help plan team selection for future championships.

**Data Driven Decision Making:** The R Shiny application and models created in this research project are useful for showing how historical athletic performance data can be utilised in an objective manner in parallel with domain expertise provided by coaches.

**Causal Inference:** As this project is mostly an observational study, a causal inference could have been incorporated into the modelling approach. Causal inference is focused on knowing what happens to Y when you change X. Prediction is focused on knowing the next Y given X. Usually, in causal inference, an unbiased estimate of the effect of X on Y is required which can involve adjusting for certain covariates as necessary.

Examples of how causal inference could be used would be to answer questions such as:

- How would Athlete A have performed if s/he had competed at the Birmingham Diamond League instead of the Gateshead Grand Prix?
- How would Athlete B have performed had s/he been from the USA instead of Oman?

A causal approach using the examples above would have involved the use of directed acyclic graphs and domain knowledge to decide which co-variables to control for and include in the model.

### 6.3 Future Research

This project could be extended in the future by pursuing the following ideas:

**Use of Current Data:** The current dataset is very large where many of the athletes may not be contributing useful information. It would be interesting to re-run the models with both aggregated data, i.e., best result for each athlete over a specified age bin or limit the data to athletes who have a 'better' spread of results, i.e., only use athletes with 20 observations over 5 years rather than 20 observations over 1 year.

**Identify Similar Trajectories:** Rather than just focus on prediction, an interesting avenue to explore would be to identify athletes with performance trajectories most like another athlete of interest. An athlete's trajectory is clearly of interest to a coach and the ability to compare trajectories (curves) between athletes would be of considerable value. This would give a coach a sense of the 'company' this athlete is keeping. One method to achieve this could involve a comparison of the Euclidean distance between trajectories to identify those athletes who have similar performances at the same age to give some further understanding as to how an athlete might progress.

**Use of Other Modelling Approaches:** The suitability and performance of other modelling approaches such as Machine Learning approaches and Generalised Additive Mixed Models could be compared to the approach used in this project to see if anything of use emerges.

**Extend to Other Events:** The current modelling approach could be extended to include other track and field events.

**Expanded Web Application:** An expanded R shiny web application could be created that models data for multiple track and field events. In an ideal world the application would:

- regularly scrape the latest performance results from the web to update all the models.
- allow users to upload their results to monitor and compare their progress against other athletes of interest.
- contain a yearly target recommendation system for a specific goal e.g., a 4–5-year target to make the Olympic standard.

## 6.4 Conclusion

In this project, relevant data were web-scraped, and an interactive R shiny application was created to allow athletic coaches to explore these data in an objective manner. Following this, LMMs were fitted to the data to model the relationship between age, country of origin and event venue on performance in the high jump. A LMM that incorporated B-splines at both the random and fixed effect level along with additional random effects, was found to be an appropriate model for modelling performance in the high jump over time.

Appropriate modelling of historical athlete performance is of great importance to athletics coaches and organisations, in order to identify potential future stars of the sport and to allocate financial support using data-driven evidence.

## 7 References

1. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*. 1996;5(3):299–314.
2. (PDF) Peak Age and Performance Progression in World-Class Track-and-Field Athletes [Internet]. [cited 2021 Sep 19].
3. Schulz R, Curnow C. Peak Performance and Age Among Superathletes: Track and Field, Swimming, Baseball, Tennis, and Golf. *Journal of Gerontology*. 1988 Sep 1;43(5):P113–20.
4. Boccia G, Moisè P, Franceschi A, Trova F, Panero D, La Torre A, et al. Career Performance Trajectories in Track and Field Jumping Events from Youth to Senior Success: The Importance of Learning and Development. *PLoS One*. 2017 Jan 27;12(1):e0170744.
5. Pinheiro JC, Bates DM, editors. *Linear Mixed-Effects Models: Basic Concepts and Examples*. In: *Mixed-Effects Models in Sand S-PLUS* [Internet]. New York, NY: Springer; 2000 [cited 2021 Sep 18]. p. 3–56. (Statistics and Computing).
6. Laird NM, Ware JH. Random-Effects Models for Longitudinal Data. *Biometrics*. 1982;38(4):963–74.
7. Schielzeth H, Forstmeier W. Conclusions beyond support: overconfident estimates in mixed models. *Behav Ecol*. 2009 Mar;20(2):416–20.
8. Grueber CE, Nakagawa S, Laws RJ, Jamieson IG. Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology*. 2011;24(4):699–711.
9. *An Introduction to Statistical Learning* [Internet]. *An Introduction to Statistical Learning*. [cited 2021 Sep 18].
10. Zuur A, Ieno EN, Walker N, Saveliev AA, Smith GM. *Mixed Effects Models and Extensions in Ecology with R* [Internet]. New York: Springer-Verlag; 2009 [cited 2021 Sep 18]. (Statistics for Biology and Health).
11. Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. *BMC Medical Research Methodology*. 2019 Mar 6;19(1):46.
12. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Statistical Science*. 1996 May;11(2):89–121.
13. Eilers PHC, Marx BD. *Practical Smoothing: The Joys of P-splines* [Internet]. Cambridge: Cambridge University Press; 2021 [cited 2021 Oct 9].

14. Keele L. Semiparametric Regression For the Social Sciences. Semiparametric Regression for the Social Sciences. 2008 Jan 22;
15. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models [Internet]. Cambridge: Cambridge University Press; 2006 [cited 2021 Oct 9]. (Analytical Methods for Social Research).
16. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. 2015 Oct 7;67(1):1–48.
17. Luke SG. Evaluating significance in linear mixed-effects models in R. Behav Res Methods. 2017 Aug;49(4):1494–502.
18. Kuznetsova A, Brockhoff PB, Christensen RHB, Jensen SP. lmerTest: Tests in Linear Mixed Effects Models [Internet]. 2020 [cited 2021 Oct 9].
19. Knowles JE, Frederick C, Whitworth A. merTools: Tools for Analyzing Mixed Effect Regression Models [Internet]. 2020 [cited 2021 Oct 9].
20. Bartoń K. MuMIn: Multi-Model Inference [Internet]. 2020 [cited 2021 Oct 9].
21. Sakamoto Y, Ishiguro M, Kitagawa G. Akaike information criterion statistics. Dordrecht, The Netherlands: D Reidel. 1986;81(10.5555):26853.
22. Liu H, Zheng Y, Shen J. Goodness-of-fit measures of  $R^2$  for repeated measures mixed effect models. Journal of Applied Statistics. 2008 Oct 1;35(10):1081–92.
23. Rice JA, Wu CO. Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics. 2001 Mar;57(1):253–9.