Better Understanding Goal Scoring in Soccer Matches

Matthew Morgan

Brigham Young University

Problem Statement and Understanding

Analysis Goals

1. Find shot-specific variables that can help better explain the number of goals scored in a match
2. To fit a model that is able to account for the team-specific variation in goal scoring after accounting for the shot-specific variables that have been identified as being able to better explain the number of goals scored in a match.

Data Set

- Obatined from FBref.com
- 2019-2022 Big 5 European Leagues (Premier League, Bundesliga, La Liga, Serie A, Ligue 1)
  - 10,754 entries of match shooting statistics
  - 120 teams

Exploratory Data Analysis (EDA)

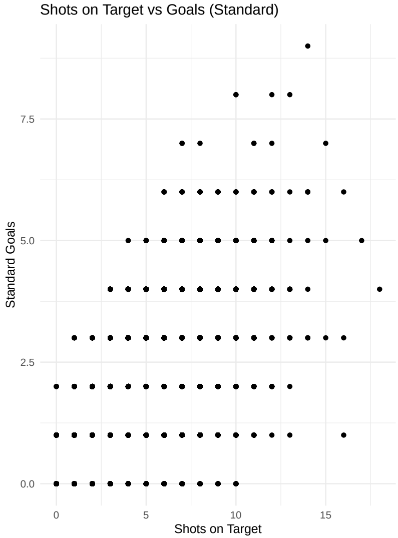

Figure 1

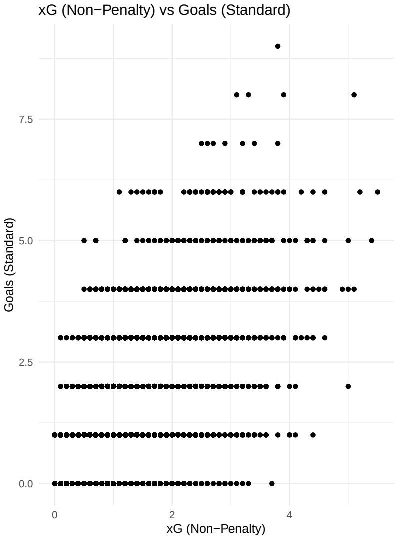

Figure 2

Proposed Model

Poisson Generalized Linear Mixed Model (GLMM) Specification

$$y_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \eta_i$$

$$\eta_i = (\beta_0 + T_{0t}) + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_{ti}$$

- y_i: number of standard goals scored in the i^{th} match
- x_{i1}: number of shots on target in the i^{th} match
- x_{i2}: non-penalty expected goals generated in the i^{th} match

Parameter Interpretation

- β_0: the log mean number of standard goals in a match where there were 0 shots on target and 0 non-penalty expected goals generated in a match is expected to be β_0 on average
- β_1: holding all else constant, the log mean number of standard goals in a match is expected to change by β_1 as the number of shots on target taken in a match increases by 1 on average
- β_2: holding all else constant, the log mean number of standard goals in a match is expected to change by β_2 as the amount of non-penalty expected goals generated in a match increases by 1 on average
- T_{0t}: random team intercept effect fot the t^{th} team which represents a variation from the overall intercept (β_0), it allows the model's estimates of standard goals scored in a match to vary by team

Model Justification - Fixed Effects Testing

Table 1: Fixed Effects Testing

Model	Resid. Dev.	Df	ΔD	AIC
Null	13441.9			32516.6
npxG	10364.4	1	3077.5	29441.1
SoT	9928.3	0	436.1	29005.0
npxG + SoT	9524.4	1	403.9	28603.1
npxG*SoT	9208.2	1	316.2	28288.9

Model Justification - Random Effects Testing

Table 2: Random Effects Testing

Model	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
GLM	3	28603.1	28624.9	-14298.5	28597.1			
GLMM	4	28593.1	28622.3	-14292.6	28585.1	11.9	1	0

Model Justification - Overdispersion Testing

Table 3: Overdispersion Testing

χ^2 Test Statistic	Ratio	Resid. Df	p-value
7711.759	0.717	10750	1

- $H_0 : \phi = 1$
- $H_1 : \phi > 1$
- For a GLMM, the usual procedure of calculating the sum of squared Pearson residuals and comparing it to the residual degrees of freedom gives an approximate estimate of an overdispersion parameter

DHARMa Package

Simulates scaled residuals that should asymptotically follow a standard uniform distribution, Uniform(0,1), for a correctly specified model in the following way:

- Simulate new response data from the fitted model for each observation
- For each observation, calculate the empirical cumulative density function for the simulated observations, which describes the possible values (and their probability) at the predictor combination of the observed value, assuming the fitted model is correct
- The residual is then defined as the value of the empirical density function at the value of the observed data, so a residual of 0 means that all simulated values are larger than the observed value, and a residual of 0.5 means half of the simulated values are larger than the observed value

Model Justification - Residuals

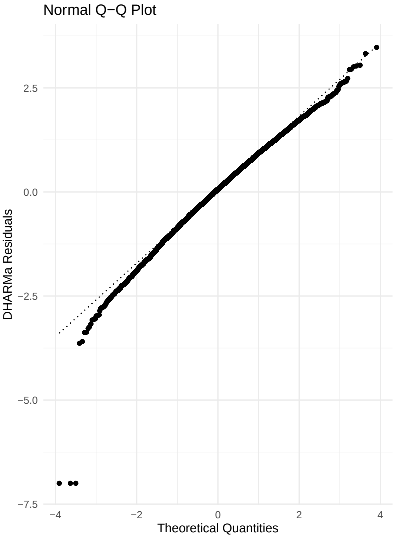

Figure 3

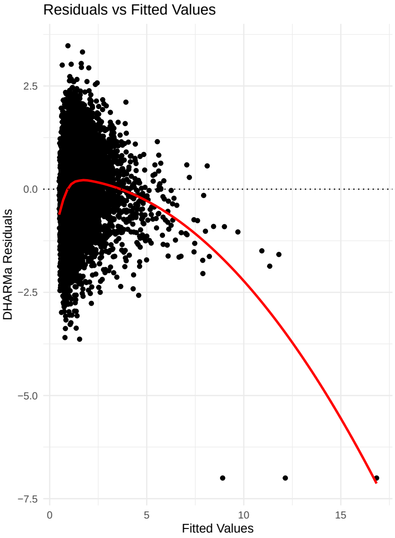

Figure 4

Results

Fixed Effects Estimates

Table 4: Fixed Effects Estimates

Intercept	SoT	npxG
-0.634	0.121	0.257

Marginal Effects Plots

Fixed Effects

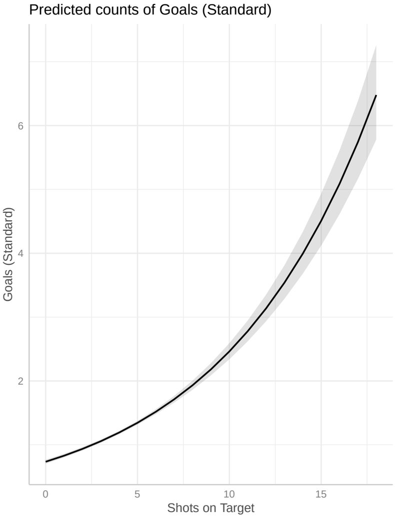

Figure 5

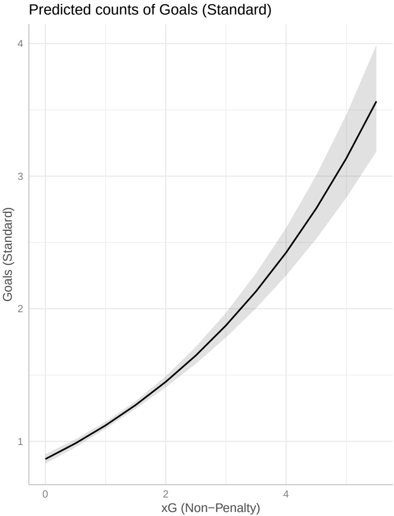

Figure 6

Random Effect Estimates

Table 5: Five Highest Random Team Intercepts

Team	Intercept
Dortmund	0.147
Lazio	0.091
Monaco	0.081
Atletico Madrid	0.071
Manchester City	0.068

Table 6: Five Lowest Random Team Intercepts

Team	Intercept
Norwich City	-0.111
Arminia	-0.076
Burnley	-0.069
Sheffield United	-0.069
Brighton and Hove Albion	-0.064

Random Effect Plots

Random Effects

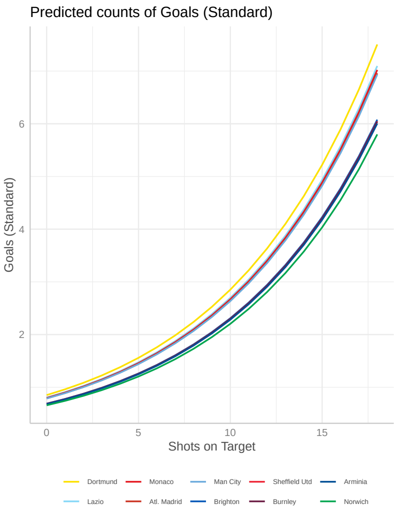

Figure 7

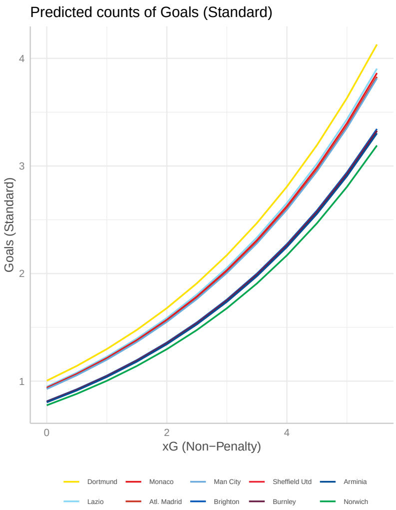

Figure 8

Conclusions

Summary

- Accounting for the quality of the shots taken throughout a match as fixed effects in the model helped to better explain the number of standard goals scored in a match
  - Shots on target (SoT)
  - Non-Penalty xG
- Both variables had positive relationships with respect to the number of standard goals scored in a match
- There is significant variability in the rate at which teams score standard goals in a match
  - Random team intercepts

Next Steps

- Next Steps
    - Model for penalty kick scoring
    - Model for own goal scoring
    - Test for significant opposing team random effect