Quantifying the effects of individual player on goal scoring through a Bayesian approach

Barba Paolo

July 3, 2023

Introduction (1)

Nowadays, football clubs have started to analyze advanced metric for player performance analysis that can lead in

- player scouting
- decision making of player's contract.

Introduction (2)

While comparing teams or player, the expected goal metric provides a good statistic in the tactical match analysis.

The xG model is a probabilistic model that assign score between 0 and 1 from any observed shot in a match. The model has development using event-level football data from StatsBomb's data.

Usually, xG models do not account for the players who take the shots, and this assumptions does not seem suitable, since the player's skill could influence the success of shot- conversion.

Aim

The aim of the analysis is to understand:

- The variables that can influence the shots result
- If there is a hierarchical structure behind the shots
- Estimate the player effect on goal scoring

Before going any further, it is needed to describe the variables that we are going to use for conducting this analysis

Dataset

The data are collected by tracking players over the course of football matchs and recording their actions such shots, passes or others. For the xG model the data were taken from the shot event. These data, collected from StatsBomb, are spatially manipulated in order to extrapolate relevant informations for the score prediction.

Data extrapolation

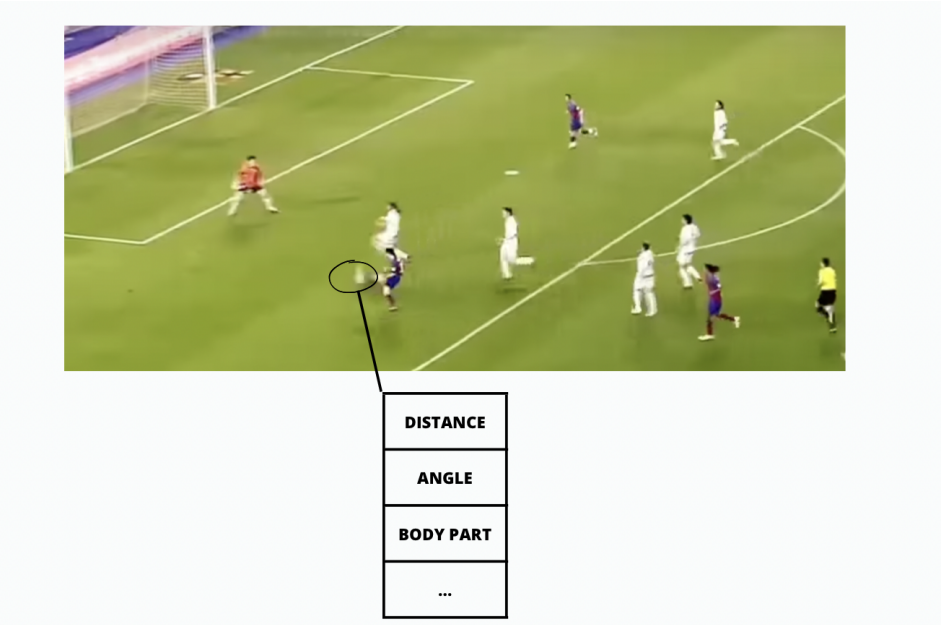


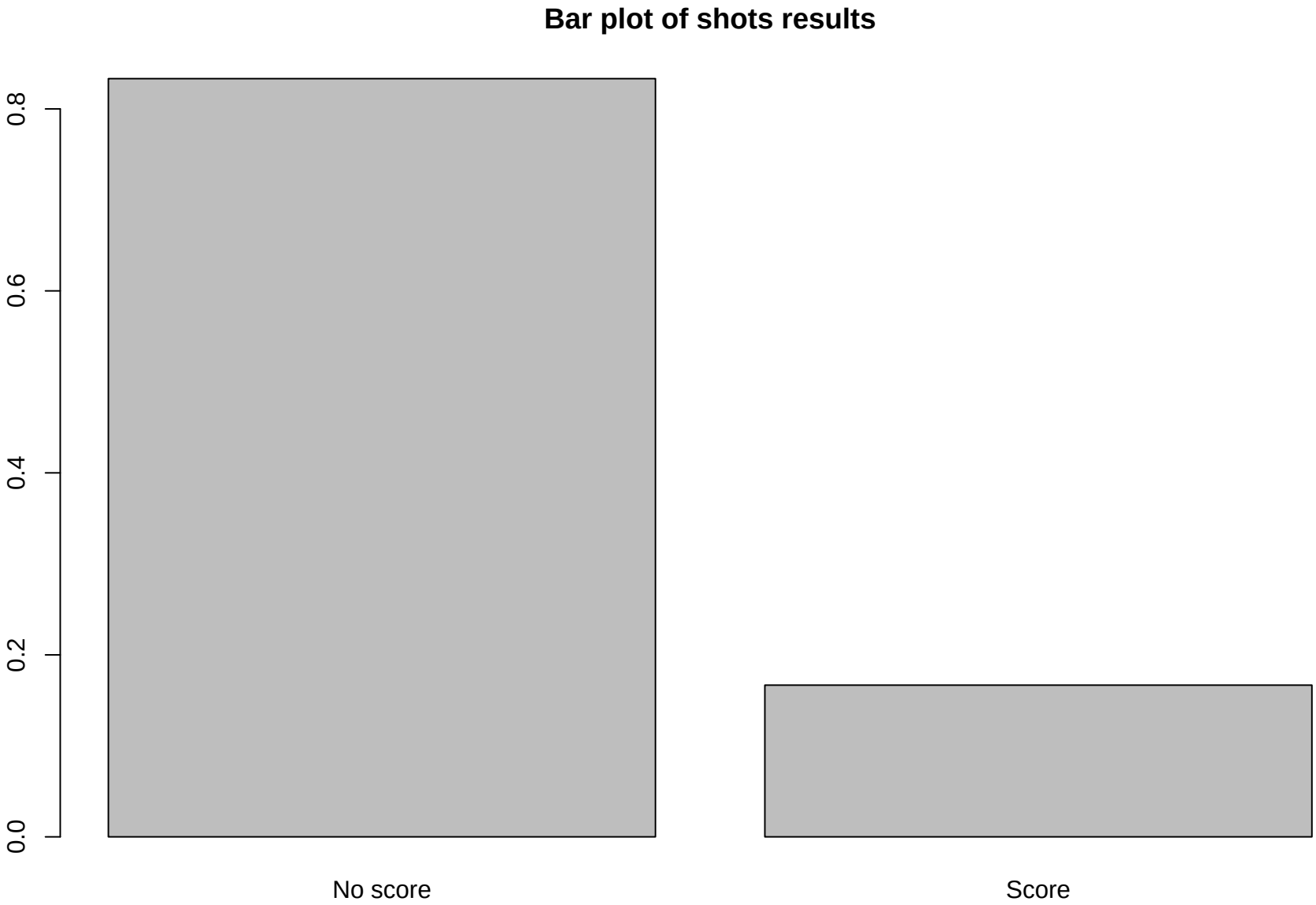
Figure 1: Messi's Goal against Albacete

Dataset description

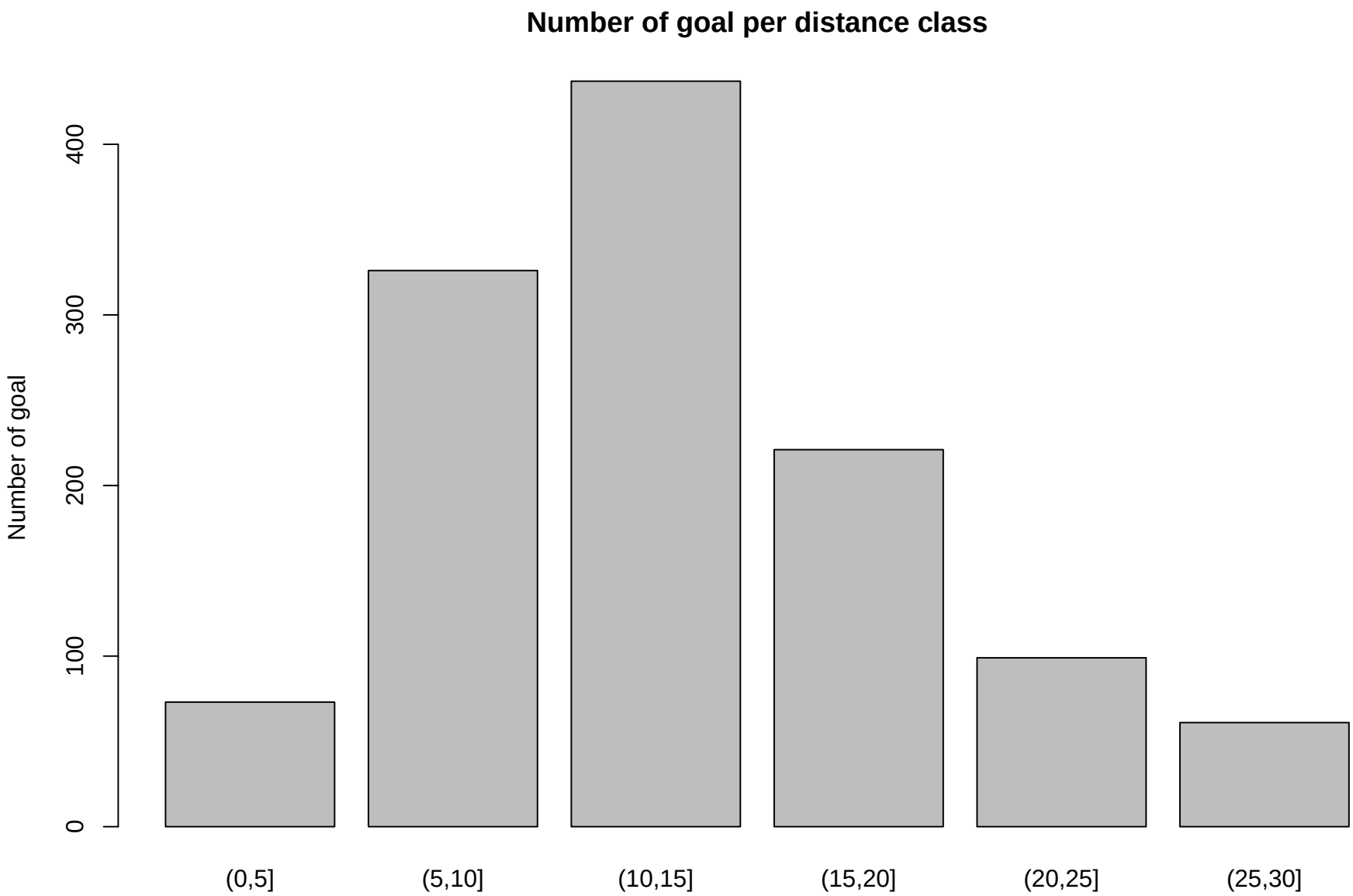
The dataset is composed by 13 variables and 7437 observations that reppresent shots of different 40 player of Barcelona F.C.

- Three continuous predictors: shot-taker's distance to goal, angle of shot and time.
- Three binary predictors: tecnique, first touch, inside 18 metres
- Two multi-class predictors: body part and preferred type

Explanatory Analysis



Explanatory Analysis (2)



Hierarchical assumptions

A reasonable assumption is that player has a unique type of shooting. Therefore, given a player, the shots are inerehntly correlate to each other and behave indipendently from the shot of another players. This introduces within player correlations. As a result, every shot is essentially grouped under a player. Therefore, models need to be fed information about the hierarchical structure, otherwise it may lead to biased inference.

Example of the hierarchical structure

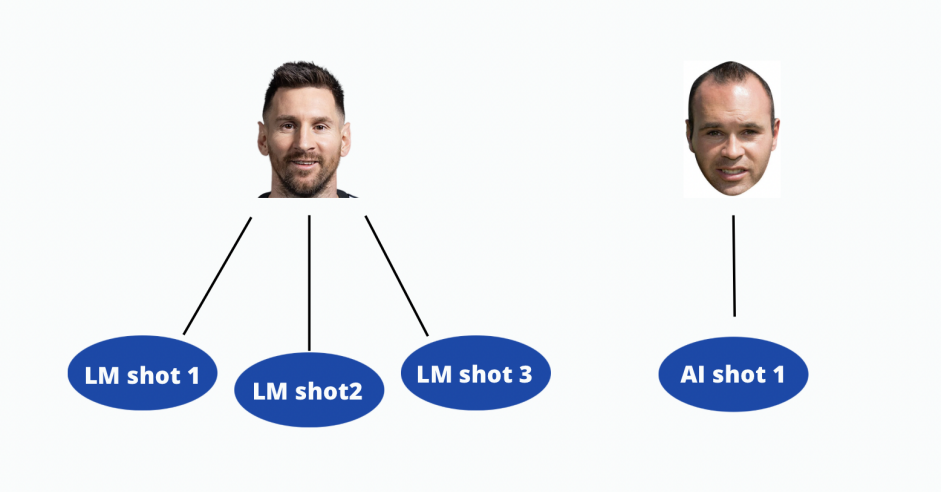


Figure 2: Example of Hierarchical structure

Model specification

For targeting the response variable, two models has been choosen.

- GLM (Generalize linear model): do not fed information about the hierarchical structure
- GLMM (Generalize linar mixed model): do fed information about the hierarchical structure

Generalized linear model specifications

$$g(\pi_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon_i$$

where:

- g is the logit link function
- π_i is the odds for a goal $\frac{Pr(Y=1)}{Pr(Y=0)}$.

Generalized linear mixed model specifications

In order to fed the information about the players structure a Linear mixed effect model only with a random intercept has been choosen.

Let us define the equation of the model and some basic notion:

$$g(\pi_{i,j}) = \beta_0 + \beta_1 x_{i,j,1} + \beta_2 x_{i,j,2} \cdots \beta_p x_{i,j,k} + \delta_j + \epsilon_{i,j}$$

- $j = 1, 2, \ldots 42$ number of players g
- $i = 1 \ldots n_g$ Number of shots done by the $g - th$ player
- $\pi_{i,j} =$ odds ratio of score from the $i - th$ shot and the $j - th$ player

Dimensionality of the parameters

Let us define the dimensionality:

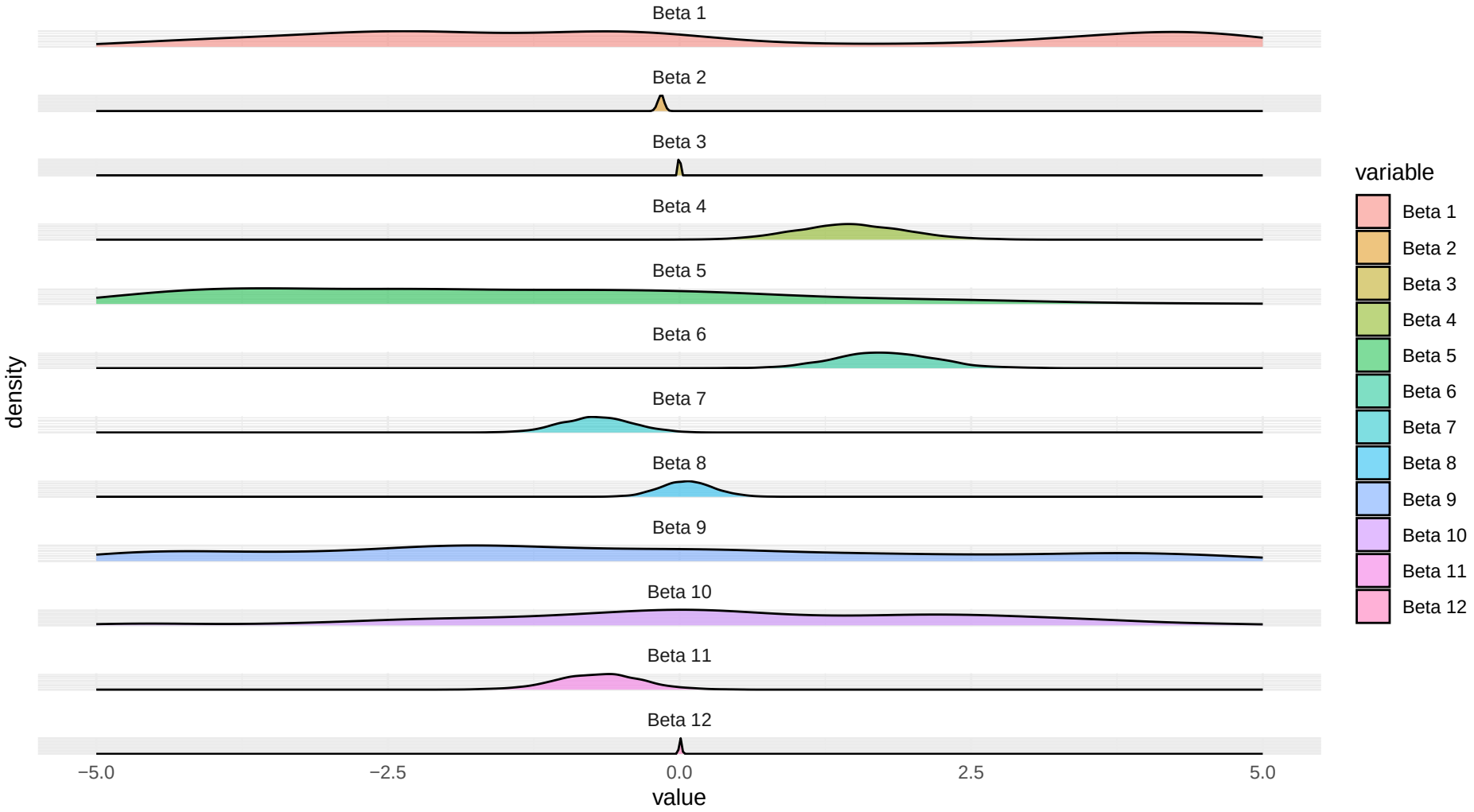
- $X_{g,i} \in (k \times 1)$ where k is the number of predictors.
- $\underline{\beta} \in (k \times 1)$ can be considered as a vector of fixed effect since it does not depend on the player.
- δj represent the random effect linked to the g-th player.
- The vector $\underline{\epsilon_g}$ is the vector of error terms.

Sampling model assumptions

- The $\epsilon_g \perp b_g \forall g = 1 \cdots 42$.
- The b_g are normally distribuited with their own variance.
- The ϵ_g are normally distribuited.

Conditionally on the player effect, the probability of scoring can be considered independent since we are deleting their common factors.

Poterior Analysis Distribution of the beta coefficient



Betas coefficient

Predictors	Mean
β_1: Intercept	-10.933
β_2: shot_distance	-0.1548
β_3: shot_angle	-0.0010
β_4: bodypartLeft Foot	1.481
β_5: bodypartOther Bodypart	-24.49
β_6: bodypartRight Foot	1.7626
β_7: techniqueVolley	-0.712
β_8: first_touchTrue	0.077
β_9: prefered_typeLeft Foot	7.394
β_{10}: prefered_typeRight Foot	7.87
β_{11}: inside_18True	-0.639
β_{12}: time	0.0081

Player impact on goal

The random effetct associated to each player can be seen as a measure of player inpact on goal, a positive random effect for player lead in a more likelihood that a shot will end up in a score.

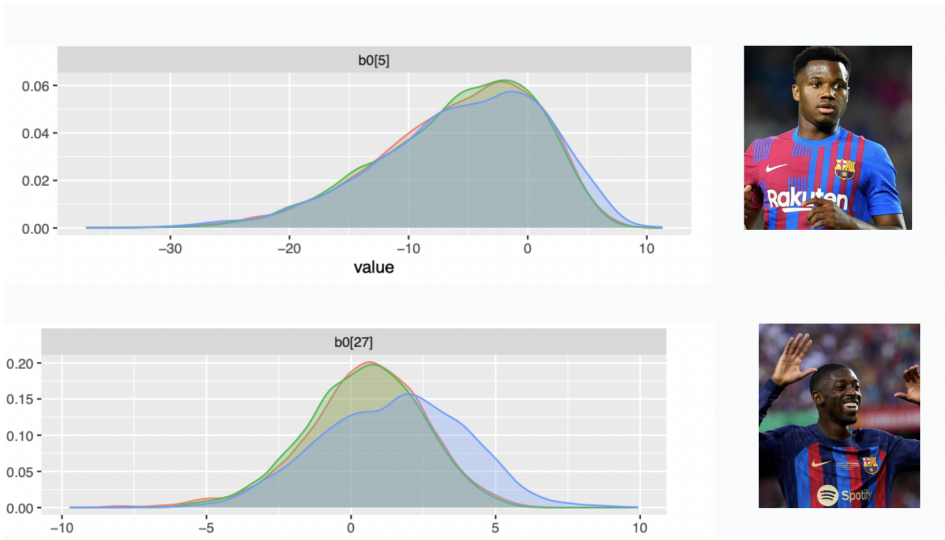


Figure 3: posteririor distribution of random effet for Ansu Fati and Dembelè

Model comparison via DIC

We are going to compare the two model via DIC (Deviance Information Criteria)

```
## [1] "DIC GLMM: 841.508199324143"
```

```
## [1] "DIC GLM: 833.717541519603"
```

Since the DIC the GLM is lower than the DIC of the GLMM , it is possible to claim that data do not show enough evidence in order to reject the null hypothesis.

So we can conclude that the assumptions of the hierarchical structure does not fit with this data

DIAGNOSTIC

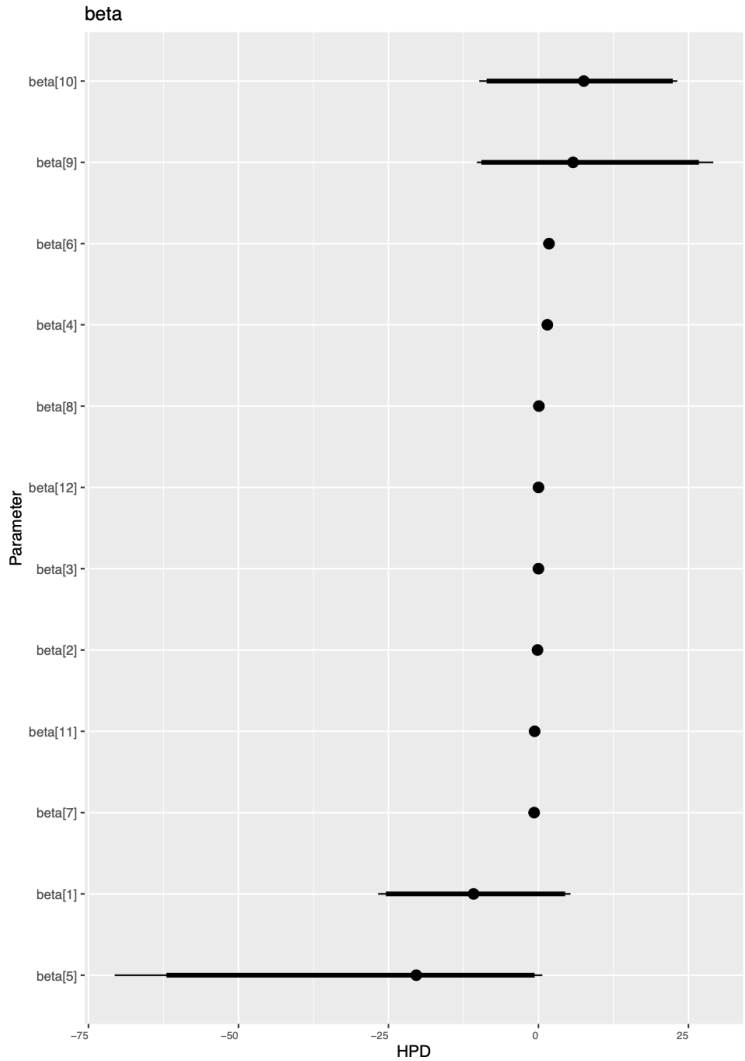


Figure 4: HPD Interval

DIAGNOSTIC (2)

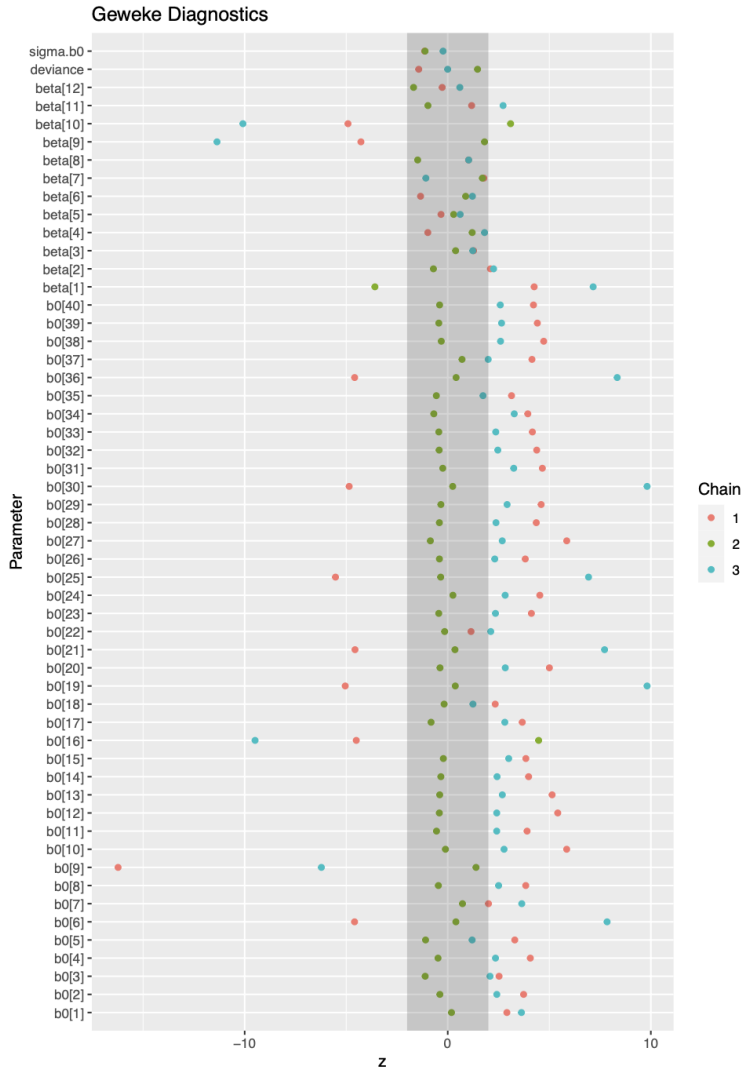


Figure 5: GEWEKE

References

- Sigrid Olthof. **Estimated Player Impact" (EPI): Quantifying the effects of individual players on football (soccer) actions using hierarchical statistical models**.
- Wes Swager, (2021). **womens soccer expected goals model for Milwaukee Rampage FC**.