

Modelling risk factors related to the performance of Two Oceans
half marathon runners between 2012 and 2015.

by

Marlise Jordaan

A thesis submitted in fulfilment of the requirements for the degree of Masters
in Statistical Science

Department of Statistics and Population Studies

University of the Western Cape



UNIVERSITY *of the*
WESTERN CAPE

Supervisor: Dr H Brydon

Co-supervisor: Mrs L Bosman

Abstract:

Limited literature is available on the effect of factors associated with the improvement in performance of Two Oceans Half Marathon (TOHM) runners. This study examined factors associated with the improvement in performance of a runner over a 4-year period for the Two Oceans Half Marathon through the application of a linear mixed model.

A subset of data was identified that according to the literature had an impact on the performance of a half marathon runner. Univariate analysis was conducted to identify factors that had an impact on the performance of the runners and a linear mixed model was applied to the model adjusted for age, gender and body mass index (BMI) containing interaction effects to determine the extent to which these factors influence the change in performance of the runners.

It was found that performance of runners improved from the first and second year the runner competed in the TOHM, but not from the second to the third or the third to the fourth year the runner participated in the TOHM. Furthermore, the study found that the training pace of a runner influences the improvement in performance of the runner.

The linear mixed model was able to identify which factors (and to what extent) influence the change in performance of a Two Oceans Half Marathon runner over a 4-year period.

Keywords: mixed modelling, chronic illness, training load, Two Oceans Half Marathon, endurance running, Strategies to reduce Adverse medical events For the Exerciser (SAFER), cardiovascular disease

Table of Contents

Abstract:	2
Table of Contents	3
List of Figures	7
List of Tables	8
Declaration	10
Acknowledgements	11
Glossary	12
1 Introduction	13
1.1 <i>Background to the study</i>	13
1.2 <i>Statement of the problem</i>	16
1.3 <i>Purpose and aim of the study</i>	16
1.4 <i>Research questions</i>	16
1.5 <i>Conclusion</i>	17
2 Literature review	20
2.1 <i>Introduction</i>	20
2.2 <i>Modelling approaches</i>	20
2.2.1 Regression	21
2.2.2 Single-factor Analysis of Variance (ANOVA)	21
2.2.3 Repeated measures ANOVA	23
2.2.4 Fixed effects model	24



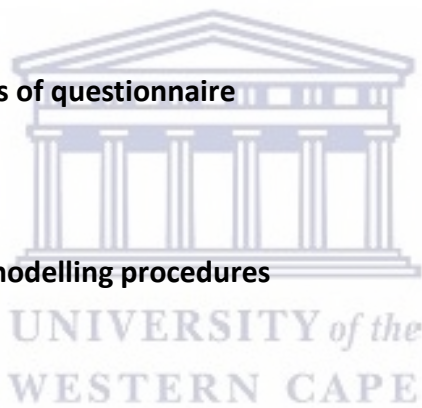
2.2.5	Random Effects Model	25
2.2.6	Conclusion	26
2.3	<i>Linear mixed models</i>	26
2.3.1	Introduction	26
2.3.2	Model structure	27
2.3.3	Covariance structures	28
2.3.4	Estimation for Linear Mixed Model	32
2.3.5	Conclusion	33
2.4	<i>Model selection</i>	35
2.4.1	Information criteria	36
2.5	<i>Confounders</i>	38
2.6	<i>Performance</i>	39
2.7	<i>Conclusion</i>	39
3	Methodology	41
3.1	<i>Introduction</i>	41
3.2	<i>Research questions</i>	41
3.3	<i>Study design</i>	42
3.4	<i>Data collection</i>	42
3.5	<i>Questionnaire</i>	43
3.6	<i>Population and participants</i>	44
3.7	<i>Data preparation</i>	44
3.7.1	Data steps overview	44
3.7.2	Detailed break-down of each step of the process	46



3.8	<i>Variables selection and creation process</i>	51
3.8.1	Selecting predictor variables	51
3.8.2	Variables created	53
3.8.3	Final variables in data set	55
3.9	<i>Pre-model analysis</i>	57
3.9.1	Descriptive statistics	57
3.9.2	Correlation	57
3.10	<i>Modelling and model selection</i>	58
3.11	<i>Conclusion</i>	59
4	Results	61
4.1	<i>Introduction</i>	61
4.2	<i>Final data set</i>	61
4.3	<i>Descriptive statistics</i>	62
4.3.1	Response variable descriptive statistics	62
4.3.2	Confounder variables descriptive statistics	63
4.3.3	Predictor variables descriptive statistics	66
4.4	<i>Checking model assumptions</i>	69
4.5	<i>Covariance structure (R and G side)</i>	72
4.6	<i>Modelling</i>	76
4.6.1	Model 1	76
4.6.2	Model 2	78
4.6.3	Model 3	79
4.6.4	Model 4	80

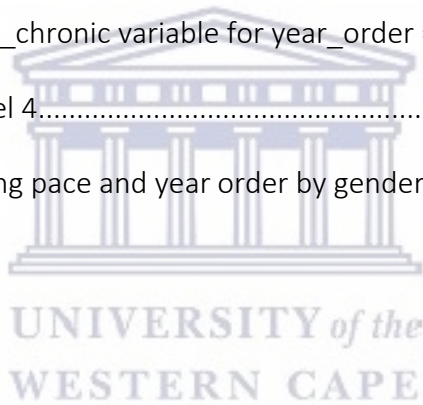


4.6.4.1	Model assumptions for Model 4	81
4.7	<i>Conclusion</i>	85
5	Conclusion and recommendations	85
5.1	<i>Introduction</i>	86
5.2	<i>Limitations</i>	86
5.3	<i>Further studies</i>	87
5.4	<i>Findings</i>	88
5.5	<i>Conclusion</i>	89
6	References	90
7	Appendix A: Main elements of questionnaire	98
8	Appendix B: Questionnaire	100
9	Appendix C: SAS code for modelling procedures	128



List of Figures

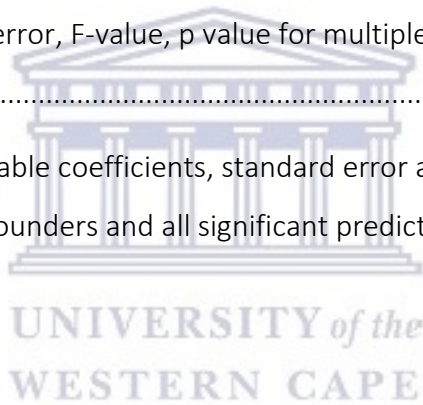
Figure 1. Data captured by SAMRC throughout the intervention period (Sewry, 2022) ...	14
Figure 2. Flow chart break-down per section per chapter	19
Figure 3. Relationship between the explanatory/covariate variables and the response variable (Bobbitt, 2020)	38
Figure 4. Residual plots and statistics	69
Figure 5. Response variable by sc_chronic variable for year_order = 1	70
Figure 6. Response variable by sc_chronic variable for year_order = 2	70
Figure 7. Response variable by sc_chronic variable for year_order = 3	71
Figure 8. Response variable by sc_chronic variable for year_order = 4	71
Figure 9. Residuals plot for model 4.....	81
Figure 10. Interaction with training pace and year order by gender	84



List of Tables

Table 1. Frequencies of entrants over the 4-year period in the original data set.....	46
Table 2. Frequencies of runners’ races for both the Ultra- and the TOHM from 2012 – 2015	46
Table 3. Frequencies of runners for the TOHM between 2012 to 2015.....	48
Table 4. Entrants that did not start the race between 2012 – 2015.....	49
Table 5. Entrants that did not finish the race between 2012 – 2015	49
Table 6. Entrants with more than one race (n = 19 847).....	49
Table 7. Example of “maxrace” variable values for runner_code 96	54
Table 8. Groups of runners.....	55
Table 9. Outline of variables included in the final data set to be analysed further	55
Table 10. Number of entrants per year (N = 19840).....	61
Table 11. Finish time in minute per year order (N=19840)	62
Table 12. Age of entrant by year	63
Table 13. Gender of entrant by year.....	63
Table 14. Overall BMI, weight and height of entrant by year.....	64
Table 15. BMI, weight and height of entrant by year and by gender	65
Table 16. Binary categorical variables descriptive statistics by year.....	66
Table 17. Chronic illness descriptive statistics by year_order for entrants that reported a chronic illness	67
Table 18. Continuous variables descriptive statistics by year	68
Table 19. Lower triangular correlation matrix of Pearson correlation coefficients between “year_order”s.....	72
Table 20. R matrix covariance structures for Model 1 with fit indices.....	73

Table 21. Predicted response coefficients, F-value, p-value and difference for consecutive years.....	77
Table 22. Predicted response coefficients, F-value, p-value and difference between consecutive years adjusted for confounding variables age, gender, BMI	78
Table 23. Coefficients, standard error, F-value and p-value for all predictor variables excluding recreationrunner	79
Table 24. Coefficients, standard error, F-value and p-value for all predictor variables excluding distancerunner	79
Table 25. Multiple model coefficients, standard error, F-value, p value results (for all significant variables and “recreationrunner”)	80
Table 26. Coefficients, standard error, F-value, p value for multiple model with significant interaction	82
Table 27 Predicted response variable coefficients, standard error and difference for consecutive years including confounders and all significant predictor variables	83



Declaration

I declare that this thesis is of my own work, that the reproduction and publication thereof by the University of the Western Cape will not infringe on any third party rights and that this thesis has not been published previously.

Marlise Jordaan

Date 2022/11/11



Signed:



Acknowledgements

Thank you to my supervisors, Dr. Brydon and Mrs. Bosman, for your guidance throughout this project. I have grown immensely throughout this project with your support!

Aan Prof Blignaut. Baie dankie vir Prof se ondersteuning deurgans my akademiese loopbaan.

Thank you to SEMLI and SAMRC for providing me the opportunity to delve into this research.

To my friend Jamie. Thank you for always encouraging me and for proofreading.

Aan my liefdevolle ouers. Sonder julle hulp sou ek nie hierdie punt in my lewe bereik het nie, vir beter ouers kon ek nie vra nie.

To my wonderful husband. To write a thank you for your support would be inadequate.



Glossary

AIC	:	Akaike Information Criterion
ANOVA	:	Analysis of Variance
BIC	:	Bayesian Information Criterion
cAIC	:	Conditional Akaike Information Criterion
CVD	:	Cardiovascular disease
IOC	:	International Olympic Committee
IQR	:	Inter quartile range
LMM	:	Linear mixed model
MLR	:	Multiple linear regression
MM	:	Mixed model
SAFER	:	Strategies to reduce Adverse medical events For the Exerciser
SAMRC	:	South Africa Medical Research Council
SAS	:	Statistical Analysis System
SEMLI	:	Sports, Exercise and Lifestyle Medicine
TOHM	:	Two Oceans Half Marathon
TOUM	:	Two Oceans Ultra Marathon

1 Introduction

1.1 Background to the study

The Two Oceans Marathon event is a series of running races held annually, typically during the month of April, in Cape Town, South Africa. The most popular event of the series is the Two Oceans Half Marathon (TOHM) race, which is 21.1km in length. Due to the 16 000 participant entry limit, it is very common for the event to reach maximum capacity every year. The first Two Oceans Marathon race was held in 1970 and was intended as a training run in preparation for the Comrades marathon. However, it quickly grew in popularity and in 1977 the race organisers changed the race into a 'Pre-race entries only' race. This meant that the participants had to register for the race in advance in order to compete due to the participation limit (Two Oceans Marathon, 2020).

The Two Oceans Marathon hosts both an Ultra (56km) and Half (21.1km) marathon race. On average for the years 2012 to 2015 there were approximately 7500 participants and 11 500 participants for the Ultra and Half marathon races respectively. Due to the nature of these types of events, the number of years a runner has been running recreationally, the intensity the runner trains at, the duration of their training sessions and the frequency of those sessions all have an impact on the performance of athletes (Foster, et al., 1996).

In 2019, the International Olympic Committee (IOC) identified the *prevention of injury and illness* in sport as one of its top research goals. The Sports, Exercise and Lifestyle Medicine Institute (SEMLI) is one of 11 IOC centres globally and by extension is tasked to research and develop effective methods to prevent injuries and illness in sport (International Olympic, 2019). SEMLI is a collaborative effort between the University of Pretoria, the University of Stellenbosch and the South African Medical Research Council (SAMRC). SEMLI is further closely involved with the Two Oceans Marathon event and is primarily concerned with the extent to which runners experience adverse medical effects during these marathon events. Data was collected by the University of Cape Town's Sport Science department from 2008 to 2011 in order to evaluate and analyse the extent of medical related problems runners face during the Half and Ultra Marathon events. The reason for such data analyses was to better

understand the adverse medical events (e.g. collapsing, cramping, heart attacks) that runners experienced during the Two Oceans Marathon event (Schwellnus, et al., 2018).

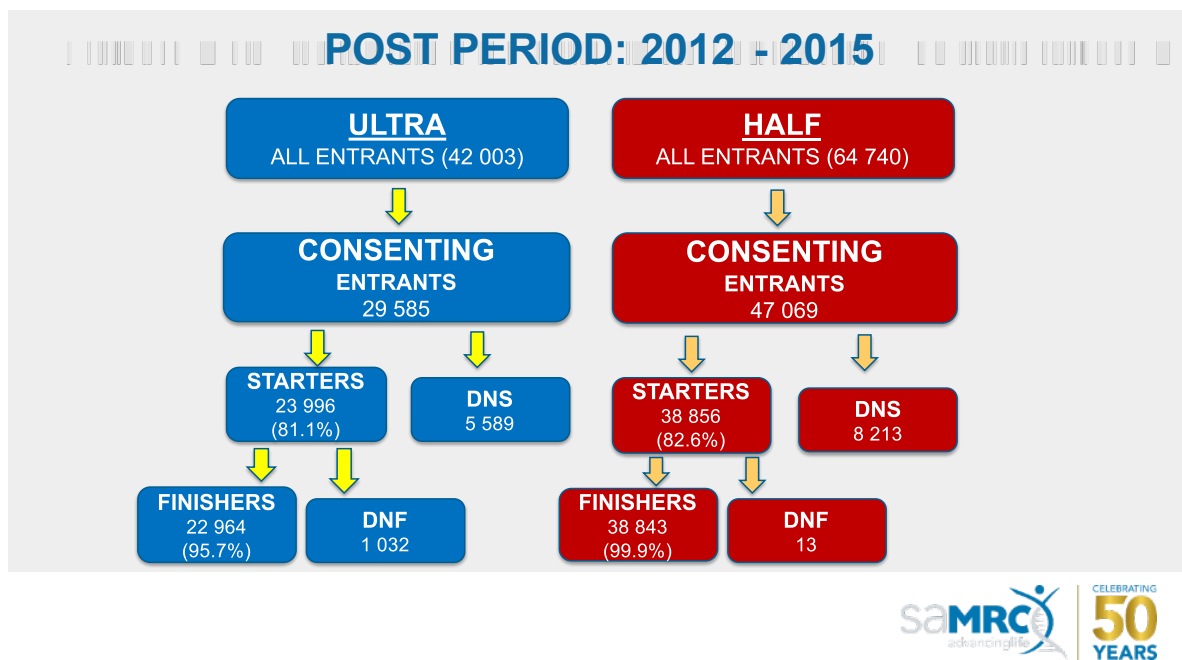


Figure 1. Data captured by SAMRC throughout the intervention period (Sewry, 2022)

Analysis of the results of the adverse medical effects experienced (from the data collected from 2012 to 2015), led to a series of publications known as the SAFER (Strategies to reduce Adverse medical events For the Exerciser) studies. Figure 1 provides a better understanding on how the study population was analysed during the 4 years, e.g. some studies' listed below only included athletes that finished the race (finishers) and excluded athletes that did not finish (DNF) the race. Similarly, for athletes that did not start (DNS) the respective races.

Some of these studies' goals included:

- To determine the need for athletes to complete, according to international guidelines for pre-participation screening of masters/leisure athletes, a medical assessment before competing in a distance running event (Schwabe, et al., 2018);
- To determine if pre-race screening and risk stratification (assigning of risk status) could predict adverse events during the Two Oceans Ultra Marathon (TOUM) running race and therefore, more broadly, in endurance running (Sewry, et al., 2020);

- To identify risk factors that could contribute to the gradual onset running-related injuries (GORRIs) in ultramarathon runners that participate in a large community-based event (Mokwena, et al., 2021);
- To determine if analgesic/anti-inflammatory medication increases the risk of medical complications during the course of endurance races (Rotunno, et al., 2018); and
- To determine if the intervention (in the form of a pre-race screening questionnaire) was successful in reducing medical complications (Schwellnus, et al., 2018)

Based on the results of the aforementioned research, the research group determined the extent to which runners are impacted by the various factors (medical encounters) that were observed during the study period. The study period results led to an intervention from 2012 -2015 in the form of a pre-race medical screening questionnaire that had to be completed by entrants that now form part of the larger data set and SAFER studies.

Apart from the information gathered through pre-race screening questionnaires, two other sources of data were also considered by the research group. Firstly, the data gathered from medical encounters when athletes visited the medical tent on race day and secondly, the performance related data gathered from each entrant's race chip. Race chips are technology used to capture times from each runner when they cross a running mat. This information was used to identify the risk factors associated with illness' and injuries participants experienced during the Two Oceans Half and Ultra marathon events (Schwellnus, et al., 2018).

None of the studies mentioned previously focussed on research and/or analysis related to the performance of the athletes. The main focus has always been on the prevention of injuries and illness. The 16 000 runners taking part in the TOHM each year might be able to extract value from the information if results pertaining to performance related factors can be quantified. Such a quantification should include all available factors that might influence performance. This could lead to an understanding as to what extent performance relates to medical factors, history of illness and injury. A performance metric could also aid runners with the prevention of illness and injury. In this study, the focus will be on only the TOHM.

1.2 Statement of the problem

From the SAFER studies it is clear that there exists a gap in the literature concerning factors that influence the performance of runners in the TOHM events. Furthermore, literature outside the scope of SAFER studies that focus on factors that influence the performance of runners in Half Marathon events, rarely includes comprehensive medical history data such as chronic diseases or injuries of the past year.

From the data gathered over the years it is also clear that missing information is common in studies of this nature due to various reasons. Part of these reasons include that although the completion of the questionnaire is compulsory, it is very lengthy, it relies on the memory of runners about their lifetime and past year running history, and the runner is able to skip certain parts of the questionnaire. Complications of handling missing observations will therefore be part of any analysis utilizing the pre-race screening data.

1.3 Purpose and aim of the study

In Section 1.2, it was established that no factors relating to the performance have been identified. Therefore, the following purpose and aims for this study were identified:

- Assess whether athletes who took part in the Two Oceans 21.1km races more than once, i.e. 2, 3 or 4 years, improve their race time.
- Assess whether the improvement is sustained after the initial improvement.
- Identify which factors (age, gender, training load, history of illness and injury, allergies), if any, contribute to or are associated with the improvement in race time.

1.4 Research questions

Given the objectives highlighted in the previous section, the following research questions have been identified:

- Can a Linear Mixed Model (LMM) be successfully implemented to evaluate to what extent training load, illness and injury impact the performance of a TOHM runner?

- Which factors contribute to an improvement in performance?
- Is a LMM robust enough to fit data with repeated measures within different factor levels?

1.5 Conclusion

Chapter 1 laid the foundation to this study by sketching the background to the study and alluded to the gap in the literature and the specific data set that will be used in this study and furthermore, specified the aforementioned through purpose, aims and research questions. Chapter 2 introduces the concept of models that act as the building blocks of the Linear Mixed Model (LMM). The LMM is the model of focus for this thesis and also the simplest form of the Mixed Model. The aforementioned models include the linear regression model, one-way ANOVA, repeated measures ANOVA, fixed effects model and finally, the random effects model. Thereafter the chapter details all the concepts of the LMM: when to use it, why to use it, the model structure, covariance structures utilised with regards to LMM and finally, the model selection tools for the LMM. The chapter closes by defining the response variable i.e. improvement in performance.

Chapter 3 explains how these aims will be answered. The chapter gives a background to the data by briefly stating how the data was collected via the questionnaire, as well as an overview of the research design of the study executed by SEMLI. Thereafter, the chapter considers any data management steps taken to assemble the final data set needed for the analysis. The process of selecting variables for the analysis, includes any new variables that were created, if variables were selected to answer the specific research questions or if any variables were transformed during the pre-model analysis process. Finally, the chapter describes the models that will be generated in the subsequent chapter to address the stated aims and discusses the selection process for the most appropriate model that will answer the research questions.

Chapter 4 discusses the results obtained. The chapter first discusses descriptive statistics of the response, confounding and predictor variables. Thereafter, the choice of covariance

structure with model fit indices are described. Finally, the chapter clarifies the results of models 1 to 4 as introduced in Chapter 3 through various statistics and figures.

The concluding chapter, Chapter 5, lists all limitations of the study and elaborates on possible future studies that can be conducted on the data set or improvements that can be made to the current analysis. The chapter concludes by summarising the results of Chapter 4 and tying the findings in to the research questions of this thesis.

The flow chart Figure 2 provides a broad overview of each chapter.



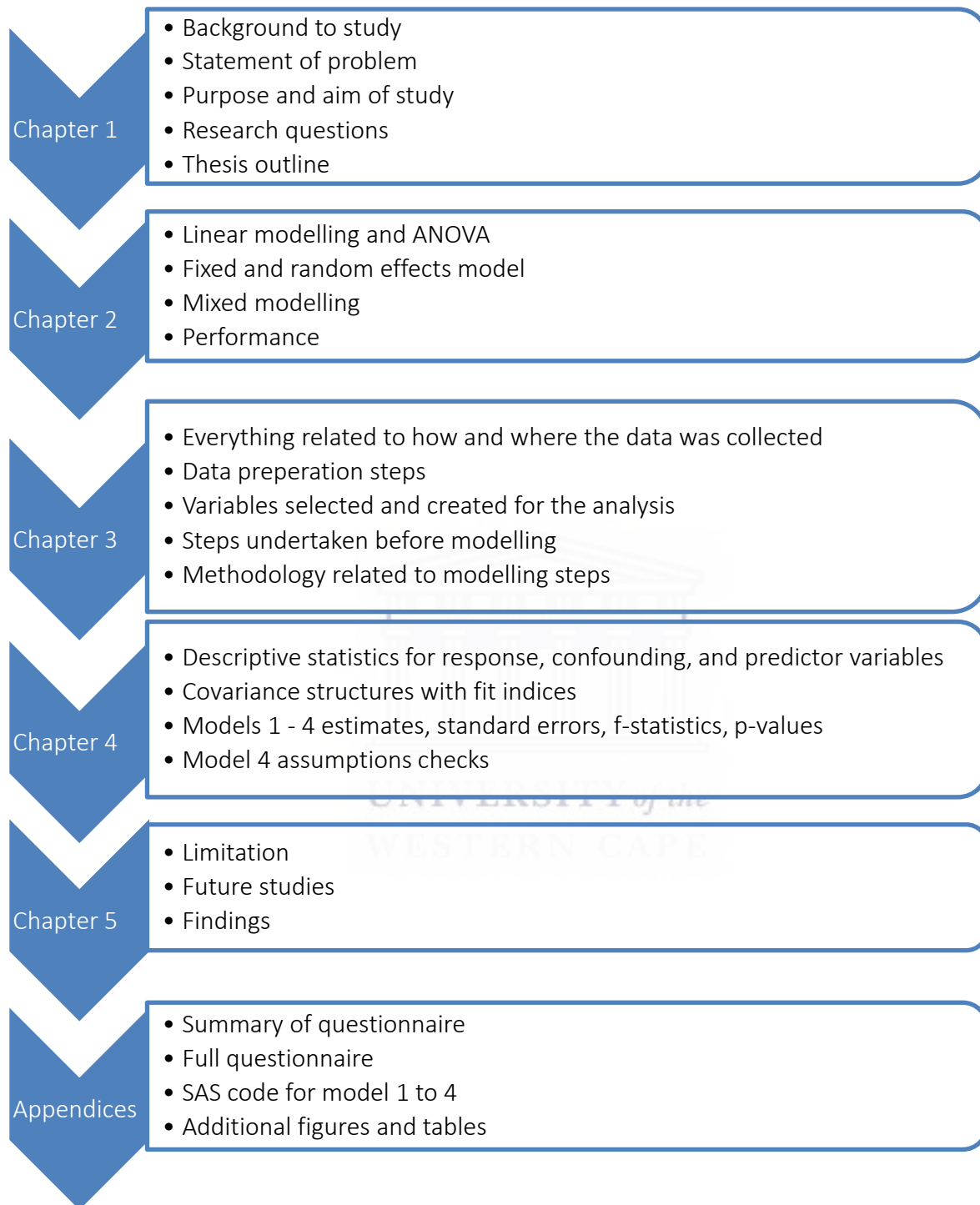


Figure 2. Flow chart break-down per section per chapter

2 Literature review

2.1 Introduction

This chapter discusses the various modelling techniques in order to find an appropriate model to fit the data at hand. The chapters introduce basic modelling approaches, simple linear regression, one-way ANOVA, and then progressively builds towards the description of the Linear Mixed Model (LMM) used in this study.

Section 2.2 describes the traditional multiple linear model, simple and repeated measures ANOVA, fixed and random effects model. Section 2.3 describes the LMM with the various components that can be used in the construction of a LMM, such as covariance structures and estimation. Section 2.4 discusses the model selection criteria that can be used to evaluate the model. Section 2.5 details more information regarding confounding variables. Finally, Section 2.6 details and discusses the measurements of performance.

2.2 Modelling approaches

Researchers often analyse data collected from multiple subjects across repeated trials, e.g. in the Two Oceans Half Marathon (TOHM) setting, repeated trials could be the different years the runners compete in. Various statistical differences might arise in these repeated trials that can be statistically explored in the modelling of this data.

These statistical differences are differences:

- between-subjects, and
- within-subjects, for example (1) the difference between male and female runners or (2) performance differences of the same subject(s) (like runners) over the years that they competed (Larumbe-Zabala, et al., 2020).

Repeated measures can be defined by more than one observation of the same variable over time, for example, when the effects of a training load on a runner are evaluated at different points in time (Kutner, et al., 2005). Data collected in this manner are known as longitudinal data (Brown & Prescott, 2015). Referring to the runner example, the longitudinal study would

be where the runners are observed over a period of time in order to track changes or effects of the training load on the runner's performance.

With this in mind, this section will aim to discuss models that will be able to detect statistical differences between- and within-subjects when dealing with repeated trials.

2.2.1 Regression

The multiple linear regression (MLR) model attempts to model the relationship between two or more independent variables and a response variable by fitting a linear function to the observed data. The MLR model can be expressed as (Kutner, et al., 2005):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad (\text{Equation 2.1})$$

where y_i is the value of the response variable in the i^{th} trial where $i = 1, \dots, n$ (n being the number of observations), $\beta_0, \beta_1, \dots, \beta_{p-1}$ are parameters, $x_{i1}, \dots, x_{i,p-1}$ are the known constants, ε_i are the error terms that are independent, identically distributed as $N(0, \sigma^2)$ and $p - 1$ the number of predictor variables. MLR is the preferred model to use when the goal is to model a statistical relationship between independent variables in order to predict the dependent variable.

Some limitations to this model include:

- Independent variables cannot be highly correlated. Correlation is the extent to which variables are related to each other.
- Error terms or residuals must be normally distributed, meaning that the distribution must be symmetric around the mean.
- It is not able to handle missing observations, unless done through some missing imputation method. Imputation methods refers to methods replacing missing data with plausible values (Chan, 2020).

2.2.2 Single-factor Analysis of Variance (ANOVA)

The Analysis of Variance (ANOVA) can be utilised when explaining the mean difference between k different group responses, i.e. to test the variation in means between different factor levels (Qualtrics, 2022). Factors can be seen as another way to describe a categorical

variable and factor levels refer the grouping of observations with certain values (Kutner, et al., 2005). ANOVA provides more information related to the levels of variability within a regression model (Barron, 1997).

The basic single-factor model formula for ANOVA according to Kutner et al. (2005) is:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad (\text{Equation 2.2})$$

where y_{ij} is the response variable value for the j th trial for the i th factor level, μ_i are population means, and ε_{ij} are independent $N(0, \sigma^2)$, $i = 1, \dots, r; j = 1, \dots, n_i$ where r is the number of factors to be considered and n_i the number of cases for the i^{th} factor.

The ANOVA hypothesis aims to test the null hypothesis:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_r .$$

In comparison to the alternative hypothesis:

$$H_a: \text{at least two means differ}$$

In other words, the null hypothesis aims to prove that the factor level means μ_i are similar and the alternative hypothesis aims to prove that there are at least two means that are different.

Some assumptions that need to be met in order to be able to use the ANOVA test are:

- the variances between the groups should be approximately equal,
- observations are independent, and
- observations are randomly obtained from the population (Anwla, 2020) (Kutner, et al., 2005).

The above assumptions pose a problem when dealing with values measured multiple times from the same subject as is often the case in longitudinal data. These values tend to be more similar than values obtained from different subjects (Zrenner, et al., 2021). Also, within-subject data tend to have correlated errors that vary over time (i.e. they are nonstationary) (West, et al., 2004). Therefore, another model that is able to handle the aforementioned should be considered.

2.2.3 Repeated measures ANOVA

An alternative approach is to use the extended standard ANOVA method, known as ANOVA with repeated measures. The ANOVA for repeated measures method can assess whether the means, on average, differ within-subjects significantly while controlling for the correlations of within-subject observations. Accordingly, when analysing data that are nested observations within-subjects, a repeated measures ANOVA model is more appropriate than one-way ANOVA and multiple regression models, which disregard data dependencies, to be discussed in Section 2.3, in the data (Brown & Prescott, 2015). Nested designs occur when a level of a factor is contained inside another factor level.

The basic formula for the single-factor repeated measures ANOVA according to Kutner et al. (2005) is:

$$y_{ij} = \mu_{..} + \rho_i + \tau_j + \varepsilon_{ij}, \quad (\text{Equation 2.3})$$

where $\mu_{..}$ is a constant or overall mean, ρ_i the fixed factor effects are independent $N(0, \sigma^2)$, τ_j the random factor effects are constants subject to $\sum \tau_j = 0$, ε_{ij} are independent and identically distributed errors with $N(0, \sigma^2)$, ρ_i and ε_{ij} are independent, and $i = 1, \dots, s; j = 1, \dots, r$.

A repeated measures ANOVA is however not ideally suited for all data sets and analyses. It can handle within-subject correlations properly and is optimised to detect changes within-subjects (Larumbe-Zabala, et al., 2020; Goulet & Cousineau, 2019; Singh, et al., 2013). In an example where, multiple runners have run the same race over a period, repeated measures ANOVA is able to detect changes in the finish times of the runners over the period. However, repeated measures ANOVA is not able to accommodate for changes within subgroups of runners or missing data on the response variable if some runners did not participate in every race during that period. Repeated measures ANOVA is sensitive to within-subject variable changes, whereas one-way ANOVA is sensitive to between-subject changes (Brown & Prescott, 2015; Verbeke, et al., 2014).

Due to the characteristics of the aforementioned modelling approaches, they might not be suitable to model longitudinal data that contains missing data or multiple clusters. Clusters

often arise in biomedical studies. A cluster in biostatistics is referred to as a group of subjects that contain similar characteristics. For example, runners that ran the same race repeatedly over 4 years are measured over this time period, but runners that ran the race once, twice or three times are also clustered into groups. Chapter 3 provides further elaboration on the clustering (Brown & Prescott, 2015). Furthermore, if data is missing on the response variable, because repeated measure ANOVA treats the measures separately, the entire subject will be deleted if one measure is missing (Grace-Martin, 2014).

Other models to consider when analysing longitudinal data with multiple levels of repeated measures includes the fixed and random effects models. Both these models are able to accommodate for clustering in a dataset as well as varying dependence within and between clusters. Recall that models discussed up to now could not account for both dependencies (Bell, et al., 2019). The next sections detail these two approaches.

2.2.4 Fixed effects model

A fixed effects model, also known as the restricted Mixed Model (MM) or the ANOVA model I with fixed factor levels, is an extended case of the single factor ANOVA model where more than one factor level is considered. In the simple one-way ANOVA, the μ_i are assumed to be the fixed effects (Demidenko, 2013). The fixed effects ANOVA model for a two-factor balanced (equal factor level sizes, $n_i = \text{constant}$) study can be seen in Equation 2.4 (Kutner, et al., 2005):

$$y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad (\text{Equation 2.4})$$

where $\mu_{..}$ is the overall mean, α_i are constants subject to restriction $\sum \alpha_i = 0$, β_j are constants subject to restriction $\sum \beta_j = 0$, $(\alpha\beta)_{ij}$ are constants subject to restrictions $\sum_i (\alpha\beta)_{ij} = 0$ where $j = 1, \dots, b$ and $\sum_j (\alpha\beta)_{ij} = 0$ where $i = 1, \dots, a$, ε_{ijk} are independent error terms distributed as $N(0, \sigma^2)$, and $i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, n$.

To describe the fixed effects in terms of an example would be to say that the fixed effects are the primary effects that the researcher is interested in. These fixed effects would remain the same if the study was repeated (Group, 2021). For example, in the TOHM, some of the fixed effects in the model would be the training load, chronic illness and history of illness including

injury variables, because if a similar study was repeated, similar results would be obtained for these variables. Using the fixed effects model, however, would not accommodate for the correlation due to the repeated measures, e.g. runners entering the race more than once over a period, in the dataset (Brown & Prescott, 2015).

2.2.5 Random Effects Model

The Random Effects Models (REM) in Kutner et al. (2005) also refers to the fixed effects model as ANOVA model II. The random effects model stated by Kutner et al. (2005) assumes that in a study with one factor, both the factor A main effects, α_i , factor B main effects, β_j , and interaction effects, $(\alpha\beta)_{ij}$, are independent random variables. The random variables can be viewed as the part of the model that causes changes in variance not explained by the fixed effects (Brown & Prescott, 2015). An assumption of a REM is that the explanatory variables have fixed relationships with the response variable over all observations, but that these fixed relationships can vary from one observation to the next (Kumar, 2021).

The random effects model or ANOVA model II for a two-factor study with equal sample sizes n is given by (Kutner, et al., 2005):

$$y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad (\text{Equation 2.5})$$

where $\mu_{..}$ is a constant, α_i , β_j , $(\alpha\beta)_{ij}$ are independent normal random variables with expectation zero and variances $\sigma_{\alpha}^2, \sigma_{\beta}^2, \sigma_{\alpha\beta}^2$, the independent error term, ε_{ijk} , is distributed as $N(0, \sigma^2)$, α_i , β_j , $(\alpha\beta)_{ij}$, and ε_{ijk} are also pairwise independent, and $i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, n$.

In terms of an example in order to better explain the above, in the TOHM, one way to incorporate the random effects would be to consider the group of runners that take part in the race two, three or four times. This type of data causes a clustered effect and can be accounted for in the model by adding a random effect that takes the correlated data into account. In this example, the researcher would therefore not be interested in measuring this specific effect, but the effect of correlation would need to be factored into the model for the

analysis to be accurate. However, the random effects model should be used with caution if the factors are not randomly sampled from the population of interest (Kutner, et al., 2005).

2.2.6 Conclusion

Section 2.2 discussed the different modelling approaches. These models, however, have some limitations to them when addressing more complex data structures. In summary, a model that can handle the aforementioned complexities is needed.

Models containing a mixture of fixed and random effects are referred to as Mixed Models (MM) (Brown & Prescott, 2015). MM is also known as mixed-effect models. The MM could serve as a solution to a data set containing both fixed and random effects as well as address the correlation introduced on various hierarchical levels when the data contains repeated measures over a period (Brown & Prescott, 2015).

The next section will describe the LMM. From this section, the thesis departs from the formulation used by Kutner et al. (2005). The dataset utilised in this study is unbalanced, meaning that not all sample sizes are equal for all factor levels. Seeing as Kutner et al. (2005) does not focus on an unbalanced MM perspective, the formulae stipulated in subsequent sections will be given as in Demidenko (2013) and Brown and Prescott (2015) which is more suitable to explain the unbalanced dataset utilised in this study.

2.3 Linear mixed models

2.3.1 Introduction

The Linear Mixed Model (LMM), the simplest form of the Mixed Model (MM), also known as the “model for repeated measures”, is utilised when measurements over a period of time (i.e. longitudinal, time-series data) are repeated more than once (Salkind, 2010).

An example of where the LMM was utilised in Sport Science, is a study by Avalos et al. (2003). In this study, Avalos et al. (2003) used a LMM to evaluate the relationship between training and performance of a group of swimmers over a period of 3 swimming seasons (repeated measurements over a period of 3 years). The training load and time are therefore the main effects of this simple repeated measures study. The analysis included clustering the

swimmers into 4 groups based on their reaction to training. The study also evaluated individual and group responses as well as between-subject responses. Although this study has a small sample size, the study is a good example of the application of a LMM when the data requires a model that allows variability within-subject and within-group, i.e. subject- and item-variability (Demidenko, 2013).

A LMM is also able to handle missing data better than a fixed effects model by controlling for the imbalance caused by the missing observations. Note that the LMM carries the assumption that the observations are missing at random. Missing data often arises in studies where there is a follow up period involved with people (Brown & Prescott, 2015). A LMM offers a way to handle the missing observations that arise from the aforementioned cases without calling for imputation or deletion of these cases (Gabrio, et al., 2022).

2.3.2 Model structure

The MM expands the fixed effects model, by including random effects, random coefficients and covariance terms in the residual variance matrix.

In general matrix notation, the LMM, the simplest of the MMs, is defined as (Demidenko, 2013):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad (\text{Equation 2.6})$$

where \mathbf{y}_i is a $n_i \times 1$ vector of responses for the i^{th} subject, $\boldsymbol{\beta}$ is the vector of fixed effects coefficients with \mathbf{X}_i the design matrix** $n \times p$ for fixed effects or the first design matrix, \mathbf{b}_i is a $k \times 1$ vector of random effects coefficients with $cov(\mathbf{b}_i) = \sigma_i^2$ with $i = 1, \dots, n$, \mathbf{Z}_i is the design matrix $n \times k$ for the random effects, $\boldsymbol{\varepsilon}_i$ is the vector for the errors with $i = 1, \dots, n$ and each having a zero mean and within-subject variance σ^2 . In the above matrix notation, \mathbf{y}_i , \mathbf{X}_i , $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}_i$ can be viewed 'fixed effects model' and \mathbf{b}_i the random effects part.

**Note that symbols in bold indicate vectors and matrices.*

***A design matrix is also known as a model matrix.*

As mentioned above, an LMM can be viewed as an expansion of a multiple linear model that allows for both fixed and random effects in the model. The LMM accommodates the data to

be both correlated and have heteroskedasticity (where variance across observations is not constant) (Brown & Prescott, 2015). Some of the assumptions of a linear model include homoscedasticity (constant variance of residuals) and independence (observations need to be independent) (James, et al., 2013). Allowance for correlated and nonconstant variability therefore violates these assumptions and adjustments need to be made.

The primary assumptions underlying the analyses performed by LMM are as follows:

- The error terms are normally distributed, have constant variance, σ^2 , and are independent.
- The means (expected values) of the predictor variables are linear in terms of a certain set of regression parameters.
- The variances and covariances of the predictor variables are in terms of a different set of regression parameters.
- All random vectors for random effects, errors and interaction terms are mutually independent (Demidenko, 2013).

2.3.3 Covariance structures

As mentioned before, the LMM requires less stringent independence criteria for repeated measures data, thereby allowing the inclusion of random effects and an appropriate covariance structure in the model (Brown & Prescott, 2015; Wolfinger, 1993). Correlations between observations can be represented explicitly within a LMM. This can be done, for example, by fitting a covariance structure for repeated observations for a subject (Zhang & Chen, 2013). Assuming the random effects follow the Gaussian (normal) distribution, observations from the same subjects can then be correlated without violating the independence assumption (Brown and Prescott, 2015; Wolfinger, 1993; Zhang and Chen, 2013).

Covariance structures are patterns of covariance matrices where the covariance is an unstandardised form of correlation and a covariance matrix represents the covariance values between pairs of variables of a vector in a square matrix form. In LMM, the specification of a covariance structure allows the researcher to introduce modelling variation between-

subjects as well as covariation between measures at different times on the same subject (Littell, et al., 2000).

The variance-covariance of \mathbf{y} , $\text{var}(\mathbf{y}) = \mathbf{V}$ (defining \mathbf{V} as the variance-covariance matrix), can be represented by the equation in matrix notation (Brown & Prescott, 2015):

$$\mathbf{V} = \text{var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}), \quad (\text{Equation 2.7})$$

where $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$ are the same as in Equation 2.6.

If it is assumed that the random effects and errors are uncorrelated, the variance of fixed effects, $\text{var}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$, and \mathbf{Z} , is a matrix of constants, then Equation 2.7 can be rewritten as:

$$\mathbf{V} = \mathbf{Z}\text{var}(\mathbf{b})\mathbf{Z}' + \text{var}(\boldsymbol{\varepsilon}), \quad (\text{Equation 2.8})$$

Now, letting $\mathbf{G} = \text{var}(\mathbf{b})$, with the random effects assumed to follow a normal distribution, $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$ and letting $\text{var}(\boldsymbol{\varepsilon}) = \mathbf{R}$ with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$, then the equation becomes

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}, \quad (\text{Equation 2.9})$$

where \mathbf{G} has the dimension $q \times q$ and q represents the total number of random effects. Note that \mathbf{G} is always diagonal because the random effects are assumed to not be correlated (Brown & Prescott, 2015).

The \mathbf{R} matrix in Equation 2.9 represents the error terms that are uncorrelated in random effects models. $\mathbf{R} = \sigma^2\mathbf{I}$ and therefore in matrix form it is:

$$\mathbf{R} = \begin{pmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{pmatrix}. \quad (\text{Equation 2.10})$$

If Equation 2.10 is considered for repeated measures, the covariance pattern is defined by random effects (Brown & Prescott, 2015). The observations within each subject (runners in the TOHM data set) are presumed to have a specific pattern of covariance (covariance structure) defined across the given time points (or 4-year period in the TOHM data set).

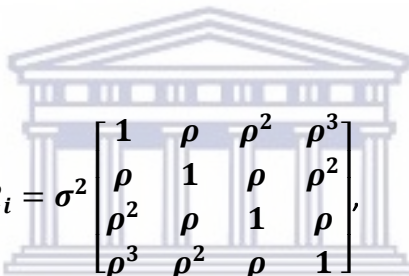
Defining, for example a covariance pattern across a 4-year period for each runner, within the residual matrix \mathbf{R} , the matrix can be written as:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{R}_4 \end{pmatrix}, \quad (\text{Equation 2.11})$$

where each \mathbf{R}_i represents a submatrix of covariances corresponding to each runner.

The covariance structure is defined by specifying a pattern for the covariance terms from the random design matrix. The pattern chosen is usually specified depending on the time component or the number of repeated trials (Brown & Prescott, 2015). The following discussion details four of these structures or patterns for 4 time points.

1. First order auto-regressive



$$\mathbf{R}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}, \quad (\text{Equation 2.12})$$

where $i = 1, \dots, n$. The variances of the auto-regressive (AR(1)) structure are homogeneous and decrease exponentially with distance between periods ($|j - k|$), with $\theta_{jk} = \rho^{|j-k|}\sigma^2$ where σ^2 is the variance of responses, θ_{jk} the covariance, and ρ indicates the correlation between elements. It also suggests that two observations measured at close intervals in time are likely to be highly correlated, but that as the samples grow further apart, they become less so (Kincaid, 2020).

2. Compound symmetry

$$\mathbf{R}_i = \begin{bmatrix} \sigma^2 & \theta & \theta & \theta \\ \theta & \sigma^2 & \theta & \theta \\ \theta & \theta & \sigma^2 & \theta \\ \theta & \theta & \theta & \sigma^2 \end{bmatrix}, \quad (\text{Equation 2.13})$$

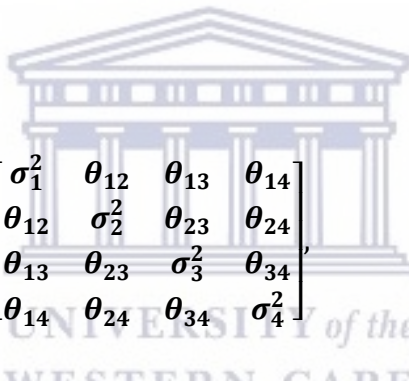
where $i = 1, \dots, n$. The compound symmetry (CS) structure is one of the more simpler covariance structures where covariances are equal (Brown & Prescott, 2015) (Kincaid, 2020).

3. Toeplitz

$$R_i = \begin{bmatrix} \sigma^2 & \theta_1 & \theta_2 & \theta_3 \\ \theta_1 & \sigma^2 & \theta_1 & \theta_2 \\ \theta_2 & \theta_1 & \sigma^2 & \theta_1 \\ \theta_3 & \theta_2 & \theta_1 & \sigma^2 \end{bmatrix}, \quad (\text{Equation 2.14})$$

where $i = 1, \dots, n$. The Toeplitz (TOEP) structure uses a different covariance for every level between the 4 time points. The TOEP structure is analogous to the autoregressive structure. The correlations, on the other hand, do not always follow a similar pattern as in the AR (1) (Kincaid, 2020).

4. Unstructured



$$R_i = \begin{bmatrix} \sigma_1^2 & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{12} & \sigma_2^2 & \theta_{23} & \theta_{24} \\ \theta_{13} & \theta_{23} & \sigma_3^2 & \theta_{34} \\ \theta_{14} & \theta_{24} & \theta_{34} & \sigma_4^2 \end{bmatrix}, \quad (\text{Equation 2.15})$$

where $i = 1, \dots, n$. The unstructured structure (UN) is sometimes also referred to as the general structure. The unstructured covariance structure is the most 'tolerant' of the aforementioned structures, allowing each term to be unique. The variance of responses, σ_i^2 , are different for each time period i . The covariances, θ_{jk} , also differ between pairs of time period j and k (Brown & Prescott, 2015).

5. Variance components

$$R_i = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}, \quad (\text{Equation 2.16})$$

where $i = 1, \dots, n$. The variance components (VC) is the simplest of all covariance matrices and is also used as the default matrix in SAS (Kincaid, 2020). The VC models unique variance components for every random or repeated effect (Kincaid, 2020).

The above-mentioned versions of the covariance structures are straightforward expansions. That is, the variances of all the diagonals of the matrices don't have to be equal (Kincaid, 2020).

There are many more covariance patterns to choose from (e.g. heterogeneous CS, heterogeneous AR(1) etc.) that are not discussed in this section because they will not be applied during the modelling procedure. Selecting the best fitting covariance pattern through fit statistics metrics, can provide additional information on the phenomenon being examined, in addition to providing suitable standard errors for fixed effect estimates. (Brown & Prescott, 2015).

2.3.4 Estimation for LMM

Several methods exist in the literature for the estimation of parameters for the LMM. Two of the most common techniques to estimate the fixed effect parameters are the maximum likelihood (ML) and restricted maximum likelihood methods (REML) techniques (Hariharan & Rogers, 2008).

The ML estimates are the maximized variance parameter estimates of the log-likelihood function (Kutner, et al., 2005). REML and ML differs based on how they estimate fixed effects and variance-covariance parameters (Kutner, et al., 2005). The REML results in a smaller bias for the variance parameter estimates in comparison to the estimates estimated with the ML.

ML and REML provide estimates for \mathbf{G} and \mathbf{R} matrices, denoted by $\hat{\mathbf{G}}$ and $\hat{\mathbf{R}}$, respectively. To obtain estimates of \mathbf{b} and $\boldsymbol{\beta}$, the standard method is to solve the LMM equations (Henderson 1984). The solutions can be written as

$$\hat{\mathbf{b}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1} \mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \quad \text{(Equation 2.17)}$$

$$\hat{\boldsymbol{\beta}} = (\hat{\mathbf{G}}\mathbf{Z}\hat{\mathbf{V}}^{-1})^{-1} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \quad \text{(Equation 2.18)}$$

These are extended normal equations from the linear model and assume that \hat{G} is non-singular, meaning the matrix determinant is non-zero (Henderson, 1984).

When the variance components (the statistic that measures random variation due to random effect only) are estimated using the ML technique, they are treated as a fixed but unknown quantity. However, the degrees of freedom lost due to estimating the fixed effects are not considered. As a result, ML estimates have smaller variances and are biased (McCulloch, et al., 2008).

The ability to compare two models in terms of their fixed and random effects terms is one of the advantages of ML over REML. It is only possible to compare two models with the same fixed effects architecture that are nested in their random effects terms if you use REML to estimate the parameters (McCulloch, et al., 2008), i.e. the REML eliminates the β parameter from the log-likelihood, so that the estimation of the variance is only defined in terms of the variance component (Brown & Prescott, 2015).

Using the REML, the variance-covariance elements are estimated with the ML averaged over all possible values of the fixed effects. Using the REML generally results in less bias concerning the fixed effects in comparison to the ML estimates (Kutner, et al., 2005).

2.3.5 Conclusion

As stated, the standard ANOVA as well as the repeated measures ANOVA have limitations regarding the handling of within- and between-subject effects and are thus not the appropriate method to analyse repeated measures data. LMM resembles classical (M)ANOVA models by replacing a subset of the fixed effects parameters capturing subject effects with random variables.

Adding random effects to classical linear models establishes MM or LMM. Mixed effects representations by definition, contain a subset of parameters associated with fixed and random effects (Brown and Prescott, 2015; Zhang and Chen, 2013). LMMs provide researchers with flexibility to capture correlations between successive measurements and the ability to analyse within-subject and between-subject variability simultaneously, facilitating results (Zhang and Chen, 2013). LMMs also remove the limitations of multiple

regression and ANOVAs for specific data sets (Brown & Prescott, 2015). Another advantage is that LMMs can handle experimental design with unequal number of observations for groups/clusters as well as missing data. In ANOVA models, if a single observation is missing, none of the measurements from that subject's trial are used in the analysis, unless an additional step is performed through imputation (Wolfinger, 1993). This can substantially reduce sample size which leads to increased estimates for standard errors and decreased statistical power (Brown and Prescott, 2015).

As mentioned, missing data causes significant problems in conventional analyses based on ANOVA or ordinary linear regression models. For MM analyses, a complete data set for each subject is not needed, and thus results in more reliable estimates of the effects and corresponding standard errors than traditional methods because it utilises both within- and between-variance estimates (Demidenko, 2013). In LMMs, compared to ANOVA/MANOVA, subjects with more missing values will generally have less influence on estimates and extreme values typically will converge toward the mean (Brown & Prescott, 2015; Barker & Shaw, 2015; Zhang & Chen, 2013).

For example, in the TOHM, if an analysis was done on the runners that started the race in the years 2012 to 2015, the missing data of concern would be the runners that did not finish the race. Another structural missing data component would be runners that did not enter the race over the entirety of the period from 2012 to 2015. If a MM is not used, runners that did not participate in all four years, would simply be deleted from the analysis. This could lead to a reduced sample size and reduced statistical power (Button, et al., 2013).

LMMs are the preferred option in many situations where ordinary linear regression, standard ANOVAs, and repeated measures ANOVAs are not recommended and will allow more sophisticated experimental designs with the capability to answer a richer set of research questions. For example, the ability to model data with multiple sources and variation and to model clustered, such as repeated measures on the same subject, or longitudinal data (Demidenko, 2013). Furthermore, unlike standard methods, mixed-effect models can provide estimates for model parameter coefficients that indicate the direction and strength of the effects (Christensen, et al., 1992; Demidenko, 2013). If the experiment design is simple (i.e.

with no missing data and normal distributions that govern the residuals) then a repeated measures ANOVA is equivalent to an LMM analysis (Brown & Prescott, 2015).

2.4 Model selection

An integral part of any modelling approach is model selection. This also is the case in LMM. The aim is often to choose the most parsimonious model which is simply choosing the simplest model that achieves the intended level of goodness from a larger set of proposed models. Since LMMs can be seen as an extension of linear regression models, many of the same methods can be used for model selection. However, in linear regression models, the observations are independent and in LMMs they are not independent (Demidenko, 2013). The data dependence impacts on model selection by reducing the effective sample size and should be considered in most model selection procedures (Schwarz, 1978). The LMM has parameters that describe the mean structure as well as variance parameters that describe the dependence structure. The selection of a LMM is therefore more complex than in linear regression models (Brown & Prescott, 2015).

The inclusion of irrelevant random effects in a model, on the other hand, would lead to a singular variance–covariance matrix of random effects, producing instability in the model (Ahn, et al., 2012).

A few points about model selection to consider:

- The subject matter and aim of the study are decisive factors in model selection (Diggle, et al., 1994),
- Graphical methods can be employed to assist in model selection (Christensen, et al., 1992)
- Various variable selection methods exist to assist model selection: Log-likelihood, information criteria, shrinkage methods (Tibshirani, 1996), fence method (Jiang, et al., 2008), Bayesian methods, etc.

2.4.1 Information criteria

Fit statistics exist to assist in model selection whereby the best model is the one that optimises some loss function. Some of the widely used criteria include the Akaike Information Criteria (AIC) (Akaike, 1973), conditional Akaike Information Criteria (cAIC) (Vaida & Blanchard, 2005) and the Bayesian Information Criteria (BIC) (Schwarz, 1978).

The aim is to minimise a function, that is the sum of a loss function plus a penalty that is usually dependent on number of parameters in the model. The criteria can be compared for models that fit the same fixed effects and where the covariance parameters are nested (Brown & Prescott, 2015). To define the loss function we can use the log likelihood function, the conditional log-likelihood or the REML.

Log Likelihood function

As mentioned, the AIC, cAIC and BIC incorporates the log-likelihood function with differences in how they utilise this penalty function. The likelihood function defines the likelihood of the model parameters given the data. The function can be used as a measure of goodness of fit with larger values indicating an improved model fit.

The log likelihood function can be given by (Brown & Prescott, 2015):

$$\log(L) = K - \frac{1}{2} [\log|V| + (Y - X\alpha)'V^{-1}(Y - X\alpha)], \quad (\text{Equation 2.19})$$

where $K = -\frac{1}{2}n\log(2\pi)$ is a constant that can be ignored in the maximization process, n = number of observations, mean vector $X\alpha$, and covariance matrix V .

Akaike Information Criterion (AIC)

The penalty function indicates model complexity, meaning that the measure will become larger as more parameters are introduced in the model (Brown & Prescott, 2015). It is based on the Kullback–Leibler distance between the true density of the distribution generating the data, and the approximating model for fitting the data (Vaida & Blanchard, 2005).

AIC can be represented by the following equation (Demidenko, 2013):

$$AIC = -2l_{max} + 2k, \quad (\text{Equation 2.20})$$

where l_{max} is the log-likelihood maximum and k the number of unknown parameters.

The idea is to combine point estimation and hypothesis testing into a single measure, thus formalising the concept of finding a good approximation of the true model in a predictive view, i.e. the smaller the AIC, the better the model fit. As stated by Demidenko (2013) and Vaida et al. (2005), AIC introduces considerable bias in the case of multicollinearity (when several variables are correlated in the model).

Conditional Akaike Information Criterion (cAIC)

cAIC was developed by Vaida et al. (2005) in order to accommodate clustered data sets. The cAIC is derived from the AIC, but assumes that the random effects are known through the variance-covariance matrix or the scaled variance-covariance matrix (Liang, et al., 2008).

The formula for the hypothesised case of the cAIC is given in Equation 2.21 (Burnham & Anderson, 2002),

$$cAIC = AIC + c, \quad (\text{Equation 2.21})$$

where $c = 2 \frac{K(K+v)}{np-K-v}$, v is the number of distinct parameters with $1 \leq v \leq p(p+1)/2$, $K = (k.p) + p(p+1)/2$ with k the number of independent variables and $p(p+1)/2$ unknown parameters. Similarly, to the AIC, a lower cAIC value indicates a better model fit. Note that with large sample sizes, cAIC and AIC will be nearly identical (Fernandez, 2006).

Bayesian Information Criterion (BIC)

On the other hand, the Bayesian information Criterion (BIC) (Schwarz, 1978) can be derived as an approximation to the Bayes factor for testing two hypotheses or from asymptotic arguments to construct criteria which lead to consistent model selection.

The BIC has the formula (Profillidis & Botzoris, 2019):

$$BIC = k \ln(n) - 2 \ln(l_{max}), \quad (\text{Equation 2.22})$$

where k is the number of unknown parameters to be estimated, n is the sample size, and l_{max} is the log-likelihood maximum.

Once again, a smaller BIC value is preferred and indicates a better model fit. Similarly, to AIC, the higher the number of parameters in the model, the larger the penalty is for the BIC. The BIC penalty is stronger than the penalty of the AIC (Wit, et al., 2012).

2.5 Confounders

Confounders are variables that affect the outcome variable. These are variables that are not of primary interest to the researcher but they can affect the response variable and therefore need to be accounted for in the model. If they are not accounted for there will be unexplained variation in the response variable and it can be more difficult to interpret the relationship between the response variable and the explanatory variables (the variables of interest) (Bobbitt, 2020). Figure 3 provides a visual representation of this relationship.

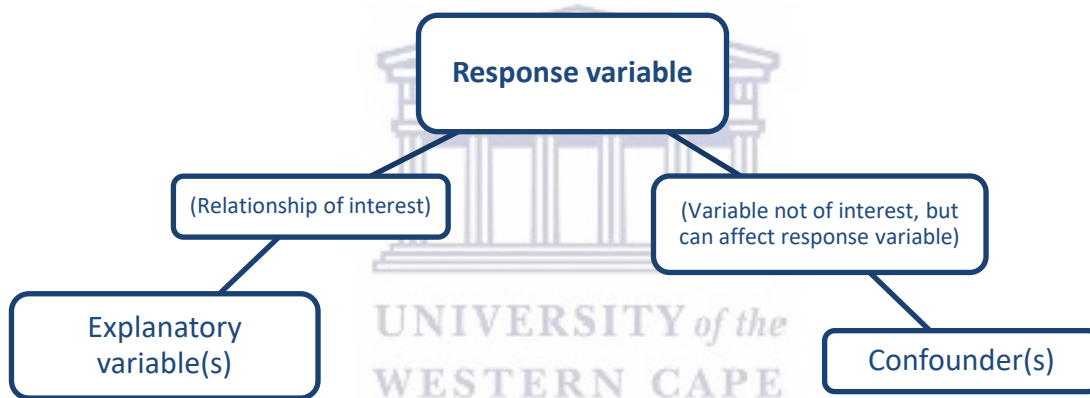


Figure 3. Relationship between the explanatory/covariate variables and the response variable (Bobbitt, 2020)

The majority of inferences such as generalized linear (mixed) regression models, assume that confounders have no explicit effects in the model. However, estimates may be skewed if the assumptions are incorrect and the confounders do correlate with other factors within the model (Eler, et al., 2019).

As previously mentioned, LMMs allows for within- and between-subject variability. It is important for the researcher to determine the variance in the response variable caused by confounders in the LMM. In the case of the LMM, the confounders' impact on the variance

of within-subjects can be seen as level 1 variability and the confounders' effect between-subject variance as level 2 variability (Hu, et al., 2010).

If performance (further discussed in Section 2.6) is measured in the TOHM data set through improvement in the entrants' race times, confounders could include variables like age and gender. By including these variables as confounders in the model, the variability of the response variable is better explained, i.e. some of the variability in the outcome variable can be explained by age and gender.

2.6 Performance

The research question aims to explain the improvement in performance as well as the factors associated with that improvement such as training load variables, history of illness and injury, chronic diseases, allergies, etc. Performance is defined in terms of the time it took the runner to finish a race (Schubert & Astorino, 2013). Improvement in running performance can therefore be defined as an improvement in the finishing time that the runner acquired from one race to the next. It can also be defined as an improvement in the amount of time the runner took to finish the race given that the route and distance is the same from, e.g. 2012 TOHM in comparison to 2013 TOHM.

Various factors influence the performance of a runner of which some are measurable and some are less so. The factors less measurable include, for example, the psychological challenges a runner might face on the day of the race (Boullosa, et al., 2020). Some of the more measurable factors influencing performance include training load (measured in terms of intensity defined as the average weekly distance the runner trained, duration defined as the average pace the runner trains at, and frequency defined as average weekly training frequency of runner), chronic illness, injuries, age, sex, body mass index (BMI) and experience of runner (Venturini & Giallauria, 2022; Boullosa, et al., 2020).

2.7 Conclusion

Chapter 2 began by introducing different modelling concepts, such as the MLR, ANOVA, fixed and random effects models. Thereafter concepts such as estimation, covariance structures and model selection were discussed in the context of LMMs. The chapter concluded by

detailing what performance is in terms of the research question and how it's defined in this study. Chapter 3 will discuss the data in more detail as well as the methodology that will be used in this study. This chapter will start by explaining the steps taken to clean the data. The data will then be described where after the process of applying the LMM and obtaining fit statistics will be discussed.



3 Methodology

3.1 Introduction

Chapter 3 details the approaches, formulae and designs used in the analysis. Section 3.2 provides a recap of the research questions. Sections 3.3 – 3.6 offer a background to the data, including the study design, data collection process and questionnaire that was utilised by the research group (SEMLI). Section 3.7 provides an account of how the final data was created from the original data set provided by the SAMRC. Section 3.8 reports the variables that were used for analysis including the subset selection, new variable creation and any investigation that was done in the pre-model analysis. Section 3.9 details the modelling process to answer the aims as outlined in Chapter 1. Finally, Section 3.10 lists the model selection procedure.

3.2 Research questions

This study is based on the following research questions:

- Can a Linear Mixed Model (LMM) be successfully implemented to evaluate if and to what extent training load, illness and injury impact the performance of a Two Oceans Half Marathon (TOHM) runner?
- Which factors contribute to an improvement in performance?
- Is a LMM robust enough to fit data containing repeated measures within different factor levels?

The main aim of this study is to determine if a mixed modelling approach can be used to evaluate to which extent the number of races, training load, illness, injury and demographic factors influence the performance of the entrant.

Applying a LMM, which allows for fitting a covariance structure to incorporate the correlated data, will provide a better model fit to the data and therefore more valid fixed effects, smaller standard errors and thus an improved power to assess the hypothesis set out in the aim. To recap, the main objective of this study is to determine possible significant differences in performance over the 4-year period and to identify the significant factors influencing improvement in performance over a 4-year period. The data collection, questionnaire, data

coding process, editing of the data and choice and encoding of variables are discussed in the following sections.

3.3 Study design

The design of the TOHM data is panelled data with a repeated measures component. Panelled data, according to Baltagi (2008) is longitudinal data or cross-sectional time-series data with observations about different cross sections, such as years across time. In the TOHM, the entities are the runners, the time period is 4 years and cross sections refers to the years.

3.4 Data collection

The data was collected between 2012 to 2015 (i.e. over 4 years). Runners used the same unique runner code from year to year so that runners participating over the 4 years could be followed up.

Data used in this study, was gathered from 3 sources:

1. A pre-race medical history questionnaire that was compulsory to complete on entry (Sewry, et al., 2020). Note that registration opened 3-5 months before the start of a race.
2. Race day data that was collected on the day of the race such as demographic information, performance related data, etc.
3. Medical complications and injury data was collected by an externally contracted team. This data was provided to SEMLI in an Excel format.

Demographic data (height, weight, previous participation, and previously completed races), medically related and training related data was collected via a compulsory online pre-race screening questionnaire that all entrants, defined as a runner registering for a race, completed from 2012 to 2015. Data from the pre-race screening questionnaire is self-reported data.

Race day data includes demographic information such as gender and age, previous participation in races and performance related data (number of starters and finishers, and

finishing times of runners). The demographic and race day data is also publicly available online.

Medical complications and injuries data consist of the entrant's name, surname, race number and all medical encounter details.

Ethical clearance was approved by the Research Ethics Committees of the Faculty of Health Sciences of the University of Cape Town (REC 009/2011 and REC R030/2013) and the University of Pretoria (REC 433/2015).

Data provided by the SAMRC are de-identified and excluded non-consenting entrants. All 3 data sources detailed above were merged using the unique race number of each entrant. This data set is referred to as the "original long data set".

3.5 Questionnaire

Overview of the questionnaire:

- The training related portion of the questionnaire relates to years of recreational running, weekly running distance, and training running speed.
- The medical portion of the pre-race screening questionnaire was sub divided into main categories namely cardiovascular disease (CVD), symptoms of CVD, risk factors for CVD, other chronic disease, general prescription medication use, medication use during racing, history of any allergies, injury and a past history of collapse during racing (Sewry, et al., 2020).
- According to Schwabe, et al. (2018), the online pre-race screening questionnaire's main elements consisted of 2 injury related questions and 13 questions relating to medical history of the entrant (cardiovascular disease (CVD), symptoms of CVD, risk factors for CVD, other chronic disease (respiratory disease, metabolic or hormonal disease, gastrointestinal disease, nervous system disease, renal or bladder disease, haematological or immune system disease, cancer, allergies). No questions were open-ended. All questions were either numerical or based upon the selection provided. Only when the entrant indicates "other" was there sometimes a space to elaborate on the response.

- A breakdown per section of the main elements of the medical screening questionnaire can be found in Appendix A.
- The full questionnaire, as provided by SEMLI, is attached in Appendix B.

3.6 Population and participants

All runners who registered for the 21.1km Two Oceans race from 2012 to 2015 were included in the data provided. Furthermore, only data for runners who completed at least two 21.1km races (excluding the 56km Two Oceans race) were used in the analysis. Runners also needed to be a minimum age of 16 years to enter the TOHM.

3.7 Data preparation

The data set provided by the SAMRC was in the format of a SAS data set. All editing and analysis were subsequently done in SAS 9.4.

3.7.1 Data steps overview

The data coding procedure is given as an overview in steps below:

Step 1: Data set provided imported into SAS.

Step 2: Data transformed from a long format to a wide format (see Step 2 for further details on formats). No cleaning or editing was conducted before this step as editing out any observations before Steps 2 - 4 could lead to the inclusion of runners that have entered for both the Two Oceans Ultra and Half Marathon.

Step 3: Once all entries of runners were represented as one observation per runner in a wide format, runners that entered for the Two Oceans Ultra Marathon (TOUM) in any of the years over the 4-year period, were excluded. This was done by transforming the data into a long format by merging into the original data set that contains all the demographic, training history, past injury and chronic illness related information.

Step 4: Years 2012 to 2015 were sorted by runner (“runnercode”) in ascending order. The year and order in which the runner entered the Two Oceans marathon was created

as variable “year_order”, with values ranging from 1 to 4 for this variable. Each entry/observation of a runner was therefore assigned a year_order value depending on if it was the first, second, third or fourth time the runner entered for the TOHM in the period. See Section 3.8.2 for more detail on how this variable was created.

Step 5: Entrants that did not have finishing times were excluded. This included entrants that did not start or did not finish the TOHM.

Step 6: Runners that only entered for one TOHM were excluded, i.e. runners were selected that have run the TOHM for at least 2 years between the years 2012 to 2015. In other words, all runners were excluded who only entered for the TOHM only once over the 4-year period. This was done in order to answer the research question relating to the improvement in performance over the 4-year period.

Step 7: The data set was provided by the SAMRC and required minimal cleaning. However, race day data and responses to the questionnaire were checked for any inconsistencies according to guidelines provided by the SAMRC and SEMLI. More details are described in Section 3.7.2 to ensure the responses are in the correct format.

Step 8: The relevant variables relating to performance and outcome variable were investigated and described.

Step 9: The “PROC MIXED” procedure was used in SAS to conduct all relevant analysis.

The below section contains more detail on some of the above summarised steps.

3.7.2 Detailed break-down of each step of the process

Step 1: Original data set contained 76654 observations from entrants of both the TOHM and TOUM (see Table 1).

Table 1. Frequencies of entrants over the 4-year period in the original data set

Race	n (%)
Ultra-marathon	29 585 (38.6%)
Half marathon	47 069 (61.4%)

Step 2: The original data set shows each entrant by year the runner has entered the TOHM. Once the data set is transformed from long to wide format, it is possible to view the amount of entries by runner over the 4 years (see Table 2). The value of 1 denotes the runners that ran a Two Oceans Ultra Marathon and the value of 2 denotes the runners that ran a Two Oceans Half Marathon in Table 2. The row of interest for this study, are the last four rows in the table.

Table 2. Frequencies of runners' races for both the Ultra- and the TOHM from 2012 – 2015

First entry	Second entry	Third entry	Fourth entry	n (%)
1	.	.	.	10078 (21.9)
1	1	.	.	3455 (7.23)
1	1	1	.	1918 (4.01)
1	1	1	1	1056 (2.21)
1	1	1	2	34 (0.07)
1	1	2	.	81 (0.17)
1	2	.	.	320(0.67)
1	2	1	.	29 (0.06)

1	2	1	1	1 (0.00)
1	2	1	2	1 (0.00)
1	2	2	.	52 (0.11)
1	2	2	1	2 (0.00)
1	2	2	2	21 (0.04)
2	1	.	.	831 (1.74)
2	1	.	1	1 (0.00)
2	1	1	.	150 (0.31)
2	1	1	1	113 (0.24)
2	1	1	2	19 (0.04)
2	1	2	.	46 (0.10)
2	1	2	2	1 (0.00)
2	2	1	.	252 (0.53)
2	2	1	1	5 (0.01)
2	2	1	2	1 (0.00)
2	2	2	1	75 (0.16)
2	2	2	2	1232 (2.58)
2	2	2	.	2723 (5.70)
2	2	.	.	6074 (12.71)
2	.	.	.	19213 (40.21)

1 = TOUM 2 = TOHM

Step 3: Only runners from TOHM and not TOUM were selected. This was done by merging the wide data set with the original long data set (see Section 3.4) in order to discard all runners that have entered for the TOUM, but still retain all information regarding performance, training and medical for the runners that have only taken part in the TOHM between 2012 and 2015.

Table 3. Frequencies of runners for the TOHM between 2012 to 2015

First entry	Second entry	Third entry	Fourth entry	n (%)
2	.	.	.	19213 (65.7)
2	2	.	.	6074 (20.77)
2	2	2	.	2723 (9.31)
2	2	2	2	1232 (4.21)
Total				29242 (100)

2 = Entry for TOHM

Note: Runners in the first row (19 213) who only entered once, were not included in the final data set (refer to step 7).

Step 4: Entrants that did not finish (DNF) or did not start (DNS) the race, were excluded. Note that the below table refers to entrants and not runners. Each runner can have many entries. The long data set represents each entry of the runners by row, therefore the total n of the below Table 4 is larger than that of Table 3. Table 4 represents the entrants that DNS and Table 5 represents the entrants that DNF the race.

Table 4. Entrants that did not start the race between 2012 – 2015

Entrants DNS vs start	n (%)
Yes (entrants that did start)	36 548 (82.21)
No (entrants that did not start)	7910 (17.79)

Table 5. Entrants that did not finish the race between 2012 – 2015

Entrant DNF vs finish	n (%)
No (entrants that did finish)	36 535 (99.96)
Yes	13 (0.04)

Step 5: Runners were disregarded that only entered the TOHM once. This study is only interested in the improvement in performance of the runners, therefore a minimum of two finish times per runner is needed. The numbers given in Table 6 are the numbers that will remain in terms of the type of race and the number of races the runners competed in, i.e. any further removal of observations will occur due to incorrect values given by the runner or the time mat.

Table 6. Entrants with more than one race (n = 19 847)

year_order	n (%)
1 st race	7945 (40.03)
2 nd race	7945 (40.03)
3 rd race	3053 (15.38)
4 th race	904 (4.55)

Step 6: Checks were done on the times the entrants crossed race mats. A race mat tracks or records the time entrants cross specific points on the route. The “time-checks” were done to ensure that the $(t + 1)^{th}$ mat time was larger than the t^{th} mat time. The finishing times (“time_f_min”) were checked to ensure that the minimum time is not smaller than the reported winner’s finishing time for the respective years and that the maximum finishing time is not greater than the official cut-off time (7 hours) for the event. One runnercode (177088) had a finishing time greater than 7 hours and was deleted from the data.

Frequency tables were created of all categorical variables to be investigated. The gender codes recorded of runner codes 137146 and 143651 were not consistent throughout the years and were deleted from the data set. Descriptive statistics such as the mean, standard deviation, median, quartile 1, quartile 3, total number of observations, total number of missing observations, were obtained for all continuous variables and the results analysed. Two training load variables, “timestrainrace” and “trainingdistance”, contained 11.09% missing observations. If the number of missing observations in a variable is greater than 10%, bias could be introduced into the results of the analysis if the variable is included (Bennett, 2001). Therefore, these two variables were not included in the study based on Bennett (2001).

Step 7: Variables selected to include in the modelling procedure were further explored. More detail on this step is discussed in Section 3.9. The results are noted in Chapter 4.

Step 8: After all variables were explored and adjusted, analysis was conducted on the final data set (see Section 2.8). Model selection took place in the form of various model fit indices (see Section 2.9).

The next section details any variables that were created to assist the modelling process.

3.8 Variables selection and creation process

3.8.1 Selecting predictor variables

The following section outlines the selection of variables hypothesised to be related to running performance. These variables were chosen as a subset and will be used in the modelling procedure in order to evaluate the impact on performance.

1. Training load factors pertaining to running can be described by the following variables (Foster, et al., 1996):
 - Intensity (average distance run per week),
 - Duration (average training pace in minutes per kilometre), and
 - Frequency (average times run per week).

The assumption is that running performance will increase with an increase in training load (Foster, et al., 1996). The variables “recreationrunner”, “distancerunner” and “training_pace” relating to training load were therefore included in the model to assess these variables’ effect on running performance. As mentioned in Section 3.7.2, variables “trainingdistance” and “timestrainrace” were initially included in the final dataset but because of the amount of missing observations contained in each variable, these variables were excluded from the analysis.

2. Running injuries lead to a decrease in training load of the runner and will influence the runner’s level of performance during races (van Mechelen, 1992). The variable “recent_run_injury”, included in the analysis, represents running injuries.
3. According to Mokwena, et al. (2021), running related injuries are influenced by history of chronic disease and history of allergies. In the dataset, the variables “sc_chronic” and “allergies” represent these factors and are included in the model construction.
4. According to Knechtle & Nikolaidis (2018) women tend to reach their fastest half marathon race times at a slightly younger age than men, but both groups tend to

reach peak performance for half marathon races between the ages of 35 and 39. Therefore, age and gender are included as confounders in the model.

5. The variable “bmi” is used to describe the body mass index of an entrant and is calculated by using the variables length and weight (contained in the data). The formula used to calculate BMI is (Messiah, 2013):

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{length}^2 \text{ (m)}}. \quad (\text{Equation 3.1})$$

The optimal BMI for peak performance for marathon racing for men is approximately $19,8 \text{ kg} \times \text{m}^2$. For women, the optimal BMI range for peak performance is $18,28 \text{ kg} \times \text{m}^2$ (Marc, et al., 2013). If a runner’s BMI changes and is then within these ranges while it was previously outside of these ranges, their performance can be influenced. The variable BMI is therefore also included in the model construction as a confounder as it could potentially influence the improvement in performance of the TOHM runner.

Variables that were captured as part of the pre-race screening questionnaires from 2012 to 2015 included training load, confounding- and chronic illness variables:

- Training load variables are described in terms of the intensity (average weekly distance runner trained), duration (average pace the runner trains at) and frequency (average weekly training frequency of runner, i.e. how many times a week do they exercise).
- Illness variables are described in terms of chronic disease variables.
- Age, gender and BMI is often included in various models as potential confounding variables in Sport Science literature (Gomez-Molina, et al., 2017). Previous SAFER studies concerning Two Oceans marathons (Rotunno, et al., 2018) adjusted for age and gender in their analysis. Therefore, age, gender and BMI are included as confounding variables.

Variables that were captured as part of the performance related data (from entrant’s timing chips on the race day) from 2012 to 2015 included all time related data from the entrant:

- The response variable to measure performance is the finishing time of the runner (in minutes).

3.8.2 Variables created

To answer the research questions, the following variables were created:

- The variable for the order of years, “year_order”. That is the number of years (1st, 2nd, 3rd, 4th) the runner entered the TOHM. This variable was created as a categorical variable where for example “year_order” has a value of 1 if it is the first time the runner has entered and “year_order” has a value of 2 if it is the second time the runner has entered for the TOHM, etc.
- A chronic disease composite score, “sc_chronic”, was created as a numeric variable ranging from 0 to 10 in order to retain as much information about this variable as possible. The composite score was based on the series of ongoing SAFER studies (see SAFER publications from 2021). The “sc_chronic” variable was created as a sum of the entrants answer out of 10 questions related to a history of chronic disease (risk factors for cardiovascular disease [CVD], history of CVD, symptoms of CVD, respiratory disease, gastrointestinal disease, nervous system/psychiatric disease, kidney/bladder disease, haematological/immune disease and cancer) (Sewry, et al., 2021). Thus the information contained in the “sc_chronic” variable encompasses a number of important chronic diseases possibly linked to the performance of runners. To include each of these chronic disease indicators separately in the model, would be a more complex analysis, therefore the variable is utilised as a numeric variable. Questions related to history of chronic diseases are questions 2 to 10 and can be found in Appendix B.
- The variable “recent_run_injury” was created as a binary variable from the variables “injury1”, “injury2” and “injury3” to indicate if the entrant was injured in the past year or not. If the entrant indicated that they were injured in the past year, a value of 1 was assigned to the variable “recent_run_injury” and if the entrant indicated that

they have not been injured in the past year, then a value of 0 was assigned to the variable.

- The variable “train_pace” was created as a continuous variable from question 5 (see Appendix B). Some respondents listed a training pace of less than 2 minutes and 30 seconds per kilometre. The World Record for 1 kilometre is 2 minutes and 11 seconds according to records kept by the World Athletics (World, 2021). Observations where the “train_pace” values were less than 2 minutes and 30 seconds, were thus deleted and assigned missing values.
- The variable “maxrace” was created to describe the maximum number of times the entrant has entered for the TOHM in the 4-year period. An example of this variable is shown in Table 7.

Table 7. Example of “maxrace” variable values for runner_code 96

Runner_code	year_order	Maximum times runner has entered (“maxrace”)
96	1	4
96	2	4
96	3	4
96	4	4

Further expanding on Table 7, to show groups of runners by number of entrants for the newly created “maxrace” variable, see below Table 8.

Table 8. Groups of runners

“maxrace”	Number of entrants	Number of runners
2	9780	4890
3	6444	2148
4	3616	904

In Table 8, the groups or clusters of runners are shown by the number of maximum races they have completed. Where “maxrace” = 4, 904 runners have completed the TOHM 4 times during the 4-year period. Where “maxrace” = 3, 2148 runners have completed the TOHM 3 times. Where “maxrace” = 2, 4890 runners have completed the TOHM 2 times.

- The response variable, “time_f_min”, was created to represent the finish time, of the entrant in minutes:

$$time_f_min = (hour(time_{finish}) * 60) + minute(time_{finish}) + (second(time_{finish})/60). \quad (\text{Equation 3.2})$$

3.8.3 Final variables in data set

Table 9 provides a detailed description of the variables that are included in the final data set.

Table 9. Outline of variables included in the final data set to be analysed further

Variable name	Description	Variable “theme”	Possible outcomes of variables
age	Age of entrant	Confounder	Continuous variable with range of (16; 85)*

gender	Gender of entrant	Confounder	1 = male, 2 = female
BMI	Body mass index	Confounder	Continuous variable with range of (15; 37)
recreationrunner	Number of years entrant has been a recreational runner	Training load variable	Continuous variables with range of (0.5 ; 70)
distancerunner	Number of years entrant has participated in distance races	Training load variable	Continuous variable with range of (0.5 ; 60)
Training_pace	In the last 12 months what is your average training speed (minutes per kilometre)?	Training load variable	Interval variable with range (2.5; 13) with step size of 0.25
sc_chronic	Chronic disease history score	History of diseases	Numeric variable with range of (0;10)
Recent_run_injury	Do you or did you suffer from any symptoms of a running injury (muscles tendons bones ligaments or joints) in the past 12 months or currently? Note: Only if an injury is/was severe enough to interfere with running or require treatment	History of injury	Binary variable with value "1" = yes and value "0" = no
allergies	Do you suffer from any allergies including a past history of allergies to medication plant	History of illness/disease	Binary variable with value "1" = yes and value "0" = no

	material or animal material?		
maxrace	Maximum number of times entrant has entered for the race	Random effect variable	2, 3, 4
time_f_min	Time entrant took to complete the race in minutes	Response variable	Continuous variable in range of (66.6 ; 517.83)

*note: age is truncated at age 16 due to the minimum age entry requirement.

3.9 Pre-model analysis

The pre-model analysis process includes examination of the predictor variables. This section will discuss the methodology used to examine descriptive statistics for continuous and categorical variables, thereafter correlation between variables will be described.

3.9.1 Descriptive statistics

As mentioned in the overview of steps, descriptive statistics was included for each variable. For continuous variables, the minimum, 1st quartile, mean, median, 3rd quartile, maximum and standard deviation as well as plots to better understand the distribution of each variable was included. Frequency tables (that included frequencies and percentages) were created for categorical variables. These results are shown and discussed in Section 4.3.

3.9.2 Correlation

Correlation is a measure to indicate the degree of association or relation between variables. The correlation between two random variables x and y can be defined by the following equation (Kutner, et al., 2005):

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}, -1 \leq r \leq 1, \quad (\text{Equation 3.3})$$

where x and y are two random variables, σ_x is the standard deviation of x , σ_y is the standard deviation of y and r the correlation coefficient.

Various different measures exist for measuring the association between variables. To measure the linear association between continuous variables, the Pearson correlation coefficient is often the most popular choice. A Pearson correlation coefficient of greater than 0.7, indicates a strong positive relationship between the variables (Sedgwick, 2012).

Correlated variables explain similar information when highly correlated, therefore including only one in the model, will explain all necessary information needed (James, et al., 2013).

Training load variables “distancerunner” and “recreationrunner” were found to be highly correlated ($r = 0.8716$). To choose the best fitting model, separate models will be analysed for the two training load variables. The fit statistics will then be utilised to choose the best fitting model and that model will be reported as the final model (see Section 4.6).

3.10 Modelling and model selection

To create the model that would be used in the construction of the mixed models, the procedure “PROC MIXED” in SAS was used.

The following steps below outline the modelling procedure:

Model 1:

Conduct univariate LMMs with response variable “time_f_min” and “year_order” as fixed effect (MODEL time_f_min = year_order). Univariate analysis explores each variable separately in the data set (Tabachnick & Fidell, 2007). The model includes the random effects (“maxrace”), three clusters (or groups) of the maximum number of years’ the runner has run (2, 3 or 4 times), this is achieved using the RANDOM statement in the “PROC MIXED” procedure.

The REPEATED statement factors in the correlation of the entrants (“runnercode”) who ran the race more than once. The REPEATED and RANDOM statements as stated here, are repeated for the following steps.

After this step, four covariance structures are applied separately, one by one, to the R side random effects (see Appendix C for SAS statements). The covariance structure that fits the

model the best according to the model selection criteria AIC, cAIC, and/or BIC, is chosen as the final model. The aim is to try and find a simpler covariance structure than the unstructured covariance structure with fewer parameters. For the G side random effect, the variance components (VC) structure is applied as this is the only structure for which the model converges. The model equations are shown in Section 4.5 and 4.6.

Model 2:

Includes confounders, in addition to those variables included in Model 1. The confounders are “gender”, “agecat” and “bmi”. If these variables “gender”, “agecat” and “bmi” are significant and improve the fit of the model, they are included in the model in order to adjust for their influence as confounders. The confounders are added to the model to eliminate their influence on the relationship between “time_f_min” and “year_order” so as to get the independent effect of “year_order” on “time_f_min”.

Model 3:

Includes predictor variables, in addition to those variables included in Model 2. The predictor variables that were assessed are: “training_pace”, “recreation_runner”, “distance_runner”, “allergies”, “schronic”, and “recent_run_injury”. Significant predictor variables will be retained and a choice will be made for the two collinear variables, “recreation_runner” and “distance_runner” as mentioned previously.

Model 4:

Includes all interaction terms, in addition to those variables included in Model 3 to investigate which factors hypothesize the change in finish time. If these interaction terms are significant, they are retained in the model. The SAS statement for Model 4 can be seen in Appendix C.

3.11 Conclusion

The chapter started by revisiting the research questions. Thereafter, a short introduction was done on the study design, data collection, questionnaire and population of the study by the

research team in order to provide more background information to the reader before the start of the data preparation.

Once the aforementioned subjects were discussed, the data preparation was discussed in Section 3.7. This included any data management or cleaning applied to the data before any descriptive statistics or modelling procedures were conducted. In the data preparation step, the thesis explains how the final data set for this thesis is selected from the original provided by the SAMRC, how relevant variables are selected to answer the research questions and how any new variables were created.

Once the final dataset was determined, an investigation was conducted on all variables included in the model through various descriptive measures such as statistics and plots.

Finally, Section 3.10, details the modelling procedure and selection conducted to answer the research questions.

The next chapter will detail the results.



4 Results

4.1 Introduction

Chapter 4 details the results according to the methodology discussed in Chapter 3. Section 4.2 describes the final data set used for analysis after data management was conducted. Section 4.3 details descriptive statistics for the response, confounding and predictor variables. Section 4.4 lists the choice of covariance matrix and how this was made for the various random effects. Section 4.5 details the results of the modelling procedure that was followed by starting with the simplest model, then including all significant predictor variables and finally detailing the results for the significant interaction terms. Section 4.6 provides a summary of the chapter.

4.2 Final data set

The final frequencies listed by “year_order”, after carrying out data management as mentioned in Chapter 3 are given by year in Table 10.

Table 10. Number of entrants per year (N = 19840)

year_order	n	%
1	7942	40.03
2	7942	40.03
3	3052	15.38
4	904	4.56

Table 10 is interpreted as follows, 904 (4.56%) runners took part in the Two Oceans Half Marathon (TOHM) race 4 times, 3052 (15.38%) runners took part in the race 3 times, and 7942 (40.03%) runners took part in the race twice.

As the data set was selected to include only runners that entered for the TOHM more than once, the frequencies for “year_order” 1 and 2 would be the same. As mentioned in Section

3.7, the runners that only competed once, were excluded from the analysis, because the focus of the analysis was on the runners that had a change in performance over two, three and four years. To measure this change, runners that only entered for the TOHM once, could not be included.

4.3 Descriptive statistics

4.3.1 Response variable descriptive statistics

Table 11. Finish time in minute per year order (N=19840)

year_order	n	Range (min; max)	Interquartile range (IQR) (Q1; Q3)	Mean	Standard deviation
1	7942	(66.6; 212.53)	(125.68; 160.15)	143.53	23.19
2	7942	(67.72; 211.23)	(123.3; 158.02)	140.98	23.64
3	3052	(75.65; 200.77)	(121.49; 158.78)	139.95	24.07
4	904	(78.2; 201.48)	(121.33; 156.18)	138.89	23.99

Table 11 shows that finish time is the fastest (138.89 minutes) for runners that took part in the race 4 times (in comparison to runners that took part 3 times, twice and once). The range is the biggest for runners that only took part in the race once.

4.3.2 Confounder variables descriptive statistics

Table 12. Age of entrant by year

year_order	n	Mean	Standard deviation
1	7942	36.89	12.26
2	7942	38.13	12.27
3	3052	40.62	12.81
4	904	43.05	12.96

In Table 12, runners that took part in the race 4 times, had an average age of 43.05 years with a standard deviation of 12.96. Seeing as this is a subset of the group of runners that took part once, twice and three times, with age increasing by 1 year for every runner, the increase in age is an expected result due to the change in time. The standard deviation across “year_order” appears consistent.

Table 13. Gender of entrant by year

year_order	Gender	n	%
1	Male	3977	50.08
	Female	3965	49.92
2	Male	3977	50.08
	Female	3965	49.92
3	Male	1632	53.47
	Female	1420	46.53
4	Male	512	56.64
	Female	392	43.36

Table 13 shows that more males than females took part in the race every year, with the greatest disparity occurring in the group of runners that took part four times. 3977 (50.08%)

of male runners took part in the race once and twice, 1632 (53.47%) of male runners took part in the race three times, and 512 (56.64%) of male runners took part in the race four times.

Table 14. Overall BMI, weight and height of entrant by year

year_order	Variable	n	n missing (%)	Mean	Standard deviation
1	BMI	7942	190 (2.99)	24.18	3.39
	Height			171.72	10.17
	Weight			71.63	13.71
2	BMI	794	196 (2.47)	24.19	3.34
	Height			171.50	10.37
	Weight			71.50	13.60
3	BMI	3052	63 (2.06)	24.37	3.27
	Height			171.66	10.61
	Weight			72.15	13.45
4	BMI	904	16 (1.78)	24.48	3.33
	Height			171.92	10.18
	Weight			72.72	13.73

In Table 14, the variables BMI, height and weight contained some missing observations, but not more than 10%. The BMI remained fairly stable for runners that took part in the race once, twice, three or four times with an overall average of $24.23 \frac{kg}{m^2}$ and overall standard deviation of 3.35. The runners that ran the race once, had the lowest average BMI score and the runners that ran the race four times had the highest average BMI score. All average BMI scores fall within the normal ranges of 18.5 to 24.9, albeit on the upper end of that range.

Table 15. BMI, weight and height of entrant by year and by gender

year_order	Gender	Variable	n	Mean	Standard deviation
1	Male	BMI	3870	25.42	3.09
		Height		178.20	8.16
		Weight		80.77	11.41
	Female	BMI	3882	22.93	3.22
		Height		165.25	7.52
		Weight		62.52	8.90
2	Male	BMI	3857	25.40	3.05
		Height		178.21	8.19
		Weight		80.71	11.16
	Female	BMI	3889	22.99	3.17
		Height		164.84	7.65
		Weight		62.37	8.81
3	Male	BMI	1594	25.44	2.99
		Height		178.07	8.27
		Weight		80.66	10.76
	Female	BMI	1395	23.15	3.14
		Height		164.34	7.92
		Weight		62.43	8.81
4	Male	BMI	503	25.52	3.07
		Height		177.68	7.84

		Weight		80.60	11.06
	Female	BMI	385	23.11	3.16
		Height		164.38	7.65
		Weight		62.42	9.36

Table 15 shows the results of BMI, weight and height by “year_order” and by gender. The highest mean BMI for men occurs in the fourth year of participation (23.11) and the highest mean BMI for women occurs in the third year of participation (23.15).

4.3.3 Predictor variables descriptive statistics

Table 16. Binary categorical variables descriptive statistics by year

Variable	year_order	n	% (of total year_order n)
Allergies (yes)	1	974	12.26
	2	841	10.59
	3	356	11.66
	4	110	12.17
Recent_run_injury (yes)	1	748	9.42
	2	637	8.02
	3	263	8.62
	4	95	10.51

There are no missing data for the categorical variables “allergies” and “recent_run_injury” as shown in Table 16. All entrants indicated on the questionnaire if they had an allergy or running injury in the past 12 months. The percentage of allergies remained fairly stable over the four races the runners competed in, with the highest percentage (12.26%) of allergies reported in the first race the runners ran. The highest percentage (10.51%) of recent injuries occurred in the fourth race the runners ran.

Table 17. Chronic illness descriptive statistics by year_order for entrants that reported a chronic illness

Variable	year_order	n**	Mean*	Standard deviation	Range (minimum; maximum)
Sc_chronic	1	2539	1	0.78	(1; 8)
	2	2165	1	0.67	(1; 6)
	3	880	1	0.69	(1; 7)
	4	291	1	0.69	(1; 4)

**rounded to nearest discrete value in range*

***n = number of people that reported a chronic illness*

Table 17 shows the descriptive statistics for entrants that reported a chronic illness. On average over the four races, approximately 70% of runners reported that they do not have any chronic illness. The highest maximum (8) number of chronic illness' occur in the first year the runners raced. The mean number (mean = 1) of chronic illnesses remained stable for entrants in the groups of runners that competed in their first, second, third or fourth race.

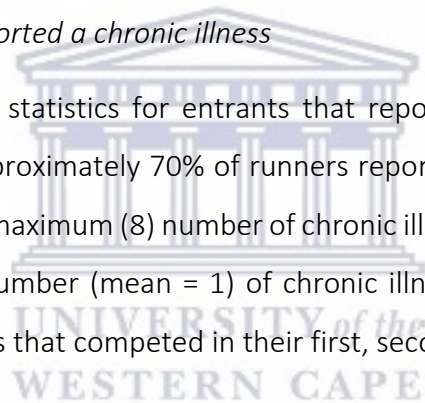


Table 18. Continuous variables descriptive statistics by year

Variable	year_order	n	Mean	Standard deviation
Training_pace	1	7808	6.15	1.13
	2	7835	6.03	1.05
	3	3018	6.04	1.03
	4	896	6.05	0.93
Recreationrunner	1	7924	7.84	8.17
	2	7883	8.49	8.10
	3	3025	10.32	8.73
	4	896	11.80	9.01
Distancerunner	1	7924	5.40	6.83
	2	7883	6.33	6.88
	3	3025	8.14	7.51
	4	896	9.74	7.72

Table 18 shows that the variable “training_pace” contains missing observations (1.52%) due to removal of incorrect input by entrants. Other training load variables also contain some missing observations, but none more than 10% that would significantly influence the analysis. Training pace had the highest average (6 minutes and 9 seconds per kilometre) in a group of runners that competed in their first race and the lowest average (6 minutes and 2 seconds per kilometre) in the group of runners that competed in their second race. The highest

average for “how many years have you been a recreational runner” and “for how many years have you participated in distance races” (variables, “recreationrunner” and “distancerunner”), occurs in runners that have run the race 4 times, which is an expected result as the time increases between these periods.

4.4 Checking model assumptions

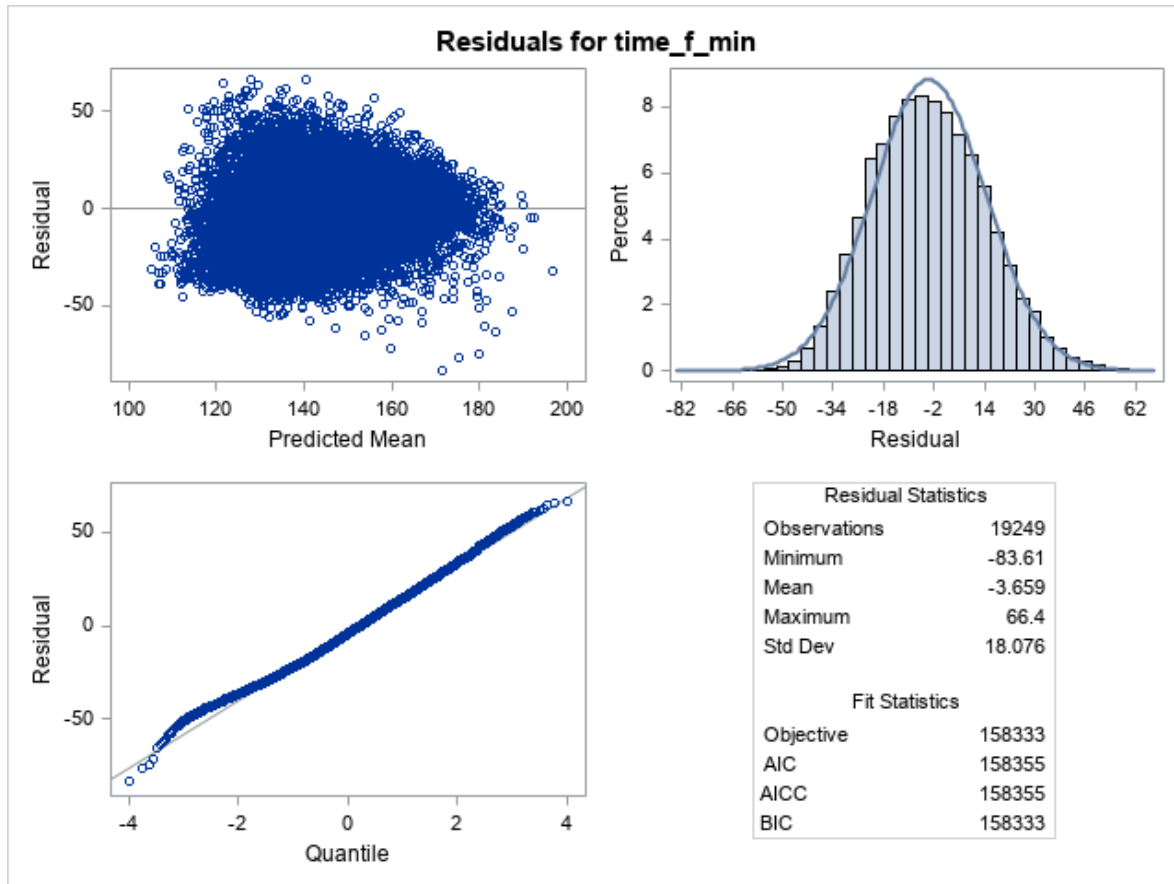


Figure 4. Residual plots and statistics

The residual plots in Figure 4 indicate that the error terms for the response variable, “time_f_min”, are normally distributed and do not violate the assumptions of the LMM. Figure 4 does not indicate departure from normality, nor any extreme outliers.

Another assumption for LMM, is that the predictor variables must be linearly related to the response variable. One method by which this assumption can be tested, is graphically. As previously mentioned, “sc_chronic” is represented as a numerical variable with approximately 70% of entrants reported no chronic illness’. Therefore, special attention is

given to ensure that representing this variable in a numerical manner does not violate any model assumptions. The below figures illustrate the linear relationship between “sc_chronic” and the response variable “time_f_min”.

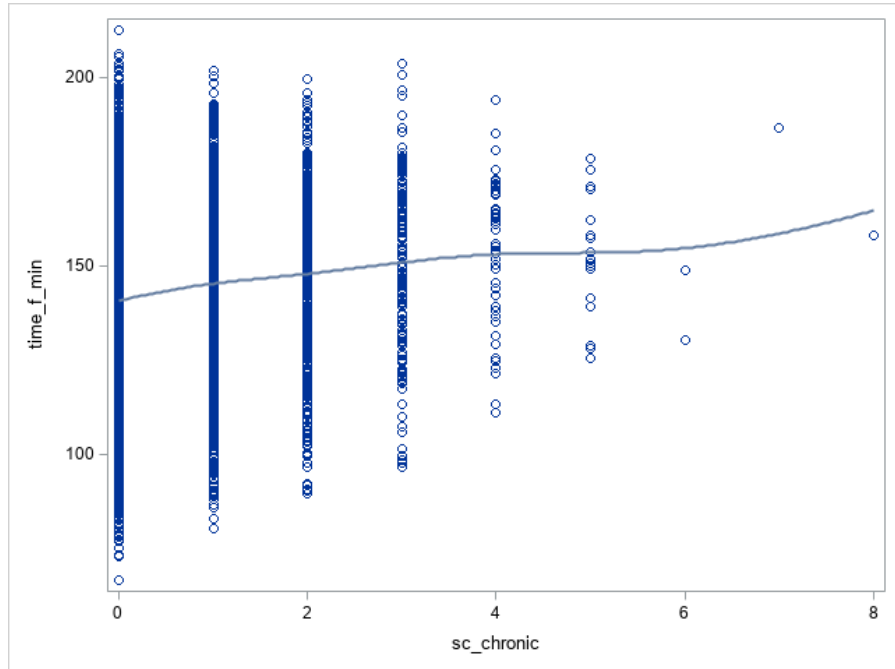


Figure 5. Response variable by sc_chronic variable for year_order = 1

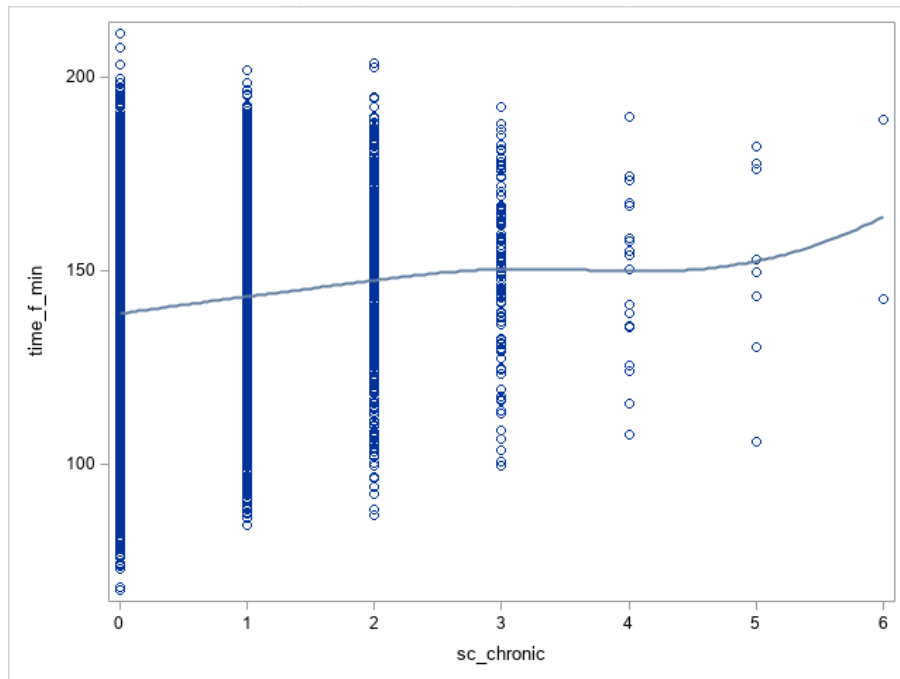


Figure 6. Response variable by sc_chronic variable for year_order = 2

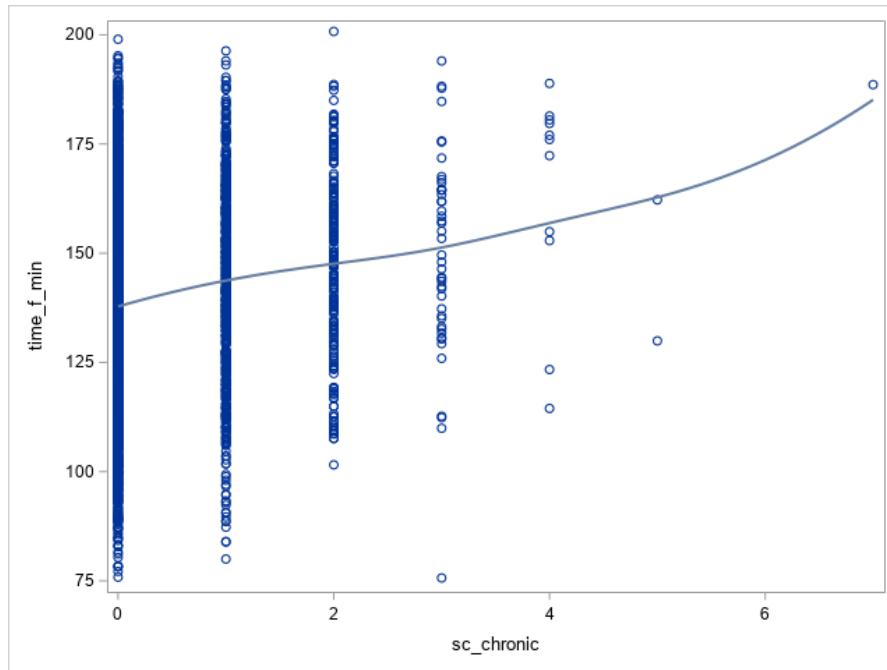


Figure 7. Response variable by sc_chronic variable for year_order = 3

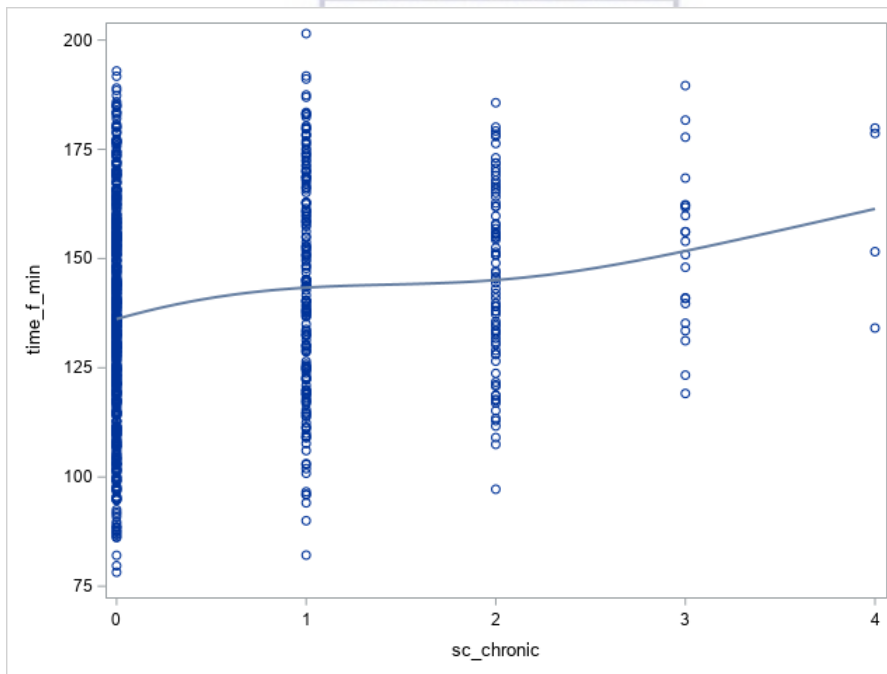


Figure 8. Response variable by sc_chronic variable for year_order = 4

Figure 5 to Figure 8 depicts the relationship between the response variable (“time_f_min”) and “sc_chronic”. Furthermore, it is noted that a trend seems to exist between the number of chronic illness’ an entrant has and their respective finishing time for the TOHM. Therefore,

retaining the variable in its current form allows the aforementioned information and the variable's possible influence on the response variable to be further investigated.

4.5 Covariance structure (R and G side)

In Section 2.3.3 Covariance structures were introduced. Recall that the choice of covariance structure can be guided by the correlation matrix of the response variable (Brown & Prescott, 2015).

Table 19. Lower triangular correlation matrix of Pearson correlation coefficients between "year_order"s

year_order	1	2	3	4
1	1			
2	0.808	1		
3	0.794	0.854	1	
4	0.788	0.837	0.860	1

When investigating the correlation matrix of the response variable, "time_f_min", shown by "year_order" in Table 19, it appears as if a simple Toeplitz covariance structure could be a good fit as the values remain relatively constant on the diagonal from left to right. However, because of the large sample size of the data set, and the complexity involved with multicollinearity, it would be worthwhile to also fit the unstructured covariance matrix as this type of matrix assumes that there is no reliable pattern.

In Section 2.3.3, the equation 2.9 for the variance-covariance matrix in matrix notation was given as follows:

$$V = ZGZ' + R$$

with **G** defined as the random effect parameters and **R** as the error terms.

As mentioned in Section 3.10, Model 1 is created to:

1. Explore the effect of the fixed effect “year_order” and random effect “maxrace” on the response variable “time_f_min”, but also
2. To find the most suitable covariance structures for the effects.

In Section 4.2, four covariance structures (Unstructured, Toeplitz, Auto Regressive, Compound Symmetry) were first applied to the R matrix of the model containing only “year_order” as fixed effects and “maxrace” as random effect. As mentioned in Section 2.4, four model fit indices (Log likelihood, AIC, cAIC, BIC) are used to evaluate the performance of the covariance structures. These results are summarised in Table 20.

Table 20. R matrix covariance structures for Model 1 with fit indices

Type covariance structure	Number of covariance parameters estimated*	'-2log(L)'	AIC	cAIC	BIC
Unstructured (UN)	9	167330.3	167352.3	167352.3	167330.3
Toeplitz (TOEP)	3	167452.9	167462.9	167462.9	167452.9
Compound symmetry (CS)	1	167531.3	167537.3	167537.3	167531.3
First order autoregressive (AR(1))	1	169268.4	169274.4	169274.4	169268.4

*number of elements in covariance matrix to be estimated

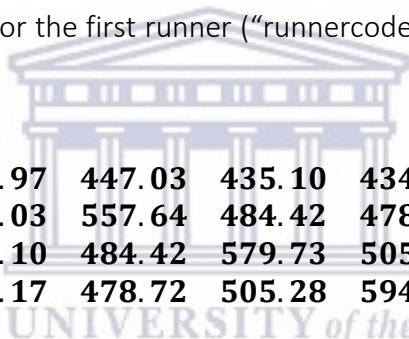
A comparison between covariance model structures for Model 1 (see Section 3.10) in Table 20, shows that the unstructured covariance matrix provides the optimal fit based on the above indices in comparison to simpler covariance structures of Toeplitz, Compound symmetry and First order autoregressive. However, this covariance structure is the most complex out of the aforementioned with 9 estimated covariance parameters. The Toeplitz

covariance structure provides the second lowest fit indices with only 3 estimated covariance parameters. The Toeplitz covariance structure is therefore much less complex than the unstructured and would be a good alternative covariance structure.

Recall from Section 2.3.3 that the unstructured covariance matrix for 4 time points can be represented as follows:

$$R_i = \begin{bmatrix} \sigma_1^2 & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{12} & \sigma_2^2 & \theta_{23} & \theta_{24} \\ \theta_{13} & \theta_{23} & \sigma_3^2 & \theta_{34} \\ \theta_{14} & \theta_{24} & \theta_{34} & \sigma_4^2 \end{bmatrix}$$

The below matrix R_1 (where $i = 1$ for the first runner for entire example below) illustrates an example of the R side matrix for the first runner (“runnercode” = 96) that participated in the TOHM for all 4 years.



$$R_1 = \begin{bmatrix} 536.97 & 447.03 & 435.10 & 434.17 \\ 447.03 & 557.64 & 484.42 & 478.72 \\ 535.10 & 484.42 & 579.73 & 505.28 \\ 434.17 & 478.72 & 505.28 & 594.27 \end{bmatrix}$$

After the unstructured covariance matrix has been chosen for the R side matrix, the same covariance structures were tested for the G side matrix. However, only the simplest covariance structure (variance components) could be fitted to the model.

Recall from Section 2.3.3 that the variance components covariance matrix for 3 clusters (see Figure 4 as reminder of the 3 clusters) can be represented as follows:

$$G_i = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

The G side variance components matrix can be seen below:

$$G_1 = \begin{bmatrix} 1.82 & 0 & 0 \\ 0 & 1.82 & 0 \\ 0 & 0 & 1.82 \end{bmatrix}$$

As mentioned previously, the model only converged using the variance components covariance structure, therefore it was not possible to test other covariance structures for the G matrix as was done for the R matrix.

Substitution of the above covariance structures in the variance of y_1 matrix notation formula of Equation 2.9 where G has the dimension qxq and represents the number of random effects parameters, results in:

$$V_1 = Z_1 \begin{bmatrix} 1.82 & 0 & 0 \\ 0 & 1.82 & 0 \\ 0 & 0 & 1.82 \end{bmatrix} Z_1' + \begin{bmatrix} 536.97 & 447.03 & 435.10 & 434.17 \\ 447.03 & 557.64 & 484.42 & 478.72 \\ 535.10 & 484.42 & 579.73 & 505.28 \\ 434.17 & 478.72 & 505.28 & 594.27 \end{bmatrix},$$

where Z_1 is a $m \times 3$ matrix and Z_1' is a $3 \times m$ matrix, with $m=4$ for this specific example.

If the above R matrix is expanded (as an example) to include the second (“runnercode” = 289 that participated 4 times) and third runner (“runnercode” = 322 that participated twice), the unstructured covariance pattern is given below:

$$R = \begin{bmatrix} \sigma_1^2 & \theta_{12} & \theta_{13} & \theta_{14} & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{12} & \sigma_2^2 & \theta_{23} & \theta_{24} & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{13} & \theta_{23} & \sigma_2^2 & \theta_{34} & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{14} & \theta_{24} & \theta_{34} & \sigma_4^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_1^2 & \theta_{12} & \theta_{13} & \theta_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{12} & \sigma_2^2 & \theta_{23} & \theta_{24} & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{13} & \theta_{23} & \sigma_2^2 & \theta_{34} & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{14} & \theta_{24} & \theta_{34} & \sigma_4^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_1^2 & \theta_{12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{12} & \sigma_2^2 \end{bmatrix}$$

4.6 Modelling

The modelling results, as discussed in Section 3.10, will be shown below for the 4 models (or steps). Models are constructed in sequential order. The results are shown in this manner so as to highlight the effects of:

1. The change in time of the runner over the 4-year period through the inclusion of the variable “year_order” in the model,
2. The importance of the inclusion of confounding variables in the analysis and the effects of confounders on the response variable,
3. The exclusion of non-significant variables from the subset of variables that could impact the change in performance of the runner according to literature,
4. And finally, the illustration of the significant variables that impact the change in performance of the runner and the extent of the significant variables impact on the change in performance.

4.6.1 Model 1

Model 1 results show the univariate LMM analysis with “year_order” included as the fixed effect and “maxrace” as the random effect after the appropriate covariance structures (for R and G side random effects) were chosen based on the results given in Section 4.5.

As seen in Section 2.3.2, the general matrix notation for the LMM, is represented as follows:

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

For the model including only the “year_order” variable, the \mathbf{X} matrix can be represented as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where \mathbf{X} is a $n \times 5$ matrix (in this case 4×5).

The fixed effects coefficients of $\boldsymbol{\beta}$, which are determined only after including fixed effects, is therefore represented as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}$$

where $\boldsymbol{\beta}$ is a 5×1 matrix. The fixed effects will change from Model 1 to 4 as more fixed effects are included.

The random effects are as described in Section 4.5 for Models 1 to 4.

The vector for residuals is a $n \times 1$ vector.

Table 21. Predicted response coefficients, F-value, p-value and difference for consecutive years

Year_order	Coefficient (Standard error ((SE)	Fixed effects F-value	Fixed effects p-value	Difference between consecutive year	
				Coefficient (SE)	p-value
1	145.83 (1.02)	54.29	<0.0001	-	-
2	143.81 (1.02)			-2.02 (0.16)	0.004
3	144.47 (1.06)			0.66 (0.23)	<0.0001
4	144.2 (1.13)			-0.28 (0.39)	0.48

The estimates given in Table 21, as well as standard errors and p-values represent those of the model without any confounders or predictor variables included, i.e. the model that only includes the dependent variable “year_order” and independent variable “time_f_min” in the

model statement. The type 3 test of fixed effects for “year_order” (p-value = <0.0001) is significant. The difference between year_order 1 and year_order 2 is also significant (p-value = 0.004) and difference between “year_order” 2 and 3 is significant (p-value = <0.0001). The estimated finish time (“time_f_min”) decreases between year_order 1 and 2, but increases slightly between years 2 and 3. The type 3 test tests the effect of excluding a term whilst retaining higher-order interactions (SAS, 1999).

4.6.2 Model 2

As mentioned in Section 3.10, model 2 includes confounders “gender”, “agecat” and “bmi”. If these variables are significant in the analysis and improve the fit of the model, they are included in the model in order to adjust for their influence as confounders.

Table 22. Predicted response coefficients, F-value, p-value and difference between consecutive years adjusted for confounding variables age, gender, BMI

Year_order	Coefficients (SE)	Fixed effects for year_order F-value	Fixed effects for year_order p-value	Difference between consecutive year	
				Coefficient (SE)	p-value
1 (ref)	146.85 (0.83)	93.14	<0.0001	-	-
2	144.22 (0.83)			-2.63 (0.17)	<0.0001
3	144.2 (0.87)			-0.02 (0.23)	0.94
4	143.49 (0.95)			-0.71 (0.40)	0.08

*p-value for age: <0.0001 , p-value for gender: <0.0001, p-value for BMI” <0.0001

Table 22 represents the statistics for when the model includes significant confounding variables that therefore need to be accommodated for. The type 3 test for the fixed effects p-value is significant (<0.0001). The difference between years 1 and 2 is significant (p-value = <0.0001). When adjusted for age, gender and BMI, the estimates now decline between years 2 and 3, where previously (in the unadjusted model), the estimates between years 2 and 3 increased. The change between years 2 and 3 is also no longer significant.

4.6.3 Model 3

In Model 3, all predictor variables are included after the construction of Model 2, but only significant predictor variables are retained. A choice is also made between the two collinear variables, “recreation_runner” and “distance_runner”.

Table 23. Coefficients, standard error, F-value and p-value for all predictor variables excluding recreationrunner

Variables	Coefficients (SE)	Fixed effects for year_order F-value	Fixed effects for year_order p-value	p-value
Training_pace	4.34 (0.12)	38.36	<0.0001	<0.0001
distancerunner	-0.23 (0.03)			<0.0001
Allergies (yes)	-0.41 (0.37)			0.26
Sc_chronic	0.77 (0.17)			<0.0001
Recent_run_injury	-1.46 (0.35)			<0.0001

Table 24. Coefficients, standard error, F-value and p-value for all predictor variables excluding distancerunner

Variables	Coefficients (SE)	Fixed effects for year_order F-value	Fixed effects for year_order p-value	p-value
Training_pace	4.42 (0.12)	42.18	<0.0001	<0.0001
recreationrunner	-0.22 (0.02)			<0.0001
Allergies (yes)	-0.41 (0.37)			0.27
Sc_chronic	0.78 (0.17)			<0.0001
Recent_run_injury	-1.43 (0.35)			<0.0001

When comparing Table 23 and Table 24, the fit indices are slightly bigger for the model including “distancerunner” (Table 23)(AIC = 158467.0) in comparison to the model including “recreationrunner” (Table 24)(AIC = 158455.5). This indicates that the model containing

“recreationrunner” is a marginally better model. Furthermore, all predictor variables, excluding allergies, are significant in both models and will be included in the final model.

Table 25. Multiple model coefficients, standard error, F-value, p value results (for all significant variables and “recreationrunner”)

Variables	Coefficients (SE)	Fixed effects for year_order F-value	Fixed effects for year_order p-value	p-value
Training_pace	4.42 (0.12)	42.08	<0.0001	<0.0001
recreationrunner	-0.22 (0.02)			<0.0001
Sc_chronic	0.75 (0.17)			<0.0001
Recent_run_injury (yes)	-1.45 (0.35)			<0.0001

Excluding allergies from the model and only including “recreationrunner” in Model 3, produces the results contained in Table 23. This model also includes the significant predictor variables “training_pace”, “recreationrunner”, “sc_chronic” and “recent_run_injury”. All aforementioned variables have a p-value of <0.0001, therefore these variables are statistically significant in the prediction of improvement in performance.

4.6.4 Model 4

In the final Model 4, interaction terms with “year_order” (e.g. “training_pace”*“year_order”) are added to the Model 3. All significant variables were added as interaction action terms to Model 3, but only significant interaction terms are retained in the model. Interpreting the main effects of the model without consideration of statistically significant interaction effects, assumes that the effect of predictor variables on the response variable is independent of other predictor variables in the model (James, et al., 2013).

The full SAS statements for Models 1 to 4 can be seen in Appendix C.

In Model 4, the X matrix is expanded to include more fixed effects. The X matrix now becomes a $n \times 18$ matrix. The β expands to a 18×1 vector. These fixed effects include 1 “training_pace” variable, 4 interaction variables for “training_pace” X “year_order”, 3 confounding variables (“age”, “gender”, “BMI”), 1 “recreation_runner” variable, 1 “recent_run_injury” variable and 1 chronic illness model variable (“sc_chronic”).

The SAS representation of the various coefficients are provided in Appendix C.

4.6.4.1 Model assumptions for Model 4

As previously mentioned in Chapter 3, an assumption of the LMM is that residuals need to be normally distributed and residual variance need to be constant across observations (Demidenko, 2004).

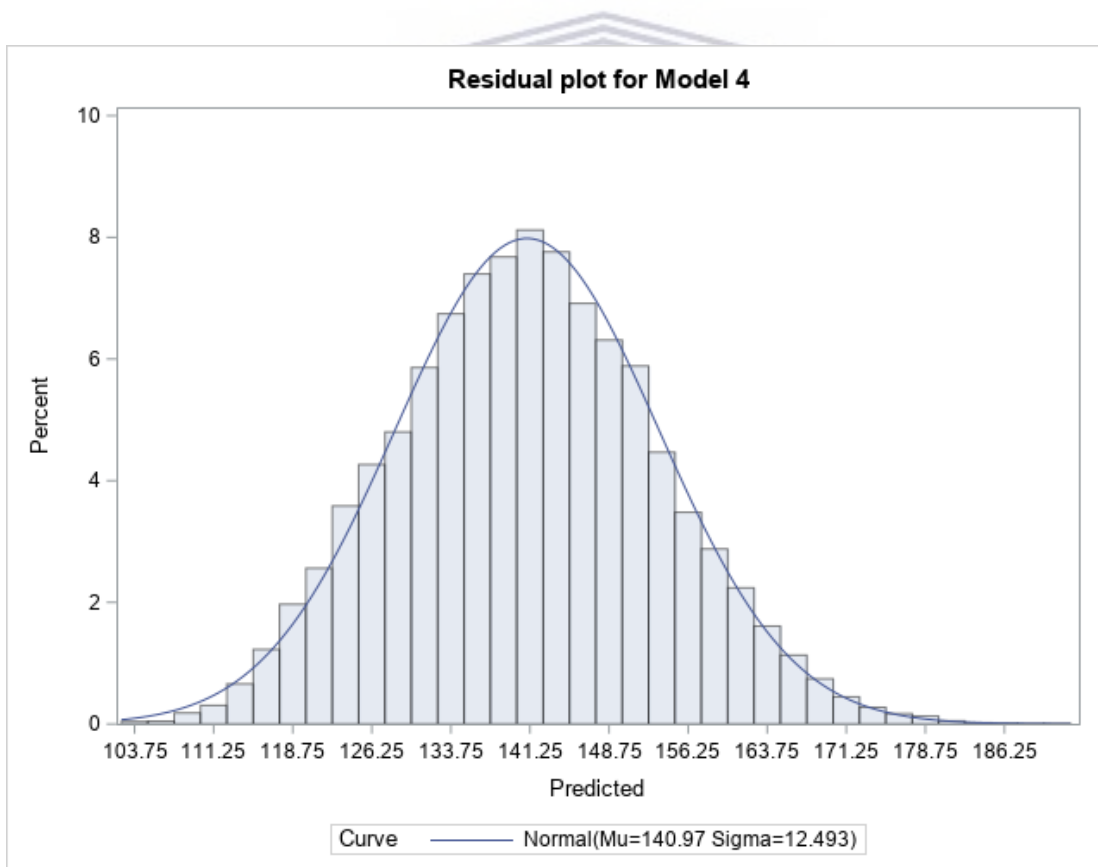


Figure 9. Residuals plot for Model 4

From Figure 9, it is clear that the residuals are normally distributed. The distribution closely follows that of the normal distribution (blue line that is superimposed over the histogram). Therefore, the model assumptions are not violated.

Table 26. Coefficients, standard error, F-value, p value for multiple model with significant interaction

Variables	Coefficients (SE)	F-value	p-value
Recreationrunner	-0.22 (0.02)	105.24	<0.0001
Sc_chronic	0.74 (0.17)	19.83	<0.0001
Recent_run_injury	-1.44 (0.35)	17.29	<0.0001
Year_order(1)*training_pace (reference)	-	4.58	-
Year_order(2)*training_pace	-0.46 (0.17)		0.0062
Year_order(3)*training_pace	-0.82 (0.25)		0.0009
Year_order(4)*training_pace	-0.25 (0.44)		0.57

Table 26 shows the significant interaction (year_order * training_pace) included in the model (reference is year_order 1). The interaction is significant between year_order 1 and year_order 2 (p-value = 0.0062) as well as between year_order 1 and year_order 3 (p-value = 0.0009).

Table 26 shows that for every 1 unit increase in chronic illness (“sc_chronic”), the finish time of the entrant increases by 0.74 minutes (approximately 44 seconds). Therefore, for an increase of 2 units in “sc_chronic”, the finish time of the entrant will increase by 1.48 minutes (approximately 1 minutes 29 seconds). The importance of retaining all information of the variable “sc_chronic” is now evident as 30% of runners indicated that they have at least one chronic illness. Furthermore, Table 26 shows that for every 1 unit increase in the years an

entrant has been running recreationally (“recreationrunner”), the finish time of the entrant will decrease by 0.22 minutes.

The coefficient for “recent_run_injury” is -1.44 (0.35) implying that the overall finish time decreases for runners reporting an injury. To further explore and understand this result, the difference of the yearly effect of this variable was obtained. It was found that the difference in finish time was only significant when a runner reported an injury in their first and second race (coefficient = 1.70, SE = 0.67, p-value = 0.001). The difference between the second and third and third and fourth race was not significant (p-value > 0.05).

Table 27 Predicted response variable coefficients, standard error and difference for consecutive years including confounders and all significant predictor variables

Year_order	Coefficients (SE)	Difference between consecutive year	
		Coefficient (SE)	p-value
1 (ref)	145.3 (0.75)	-	-
2	143.3 (0.76)	-1.86 (0.17)	<0.0001
3	143.6 (0.81)	0.18 (0.24)	0.455
4	143.1 (0.90)	-0.49 (0.41)	0.225

Table 27 confirms the results found in Table 22, but unlike Model 2 that was just adjusted for confounding variables, Model 4 is also adjusted for all significant predictor variables associated with the response variable. Model 2 fit indices (AIC=160830.2, AICC=160830.2, BIC=160830.2) are higher than Model 4 fit indices (AIC=158425.3, AICC=158447.3, BIC=158447.3), indicating Model 4 to be an improved fit. Table 27 therefore confirms that there is a significant (p<0.0001) improvement in finish time between the first and second races.

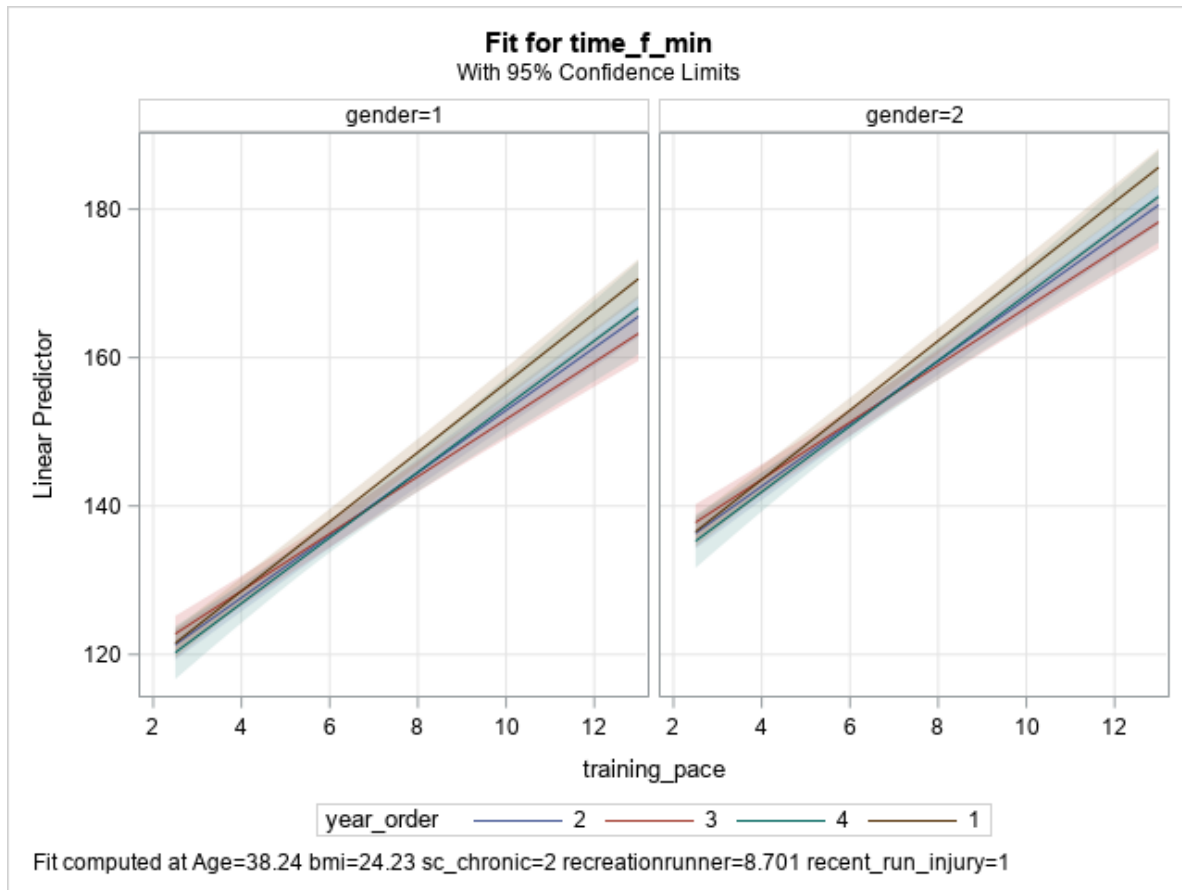


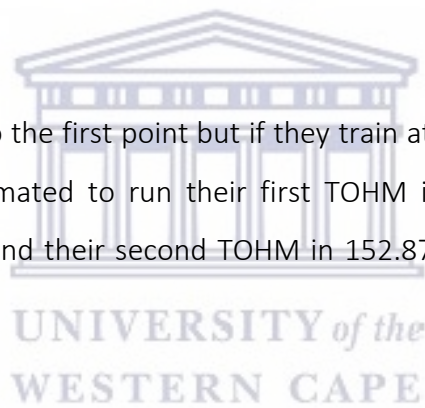
Figure 10. Interaction with training pace and year order by gender

Figure 10 indicates that if an entrant trains at an approximate pace of 6 minutes per kilometre for first year's TOHM and then increases training pace to below 6 minutes per kilometre, the entrants performance will improve the second time they compete in the TOHM. Figure 10 also shows the differentiation in gender where gender=1 is male and gender=2 is female. From this, it is clear that the intercept of the male runners are lower than for the female runners, but the slope is approximately the same for both genders. Originally, the model indicated that there is an improvement in performance between the first and the second time a runner competes in the TOHM. However, the interaction indicates that this improvement is only valid for the slower runners between their first and second race.

Because the plots are difficult to read, below are a few estimates highlighted from the graph to make the result clearer (note that the choice of estimates in Figure 10 and below for variables "age", "bmi", and "recreationrunner" are at their respective overall averages.

Furthermore, the estimates are reported for the male gender, with 2 chronic illness', and a runner that had a running injury in the past 12 months):

- If a runner is male, had an injury in the past 12 months, has an age of 38.26 years, has 2 chronic illness's, has been a recreation runner for the past 8.7 years, has a BMI of 24.2, and trains at a pace of 3 minutes/km, then they are estimated to run their first TOHM in 123.83 minutes with a standard error of 0.98, and their second TOHM in 123.40 minutes with a standard error of 1.0.
- With all criteria similar to the prior point but if they now train at a pace of 6 minutes per kilometre, they are estimated to run their first TOHM in 137.86 minutes with a standard error of 0.88, and their second TOHM in 136.03 minutes with a standard error of 0.88.
- With all criteria similar to the first point but if they train at a pace of 10 minutes per kilometre, they are estimated to run their first TOHM in 156.56 minutes with a standard error of 1.06, and their second TOHM in 152.87 minutes with a standard error of 1.10.



4.7 Conclusion

Chapter 4 detailed all results for the methods discussed in Chapter 3. The chapter discussed descriptive statistics such as n, %, range, IQR, mean, and standard deviation of the response, confounding and predictor variables. Thereafter, the choice of covariance structure with model fit indices (AIC, cAIC, BIC) were described. The unstructured covariance structure was chosen as best choice because the goodness of fit statistics indicated that it fits the initial model best but also because of its flexibility. Finally, the chapter detailed the findings of the Models 1 to 4 through estimates such as standard errors, the F-values, and the p-values. The final model included significant variables "recreationrunner", "sc_chronic", "recent_run_injury" and "training_pace" and explored the interaction effect between "yea_order" and "training_pace" more in depth.

5 Conclusion and recommendations

5.1 Introduction

The study aimed to investigate whether runners taking part in the TOHM multiple times over four years improved their race time from year to year, and to assess which factors contributed to the improvement in race time. In summary, it was found that runners did improve their performance (race time), but only from their first race to their second race, on average by 1 minute 52 seconds, and that 4 factors contributed to an improvement in race time namely, chronic illness, training pace, recent running injuries and for how many years runner has been running recreationally.

Section 5.2 provides an account of the limitations to this thesis. Section 5.3 lists further studies or research opportunities that can be conducted on this data set. Section 5.4 concludes the thesis with a short summary of the results.

5.2 Limitations

A few of the limitations to this study included:

- No weather-related data was included in the analysis to see how that affected the entrant on the day of the race.
- Only chronic illness variables were included in the analysis. No variables were included in the analysis about the health of the athlete on race day (illness on or prior to race day that could have affected performance).
- The analysis did not investigate all possible simpler covariance structures for R side random effects matrix (such as Toeplitz) in order to fit other types of covariance structures to the G side random effects matrix. If a simpler covariance structure was found to be just as effective as the unstructured covariance matrix and was fitted to the R side random effects matrix, the model might have allowed other covariance structures to converge when fitted on the G side random effects matrix.

- The years that the runner took part in their first, second, and third race, does not necessarily indicate consecutive years of participation for the runners. The analysis did not investigate if the runner 'skipping a year' of racing influences the result.

5.3 Further studies

Further studies or research opportunities for this data can include:

- Including more performance related factors (factors that according to literature will have an impact on the performance of the entrant).
- Include more confounding variables that will have an impact on the performance of the entrant on race day, such as weather on the day of the race. Scores such as Universal Thermal Comfort Index (UTCI) can be incorporated as confounding variables to adjust for the impact on performance on race day.
- Investigate the mat times runners crossed certain points in route. During the TOHM, the runners line up in blocks at the start of the race. Runners that stand at the back are influenced by the pace of the "pack" when they are at the start as there is often not space for them to run at their own speed from the "pack" as the front runners (that have more space) are able to do. Therefore, in order to eliminate the aforementioned, the entrants race pace between two mat times can be investigated as performance instead of the time across the entirety of the course.
- In order to confirm the results attained in Chapter 4, the model can be applied to a different population, like 56km TOUM and 90km Comrades Ultra marathon.
- The study included some time varying covariates, e.g. "recent_run_injury", which needs to be further explored to fully understand the contribution of this variable on the improvement in finish time.

- If the focus of the “sc_chronic” variable is on the presence of chronic illness, the chronic illness variable can be investigated as a binary variable.

5.4 Findings

This study aimed to answer the following aims:

- Do runners who took part in the Two Oceans 21.1km races more than once, i.e. 2, 3 or 4 years, improve their race time?
- Assess whether the improvement is sustained after the initial improvement.
- Which factors (age, gender, training load, history of illness and injury, allergies), if any, contribute to or are associated with the improvement in race time?

The study has found that there is a significant improvement in performance for runners that took part in their second TOHM (in comparison to their first TOHM). Thereafter, there is no significant difference between runners taking part three or four times.

Confounding factors that are related to performance includes gender, age and BMI. Predictor variables that influence performance include training load variables, training pace and recreation runner, as well as the chronic illness variable and recent running injuries.

The modelled factors in the LMM reported training pace to have an association with increase in performance. Therefore, this study found that if an entrant has a training pace of 6 minutes per kilometre in the first year that they partake in the TOHM, and then increase the training pace to below 6 minutes per kilometre in the second year they partake in the TOHM, there is a significant difference in their performance between the first and second year the runner partook in the TOHM. Therefore, the pace the runner trains at is the most important factor when they are aiming to improve their performance in the TOHM between the first and second time they partake.

Furthermore, the study confirmed an expected result, that male runners have a faster finishing time in comparison to female runners. However, in addition to this result and what

was perhaps not expected, is that both genders have the same approximate improvement in performance between their first and second year of competing in the TOHM.

5.5 Conclusion

Chapter 5 discussed further studies that could build upon and improve the results found in this thesis, limitations to this analysis that needs to be taken into consideration when viewing the results, and provided a summary of the results from the analysis outlines in Chapters 3 and 4.

As mentioned in Section 5.4, the main finding of this study, was that an increase in training pace (i.e. runner training at a faster pace) between the first and second year of competing in the TOHM, leads to an improvement in performance between the first and second year of competing in the TOHM.

This result will not come as a surprise to most runners, as runners believe that their 10km training pace is an indicator of their Half Marathon pace and therefore finish time, i.e. runners prioritise speed workouts as they have long since realised the association between an increase in training speed and an increase in Half Marathon performance (Hamilton, 2017). This is however a novel finding in the literature with regards to accessing training pace to the improvement in performance instead of, as found in many other studies, training pace being associated to performance (not specifically improvement in performance) (Knechtle, et al., 2010).

6 References

Ahn, M., Zhang, H. H. & Lu, W., 2012. Moment-based method for random effects selection in linear mixed models. *Statistica Sinica*, Volume 22, pp. 1539-1562.

Akaike, H., 1973. *Information theory and an extension of the maximum likelihood*. Budapest, Petrov, B., Csák, pp. 267-281.

Anwla, P. K., 2020. *Introduction to ANOVA for Statistics and Data Science*. [Online] Available at: <https://www.analyticsvidhya.com/blog/2020/06/introduction-anova-statistics-data-science-covid-python/> [Accessed 12 June 2022].

Avalos, M., Hellard, P. & Chatard, J., 2003. Modeling the training-performance relationship using a mixed model in elite swimmers. *Med Sci Sports Exerc*, 35(5), pp. 838-846.

Baltagi, B. H., 2008. *Econometric Analysis of Panel Data*. 4th ed. Chichester: John Wiley & Sons.

Barker, L. E. & Shaw, K. M., 2015. Best (but oft-forgotten) practices: checking assumptions concerning regression residuals. *The American journal of clinical nutrition*, 102(9), pp. 533-539.

Barron, A., 1997. *Introduction to Statistics*, Yale: Yale University Department of Statistics.

Bennett, D., 2001. How can I deal with missing data in my study?. *Australian and New Zealand Journal of Public Health*, 25(5), p. 464.

Bobbitt, Z., 2020. *What is a Covariate in Statistics?*. [Online] Available at: <https://www.statology.org/covariate/> [Accessed 14 June 2022].

Boullosa, D. et al., 2020. Factors Affecting Training and Physical Performance in Recreational Endurance Runners. *Boullosa D, Esteve-Lanao J, Casado A, Peyré-Tartaruga LA, Gomes da Rosa R, Del Coso J. Factors Affecting Training and Physical Performance in Recreational*

Endurance Runners. Sports (Basel). 2020;8(3):35. Published 2020 Mar 15. doi:10.3390/sports8030035, 8(3), p. 35.

Brown, H. & Prescott, R., 2015. *Applied mixed models in medicine*. 3 ed. West Sussex: John Wiley & Sons Ltd.

Burnham, K. P. & Anderson, D. R., 2002. 7. Statistical Theory and Numerical Results. In: *Model selection and multimodel inference: A Practical Information-Theoretic Approach (2nd ed)*. New York: Springer, pp. 424-426.

Button, K. S. et al., 2013. Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience*, 14(5).

Chan, G., 2020. *Linear regression with missing data*. [Online] Available at: https://statsnotebook.io/blog/analysis/regression_missingdata/ [Accessed 26 August 2022].

Christensen, R., Pearson, L. M. & Johnson, W., 1992. Case-deletion diagnostics for mixed models. *Technometrics*, Volume 34, pp. 38-45.

Demidenko, E., 2013. *Mixed Models Theory and Applications with R*. 2nd ed. New Jersey: John Wiley & Sons, Inc..

Diggle, P. J., Liang, K. Y. & Zeger, S. L., 1994. *Analysis of longitudinal data*, Oxford: Clarendon Press.

Erler, N. S., Rizopoulos, D. & Jaddoe, V. W., 2019. Bayesian imputation of time-varying covariates in linear mixed models. *Statistical methods in medical research*, 28(2), pp. 555-568.

Fernandez, G. C., 2006. *All possible model selection in PROC MIXED - a SAS macro application*. Kansas, New Priarie Press.

Foster, C. et al., 1996. Athletic performance in relation to training load. *Wisconsin medical journal*, Volume 95, pp. 370-374.

Gabrio, A., Plumpton, C., Banerjee, S. & Leurent, B., 2022. Linear mixed models to handle missing at random data in trial-based economic evaluations. *Health Economics*, 31(6), pp. 1276-1287.

Gomez-Molina, J. et al., 2017. Predictive Variables of Half-Marathon Performance for Male Runners. *Journal of Sports Science & Medicine*, 16(2), pp. 187-194.

Goulet, M. & Cousineau, D., 2019. The power of replicated measures to increase statistical power. *Advances in Methods in Practices in Psychological Science*, 2(3), pp. 199-213.

Grace-Martin, K., 2014. *the analysis factor*. [Online] Available at: <https://www.theanalysisfactor.com/when-repeated-measures-anova-not-work-for-repeated-measures-data/> [Accessed 17 October 2022].

Hamilton, A., 2017. *Sports Performance Bulletin*. [Online] Available at: <https://www.sportperformancebulletin.com/endurance-training/training-structure-and-planning/half-marathon-training/> [Accessed 1 November 2022].

Hariharan, S. & Rogers, J. H., 2008. Estimation Procedures for Hierarchical Linear Models. In: A. A. C. a. D. B. McCoach, ed. *Multilevel Modeling of Educational Data*. Charlotte, NC: Information Age Publishing, Inc., pp. 104-139.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning*. 2nd ed. California: Springer.

Henderson, C. R., 1984. *Application of linear models in animal breeding*, Guelph, ON: University of Guelph.

Hu, B., Shao, J. & Palta, M., 2010. Variability explained by covariates in linear mixed-effect models for longitudinal data. *The Canadian Journal of Statistics*, 38(3), pp. 352-368.

International Olympic, C., 2019. *IOC recognises 11 Research Centres worldwide for prevention of injury and protection of athlete health*. [Online] Available at: <https://olympics.com/ioc/news/ioc-recognises-11-research-centres-worldwide->

for-prevention-of-injury-and-protection-of-athlete-health

[Accessed 31 March 2022].

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. In: *An introduction to statistical learning with applications in R*. London: Springer, pp. 60-120.

Jiang, J., Rao, J. S., Gu, Z. & Nguyen, T., 2008. Fence methods for mixed model selection. *Annals of Statistics*, Volume 36, pp. 1669-1692.

Kincaid, C., 2020. Guidelines for selecting the covariance structure in mixed model analysis. *COMSYS Information Technology Services, Inc.*, pp. 1-8.

Knechtle, B., Knechtle, P., Roseman, T. & Senn, O., 2010. Sex Differences in Association of Race Performance, Skin-Fold Thicknesses, and Training Variables for Recreational Half-Marathon Runners. *Perceptual and Motor Skills*, 111(3), pp. 653-668.

Knechtle, B. & Nikolaidis, P. T., 2018. Sex- and age-related differences in half-marathon performance and competitiveness in the world's largest half-marathon – the GöteborgsVarvet. *Research in Sports Medicine*, 26(1), pp. 78-85.

Kumar, A., 2021. *Fixed vs Random vs Mixed Effects Models*. [Online] Available at: <https://vitalflux.com/fixed-vs-random-vs-mixed-effects-models-examples/#:~:text=Fixed%20effects%20models%20are%20recommended,has%20constant%20variance%20across%20units.>

[Accessed 12 June 2022].

Kumar, A., 2021. *Fixed vs Random vs Mixed Effects Models – Examples*. [Online] Available at: <https://vitalflux.com/fixed-vs-random-vs-mixed-effects-models-examples/#:~:text=Fixed%20effects%20models%20are%20recommended,has%20constant%20variance%20across%20units.>

[Accessed 29 April 2022].

Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, W., 2005. *Applied Linear Statistical Models*. 5th ed. New York: McGraw-Hill Irwin.

Larumbe-Zabala, E. et al., 2020. *Longitudinal Analysis of Marathon Runners' Psychological State and Its Relationship with Running Speed at Ventilatory Thresholds*, s.l.: Frontier Psychology.

Liang, H., Wu, H. & Zou, G., 2008. A Note on Conditional AIC for Linear Mixed-Effects Models. *Biometrika*, 95(3), pp. 773-778.

Littell, R. C., Pendergast, J. & Natarajan, R., 2000. Modelling covariance structure in the analysis of repeated measures data. *Statistics in medicine*, Volume 19, pp. 1793-1819.

Marc, A. et al., 2013. Marathon progress: demography, morphology and environment. *Journal of Sports Sciences*, 32(6), pp. 37-41.

McCulloch, C. E., Shayle, R. S. & Neuhaus, J. M., 2008. Estimation. In: *Generalized, Linear, and Mixed Models*. New York: Wiley, pp. 20-23.

Messiah, S., 2013. Body Mass Index. *Encyclopedia of Behavioral Medicine*.

Mokwena, P. et al., 2021. Chronic disease, allergies and increased years of running are risk factors predicting gradual onset running related injuries in ultramarathon runners - SAFER XIX study in 29 585 race entrants. *Clinical Journal of Sport Medicine*, p. Advance online publication.

Molenberghs, G., Bijlens, L. & Shaw, D., 1997. Linear Mixed Models and Missing Data. In: *Linear Mixed Models in Practise*. New York: Springer, pp. 191-274.

Profillidis, V. A. & Botzoris, G. N., 2019. Chapter 6 - Trend Projection and Time Series Methods. In: V. A. Profillidis & G. N. Botzoris, eds. *Modeling of Transport Demand*. Online publication: Elsevier, pp. 225-270.

Qualtrics, 2022. *What is ANOVA (Analysis Of Variance) and what can I use it for?*. [Online] Available at: [https://www.qualtrics.com/experience-management/research/anova/#:~:text=You%20would%20use%20ANOVA%20to,are%20unequal%20\(or%20different\).](https://www.qualtrics.com/experience-management/research/anova/#:~:text=You%20would%20use%20ANOVA%20to,are%20unequal%20(or%20different).)

[Accessed 12 June 2022].

Rotunno, A. et al., 2018. Novel Factors Associated With Analgesic and Anti-inflammatory Medication Use in Distance Runners: Pre-race Screening Among 76654 Race Entrants-SAFER Study VI. *Clinical Journal of Sports Medicine*, 28(5), pp. 427-434.

Salkind, N. J., 2010. *Encyclopedia of Research Design*, California: SAGE Publications, Inc..

SAS, 1999. *Type III Tests*. Version 8 ed. Cary, NC: SAS Institute Inc..

SAS, I. I., 2015. *SAS/STAT® 14.1 User's Guide The MIXED Procedure*, NC. USA: SAS Institute Inc..

Schubert, M. M. & Astorino, T. A., 2013. A Systematic Review of the Efficacy of Ergogenic Aids for Improving Running Performance. *Journal of Strength and Conditioning Research*, 27(6), pp. 1699-1707.

Schwabe, K. et al., 2018. Leisure athletes at risk of medical complications: outcomes of pre-participation screening among 15778 endurance runners- SAFER VII. *The Physician and Sportsmedicine*, 46(2), pp. 405-413.

Schwarz, G., 1978. Estimating the Dimension of a Model. *Annals of Statistics*, Volume 6, pp. 461-464.

Schwellnus, M. et al., 2018. Prerace medical screening and education reduce medical encounters in distance road races: SAFER VIII study in 153 208 race starters. *British Journal of Sports Medicine*, 53(10), pp. 634-639.

Sedgwick, p., 2012. Pearson's correlation coefficient. *The British Medical Journal*, p. 345.

Sewry, D. N., 2022. *SEMLI's involvement with the Two Oceans Marathon Events* [Interview] (30 March 2022).

Sewry, N. et al., 2020. Pre-race screening and stratification predicts adverse events—A 4-year study in 29585 ultra-marathon entrants, SAFER X. *Scandinavian Journal of Medicine & Science in Sports*, Volume 30, pp. 1205-1211.

Sewry, N. et al., 2021. Risk factors for not finishing an ultramarathon: 4-year study in 23996 race starters, SAFER XXI. *The Journal of Sports Medicine and Physical Fitness*.

Singh, V., Rana, R. & Singhal, R., 2013. Analysis of repeated measurement data in the clinical trials. *Journal of Ayurveda and integrative medicine*, 4(2), pp. 77-81.

Tabachnick, B. G. & Fidell, L. S., 2007. *Using multivariate statistics*. Boston: Pearson.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.*, 58(1), pp. 267-288.

Two Oceans Marathon, 2020. *History of Two Oceans Marathon*. [Online] Available at: <https://www.twooceansmarathon.org.za/about-two-oceans/history/#:~:text=The%20Two%20Oceans%20Marathon%20is,to%20face%20the%20unknown%20challenge>.

[Accessed 17 April 2022].

Vaida, F. & Blanchard, S., 2005. Conditional Akaike Information for mixed-effects models. *Biometrika*, Volume 92, pp. 351-370.

van Mechelen, W., 1992. Running injuries. A review of the epidemiological literature. *Sports Medicine*, 14(5), pp. 320-325.

Venturini, E. & Giallauria, F., 2022. Factors Influencing Running Performance During a Marathon: Breaking the 2-h Barrier. *Front Cardiovascular Med*, Volume 9.

Verbeke, G., Fieuws, S., Molenberghs, G. & Davidian, M., 2014. The analysis of multivariate longitudinal data: a review. *Statistical Methods Med Res*, 23(1), pp. 42-59.

West, S. G., Biesanz, J. C. & Kwok, O. M., 2004. In: C. M. & A. P. C. Sansone, ed. *Within-Subjects and Longitudinal Panter: Design and Analysis Issues*. s.l.:Sage Publications, Inc., pp. 287-312.

Wit, E., van den Heuvel, E. & Romeijn, J. W., 2012. 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3), pp. 217-236.

Wolfinger, R., 1993. Covariance structure selection in general mixed models. *Communications in Statistics - Simulation and Computation*, 22(4), pp. 1079-1106.

World, A., 2021. *World Records*. [Online]
Available at: <https://www.worldathletics.org/records/by-category/world-records>
[Accessed 27 August 2022].

Zhang, G. & Chen, J. J., 2013. Adaptive fitting of linear mixed-effects models with correlated random-effects. *Journal of Statistical Computational Simulation*, 83(12).

Zou, H., 2006. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, Volume 101, p. 1418.

Zou, H. & Hastie, T., 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*, 67(2), pp. 301-320.

Zrenner, M. et al., 2021. Retrospective analysis of training and its response in marathon finishers based on fitness app data. *Frontier Physiology*, Volume 12.



7 Appendix A: Main elements of questionnaire

(1) Have you ever suffered from any heart or blood vessel conditions including heart attack, undiagnosed chest pain, coronary artery bypass operation, angioplasty (balloon), heart failure, heart transplant, cardiac arrhythmia (abnormal heart beat), rheumatic fever, heart murmur, cardiomyopathy, myocarditis, use of a pacemaker, or inherited heart defect?

(2) Do you currently suffer from any symptoms of heart or blood vessel disease, including any of the following: shortness of breath when sitting or lying down, shortness of breath with mild exercise, waking up with shortness of breath at night, palpitations that make you dizzy, chest pain when sitting or performing exercise or when you are emotionally stressed, pain (or discomfort) in the neck jaw arms at rest or during exercise, dizziness during exercise or fainting spells)?

(3) Are you aware or have you ever been diagnosed with any risk factors for heart or blood vessel disease including high blood cholesterol, a family member with heart disease, cigarette smoking, lack of physical activity, high blood pressure, being overweight, or having diabetes mellitus (sugar sickness)?

(4) Do you currently suffer from any metabolic or hormonal disease including diabetes mellitus thyroid gland disorders hypoglycemia (low blood sugar) hyperglycemia (high blood sugar), or heat intolerance?

(5) Do you suffer from any respiratory (lung) disease including asthma, emphysema (COPD), wheezing, cough, postnasal drip, hay fever, or repeated flu like illness?

(6) Do you suffer from any gastrointestinal disease including heartburn, nausea, vomiting, abdominal pain, weight loss or gain (> 5kg), a change in bowel habits, chronic diarrhea, blood in the stools, or past history of liver or gallbladder disease?

(7) Do you suffer from any diseases of the nervous system including past history of stroke or transient ischemic attack (TIA), frequent headaches, epilepsy, depression, anxiety attacks, muscle weakness, nerve tingling, loss of sensation, or chronic fatigue?

(8) Do you suffer from any disease of the kidney or bladder including past history of kidney or bladder disease, blood in the urine, loin pain, kidney stones, frequent urination, or burning during urination?

(9) Do you suffer from any disease of the blood or immune system including anemia, recurrent infections, HIV/AIDS, leukemia, or are you using any immunosuppressive medication?

(10) Do you suffer from any growths or cancer, including a past history of cancer?

(11) Do you suffer from any allergies including a past history of allergies, to medication, plant material, or animal material?

(12) At the moment do you use any prescribed medication on a daily weekly or monthly basis to treat chronic (long-term) medical conditions or injuries?

(13) Have you ever collapsed (fell down not because of an accident needing medical attention) during at the finish or after a race or training session?

(14) Do you, or did you suffer from any symptoms of a running injury (muscles tendons bones ligaments or joints) in the last 12 months?

(15) Have you ever in your running career suffered from muscle cramping (painful spontaneous sustained spasm of a muscle) during or immediately (within 6 h) after running (in training or competition)?

*: Once a participant answered “yes” to any of the main screening questions, further details were obtained using “dropdown” boxes with additional questions

8 Appendix B: Questionnaire

Page 1 questions (all compulsory fields)

Please note that we require you to provide answers to all the questions

General running and training information

For how many years have you been a recreational runner* (Please select from the dropdown box)

years

For how many years have you participated in distance running events > 2 hours?* (Please select from the dropdown box)

years

In the last 12 months, on average, how many times a week do you run (train and race) (Please select from the dropdown box)?*

per week

In the last 12 months, what is your average weekly running distance in km?* (Please select from the dropdown box)

km/week

In the last 12 months, what is your average training speed? (Please select from the dropdown boxes – km box and hour box) *

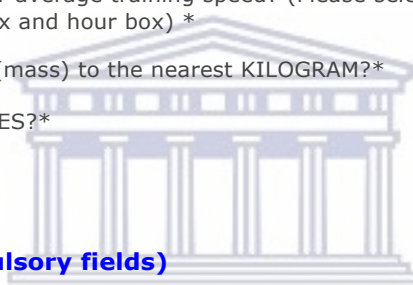
min/km

What is your current body weight (mass) to the nearest KILOGRAM?*

kg

What is your height in CENTIMETRES?*

cm



Page 2 questions (all compulsory fields)

General running training information

In the past 12 months, please indicate the average percentage time that you cycle on a treadmill?

% time on treadmill

In the past 12 months, please indicate the average percentage time that you spent running on roads (tar/concrete/brick)?

% time on roads

In the past 12 months, please indicate the average percentage time that you do trail/mountain running on gravel roads (e.g. jeep tracks)?

% time running on gravel roads

In the past 12 months, please indicate the average percentage time that you do trail/mountain running on footpaths/single tracks?

% time running on footpaths / single tracks

- The information is **NOT** intended to prevent you from taking part on race day
- Please be as accurate and comprehensive as you can in providing this information

Are you aware or have you ever been diagnosed with any risk factors for heart or blood vessel disease, including high blood cholesterol, a family member with heart disease, cigarette smoking, lack of physical activity, high blood pressure, being overweight or having diabetes mellitus (sugar sickness)?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from

You may tick more than one box if needed

- High blood pressure
- High blood cholesterol
- Cigarette smoking
- Obesity (overweight)
- Diabetes mellitus
- Family history of heart disease (< 50 years)

Page 4 questions (yes/no compulsory)

Have you ever suffered from any heart or blood vessel conditions, including heart attack, undiagnosed chest pain, coronary artery bypass operation, angioplasty (balloon), heart failure, heart transplant, cardiac arrhythmia (abnormal heart beat), rheumatic fever, heart murmur, cardiomyopathy, myocarditis, use of a pacemaker or inherited heart defect?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from (you may tick more than one box if needed)

- Myocardial infarct (heart attack)
- Chest pain that has been diagnosed as "angina"
- Coronary artery bypass graft (CABG)
- Angioplasty (no stent)
- Angioplasty (with stent)
- Heart failure
- Heart transplant
- Arrhythmia
- Rheumatic fever
- Heart murmur
- Cardiomyopathy
- Myocarditis
- Use of a pacemaker
- Inherited conditions of the heart or blood vessels
- Any other form of heart or blood vessel disease (please specify)

Page 5 questions (yes/no compulsory)

Do you currently suffer from any symptoms of heart or blood vessel disease including swollen ankles, abnormal shortness of breath (with exercise), chronic dry cough, palpitations, chest pain, pain (or discomfort) in the neck, jaw, or arms at rest or during exercise, dizziness, fainting spells, and/or calf pain when cycling/running/walking/swimming?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from (you may tick more than one box if needed)

- Swollen ankles
- Water retention
- Shortness of breath when sitting or lying down
- Shortness of breath with mild exercise
- Waking up with shortness of breath at night
- Palpitations with no dizziness
- Palpitations that make you dizzy
- Chest pain when sitting
- Chest pain when performing exercise
- Chest pain when you are emotionally stressed
- Pain (or discomfort) in the neck, jaw, arms at rest or during exercise
- Dizziness during exercise
- Fainting spells
- Chronic dry cough
- Painful calves when walking

Page 6 questions (yes/no compulsory)

Have you ever collapsed (fell down-NOT because of an accident) needing medical attention during, at the finish or after a race or training session?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following questions on the same page (compulsory to select / complete fields)

Have you ever collapsed during training or racing?

- Training
- Racing
- Training and racing

How many times have you collapsed in training session or races during the last five years?

Races:

Training session:

How many times have you collapsed in training session or races during the last 12 months (1 year)?

When you collapse, does it mostly occur before or after the finish line / completion of the training session?

- Before the finish
- After the finish

What is the cause of your collapse?

- Dehydration
- Heat illness
- Hyponatraemia(low salt levels confirmed by a blood test)
- Low blood pressure
- Low blood sugar
- Other condition, please specify

Have you ever been ill enough during a race to require an intravenous (IV) drip?

- **YES**
- **NO**

Page 7 questions (yes/no compulsory)

Have you ever in your running career suffered from muscle cramping (painful, spontaneous, sustained spasm of a muscle) during or immediately (within 6 hours) after running (in training or competition)?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following questions on the same page (compulsory to select / complete fields) – may need to split over two pages as there are a number of questions

For how many years have you suffered from cramping?

Did you suffer from cramping during or after running in the last 12 months?

- Yes
- No

In the last 10 races or training sessions, how many times have you experienced cramping?

Races /10:

Training sessions /10:

What treatment/s have you had that successfully relieved an acute cramp?

You can tick more than one

- Stretching
- Resting
- Drinking fluid
- Ice application

- Massage
- Magnesium
- Salt (tablets or solution)
- Pickle juice in your mouth

- Other, please specify

At what point in the race or training session do you usually first experience cramping?

- First quarter
- Second quarter
- Third quarter
- Fourth quarter
- After the race
- No pattern
- Other, please specify

In which muscle do you usually cramp?

Please tick the muscle in which cramps most frequently occur

- Calves
- Hamstrings
- Quadriceps (thigh)
- Foot muscles
- Other, please specify

Have you ever suffered from cramping in your whole body (arms and legs)?

- Yes
- No

Have you ever been admitted to hospital following cramping?

- Yes
- No

Have you ever been confused or in a coma during or after a cramping episode?

- Yes
- No

Have you ever had "dark urine" in the 3 days following a cramping episode?

- Yes
- No

If you cramp, how severe is the cramp usually?

Please tick one box

- Mild: < 5 minutes and you are able to continue exercising
- Moderate: 5-15 minutes and you are able to continue exercising
- Severe: >15 minutes or if you have to STOP exercising

Page 8 questions (yes/no compulsory)

Do you currently suffer from any metabolic or hormonal disease including diabetes mellitus, thyroid gland disorders, hypoglycaemia (low blood sugar), hyperglycaemia (high blood sugar), or heat intolerance?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from

You may tick more than one box if needed

- Hyperglycaemia (high blood sugar) (Pre-diabetes)
- Type 1: Insulin dependent (Diabetes Mellitus)

- Type 2: Non insulin dependent (Diabetes Mellitus)
- Underactive thyroid (hypothyroidism)
- Overactive thyroid (hyperthyroidism)
- Hypoglycaemia (low blood sugar)
- Heat intolerance

Page 9 questions (yes/no compulsory)

Do you suffer from any respiratory (lung) disease including asthma, emphysema (COPD), wheezing, cough, postnasal drip, hay fever, or repeated flu like illness?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from

You may tick more than one box if needed

- Asthma (Non exercise-induced)
- Asthma (Exercise-induced)
- Wheezing during exercise
- Cough during exercise
- Post nasal drip
- Allergies/hay fever (ear, nose, throat)
- Repeated infections in the respiratory tract (> 3 per year)
- Previous lung complaints
- COPD (Chronic obstructive pulmonary disease)
- Interstitial lung disease
- Cystic fibrosis
- Other respiratory complaints

Page 10 questions (yes/no compulsory)

Do you suffer from any gastrointestinal disease including heartburn, nausea, vomiting, abdominal pain, weight loss or gain (> 5kg), a change in bowel habits, chronic diarrhoea, blood in the stools, or past history of liver or gallbladder disease?

- Yes
 No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from

You may tick more than one box if needed

- Heartburn
 Nausea/vomiting
 Abdominal pain
 Weight loss (>5kg) in the last 2 years
 Weight gain (>5kg) in the last 2 years
 A change in bowel habits over the last year
 Chronic diarrhoea
 Blood in stool
 Abdominal complaints during exercise
 Liver/gallbladder disease
 Other gastrointestinal complaints

Page 11 questions (yes/no compulsory)

Do you suffer from any diseases of the nervous system including past history of stroke or transient ischaemic attack (TIA), frequent headaches, epilepsy, depression, anxiety attacks, muscle weakness, nerve tingling, loss of sensation, or chronic fatigue?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from

You may tick more than one box if needed

- Stroke or transient ischaemic attack
- Frequent headaches
- Epilepsy
- Depression
- Anxiety attacks
- Other psychological/psychiatric conditions
- Muscle weakness
- Nerve tingling/loss of sensation
- Chronic fatigue
- Other nervous system complaints

Page 12 questions (yes/no compulsory)

Do you suffer from any disease of the kidney or bladder including past history of kidney or bladder disease, blood in the urine, loin pain, kidney stones, frequent urination, or burning during urination?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from

You may tick more than one box if needed

- Past history of kidney disease
- Past history of bladder disease
- History of blood in the urine
- Chronic loin pain
- History of kidney stones
- Frequent urination
- Burning during urination

Page 13 questions (yes/no compulsory)

Do you suffer from any disease of the blood or immune system including anaemia, recurrent infections, HIV/AIDS, leukaemia, or are you using any immunosuppressive medication?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from

You may tick more than one box if needed

- Past history of anaemia
- Past history of cancer of the blood cells (leukaemia)
- Past history of cancer of the lymphatic system (lymphoma)
- Past history of blood disorders
- History of HIV/AIDS
- History of a depressed immune system

Page 14 questions (yes/no compulsory)

Do you suffer from any growths or cancer including a past history of cancer?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from

You may tick more than one box if needed

- Past history of cancer
- Current undiagnosed growth

Page 15 questions (yes/no compulsory)

Do you suffer from any allergies including a past history of allergies to medication, plant material or animal material?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate condition/s that you suffer/ed from

You may tick more than one box if needed

- Past history of allergies to medication
- Past history of allergies to plant material
- Past history of allergies to animal material
- History of any other allergies

Have you ever needed to use adrenalin to control your allergic symptom or have you been advised to carry an adrenalin pen (Epipen®)

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Please tick the appropriate box

You may tick more than one box if needed

- I used adrenalin to control allergic symptoms
- I have been advised to carry an adrenalin pen (Epipen ®) I

Page 16 questions (yes/no compulsory)

At the moment, do you use any prescribed medication on a daily, weekly or monthly basis to treat chronic (long-term) medical conditions or injuries?

- Yes
- No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

Pease tick the type of medication/s that you are taking from the list below:

You may tick more than one box if needed. If your medication type is not on the list please enter it in the free text box that is below the list.

- Cholesterol lowering medication
- Blood pressure lowering medication
- Medication to control heart rhythm
- Medication to treat heart failure
- Medication to prevent blood clots (blood thinners)

- Other medication to treat heart disease
- Medication (tablets) to treat type 2 diabetes
- Insulin for diabetes
- Medication to treat anxiety
- Anti-depressant medication
- Medication to improve concentration
- Medication to stop smoking

- Anti-asthma medication
- Medication to treat long standing joint problems (rheumatoid arthritis)
- Medication to treat thyroid disease
- Hormone therapy ie. oestrogen, progesterone, oral contraception
- Medication to control long standing inflammation ie. cortisone, immune suppressants
- Medication to control gastrointestinal related problems ie. spastic colon, reflux, Chrohns disease, Ulcerative Colitis
- Herbal medication to treat joint or muscle ailments

- Other medication (please list in box below)



Page 17 questions (yes/no compulsory)

Have you ever in your running career used medicines to treat injuries in the week before or during a race – including anti-inflammatory drugs, cortisone (pills, or injection), or pain killers?

- Yes
- No

If no response, go to next page

**If yes response, then drop down the following boxes on same page
(compulsory to select at least one)**

Which of the following medicines have you used in the past to treat an injury in the week just BEFORE a race?

- Paracetamol (e.g. Panado, Tylenol)
- Non-steroidal anti-inflammatories (e.g. Voltaren, Cataflam)
- Cortisone (pills)
- Cortisone injection
- Codeine
- Anti-inflammatory gels/creams/patches
- Any other pain killers
- Herbal medication

Which of the following medicines have you used in the past to treat an injury DURING a race?

- Paracetamol (e.g. Panado, Tylenol)
- Non-steroidal anti-inflammatories (e.g. Voltaren, Cataflam)
- Cortisone (pills)
- Cortisone injection
- Codeine
- Anti-inflammatory gels/creams/patches
- Any other pain killers

Page 18a and b questions (yes/no compulsory)

If no, proceed to Page 19

If yes, proceed to 18b

Do you or did you suffer from any symptoms of a CHRONIC (no accident) running injury (muscles, tendons, bones, ligaments or joints) IN YOUR RUNNING CAREER?

(NB: Only if an injury is/was severe enough to interfere with cycling, or require treatment e.g. use medication, or require you to seek medical advice from a health professional)

- Yes
- No

Injury 1

Page 18b questions (yes/no compulsory)

Do you or did you suffer from any symptoms of a CHRONIC running injury (muscles, tendons, bones, ligaments or joints) IN THE PAST 12 MONTHS OR CURRENTLY?

(NB: Only if an injury is/was severe enough to interfere with running, or require treatment e.g. use medication, or require you to seek medical advice from a health professional)

- Yes
- No

If no response to 18b, go to next page

If yes response to 18b, then drop down the following box on same page (compulsory to select at least one)

Pease tick if past or current:

- Past
- Current

How long ago did you first become aware of the CHRONIC injury? (months)

months

Please indicate which side of your body is injured (if applicable)

- Right
- Left
- Both

Please indicate which anatomical area is/was injured (single select)

- Head
- Neck
- Face
- Front chest
- Back chest
- Shoulder
- Upper arm
- Elbow
- Forearm
- Wrist
- Finger
- Lower back
- Hip
- Groin
- Hip muscle (including gluteus / buttock muscles)
- Hamstring muscle
- Quadriceps muscle
- Calf muscle
- Knee
- Shin / Lower leg
- Achilles
- Ankle
- Foot
- Other, please specify

Please indicate the type of structure that was injured (single select)

- Muscle (e.g. strain)
- Ligament (e.g. sprain)
- Tendon
- Joint (e.g. arthritis)

Nerve (e.g. numbness during or after cycling)

Bone (e.g. bruise or stress fracture)

Other, please specify

Please indicate if your injury was any of the following more common running injuries (single select)

Knee - Patellofemoral pain / anterior knee pain

Knee - Iliotibial band (ITB)

Neck pain

Lower back pain

Groin / genital numbness

Saddle sores

Hip joint pain

Hip muscle injury (including gluteus / buttock muscles)

Hamstring injury

Quadriceps muscle injury

Achilles tendon injury

Calf muscle injury

Foot or heel pain

Shoulder pain

Elbow pain

Wrist pain

Numbness in the hand / fingers

Other, please specify

Please indicate the severity of the injury

I only experience symptoms after exercise

I experience symptoms during exercise, but it does not interfere with exercise

I experience symptoms during exercise that may interfere with my training/ competition

I am so painful that I may not be able to train or compete

Please indicate how your injury was treated to date (you can tick more than one)?

- Rest
- Tablets
- Stretches
- Cortisone injection
- Physiotherapy
- Other injection
- Surgery
- Bicycle set-up
- Strengthening exercises
- Equipment change (cycling pants)
- Equipment change (cycling gloves)
- Equipment change (cycling shoes)
- Other, please specify

Would you like to list another important CHRONIC running injury?

- Yes
- No

If no response, go to next page 19

If yes response, then drop down the following box on same page (compulsory to select at least one)

(At this point, there is an option to complete details for more than one injury using the same data capture procedure for the first injury)

Page 19a questions (yes/no compulsory)

Do you or did you suffer from any symptoms of an ACUTE ACCIDENTAL injury (muscles, tendons, bones, ligaments or joints) IN THE PAST 12 MONTHS OR CURRENTLY?

(NB: Only if an injury is/was severe enough to interfere with cycling, or require treatment e.g. use medication, or require you to seek medical advice from a health professional)

- Yes
- No

If no response to 19a, go to next page 20

If yes response to 19a, then drop down the following box on same page (compulsory to select at least one)

Pease tick if past or current:

- Past acute injury (accident)
- Current acute injury (accident)

How long ago did your ACUTE ACCIDENTAL injury occur? (weeks)

weeks

Please indicate which side of your body is injured (if applicable)

- Right
- Left
- Both

Please indicate which MAIN anatomical area is/was injured in the accident (single select)

- Head
- Neck
- Face
- Front chest
- Back chest
- Front abdomen

F flank -back

- Shoulder
- Upper arm
- Elbow
- Forearm
- Wrist and hand
- Finger
- Lower back
- Hip
- Groin
- Hip muscle (including gluteus / buttock muscles)
- Hamstring muscle
- Quadriceps muscle
- Calf muscle
- Knee
- Shin / Lower leg
- Achilles
- Ankle
- Foot
- Other, please specify

Please indicate the type of injury occurred in the MAIN anatomical area that was affected (single select)

- Muscle strain
- Muscle rupture
- Tendon rupture
- Ligament sprain
- Ligament rupture
- Joint cartilage tear / injury
- Joint dislocation
- Significant skin injury (laceration or graze)

Nerve (e.g. numbness or loss of muscle power)

Bone (fracture)

Bone (bruise)

Head injury (fracture)

Head injury (concussion)

Internal organ injury i.e. lung, spleen , liver, heart,
intestines, kidney ,bladder

Other, please specify

Please indicate the severity of the ACUTE injury by indicating how many days were you not able to run following the injury

1-7 days

7-14 days

14-28 days

> 28 days

Please indicate how your ACUTE injury was treated to date (you can tick more than one)?

Rest

Tablets

Injections

Physiotherapy

Rehabilitation

Surgery

Other, please specify

Would you like to list another important ACUTE ACCIDENTAL running injury?

Yes

No

If no response, go to next page 20

If yes response, then drop down the following box on same page (compulsory to select at least one)

(At this point, there is an option to complete details for more than one ACUTE injury using the same data capture procedure for the first injury)

Page 20 questions (yes/no compulsory) (Can this question only come up after certain questions – listed in email)

Have you consulted with a medical doctor in the last 12 months to obtain medical clearance that you can safely participate in endurance running?

- Yes
 No

If no response, go to next page

If yes response, then drop down the following box on same page (compulsory to select at least one)

If yes, please indicate which of the following procedure formed part of the medical assessment for clearance to participate in endurance running? (you may tick more than one box if needed)

- Your doctor spoke to you only
- Your doctor spoke to you and examined you physically
- You performed an exercise test but no ECG (electrical leads attached to your chest to measure the hearts response to exercise)
- You performed an exercise test with an ECG (electrical leads attached to your chest to measure the hearts response to exercise)
- You had an echocardiogram (a sonar of the heart to examine the structure of the heart)
- You had blood tests for cholesterol
- You had other blood tests
- You had other tests (please specify)

After seeing your medical practitioner please indicate which of the following applied?

- My doctor did not give clearance for me to run

- My doctor did give clearance for me to run but with some restrictions and guidelines on safe participation
- My doctor did give clearance to run with no restrictions

Page 21 questions (yes/no compulsory)

Consent for medical information to be used for research purposes

You do also have the opportunity to volunteer that the information on these medical questionnaires can be used for ongoing medical and scientific research to improve race safety and medical care.

The Institute for Sport, Exercise Medicine and Lifestyle Research of the University of Pretoria, in collaboration with the race organizers and the [Medical service Provider] medical team conducts on-going research to improve race safety (protecting the health of the athlete and reducing injury risk). Your participation in this research effort is to improve safety and is entirely voluntary. Please read through the Participant information and then you will be given the opportunity to consent that your information in the medical questionnaires can be included in research studies, and that you can be contacted about participating in other components of the research project that relate to muscle cramps and injuries.

Participant information of the research studies:

The main aim of these studies is to determine if there are any factors that can be identified before the race that will predict whether an athlete is likely to develop a medical problem (including cramps and injuries) during or after the race. The details of the studies are as follows:

- At the race entry and registration, a web-based (or a paper-based) questionnaire detailing personal particulars and medical information, will be completed as part of the race entry and race registration requirements.
- The completion of a questionnaire is not associated with any risk. Questionnaire and other clinical data (paper and electronic) will be kept confidential, will be kept secure, and will not be made available to any party other than the medical and research team without the consent of the individual participant.
- You may be contacted before or after the race (by telephone or email), for further information, advice and participation in research related to injuries or a medical condition (such as cramps) that you developed before, during or after the race.
- Volunteering to make medical information available for on-going research has no direct benefit to an individual athlete. However, the long term anticipated benefits of this research are to identify factors that may predispose an increased risk of medical consequences and injury in endurance athletes. This information will eventually assist athletes in decreasing their risk of medical complications and injuries during racing and training.

Consent to participate in the research study

- I understand that I am free to volunteer to participate in the study on pre-race predictors (including medical history, medication use, and injuries) of medical complications that may occur in runners before, during and immediately after the race
- I understand that my participation in this research project may have no direct benefits to me during the race. However, I understand that my participation in the research project will advance the medical and scientific knowledge related to endurance sports. Therefore, information gathered through my participation in this project could advance the future medical care, training advice and performance of endurance athletes.
- I have read the participant information and am satisfied that the procedures and concepts

- have been explained to me in full.
- I agree that all the questionnaire information, my performance during the race, together with all the other data collected from the various components of this study may be used to answer scientific questions about the medical conditions, injuries, physiological responses and measures of performance associated with the preparation, participation in and completion of a race.
 - I have been informed that the individual data derived from my participation will remain confidential
 - I understand that the data obtained from this study may be used for the research components of higher degrees at the University of Pretoria.
 - I understand that the Research Ethics Committee of the Faculty of Health Sciences at the University of Pretoria has approved the protocol for this research study (REC number XXXXXX).
 - I understand that each of the medical practitioners involved in the research study on athletes will have up to date professional medical insurance.
 - I understand that I can contact members of the research team should I have any questions related to the study. Contact details of the research team are as follows: +27 12 CCCCCCCC
 - I hereby consent to participate in this study, and that I can be contacted for information about research studies on injuries and medical conditions.
 - I understand that I may withdraw from this study at any time without further question.

Consent to allow medical information in this questionnaire to be used in ongoing research

Yes, I give consent that the information from the medical questionnaires can be used in ongoing research

No, I do not give consent that the information from the medical questionnaires can be used in ongoing research



Medical questionnaire at the time of registration (implementation to be discussed)

Exercise and symptoms of an acute infection

Symptoms of acute illness and infections such as flu, gastro-enteritis (upset stomach) and other infections (e.g. bladder) are more common in athletes just before a race (after periods of peak training). Exercising with symptoms of an infection can increase the risk of medical complications during the race.

The symptoms of infections vary but include the following: generally not feeling well, fever, general muscle pain, general joint pain, general tiredness, headache, sore throat, blocked or runny nose, sore ears, cough, wheeze, diarrhoea, nausea, vomiting, or abdominal cramps/pain.

Please answer the following question so that we can give you advice:

Question 1:

Do you have any of these symptoms of acute illness (today or in the last 7 days)?

No

Yes

Question 2: Symptoms of an acute infection or illness (if yes to question 1)

9 Appendix C: SAS code for modelling procedures

Model 1

```
proc mixed data=c;
class runnercode year_ORDER(ref="1");
model time_f_min=year_order/ s ddfm=kenwardroger e3 cl;
repeated year_order/ type=un subject=runnercode r rcorr;
lsmeans year_order/ diff=all;
random maxrace/type=vc g gcorr;
run;
```

Model 2

```
proc mixed data=c;
class runnercode year_ORDER(ref="1") gender;
model time_f_min=year_order age gender bmi/s ddfm=kenwardroger e3 cl;
repeated year_order/ type=un subject=runnercode r rcorr;
lsmeans year_order/ diff=all;
random maxrace/type=vc g gcorr;
run;
```

Model 3

```
proc mixed data=c;
class runnercode year_ORDER(ref="1") gender recent_run_injury(ref='0');
```



```

model time_f_min=year_order age gender bmi training_pace recreationrunner
recent_run_injury sc_chronic/ s ddfm=kenwardroger e3;

repeated year_order/ type=un subject=runnercode;

lsmeans year_order/ diff=all;

random maxrace/ g gcorr;

run;

```

Model 4

```

proc mixed data=c plots(maxpoints=none);

class runnercode year_ORDER(ref='1') gender recent_run_injury(ref='0') ;

model time_f_min=year_order|training_pace age bmi gender recreationrunner
recent_run_injury sc_chronic/ s ddfm=kenwardroger e3 residual;

repeated year_order/ type=un subject=runnercode;

lsmeans year_order/ diff=all;

random maxrace/ g gcorr solution;

run;

```

