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

Barba Paolo
Sapienza, Univeristy of Rome

Abstract

The goal is to estimate the effect of Barcelona players on goal scoring using a Generalized Linear Mixed Model. Model comparison will be performed via Deviance Information Criterion (DIC) using Markov Chain Monte Carlo (MCMC) algorithms, such as the Gibbs Sampling method, to obtain the posterior distribution of the beta coefficients

1. Introduction

Football clubs nowadays utilize advanced metrics for player performance analysis, which aids in player scouting and decision-making during player market transfers. When comparing teams or players, the expected goal (xG) metric serves as a valuable statistic in tactical match analysis. The xG model is a probabilistic model that assigns a score between 0 and 1 to any observed shot in a match. The xG model's development involves using event-level football data from StatsBomb's dataset. This data is collected by tracking players throughout a football match and recording their actions, such as shots, passes, and other events. For the xG model, only shot events are used as data points. Typically, xG models do not take into account the players who take the shots. However, this assumption might not be suitable, as a player's skill can influence the success of shot conversion. The dataset contains repeated measures for multiple players, as they can make multiple shots during a match. As a result, the same player may appear multiple times in the dataset. The events in the data are nested under a player hierarchy, illustrating the relationship between players and the events they generate.

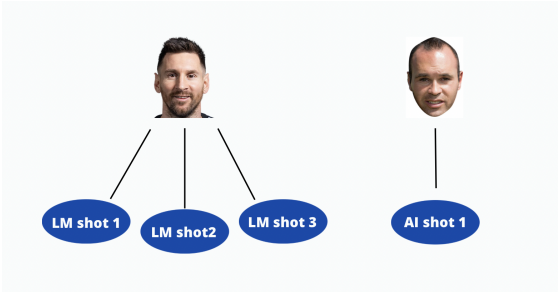

Figure 1: An example of hierarchical structured data for Messi and Iniesta in a match.

The data hierarchy described above breaks the fundamental assumption of independence of observations used in many statistical machine learning models. Therefore, it becomes crucial to employ hierarchical statistical models that can properly account for the repeated features in event data. These models are designed to fit on data with inherent hierarchy, and the model itself incorporates a hierarchical structure in terms of its parameters.

2. Dataset

The dataset originates from StatsBomb and has undergone necessary processing. It comprises records of matches played by Barcelona FC. Below is an illustrative example of a football event, accompanied by the relevant type of data extracted from that particular action.

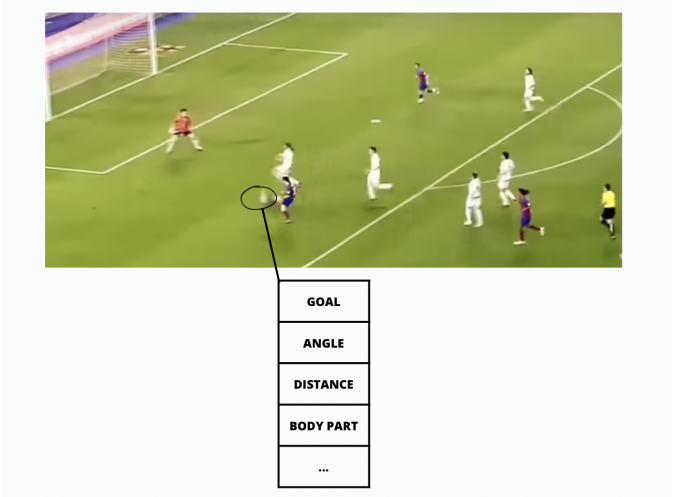

Figure 2: Messi's Goal against Albacete

A reasonable assumption is that each player has a unique type of shooting. Consequently, given a specific player, their

shots inherently correlate with each other and behave independently from the shots of other players. This introduces within-player correlations, as event-level data represents longitudinal data. As a result, each shot is essentially grouped or nested under a specific player. Therefore, models used for analysis must be provided with information about this hierarchical structure; otherwise, it may lead to biased inferences. Properly accounting for this hierarchy allows for accurate parameter estimation for each player, enabling the generation of a meaningful comparison metric between players.

3. Model

3.1. Predictors

The predictor variables are selected and derived from the dataset. To mitigate multicollinearity issues, certain predictors are excluded from the final model. The following predictors are utilized in the models:

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

3.2. GLM and GLMM

Usually for targeting binary variables, the model used is the generalized linear model. The equation for this model for every shot is illustrated below:

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

where the g is the logit link function and the π_i is the odds for a goal $\frac{Pr(Y=1)}{Pr(Y=0)}$. The GLM does not fed any information about the players. This implies that the model fails to adjust for the within-player correlation.

To adequately account for the players' influence in the data, it is feasible to extend the Generalized Linear Model (GLM) framework from a single level to a multilevel framework, thereby creating a Generalized Linear Mixed Model.

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

This framework is similar to GLM except for the j index and the δ_j term. The j index refers to the j-th player in the data. The δ_j term is a statistical parameter that represents the random effect associated with the $j{-}th$ player. As the model is trained, it computes the values of the parameters in the equation including the δ_j. As a result, each player in the data set will have their own estimated δ measure. This implies that each player will have their own unique intercept or baseline for a shot that they take. This player-specific baseline can be derived by calculating the equation, $\beta_0 + \delta_j$. The particular GLMM framework used here, is a mixed model only with a random intercept for each player.

3.3. Model specifications

First of all, let us define some basic notion for the GLMM model.

- $g = 1, 2, \ldots 40$ number of players selected.
- $i = 1 \ldots n_g$ Number of shots done by the $g - th$ player
- π_i, g = odds ratio for a goal of the $i - th$ shot and the $g - th$ player.
- π_g = odds ratio for goal of the $g - th$ player. $\in (n_g \times 1)$

3.4. 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.

For model convergence reason, only player with more than 30 shots are taken for the analysis and continuous predictor are scaled.

For parameter estimation, since there are not closed form solutions for GLMMs, MCMC algorithm such as Gibbs sampling has been used.

3.5. Assumptions

While adopting the GLMM model with random effect only on the intercept it is import to make few assumptions:

- The errors are independent from the random effect
- The β_s are normally distributed between the sampling model, the errors are normally distributed.
- The variability inside each group is the same so the random effects are able to explain the dependence structures between units in the same group.

For computing the model non-informative prior for the parameters has been chosen.

4. Posterior Analysis

4.1. Experiment Results

The predictor effects obtained from the GLMM are presented in the table below. The results are expressed as interpretable odds ratios, indicating incremental or decremental changes. Effects greater than zero signify a positive multiplicative impact on the odds of shot conversion, while effects less than zero indicate a negative multiplicative effect.

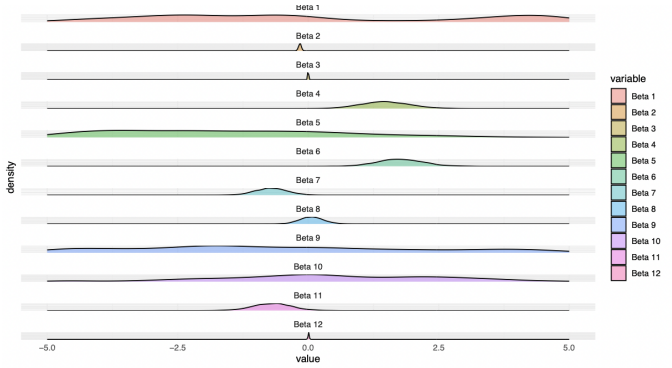

Figure 3: posterior distribution of the beta coefficients

Predictors	Mean
β_1: Intercept	-10.933
β_2: shot_distance	-0.1548
β_3: shot_angle	-0.001
β_4: bodypart Left Foot	1.481
β_5: bodypart Other	-24.49
β_6: bodypart Right 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

Table 1: Mean effect for the fixed parameters used in the GLMM model

4.2. Parameter interpretation

It is evident that some effects are positive, while others are negative, and some are very close to zero. Specifically, the shot distance appears to have a negative influence on the probability of scoring. Moreover, it is noteworthy that a positive impact on scoring is directly related to the preferred type parameters; for instance, if the player shoots with their preferred foot, it results in a higher likelihood of goal scoring. On the other hand, a significant negative impact is associated with using other body parts for the shot. However, it is important to consider that these conclusions rely on a low number of shot observations, leading to non-convergence and potentially unreliable results.

For each player can be added the player effect on goal as the intercept; if is greater than 0 it will result in a more likelihood score and at the contrary if lower than 0.

Player	Random Effect
Lionel Andrés Messi	0.135
Antoine Griezmann	-0.034
Seydou Kéita	0.0008

Displayed are the 90% Highest Posterior Density (HPD) intervals for the fixed effect parameters. The parameters is considered statistically significant if the HPD interval associated does not include 0.

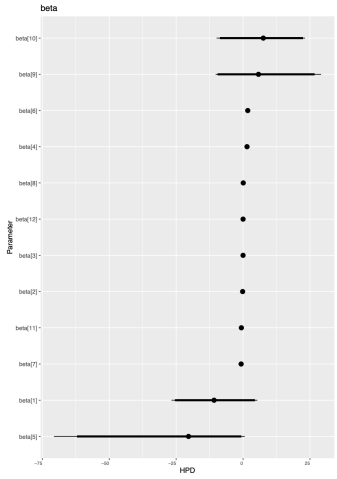

Figure 4: HPD intervals of fixed effect parameters

4.3. Model comparison

For model comparison, the Deviance Information Criteria (DIC) has been chosen. It is important to note that DIC is applicable only when the posterior distribution is approximately normal. The DIC value for the Generalized Linear Mixed Model (GLMM) classifier is approximately 841.5081, while the DIC value for the Generalized Linear Model (GLM) classifier is approximately 833.717.

The data analysis does not provide sufficient evidence to support accepting the null hypothesis of the deviance between the random effect being greater than 0. Consequently, it is reasonable to conclude that the hierarchical structure proposed initially does not align well with the observed data. There are no top-level components that exert influence on each other, and no statistically significant structural variables are identified.

For future investigations, it is recommended to consider the inclusion of an additional level in the data, such as the team to which the players belong. This extended hierarchical structure could offer valuable insights and enhance the understanding of how team-level factors may influence player performance.

Despite, it can be observed that the selected Barcelona players exhibit similar football skills, leading to a non-significant difference between players within the team. However, it is important to acknowledge that expanding the analysis to include a sample of players from different teams (e.g., all players in the Liga) may yield different results and potentially reveal significant variations among players.

References

[1] Estimated Player Impact" EPI Quantifying the effects of individual players on football (soccer) actions using hierarchical statistical models
[2] Womens soccer expected goals model for Milwaukee Rampage FC