Analyzing Football Player Performance: A Bayesian Approach

— Project Report —

Advanced Bayesian Data Analysis

Bilal Tanvir Bhatti

Rafay Maqsood

March 17, 2024

TU Dortmund University

Contents

1 Introduction 2

2 Data 2

2.1 Data Filtering 3

3 Models 3

3.1 Poisson Regression 4

3.2 Model 1 5

3.3 Model 2 6

3.4 Model 3 6

3.5 Model 4 7

3.6 Model 5 8

4 Convergence Diagnostics 8

5 Limitations and Potential improvements 9

6 Conclusion 10

7 Reflection on own Learnings 10

A Additional figures 11

1 Introduction

In the dynamic world of football, understanding the factors affecting player performance is vital for teams striving to succeed. From exciting goals to the subtle effects of age, many elements influence how well a player performs on the field. In our report titled "Examining Football Player Performance: Using Bayesian Analysis," we delve into sports analytics to uncover the intricate link between age and various performance metrics, particularly focusing on goals scored.

The motivation behind this study stems from the growing reliance on data-driven decision-making in football clubs worldwide. With the sport constantly evolving, teams seek to grasp the critical factors shaping player performance to gain a competitive edge. Age, a fundamental aspect of a player's career, emerges as a significant factor warranting thorough investigation. By elucidating how age impacts performance measures like goals scored, clubs can make informed choices regarding player recruitment, development strategies, and tactical planning.

Given these challenges, we propose employing Bayesian methods as a solution. Bayesian techniques offer a robust statistical approach capable of handling uncertainty and integrating prior knowledge into the analysis. Through Bayesian modeling, we can create adaptable models that not only quantify the influence of age on performance metrics but also provide credible intervals to assess the uncertainty surrounding these estimates. Furthermore, Bayesian methods allow us to effectively consider confounding variables, enhancing the accuracy of our understanding of the age-performance relationship.

In Section 2, we give insights about data filtering and data sources. Section 3, explains Poisson Regression and delves into the insights about the models, and their interpretations. Then, in Section 4, we look at how well the models converged by examining trace plots and checking Rhat values. Furthermore, in Section 5, we define the limitations of the project and talk about how future experiments could be improved. In Section **??** we discuss the effects on goal-scoring, informing strategic decisions in player development and team management. Lastly, in the section 7, we provide the self-learnings from the project.

2 Data

The original dataset comprises multiple CSV files containing information on various aspects of football competitions, games, clubs, players, and appearances spanning from 2012 to 2023. This extensive dataset encompasses over 60,000 games across numerous seasons in major competitions, data from over 400 clubs participating in these competitions, details on more than 30,000 players affiliated with these clubs, historical records of over 400,000 player market valuations, and over 1,200,000 player appearance records from all games, among other data points. The dataset exhibits high data integrity, with no missing values, indicating its overall quality. The dataset was sourced from Kaggle and is readily accessible on `https://www.kaggle.com/datasets/davidcariboo/player-scores/data`.

For our project, we focused on extracting data specifically from the English Premier League (EPL), spanning from 2012 to 2023. This subset of data includes information on all players who participated in the EPL during this period, resulting in $n = 1,170$ observations including the variables *name* (character), *age* (numeric), *goals* (numeric), *cum_goals* (numeric) [total number of goals], *player_id* (numeric), *position* (character) [*Midfielder, Defender, Goalkeeper, Attack*]. Our primary objective was to extract data related to the number of goals scored by each player in each respective year of their tenure in the league.

2.1 Data Filtering

To begin our data analysis process, we first load the dataset by importing five CSV files into the R environment such as competitions, games, game_lineups, players, and game_events. These files contain comprehensive information regarding football competitions, games, players, and game events.

Once the data is loaded, we implement a series of filtering steps to refine our dataset. Initially, we filter competitions to retain only those identified by the sub_type ($GB1$) denoting the English Premier League. This step involves extracting the competition_id corresponding to ($GB1$) sub_type for further use in filtering.

Subsequently, we filter games based on the selected competition_id, narrowing down the dataset to include only games relevant to the English Premier League. Further refinement involves filtering game events to focus solely on events pertinent to goal scoring, ensuring our analysis centers on crucial aspects of player performance. Additionally, we filter player lineups to extract unique player IDs associated with the selected games, eliminating any duplicate entries. Finally, player details are filtered to retrieve comprehensive information based on the unique player_ids selected, including player_id, name, country_of_citizenship, date_of_birth, position, foot preference, and height_in_cm.

Through these meticulous filtering steps, we ensure the dataset is cleaned and tailored specifically for our analysis. By focusing on the English Premier League, goal-scoring events, and relevant player details, we enhance the accuracy and relevance of our subsequent analysis, thereby addressing the research question more effectively.

3 Models

This section introduces the Poisson Regression followed by the statistical models used to analyze football player performance. Each model is explained simply, along with its interpretation. By examining different approaches, we gain insights into how player attributes affect on-field performance. This section serves as a guide to understanding the relationship between player characteristics and success in football, offering valuable insights for sports analytics and future research.

3.1 Poisson Regression

Poisson regression is a statistical method utilized for modeling count data, particularly when the outcome variable represents the number of occurrences of an event within a fixed interval or region. It is particularly useful when the number of trials n in a binomial distribution is unknown or excessively large.

In a Poisson regression model, the shape of the distribution is characterized by a single parameter λ(lambda), representing the expected value or rate of occurrence of the event of interest. This parameter is crucial for defining the likelihood of observing a specific count value y. The Poisson likelihood is specified as:

$$y \sim \text{Poisson}(\lambda)$$

The parameter λ denotes the expected value of the outcome y.

To construct a Generalized Linear Model (GLM) with a Poisson likelihood, a link function is required to relate the linear predictor to the expected value λ. The conventional choice for the link function in Poisson regression is the log link, which ensures that the expected value λ remains positive. Thus, the GLM is formulated as:

$$\begin{aligned} y &\sim \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \end{aligned}$$

Here, y_i, represents the count outcome for observation, i, $x_i 1, x_i 2, ..., x_i k$, denote the predictor variables, $\beta_0, \beta_1, ..., \beta_k$ are the corresponding coefficients, and λ_i is the expected value of y_i determined by the linear combination of predictors. The parameter λ can be interpreted as either the expected value or the rate of occurrence, allowing for flexibility in modeling scenarios where exposure varies across cases. This flexibility enables the incorporation of exposure as a predictor variable in the model, facilitating the analysis of count data aggregated over different periods or exposures. McElreath 2016, p. 311.

In the given model, additional regularizing priors with a mean of 0 and standard deviation of 3 (Normal(0, 3)) are applied to the regression coefficients $\beta_0, \beta_1, ..., \beta_k$ to account for the limited sample size and to prevent overfitting. These priors represent weak prior beliefs about the parameter values, acknowledging the need for more data to estimate them accurately.

$$\log(\lambda) \sim \text{Normal}(0, 3)$$

In the analysis of our football player performance data, we employed a Poisson regression model to investigate the relationship between player age and the number of goals scored in matches. The Poisson regression model was chosen due to its suitability for analyzing count data, such as the number of goals scored in football matches.

3.2 Model 1

In our first Poisson regression model, we didn't consider individual differences among players. This means we assumed a constant intercept for all players, ignoring any variation in goal-scoring performance based on individual player traits.

The model simply looks at how the number of goals scored (goals) relates to the predictor variable age.

$$\text{goals} \sim \text{age}$$

Understanding the outcomes of the model in Table 1, the estimated intercept is positioned at 1.34, theoretically denoting the anticipated logarithmic count of goals when the player's age is zero. In practical terms, this implies that if an individual has engaged in football for a year, their intercept would be 1.34. However, this interpretation lacks practical significance. The coefficient for age is close to zero (-0.00), suggesting a negligible relationship between age and the number of goals scored. Despite seemingly precise estimates with small standard errors, the wide credible interval (-0.01 to 0.01) indicates significant uncertainty about the actual association between age and goal-scoring performance.

Table 1: Model 1 - Linear regression parameter estimates.

Covariate	**Estimate**	**95% Interval)**	**Est.Error**	**Rhat**
intercept	1.35	[1.13, 1.57]	0.11	1
age	-0.00	[-0.01, 0.01]	0.00	1

Our model suggests that age may not significantly impact the number of goals scored by football players in our dataset. However, it's important to interpret these findings cautiously due to the uncertainty surrounding the estimates. The lack of a clear relationship between age and goal-scoring performance highlights the need for further investigation. In this model, the relationship between age and goals isn't clear because it treats all players the same, regardless of their position like goalkeeper, defender, or midfielder. This means it doesn't consider each player's unique characteristics, especially the position they play, which is crucial in scoring goals in football.

Future research endeavors should prioritize the inclusion of random effects in the modeling framework to account for individual player differences effectively.

3.3 Model 2

The second model accounts for random effects, meaning it considers that there are inherent differences between players that can't be captured by the model's variables.

This formula indicates that the number of goals scored is influenced by age while considering individual player differences captured by the random intercept specified for each player.

$$\text{goals} \sim \text{age} + (1|\text{player\_id})$$

In Table 2, The estimated intercept in our model tells us the average number of goals a player would score if they were a football player. We found this to be around 0.72, but with a bit of uncertainty - the standard error is 0.16. The 95% credible interval for the intercept, which is [0.40, 1.05], means we're pretty confident (95% confident) that the real value of the intercept falls somewhere in this range. A smaller standard error means we're more certain about our estimate.

Table 2: Model 2 - Linear regression parameter estimates.

Covariate	**Estimate**	**95% Interval)**	**Est.Error**	**Rhat**
intercept	0.72	[0.40, 1.05]	0.16	1
age	0.01	[-0.01, 0.02]	0.01	1

Moving on to age, our estimated coefficient is 0.01. This suggests that for every one-year increase in age, we'd expect a player to score about $e^{0.01} \approx 1.01$ more goals. The standard error for this coefficient is also 0.01, so we're quite precise in our estimate. The 95% credible interval for the age coefficient is [-0.01, 0.02], meaning there's a 95% chance that the real effect of age on goal scoring is somewhere within this range. The interval is narrow, indicating high confidence in our estimate. A potential scale reduction factor (Rhat) close to 1 shows good convergence, meaning our model's estimates are stable.

After accounting for each player's individual characteristics, we notice a minor influence of age on the number of goals scored. However, this effect isn't very significant because our dataset includes players from positions like goalkeeper, defender, and midfielder, who typically don't score many goals. Consequently, this has notably reduced the coefficient for goals.

3.4 Model 3

The third model incorporates random effects by introducing individual intercepts for each player, specifically focusing on those in attacking positions. The formula suggests that the goal-scoring performance is not only affected by age but also takes into account unique player variations, as represented by the random intercepts assigned to each attacking player.

$$\text{goals} \sim \text{age} + (1|\text{player\_id})$$

The model results in Table 3 suggest that, on average, players in attacking positions are estimated to score $e^{0.73} \approx 2.07$ goals when they are at the age of 0 (interpreted as the intercept). Additionally, for every one-unit increase in age, there is a predicted increase of 0.03 goals scored. The 95% interval for the intercept ranges from 0.31 to 1.15, indicating the uncertainty around this estimate, while the age coefficient has a 95% interval from 0.01 to 0.04. The estimated standard error for the intercept is 0.21, and for age, it's 0.01. The Rhat statistic of 1 suggests convergence of the model.

Table 3: Model 3 - Linear regression parameter estimates.

Covariate	**Estimate**	**95% Interval)**	**Est.Error**	**Rhat**
intercept	0.73	[0.31, 1.15]	0.21	1
age	0.03	[0.01, 0.04]	0.01	1

3.5 Model 4

This model recognizes that players can have different starting points for their goal-scoring rates and that the impact of age on their goal-scoring rates can vary although player position is not included as a fixed effect in this model.

$$\text{goals} \sim \text{age} + (1 + \text{age}|\text{player\_id})$$

As shown in Table 4, The model finds a negative correlation between intercepts and age, which means that players with a higher starting goal-scoring rate tend to see a steeper decline in their goal-scoring rate as they age.

The model estimates that the average goal-scoring rate for a player at the start of their career is $e^{0.96} \approx 2.60$ goals. Additionally, the coefficient for age is close to zero, indicating that, on average, age doesn't significantly affect goal scoring. This conclusion holds even after considering the individual differences among players that the model captures.

Table 4: Model 4 - Linear regression parameter estimates.

Covariate	**Estimate**	**95% Interval)**	**Est.Error**	**Rhat**
intercept	0.97	[0.51, 1.39]	0.22	1
age	-0.00	[-0.02, 0.02]	0.01	1

Furthermore, the standard errors for the intercept and age are 0.22 and 0.01, respectively. The 95% credible intervals for the intercept and age are [0.51, 1.39] and [-0.02, 0.02], respectively. These intervals indicate the range within which the true values of the intercept and age coefficients are likely to lie with 95% probability. The convergence and sampling diagnostics indicate that the model has converged and the estimates are reliable.

3.6 Model 5

This model considers that players start with different goal-scoring rates and that the impact of age on their goal-scoring rates varies. It compares players in different positions to attackers, recognizing that some players start with higher goal-scoring rates. It found that players with higher initial goal-scoring rates tend to see a faster decline in their goal-scoring rates as they get older.

$$\text{goals} \sim \text{age} + \text{position} + (1 + \text{age}|\text{player\_id})$$

According to the values in Table 5, The model estimates that attackers are expected to score an average of 1.00 goals at the beginning of their football career, assuming all other factors remain constant. Additionally, the coefficient for age is 0.02, indicating a positive impact on goal scoring as players get older. The credible interval for this coefficient is between 0.00 and 0.03, suggesting that this effect is statistically supported.

Table 5: Model 5 - Linear regression parameter estimates.

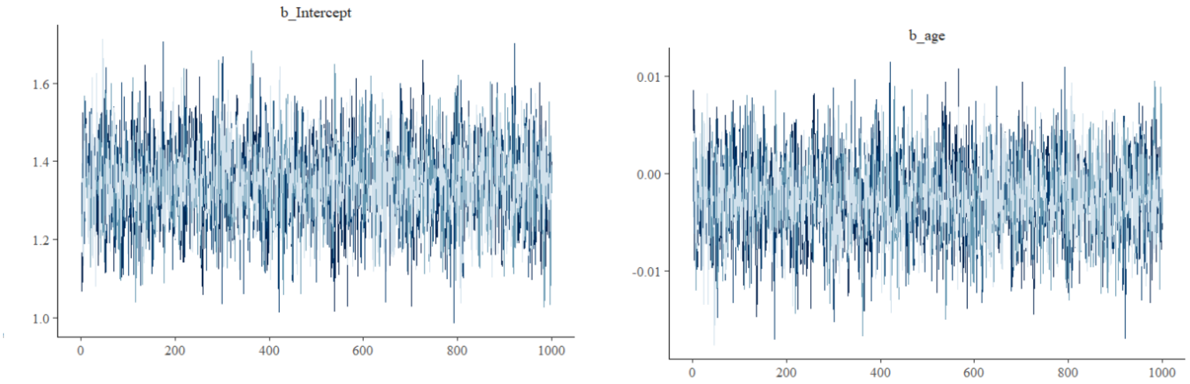
Covariate	**Estimate**	**95% Interval)**	**Est.Error**	**Rhat**
intercept	1.01	[0.61, 1.41]	0.20	1.01
age	0.02	[0.00, 0.03]	0.01	1.01
Defender	-1.02	[-1.18, -0.85]	0.08	1.00
Goalkeeper	-1.63	[-2.32, -0.99]	0.34	1.00
Midfielder	-0.46	[-0.62, -0.30]	0.08	1.00

As age remains constant, the model estimates that Defenders are expected to score approximately $e^{-1.02} \approx 0.36$ times the goals scored by Attackers, holding other factors constant. Similarly, Midfielders are estimated to score approximately $e^{-0.46} \approx 0.63$ times the goals scored by Attackers, and Goalkeepers are estimated to score approximately $e^{-1.67} \approx 0.18$ times the goals scored by Attackers. The diagnostic tests for convergence and sampling suggest that the model has converged, and the estimates are trustworthy as the $\hat{R}$ value is 1.

4 Convergence Diagnostics

Convergence diagnostics check if Markov Chain Monte Carlo (MCMC) simulations have properly checked the posterior distribution. The trace plots display how parameter values change over simulation steps. A good trace plot should look random and evenly spread out after an initial warm-up phase, indicating the MCMC chain has reached its target distribution. Despite the helpful insights from trace plots, it's advisable to also rely on quantitative measures like Rhat, a common metric for confirming convergence. These tools are important for trustworthy Bayesian analysis.

The trace plot illustrates how a parameter (such as b_intercept and b_age) changes throughout the iterations of an MCMC simulation. The absence of trends or patterns in the trace plot, with the parameter values covering the full range, indicates favorable convergence.



(a) Trace Plot - Intercept

(b) Trace Plot - Age

Figure 1: Trace Plots for Model 1 - Goals $\sim Age$

Although trace plots offer valuable insights, it's advisable to complement visual examination with quantitative convergence diagnostics. The potential scale reduction factor (Rhat) is a widely used metric for evaluating convergence. In this case, the Rhat value of 1 for the model indicates satisfactory convergence, as demonstrated in Table 1.

The convergence of all our models underscores the reliability of our findings. As exemplified by the first model, which we have explained thoroughly in the main body of this report, all subsequent models exhibit similar convergence patterns. Detailed trace plots illustrating this convergence for each model can be found in Figures 2, 3, 4, 5 in the appendix. These trace plots provide compelling evidence of the stability and consistency of our modeling approach across different iterations, affirming the validity of our analyses.

5 Limitations and Potential improvements

In our analysis, we've primarily focused on age and player position as predictors for goal scoring, acknowledging their significance in understanding player performance. However, it's important to recognize that our study's scope is limited, particularly in terms of sample size and the exclusion of factors like team tactics and individual skills. While our findings suggest correlations between age, player position, and goal scoring, it's essential to interpret them with caution, as causality cannot be definitively established. Moreover, the generalizability of our models may vary across different soccer leagues and playing environments. To enhance the robustness of our analysis, future research could consider incorporating additional predictor variables and expanding the dataset to improve accuracy and applicability.

6 Conclusion

In conclusion, our project utilized a Bayesian approach to analyze football player performance, focusing on the influence of age on key metrics such as goals scored. By leveraging data from Transfermarkt via Kaggle, we meticulously cleaned and pre-processed the dataset, calculating player ages at the time of goal scoring and aggregating goal data to create a comprehensive data frame. Employing Poisson regression models with varying complexities, including random effects for individual player intercepts and slopes of age, we uncovered valuable insights into how age impacts goal-scoring performance. Our findings suggest that while there may be minimal overall change in goals scored with increasing age, there are nuanced differences among player positions, with attackers exhibiting a positive association between age and goals scored. Furthermore, our Bayesian framework not only enhanced our ability to draw meaningful conclusions but also accounted for uncertainties in parameter estimates, offering a more robust analysis. These insights hold significance for player development, team management, and decision-making processes within the realm of football analytics.

7 Reflection on own Learnings

The report provides a comprehensive understanding of the use of Bayesian methods and Poisson regression models in analyzing count data, such as the number of goals scored in football matches. It highlights the importance of considering individual differences among players when analyzing their performance. The study underscores the need for further research to incorporate additional predictor variables and increase sample size to enhance the accuracy and applicability of the models. It also emphasizes the importance of convergence diagnostics in Bayesian analysis to ensure the reliability of the model estimates.

Appendix

A Additional figures

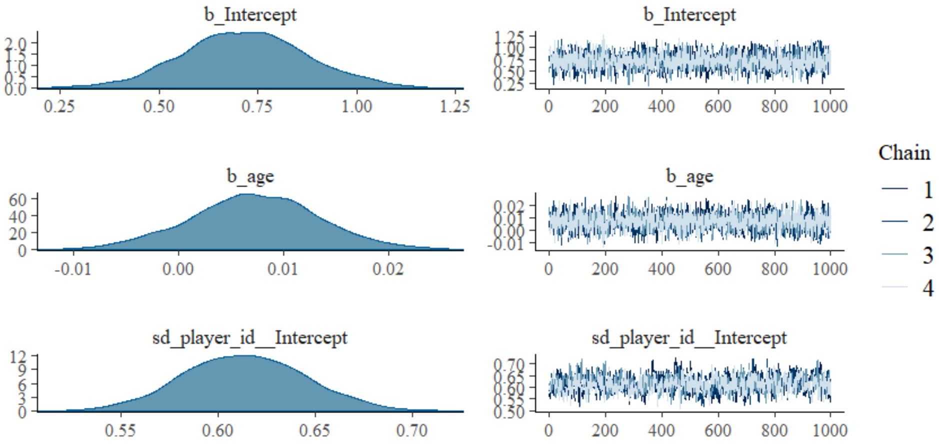


Figure 2: Trace Plots for Model 2 - goals ∼ age + (1|player_id)

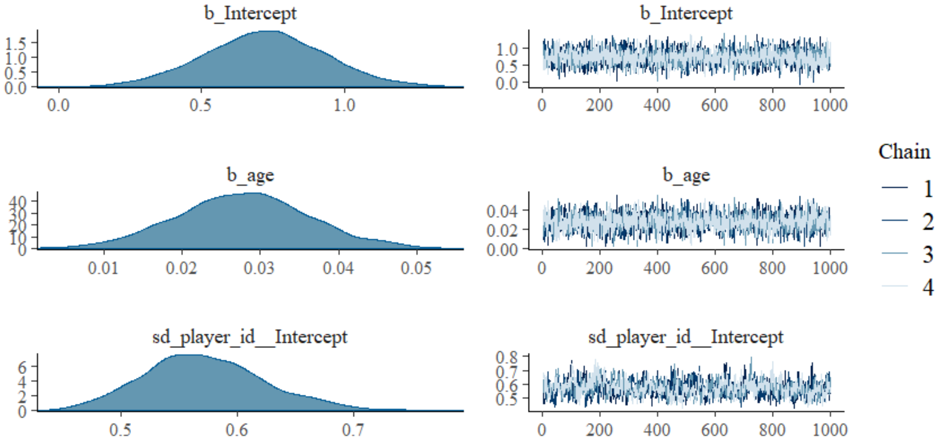


Figure 3: Trace Plots for Model 3: Players(Attackers) - goals ∼ age + (1|player_id)

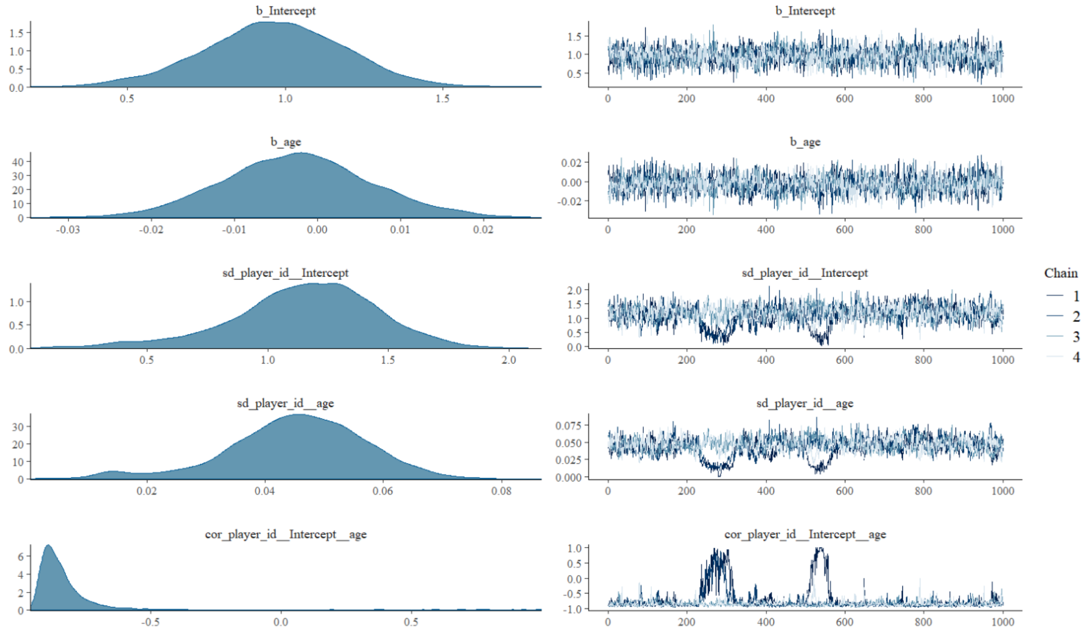


Figure 4: Trace Plots for Model 4: goals $\sim$ age + (1 + age|player_id)

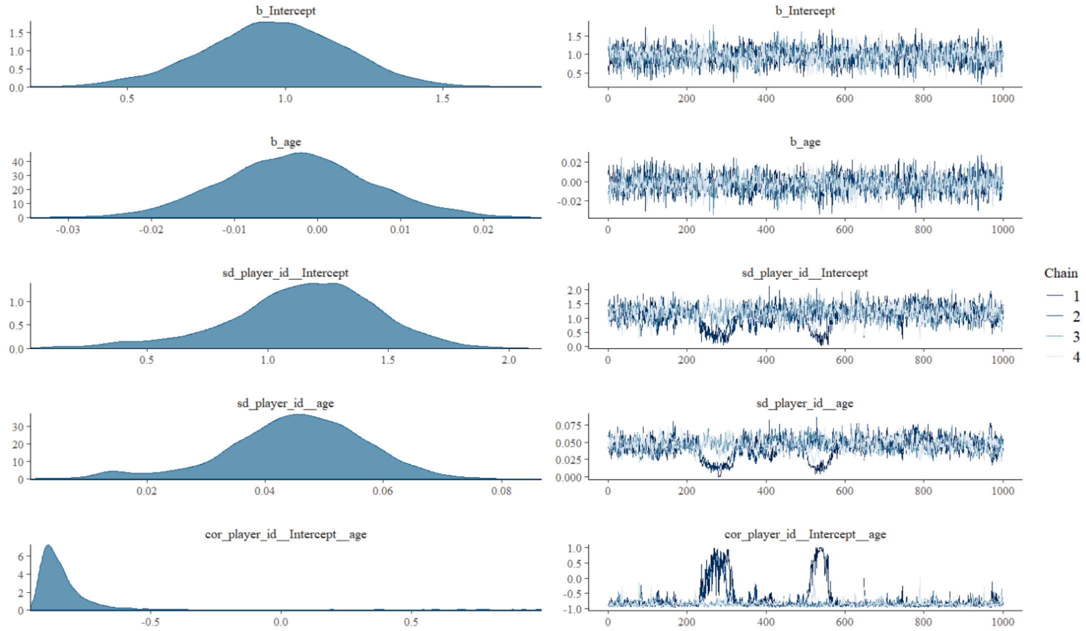


Figure 5: Trace Plots for Model 5: goals $\sim$ age + position + (1 + age|player_id)

References

McElreath, Richard (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press/Taylor & Francis Group. ISBN: 9781482253443.