Bayesian Analysis Report

Davies Luo (Zheng Luo)

2024-02-22

Introduction

Cross-country running is a sport that amalgamates endurance, strategy, and adaptability, challenging athletes across varied terrains and conditions. In this report, I will delve into the North East Harrier League dataset to unearth the multifaceted influences on race performance. Through multiple Bayesian analytical methods, I will scrutinise race times in relation to athletes' age, pack classification, and the different characteristics of each course, alongside various environmental factors. The exploratory journey seeks to distill these complexities into a coherent narrative, elucidating the subtle and overt forces that affect the finishing time the most in this competitive sports.

Data Exploratory

Before diving into the complexities of Bayesian Analysis, it is essential to conduct a thorough exploratory data analysis (EDA). EDA is a critical step in the data analysis process, as it allows me to uncover underlying structures, extract important variables, detect anomalies, test assumptions, and develop an intuitive understanding of the dataset. The dataset encompasses a myriad of factors including athlete age, speed pack, course name and environmental conditions with physical attributes of each race. The following table of the data head offers a preliminary view into the complex interplay of variables that I aim to unravel:

	Number	Pack	Age	Course	Year	Temperature	Windspeed	Distance	Elevation	Response
6299	629	S	MV45	Druridge	2016	12	11	6.11	164	40.25000
1697	454	S	MV45	Alnwick	2018	0	25	6.23	NA	50.38333
2651	1272	F	MV40	Aykley	2017	7	12	6.13	277	35.81667
4829	1310	S	MV45	Aykley	2018	11	8	6.13	277	49.20000
276	455	M	Msen	Gosforth	2018	7	20	6.03	244	40.40000
6555	210	S	MV55	Druridge	2016	12	11	6.11	164	45.00000

From the initial look of the data, there's lack of record for the 'Elevation' on the course 'Alnwick' so I remove the 'Elevation' for consistency. Furthermore, I noticed there are "guest" and "n/c" age groups in 'Age' variable, which do not contribute to the analysis, so I get them removed as well.

Then it is imperative to cleanse the dataset, ensuring a foundation built on accuracy and completeness. This process involves the removal of any incomplete records, as denoted by NA values, which might otherwise skew the interpretations. Moreover, I convert categorical variables into a numeric codex that facilitates computational analysis. 'Pack', 'Age', and 'Course' transmute from mere labels to quantifiable entities, made suitable for the explorations ahead. Furthermore, I notice multiple duplication in 'Number', suggesting I need to consider individuality of athletes in further analysis.

Number	Pack	Age	Course	Year	Temperature	Windspeed	Distance	Response
629	3	5	3	2016	12	11	6.11	40.25000
454	3	5	1	2018	0	25	6.23	50.38333
1272	1	4	2	2017	7	12	6.13	35.81667
1310	3	5	2	2018	11	8	6.13	49.20000
455	2	1	4	2018	7	20	6.03	40.40000
210	3	7	3	2016	12	11	6.11	45.00000

Correlation Heatmap

With the data primed, now I proceed to weave a correlation matrix, a web that captures the essence of interactivity between variables. This matrix illuminates the strengths of relationships within the dataset, where I can discern patterns, identify the strands of strongest connection, and subtle influence among them.

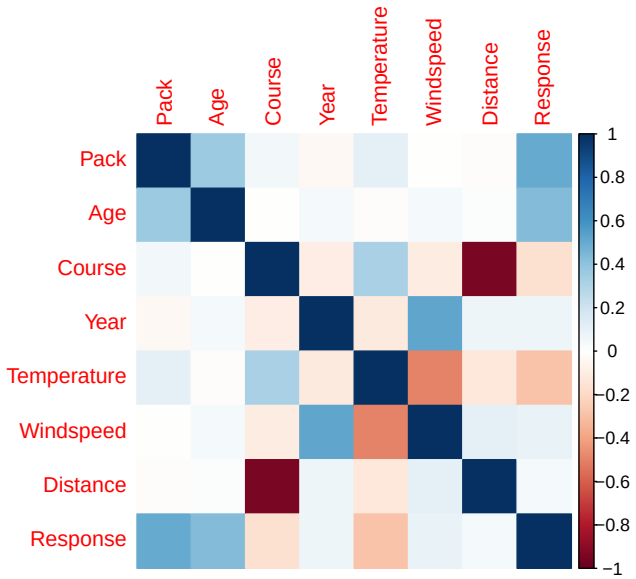


Figure 1: Correlation Heatmap

The correlation matrix has revealed several interesting relationships that worth further investigation. I observe a positive correlation between 'Response' and 'Age', indicating that older age groups might have longer race times. Additionally, 'Pack' also shows a strong correlation with 'Response', suggesting that different pack groups could be associated with faster times. To understand how these factors may influence race times, I will conduct a series of detailed exploratory analyses.

Exploration of Age and Response

To delve deeper into the relationship between the athlete's age group and their race times, I will create a box plot with a regression line. This will help visualise whether older age groups tend to have longer race times, as suggested by the correlation matrix.

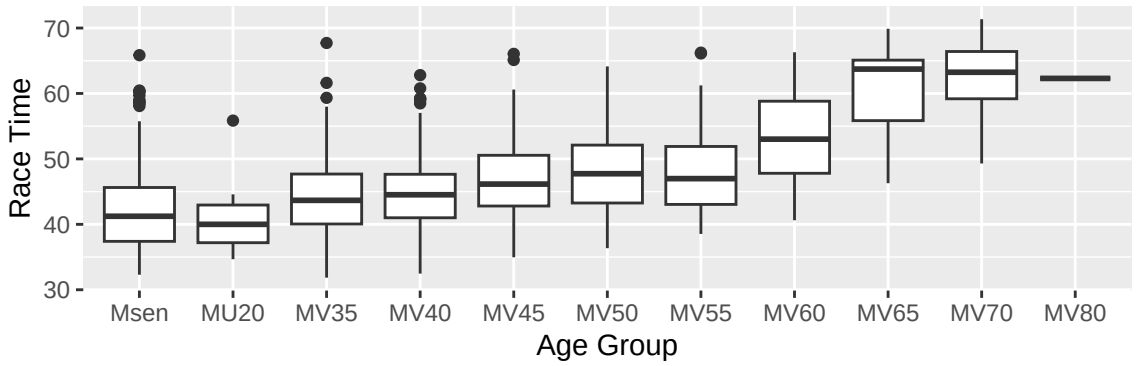


Figure 2: Race Times across Different Age Groups

The boxplot of race times across different age groups reveals a clear trend in performance with respect to age. Younger runners, particularly those under 20 (MU20), tend to have quicker race times, as indicated by the lower median and smaller interquartile range. As age increases, there is a noticeable increase in both the median race times and the variability of those times, which is evident in the widening of the boxplots for older age groups. The veteran categories, particularly those over 60 (MV60 and above), display the highest race times with increased spread, suggesting a wider range of performance within these groups. These findings prompt further investigation into the impact of age on race performance, controlling for other variables such as course difficulty and weather conditions, to better understand the extent of its influence on the race times.

Exploration of Age, Pack and Response

Given a positive correlation between 'Pack' and 'Response', I will analyse the distribution of response time within pack categories to see if there is a trend towards runners being in slower or faster packs. And in order o understand the interplay between age and pack classification, I will add age into the plot as well.

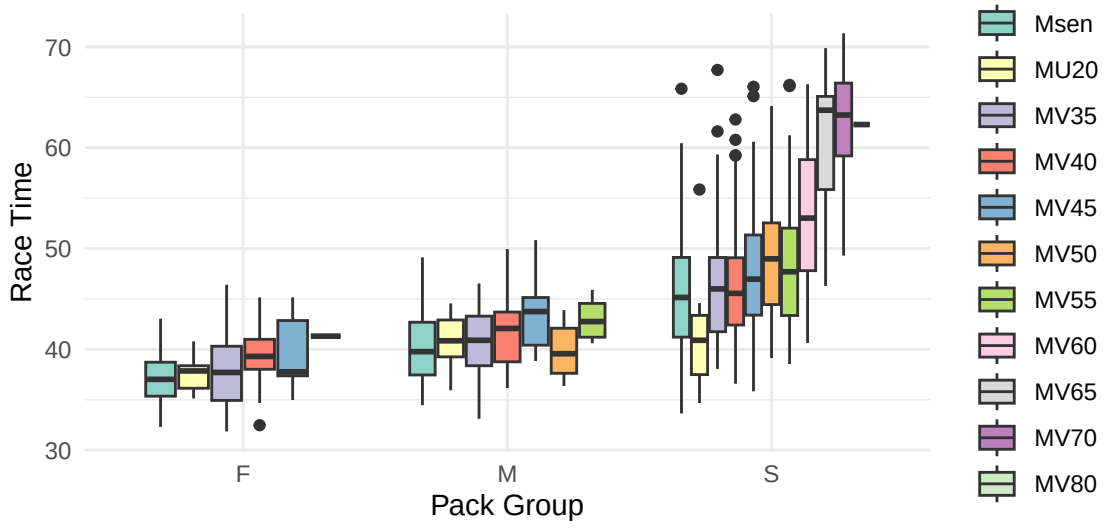


Figure 3: Race Times across Different Pack Groups

Analysing race times across the different pack groups — fast (F), medium (M), and slow (S) — indicates a stark differentiation in performance. Runners in the fast pack show notably quicker race times with a lower median and a tight interquartile range, implying a consistent performance among these athletes. And within the fast pack, the spread of race times is relatively narrow across the younger age groups, signifying

top performance are almost only among these younger athletes. The medium pack displays a broader range, with middle-aged groups (MV50 to MV55) being well-represented. Intriguingly, in the slow pack, there's a visible trend where older age groups, especially those above 60 (MV60 and beyond), tend to have not only higher race times but also a more considerable spread, indicating a diverse range of performance. These initial observations will be further examined through more complex models to assess the true impact of pack and age classification on race times while considering other potential confounding factors.

Exploration of Course and Response

In the realm of cross-country running, the course plays a pivotal role in athlete performance, each race venue may present its unique set of challenges. The variable 'Course' offers a natural experiment to examine how geographic and environmental factors impact race times.

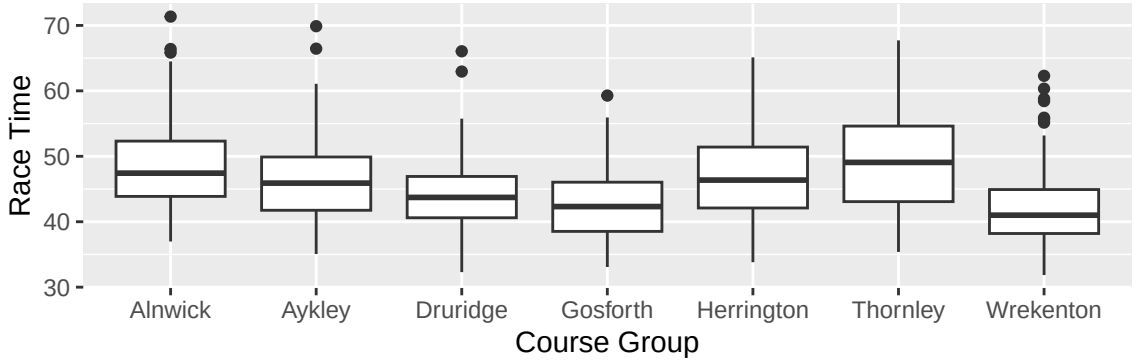


Figure 4: Race Times across Different Course Groups

The comparison of race times across different course groups presents a nuanced picture of the impact of race location on performance. Certain courses, such as Wrekenton and Thornley, show higher medians and greater variability in race times, suggesting these courses may present more challenging conditions or profiles that could affect race performance. Conversely, courses like Alnwick and Aykley demonstrate lower medians, which could indicate less challenging terrain or more favorable conditions for faster times. The insights gained from this visualisation serve as a foundation for subsequent analyses where course effects will be quantified more precisely using multilevel models that take into account the hierarchical nature of the data, alongside other race-specific factors like weather profiles.

Exploration of Response Distribution

Before delving into complex model building, it is crucial to examine the underlying distribution of the 'Response' variable — the race times. A quintessential step in this exploratory phase is the construction of a Quantile-Quantile (QQ) plot, which serves to compare the distribution of race times against a theoretical normal distribution. The normality assumption underpins many statistical models, including the conventional linear regression.

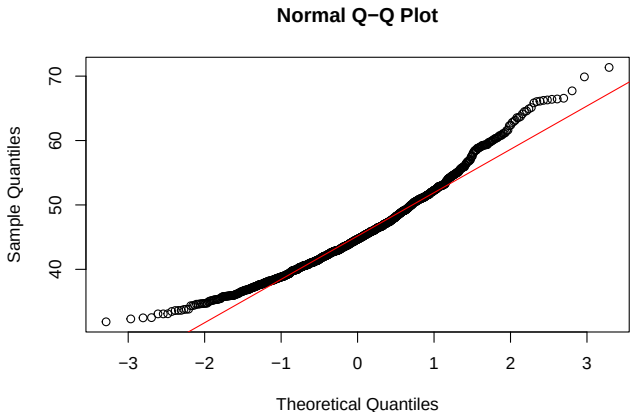


Figure 5: QQ Plot for Response

The alignment of the data points along the reference line indicates that the race times are reasonably well-modeled by a normal distribution, especially in the central quantiles. However, slight deviations at the tails suggest potential outliers or extreme values that are not entirely captured by a normal model. Despite these minor discrepancies, the overall pattern justifies the consideration of models that assume normality, such as the Normal Linear Regression Model. Moreover, the presence of slight deviations encourages the exploration of Non-conjugate Models, which do not strictly adhere to normality assumptions and can offer more flexibility. Additionally, to uncover the synergy between variables, such as how age and pack dynamics jointly influence race times, aturally leads us to consider Interaction model as well. Furthermore, the layered complexity of our data — with its multitude of races, courses, and athletes — makes Hierarchical models an especially attractive option for capturing the nested structures within. I believe these models will allow me to have a nuanced understanding of the race times that accounts for both individual and group-level variability.

Bayesian Analysis

Non-Conjugate Model

The Non-Conjugate model represents a sophisticated approach to understanding the intricacies of athlete performance data. Unlike conjugate models where the posterior distributions belong to the same family as the prior, non-conjugate models do not restrict the form of the posterior. This flexibility allows me to accommodate more realistic and intricate relationships between the variables. The mathematical representation of the non-conjugate model is given by:

$$\text{Response}_i \sim \mathcal{N}(\mu_i, \tau)$$

where the mean μ_i is a linear combination of the predictors and random effects:

$$\mu_i = \beta_0 + \beta_{\text{Course}} \times \text{Course}_i + \beta_{\text{Age}} \times \text{Age}_i + \beta_{\text{Temp}} \times \text{Temperature}_i + \beta_{\text{Pack}} \times \text{Pack}_i$$

$$+\beta_{\text{Year}} \times \text{Year}_i + \beta_{\text{Wind}} \times \text{Windspeed}_i + \beta_{\text{Dist}} \times \text{Distance}_i + u_{\text{Number}_i}$$

In this equation, β coefficients represent the fixed effects of the respective variables, while u_{Number_i} denotes the random effects associated with each athlete, capturing individual variability beyond the fixed effects. The precision of the response variable is denoted by τ, and its reciprocal is the variance, known as σ^2.

Random effects u are normally distributed with mean 0 and precision τ_u, allowing the model to account for the repeated measures within athletes.

The priors for the model parameters were chosen to be weakly informative, (0,1) sets a neutral, centered starting point for the regression coefficients, suggesting no initial bias towards positive or negative effects. The (1,0.001) prior for precision terms reflects an expectation of variability in the data, allowing the model to be informed primarily by the observed data rather than strong prior beliefs.

Strategically, the fitting process will involve running multiple Markov Chain Monte Carlo (MCMC) chains to ensure a thorough exploration of the parameter space and to assess convergence using diagnostics such as the Gelman-Rubin statistic, the effective sample size and visual inspections of trace plots will further corroborate the stability and reliability of the estimates. After assuring convergence, I will then use the 95% credible intervals to check if each variable is useful for this model or not.

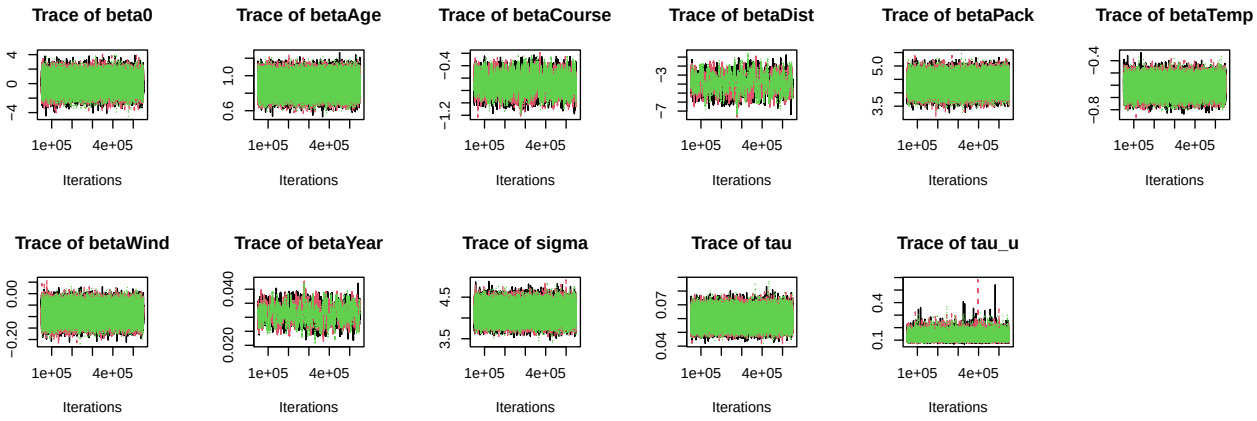


Figure 6: Trace Plots and Density Plots of Non-Conjugate Model

Table 3: Gelman-Rubin statistic of Non-Conjugate Model

	beta0	betaAge	betaCourse	betaDist	betaPack	betaTemp	betaWind	betaYear	sigma	tau	tau_u
Point est.	1	1	1.00	1.01	1	1	1	1.01	1	1	1
Upper C.I.	1	1	1.01	1.02	1	1	1	1.02	1	1	1

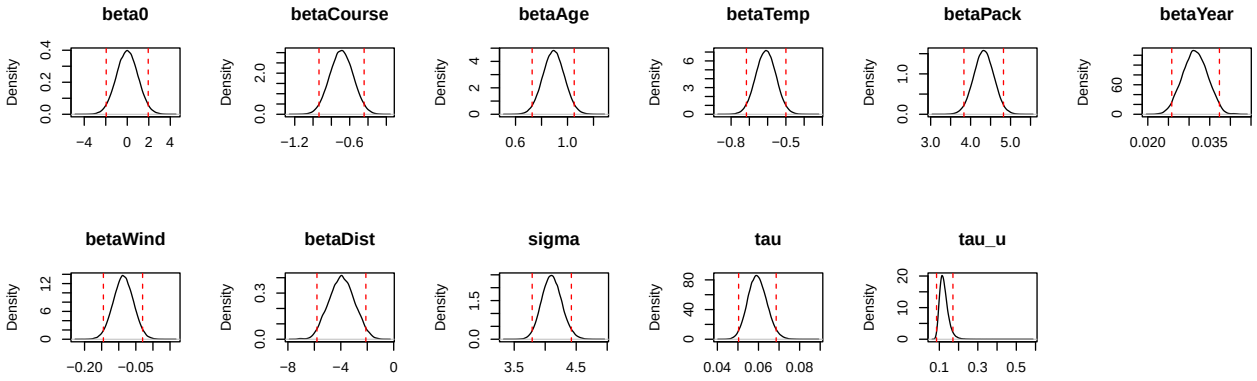
Table 4: Effective Sample Sizes of Non-Conjugate Model

beta0	betaAge	betaCourse	betaDist	betaPack	betaTemp	betaWind	betaYear	sigma	tau	tau_u
13018.57	95563.58	597.94	406.41	33623.65	12850.24	52987.35	376.91	58832.81	59686.01	52995.96

As depicted in Figure 6, the trace plots The trace plots demonstrate satisfactory convergence characteristics. The trace plots, which visualise the MCMC chains for each parameter, exhibit the desired "fuzzy caterpillar" appearance, with the chains mixing well and covering the parameter space evenly. This suggests that the chains are stable and have reached equilibrium. The Gelman-Rubin statistic, as shown in Table 3, corroborates this by presenting values close to 1 for all parameters, indicating no significant between-chain variability and thus supporting the assumption of convergence. Furthermore, the effective sample sizes for the parameters, detailed in Table 4, are considerably high, with values well into the thousands, which suggests that the MCMC samples are representative and reliable for posterior inference.

Table 5: Posterior distribution of Non-Conjugate Model Summary

	Mean	SD	X2.5.	X25.	X50.	X75.	X97.5.
beta0	-0.01	1.00	-1.96	-0.68	-0.01	0.67	1.95
betaAge	0.89	0.08	0.73	0.84	0.89	0.95	1.06
betaCourse	-0.69	0.13	-0.94	-0.78	-0.69	-0.60	-0.44
betaDist	-3.93	0.96	-5.77	-4.60	-3.93	-3.27	-2.07
betaPack	4.33	0.25	3.83	4.16	4.33	4.50	4.82
betaTemp	-0.61	0.06	-0.72	-0.65	-0.61	-0.57	-0.50
betaWind	-0.09	0.03	-0.14	-0.11	-0.09	-0.07	-0.03
betaYear	0.03	0.00	0.03	0.03	0.03	0.03	0.04
sigma	4.11	0.16	3.80	4.00	4.10	4.21	4.44
tau	0.06	0.00	0.05	0.06	0.06	0.06	0.07
tau_u	0.12	0.02	0.09	0.11	0.12	0.14	0.18



The posterior distribution of the parameters, summarised in Table 5, provides insights into the relationship between the predictors and the race times. The mean values of beta coefficients reflect the average influence of each predictor on race times, with 'betaDist' showing a notably strong negative association, suggesting longer distances are correlated with longer race times.

The high effective sample sizes for all parameters indicate a high degree of confidence in these estimates. The density plots for the beta coefficients further reinforce these findings, with the majority of the posterior mass centered around the mean values. The 95% credible intervals (dashed red lines) providing a range within which the true parameter values likely fall with high probability, and all variables' posterior density does not includes 0, all the variables will be retained from the model.

Normal Linear Regression Model

Bayesian Normal Linear Regression extends the framework of linear regression into the Bayesian paradigm. By assuming a linear relationship between the response and the predictors, the model provides estimates of the average effect of each predictor, holding the others constant. This linearity assumption, coupled with the normality of errors, allows me to infer the significance and magnitude of each predictor's influence on race times. The Normal Linear Regression Model used in the analysis can be expressed by the following equation:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

where Y_i is the response variable, representing the race time for the i^{th} participant. X_{ij} represents the j^{th} predictor variable for the i^{th} participant. The intercept term, β_0, captures the expected value of Y when

all predictors are at their reference levels. The coefficients β_j for $j = 1, \ldots, k$ correspond to each predictor, signifying the expected change in the response variable per unit change in the predictor. Finally, ϵ_i is the error term for the i^{th} observation, which is presumed to be normally distributed with a mean of zero and constant variance σ^2.

In setting the prior for the Bayesian Normal Linear Regression Model, I employed the empirical Bayes method. The variance of the response variable, which reflects the variability of race times, was calculated from the data and found to be approximately 47.5, thereby anchoring the prior belief to the observed data. To allow the data to substantially inform the posterior distribution, I set the prior degrees of freedom d to 1. This choice represents a weakly informative prior, ensuring that the posterior inferences are driven more by the data than by strong prior assumptions. Then the prior scale v and the precision were caculated out to be 0.021.

I then use the linmod function, which takes the prior and data to compute the posterior distribution. This function combines the likelihood of the observed data with the prior distribution using Bayes' theorem and returns the posterior mean and variance of the regression coefficients.

Table 6: Coefficents Summary of Normal Linear Regression Model

	(Intercept)	Course	Age	Pack	Temperature	Windspeed	Distance
Value	19.12	0.07	0.87	4.79	-0.74	-0.12	3.08

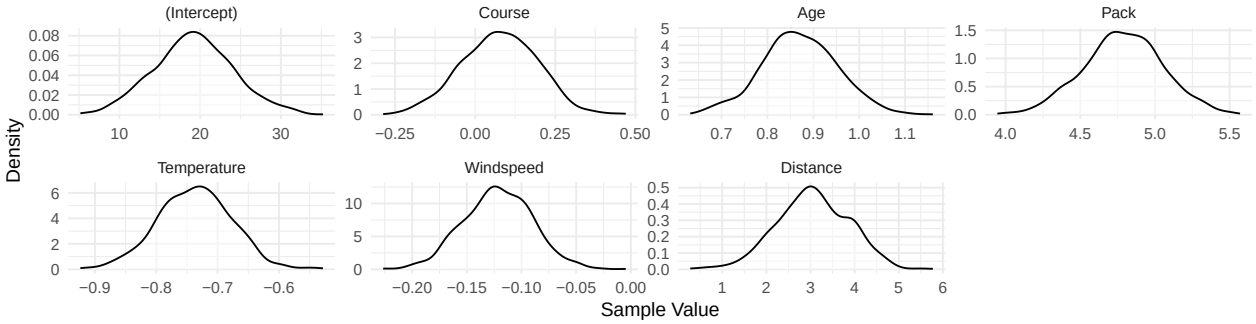


Figure 7: Density Plots for Posterior Distributions of Coefficients

The density plots for the posterior distributions of coefficients from the Normal Linear Regression Model provide key insights into the influence of each variable on the race times. The intercept's central value suggests a baseline for race times when other predictors are at reference levels. The coefficients for Pack and Age have relatively quite high value of means, indicating notable influences on race times, with Pack being particularly impactful. It is a bit lower for the Age variable, but still is a critical determinant, suggesting that both pack and age classification significantly affects performance.

Temperature displays an inverse relationship with race times, where a unit increase in temperature is associated with a decrease in race time. This implies better performance in cooler weather. Windspeed's slight negative coefficient suggests minimal but noticeable effects on race times. Distance has the most substantial coefficient, clearly indicating that as race distances increase, so do race times. This relationship emphasises the importance of endurance training for longer races.

Collectively, these coefficients highlight the multifactorial nature of race performance. Age and pack level are pivotal, but environmental factors also significantly influence outcomes.

Interaction model

To continue distilling nuanced insights from the data, I introduce the interaction model, a statistical method designed to unravel the synergistic effects between variables. This model is constructed to capture the

combined effects of different variables on the response variable. The interactions between variables such as course and temperature, pack and age, among others, are included to understand how these combined factors influence the outcome beyond their individual contributions.The mathematical formulation of the interaction model is as follows:

$$\mu_i = \beta_0 + \beta_{\text{Course}} \times \text{Course}_i + \beta_{\text{Age}} \times \text{Age}_i + \beta_{\text{Temp}} \times \text{Temperature}_i$$

$$+ \ldots + \beta_{\text{PackTemp}} \times (\text{Pack}_i \times \text{Temperature}_i) + \ldots + u_{\text{Number}_i}$$

where β_0 is the overall intercept, $\beta_{\text{Course}}, \beta_{\text{Age}}, \ldots$ are coefficients for the main effects, $\beta_{\text{PackTemp}}, \ldots$ are coefficients for the interaction terms, and u_{Number_i} represents the random athlete-specific effects. This model allows for a deeper understanding of how the variables do not just contribute individually but also how their combinations affect the response.

Table 7: Gelman-Rubin statistic of interaction terms only(hid others to save space)

	beta0	betaCourseTemp	betaCourseWind	betaPackAge	betaPackTemp	betaPackWind	sigma	tau	tau_u
Point est.	1	1.00	1	1.00	1.00	1.00	1	1	1
Upper C.I.	1	1.01	1	1.01	1.01	1.01	1	1	1

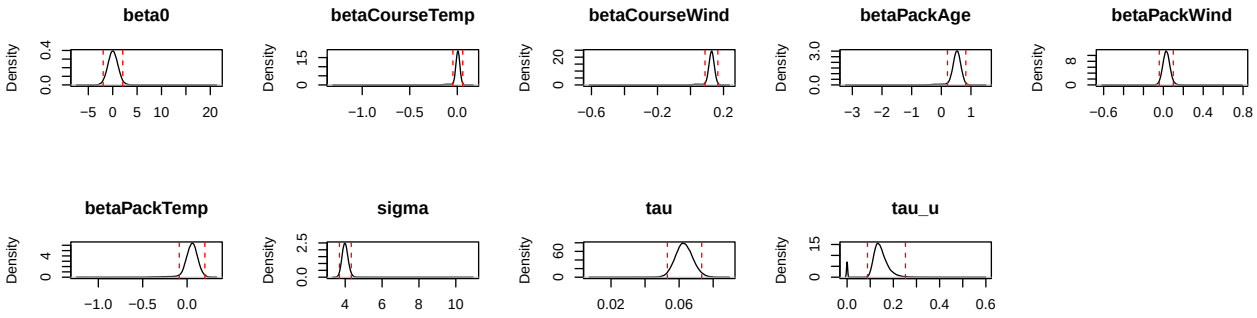
Table 8: Effective Sample Sizes of interaction terms only(hid others to save space)

beta0	betaCourseTemp	betaCourseWind	betaPackAge	betaPackTemp	betaPackWind	sigma	tau	tau_u
6103.92	1978.45	2220.42	1933.09	1427.52	2318.69	26466.23	27058.36	26641.08

The Bayesian analysis indicated satisfactory convergence for the model parameters, as evidenced by the Gelman-Rubin statistic values nearing 1, suggesting that the MCMC chains have stabilised. Same as the effective sample sizes, crucial for accurate parameter estimation, are robust across the board.

Table 9: Posterior distribution of interaction terms only(hid others to save space)

	Mean	SD	X2.5.	X25.	X50.	X75.	X97.5.
beta0	0.05	1.04	-1.94	-0.64	0.04	0.72	2.08
betaCourseTemp	0.00	0.04	-0.09	-0.01	0.01	0.02	0.05
betaCourseWind	0.12	0.03	0.05	0.12	0.13	0.14	0.16
betaPackAge	0.50	0.20	0.00	0.43	0.52	0.61	0.77
betaPackTemp	0.05	0.09	-0.16	0.01	0.05	0.10	0.18
betaPackWind	0.03	0.04	-0.04	0.01	0.03	0.05	0.10
sigma	4.00	0.19	3.69	3.88	3.99	4.10	4.34
tau	0.06	0.01	0.05	0.06	0.06	0.07	0.07
tau_u	0.14	0.04	0.00	0.12	0.14	0.16	0.22



The posterior density plots for the interaction terms are particularly telling. Such as 'betaCourseTemp' and 'betaPackWind', offer nuanced insights into how combinations of factors interplay to affect race outcomes. Although some interactions do not show a strong effect, with 95% credible intervals spanning zero, this does not diminish their importance. These interactions could still provide valuable information when considering specific conditions or subgroups within the data.

Furthermore, the interaction between pack level and age (betaPackAge) may indicate that the effect of an athlete's pack on their finishing time could vary significantly with age, indicating the distribution of different age groups varies in different Pack group, which meets the observation from the EDA phase. It is possibly due to differences in experience or stamina despite their age difference.

Similarly, interactions involving environmental factors like temperature and wind (betaCourseTemp, betaCourseWind, betaPackWind, betaPackTemp) suggest that the race conditions could amplify or diminish the effect of the course or pack level, which might be due to the unique geographical or environmental conditions of each course, and these conditions affect different pack groups differently as the athletes in the faster pack may be more experienced or have higher resilience with harsh condition.

Hierarchical model and mixture distributions

The Hierarchical model stands out as an advanced approach in statistical modeling, particularly suitable for complex datasets like this one, where data points are naturally clustered into different levels such as age groups, pack groups, and courses. By using this inherent structure, the Hierarchical model allows for the modeling of individual and group-level effects simultaneously, offering a rich, multilayered understanding of the data. This method not only captures the fixed effects of observed variables but also accommodates random effects to account for unobserved heterogeneity among participants and locations. It is mathematically represented as follows:

$$\text{Response}_i \sim \mathcal{N}(\mu_i, \tau)$$

where the mean μ_i is a linear combination of the predictors and random effects:

$$\mu_i = \beta_0 + \beta_{\text{Course}}[\text{Course}_i] + \beta_{\text{Pack}}[\text{Pack}_i] + \beta_{\text{Age}}[\text{Age}_i] + \beta_{\text{Temp}} \times \text{Temperature}_i + \beta_{\text{Wind}} \times \text{Windspeed}_i$$

$$+\beta_{\text{Dist}} \times \text{Distance}_i + \beta_{\text{Year}} \times \text{Year}_i + u_{\text{Number}_i}$$

In this model, β_0 is the overall intercept. The vectors of random effects, β_{Course}, β_{Pack}, and β_{Age}, capture the unique influence of each course, pack, and age group respectively. The fixed effects coefficients for the environmental and temporal variables are denoted by β_{Temp}, β_{Wind}, β_{Dist}, and β_{Year}. The term u_{Number_i} represents the random athlete-specific effects. The precision of the response distribution is given by τ, and τ_u is the precision for the random athlete-specific effects.

Due to the length limitation of the report, the numerical results from the Hierarchical model is shown in the Appendix 1, which provides me strong evidence for convergence of the model. The Gelman-Rubin statistic results, ideally close to 1 for all parameters, suggesting that the different MCMC chains are consistent with each other. This is supported by substantial effective sample sizes in the same table, indicating reliable estimates of the posterior distributions. The large effective sample sizes, particularly for the age group parameters, ensure that the posterior estimates are based on a sufficient amount of information from the MCMC chains.

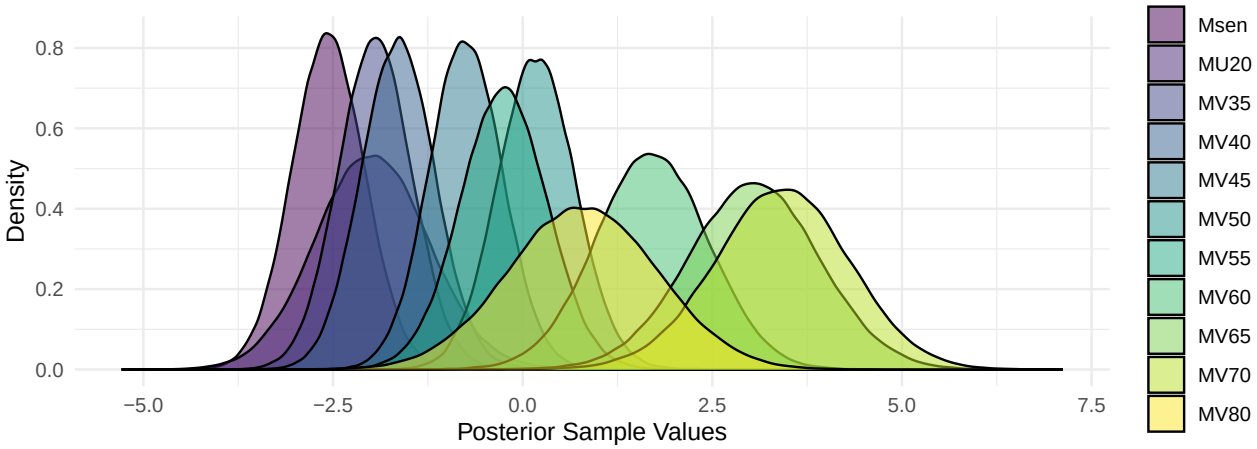


Figure 8: Density Plots for Different Age Groups

Figure 8 presents a compelling visual narrative of the relationship between age groups and race performance. The density plot for the youngest runners, labeled MU20, is notably skewed to the left, suggesting faster race times than the grand mean, indicative of their youthful advantage. In stark contrast, the plot for the senior group, Msen, sits just right of center, hinting at a slight delay in race times when controlling for other variables. The age groups from MV35 to MV65 exhibit an intriguing gradual shift to the right, with each successive category indicating a subtle but noticeable increase in race times. The MV55 group, hovering near the mean, serves as a transitional point before the more pronounced age-related deceleration becomes apparent in the MV60 and MV65 groups. Meanwhile, the MV70 and MV80 groups, with their broader and more dispersed distributions, speak to the heightened variability and generally slower race times within these elder cohorts. This pattern across the plots not only confirms the significant influence of age on racing prowess, as postulated in the EDA, but also captures the intrinsic diversity of performance within each age group, particularly the oldest competitors.

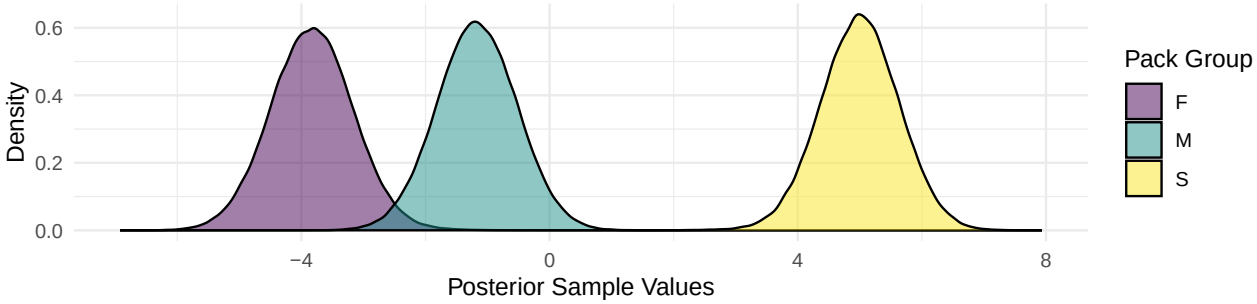


Figure 9: Density Plots for Different Pack Groups

In Figure 9, the density plots for the pack groups unfold a compelling story of race dynamics. The ‘F’ pack, representing the quick runners, shows a pronounced leftward skew, demonstrating its members’ propensity for speed and their tendency to clock in faster race times. This distinct separation from the rest reinforces

the pack's characteristic swiftness. Central to the graph, the 'M' pack's density hovers around the zero mark, portraying a median performance with times that are representative of the broader competitor field. At the far right, the 'S' pack's density plot, with its peak markedly right-shifted, is emblematic of more leisurely paces, encapsulating those runners whose race times are generally longer. The graphical stratification displayed here not only validates the pack categorisation employed by the race organisers but also vividly illustrates the substantial impact that pack alignment has on racing performance.

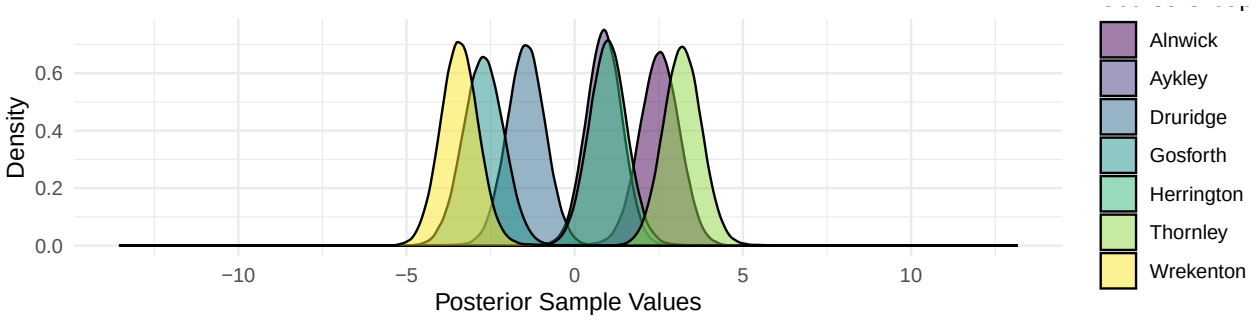


Figure 10: Density Plots for Different Course Groups

Figure 10's density plots reflect the distinct characteristics of each race course and their impact on performance. The plots for Alnwick and Aykley present near the zero mark, suggesting that times here are representative of the overall mean, indicating these courses may offer standard racing conditions. In notable contrast, the plot for Wrekenton is markedly shifted to the right, alluding to slower race times that may reflect the increased difficulty or complexity of this course. Meanwhile, courses like Druidge and Gosforth show wider distributions, hinting at a higher variability in race times which could point to a mix of conditions or a broader spectrum of participant performance levels. The visual juxtaposition of these courses demonstratesthe variable influence of different terrains and settings on racing outcomes.

Summary

Upon the analysis, the hierarchical Bayesian model emerged as the most adept at capturing the nuanced interplay of variables influencing race times. It adeptly accommodates the nested data structure, integrating both individual athlete variations and collective group dynamics.

My exploration reinforces the notion that age significantly delineates athletic performance, drawing a vivid line between the vitality of youth and the experience of age, with younger participants outpacing their older counterparts. Pack classification emerged as a defining axis of performance, stratifying athletes in a manner that resonates with their speed capabilities, and varies by the age group at the same time. Moreover, the character of the race course was illuminated as a significant factor influencing outcomes, with specific locations contributing to significant deviations in race times. Alongside these elements, environmental factors such as temperature and windspeed were found to subtly yet measurably affect race times, with cooler temperatures and milder windspeeds correlating with improved performance. Additionally, the increasing race distances underpin the role of endurance, as longer distances significantly lengthen race times, demonstrating the multifaceted nature of athletic achievement in cross-country running.

In conclusion, this report illuminates the intricate inter-dependencies of age, pack dynamics, and course characteristics, showing that each stroke contributes to the overall performance in the league. The findings, distilled from a Bayesian perspective, present a narrative where the individual athlete's abilities are intertwined with the group identity and environmental factors, all coalescing to shape the outcomes on the race track.

Appendix

1. Gelman-Rubin Statistic and Effective Sample Sizes of Hierarchical Model

Parameter	Gelman-Rubin	NA	Effective Sample Size
beta0	1	1.00	8464.12
betaAge[1]	1	1.00	70528.77
betaAge[10]	1	1.00	122148.35
betaAge[11]	1	1.00	101685.12
betaAge[2]	1	1.00	61751.20
betaAge[3]	1	1.00	59663.01
betaAge[4]	1	1.00	81043.11
betaAge[5]	1	1.00	67271.60
betaAge[6]	1	1.00	70795.25
betaAge[7]	1	1.00	101704.79
betaAge[8]	1	1.00	90713.35
betaAge[9]	1	1.00	101839.32
betaCourse[1]	1	1.00	10067.77
betaCourse[2]	1	1.00	6116.02
betaCourse[3]	1	1.00	12460.36
betaCourse[4]	1	1.00	74116.05
betaCourse[5]	1	1.00	1656.74
betaCourse[6]	1	1.00	7512.16
betaCourse[7]	1	1.00	6147.97
betaDist	1	1.01	287.89
betaPack[1]	1	1.00	21390.14
betaPack[2]	1	1.00	30855.57
betaPack[3]	1	1.00	36137.51
betaTemp	1	1.00	9415.03
betaWind	1	1.00	12639.48
betaYear	1	1.01	252.44
tau	1	1.00	57915.43
tau_u	1	1.00	39136.18

2. Posterior distribution Summary of Hierarchical model

	Mean	SD	X2.5.	X25.	X50.	X75.	X97.5.
beta0	0.00	1.04	-1.97	-0.69	-0.01	0.68	2.04
betaAge[1]	-0.75	0.49	-1.70	-1.08	-0.75	-0.42	0.21
betaAge[2]	-1.64	0.50	-2.59	-1.97	-1.64	-1.31	-0.66
betaAge[3]	-2.56	0.48	-3.50	-2.88	-2.56	-2.24	-1.61
betaAge[4]	-0.26	0.58	-1.39	-0.64	-0.25	0.13	0.86
betaAge[5]	-1.95	0.49	-2.90	-2.27	-1.95	-1.62	-0.99
betaAge[6]	0.19	0.52	-0.82	-0.16	0.18	0.53	1.20
betaAge[7]	3.04	0.86	1.33	2.46	3.04	3.62	4.73
betaAge[8]	-1.98	0.75	-3.42	-2.48	-1.99	-1.48	-0.49
betaAge[9]	1.70	0.74	0.24	1.20	1.70	2.20	3.15
betaAge[10]	0.79	0.98	-1.12	0.13	0.79	1.46	2.70
betaAge[11]	3.41	0.89	1.64	2.82	3.42	4.01	5.13
betaCourse[1]	-1.46	0.61	-2.61	-1.83	-1.45	-1.07	-0.33

	Mean	SD	X2.5.	X25.	X50.	X75.	X97.5.
betaCourse[2]	2.50	0.65	1.26	2.12	2.52	2.92	3.67
betaCourse[3]	0.86	0.55	-0.21	0.51	0.87	1.23	1.91
betaCourse[4]	-2.71	0.61	-3.91	-3.12	-2.71	-2.30	-1.51
betaCourse[5]	3.22	0.63	2.09	2.82	3.20	3.59	4.38
betaCourse[6]	-3.41	0.59	-4.51	-3.80	-3.42	-3.05	-2.26
betaCourse[7]	1.00	0.59	-0.11	0.61	0.99	1.37	2.15
betaDist	0.09	1.52	-1.73	-0.65	-0.04	0.60	1.94
betaPack[1]	4.99	0.64	3.74	4.57	5.00	5.42	6.23
betaPack[2]	-3.83	0.67	-5.13	-4.28	-3.83	-3.38	-2.51
betaPack[3]	-1.17	0.65	-2.44	-1.61	-1.17	-0.74	0.09
betaTemp	-0.14	0.09	-0.30	-0.19	-0.14	-0.09	0.02
betaWind	0.05	0.04	-0.02	0.03	0.05	0.07	0.11
betaYear	0.02	0.02	0.02	0.02	0.02	0.02	0.03
tau	0.07	0.01	0.06	0.07	0.07	0.08	0.08
tau_u	0.14	0.03	0.10	0.12	0.13	0.15	0.19