Better Understanding Goal Scoring in Soccer Matches

Matthew Morgan

Introduction

In soccer, the obvious main objective is to score more goals than the opponent. Perhaps just as obvious is that in order to score goals, a team must take shots. Although, what characteristics about the shots that a team takes throughout a match determine how many goals are scored? This question forms the basis of the first goal of this analysis which will be to find shot-specific variables that can better explain the number of goals scored in a match.

Additionally, there is often much variation in goal scoring due to the teams that are playing. Some of this variation is due to tactics. Teams may take a possession-based approach and are constantly taking shots throughout the game while other teams may focus on maintaining a low defensive block, soaking up pressure from the opposition, and taking their shots on counterattacks. Additionally, some teams have higher quality players throughout their squad and as a result, the team does better at taking higher quality shots throughout a match. Furthermore, the expected scoring output of a team often varies from the actual scoring output of a team. For example, a team could have taken 20 shots in a game, have been expected to score 3 goals, and actually scored 1. Another team may have only taken 7 shots, have been expected to score 1 goal, and actually scored 2. Given all these considerations that will have to be made for teams, the second goal of this analysis will be 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.

The rest of the paper will have the following format. The source of the data will be identified and the data sets used in this analysis will be explained. Next the model utilized to accomplish the goals of this analysis will be outlined mathematically and justified by model comparison and diagnostics. Finally, the results of the model will be shown as well as explaining the strengths and weaknesses of the model in accomplishing the goals of this analysis.

Data and Exploratory Data Analysis

The data for this analysis was obtained from FBref.com, a comprehensive website for football/soccer statistics from all over the world. Team-specific shooting data for each match in the 2019-2020, 2020-2021, and 2021-2022 English Premier League seasons as well as the 2019-2022 German Bundesliga, Spanish La Liga, Italian Serie A, and French Ligue 1 seasons were obtained. This team-specific shot data was then combined into a single data set of the “Big 5” leagues in Europe. These three seasons of the Big 5 European leagues contained 10,754 entries of match shooting statistics for 120 different teams.

While there were multiple shooting variables available in each data set, only three were considered for this analysis. The number of standard goals scored by the team, the number of shots on target (SoT) taken by the team, and the amount of non-penalty expected goals generated by the team. Standard goals are goals the team scored that were not penalty kicks or own goals. Shots on target are shots that either go into the net or are saved by the goalkeeper. Expected goals (xG) measures the quality of a shot based on several variables such as assist type, shot angle and distance from goal, whether it was a headed shot and whether it was defined as a big chance.

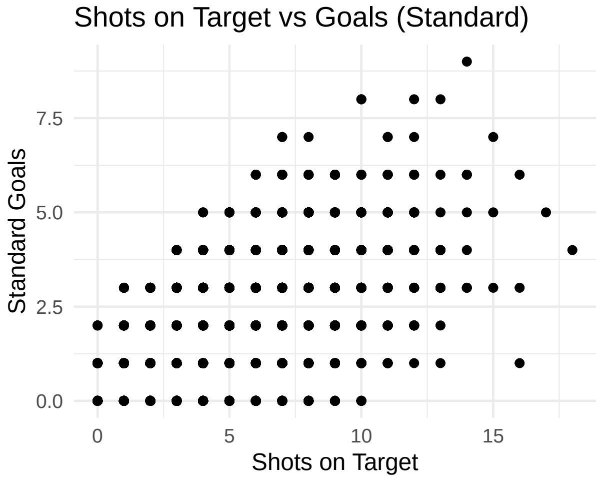

Figure 1

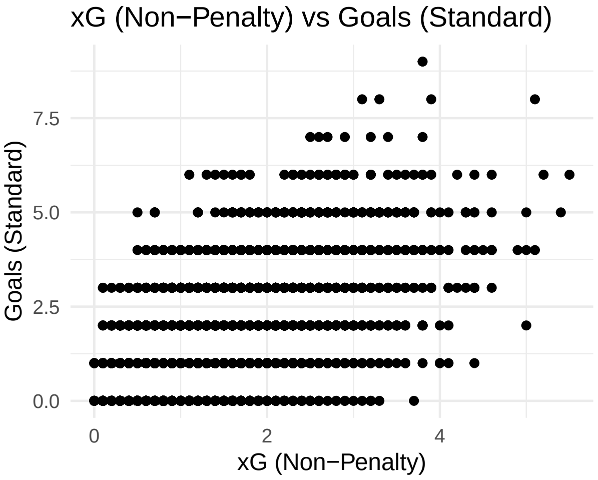

Figure 2

Figure 1 is a scatter plot of shots on target vs standard goals and it suggests an increasing relationship between the number of shots on target taken in a match and the number of standard goals scored in a match. Figure 2 is a scatter plot of non-penalty xG vs standard goals and it also suggests an increasing relationship between the amount of non-penalty xG generated in a match and the number of standard goals scored in a match.

Model

Poisson Generalized Linear Mixed Model (GLMM)

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

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

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

- 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
- 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

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

Table 1 provides justification for including shots on target and non-penalty xG as fixed effects in the model. It can be seen that individually, both variables had significant changes in deviance from a null model and that the AIC decreased. Additionally, when both variables were included together, there was still a significant change in deviance and the AIC further decreased.

Finally, while there was still a significant change in deviance and further reduction of AIC when the interaction between shots on target and non-penalty xG was considered in addition to both of the variables separately, when a random intercept team effect was added along with those three fixed effects, the model failed to converge and was nearly unidentifiable so it was decided to leave the interaction out of the model.

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

Table 2 provides justification for the inclusion of a random team intercept effect in the model. The results of the χ^2 likelihood ratio test that was performed for a model without a random intercept team effect (GLM) vs a model with a random intercept team effect (GLMM) where both models also had fixed effects for shots on target and non-penalty xG show that the GLMM had a significant change in deviance compared to the GLM. Additionally, the AIC and BIC were both lower for the GLMM than the GLM.

Table 3: Overdispersion Testing

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

Table 3 contains the results of the χ^2 likelihood ratio test of the overdispersion parameter (ϕ) for the model and provides justification as to why the model did not include an overdispersion parameter. 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. Given that $H_0 : \phi = 1$ and $H_1 : \phi > 1$, it can be concluded that there is not enough evidence to reject the null hypothesis that there is no overdispersion.

Standard residuals for a GLMM are hard to assess as the plots often seem to show problems such as non-normality and heteroscedasticity, even if the model is correctly specified. As a workaround to this common issue, the `DHARMa` package (Hartig, 2022) was utilized to create residuals for GLMMs that can be interpreted as intuitively as residuals for a linear model. The `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

For this analysis, the simulated scaled residuals from the `DHARMa` package were transformed to follow a normal distribution and then residual plots were created with these simulated normal residuals.

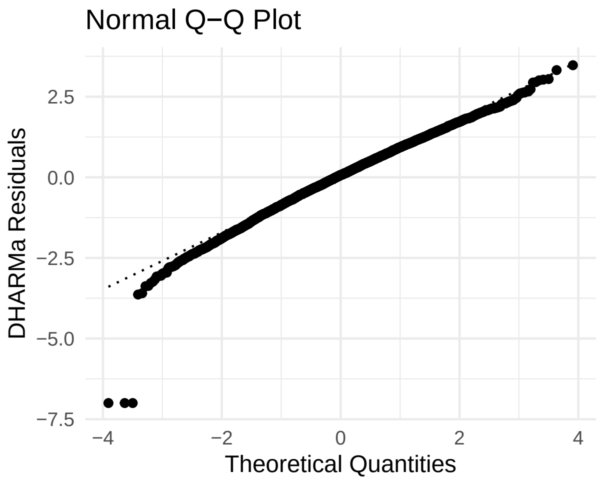

Figure 3

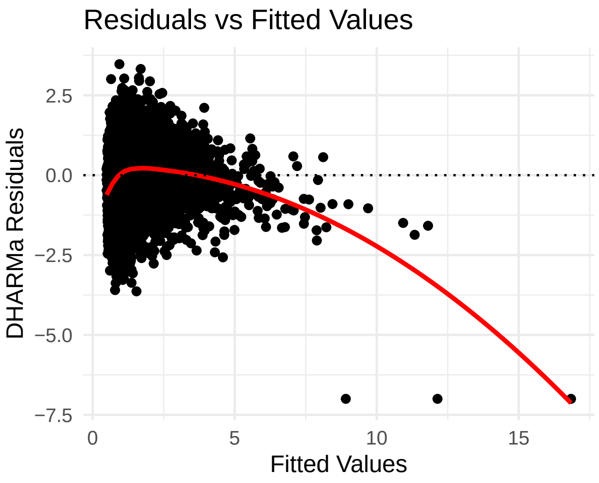

Figure 4

Figure 3 and Figure 4 are residual plots that provide justification that the model fit the data set well. A majority of the residuals fell along or were close to the straight line save three clear outliers which suggested that the data plausibly came from a Poisson distribution. Furthermore, the residuals vs fitted vales plot showed a fairly even spread of residuals above and below the $y = 0$ line save the three clear outliers mentioned previously. Overall, this suggested that there were no significant patterns in the residuals and that the model is again reasonable for the data.

To ensure that the outliers were not having a significant impact on model estimates, the same model was fit to a new data set where the three outliers had been removed and the resulting model estimates did not significantly change from the model that was fit on all of the data.

Results

Table 4: Fixed Effects Estimates

Intercept	SoT	npxG
-0.634	0.121	0.257

Table 4 provides the model estimates of the fixed effects for shots on target and non-penalty xG. Overall, it can be concluded that as the number of shots on target taken or the amount of non-penalty xG generated in a match increases, the number of standard goals scored in a match is expected to increase as well. More specifically, it appears that the number of standard goals scored in a match increases more with respect to increases in the amount of non-penalty xG generated than with respect to the increase in the number of shots on target taken.

Fixed Effects

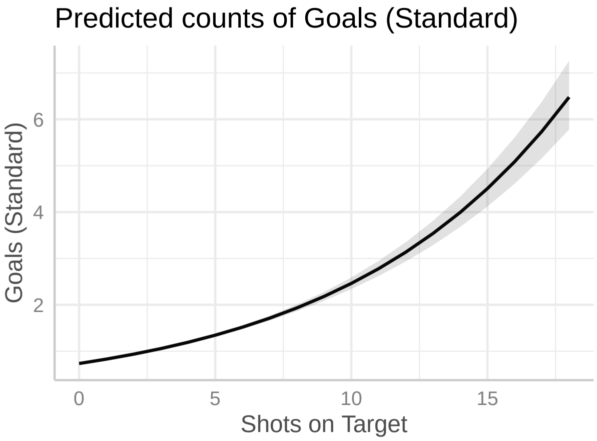

Figure 5

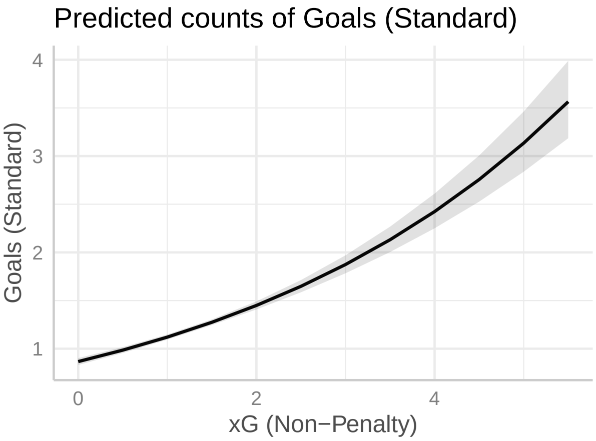

Figure 6

Figure 5 provides a visual representation of the estimated marginal effect that the number of shots on target taken in a match has on the number of standard goals scored in a match as well as the confidence interval of that effect. Figure 6 provides a visual representation of the estimated marginal effect that the amount of non-penalty xG generated in a match has on the number of standard goals scored in a match as well as the confidence interval of that effect. It should also be noted that there is more uncertainty in what the marginal effects of shots on target and non-penalty xG is for the number of standard goals scored for higher values of shots on target and non-penalty xG. Overall, these visualizations reinforce the earlier stated conclusions that as the number of shots on target taken or the non-penalty xG generated in a match increases, the number of standard goals scored in a match is expected to increase as well.

Based on these results for the fixed effects, it appears that the model effectively accomplishes the first goal of this analysis as the number of shots on target taken in a match as well as the non-penalty xG generated in a match help to better explain the number of standard goals scored in a match.

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

Table 5 provides the five highest random team intercepts and Table 6 provides the five lowest random team intercepts from the model. The simple way to think about these random team intercepts is that the higher the random team intercept, the higher the estimated number of standard goals scored in a match will be for that team.

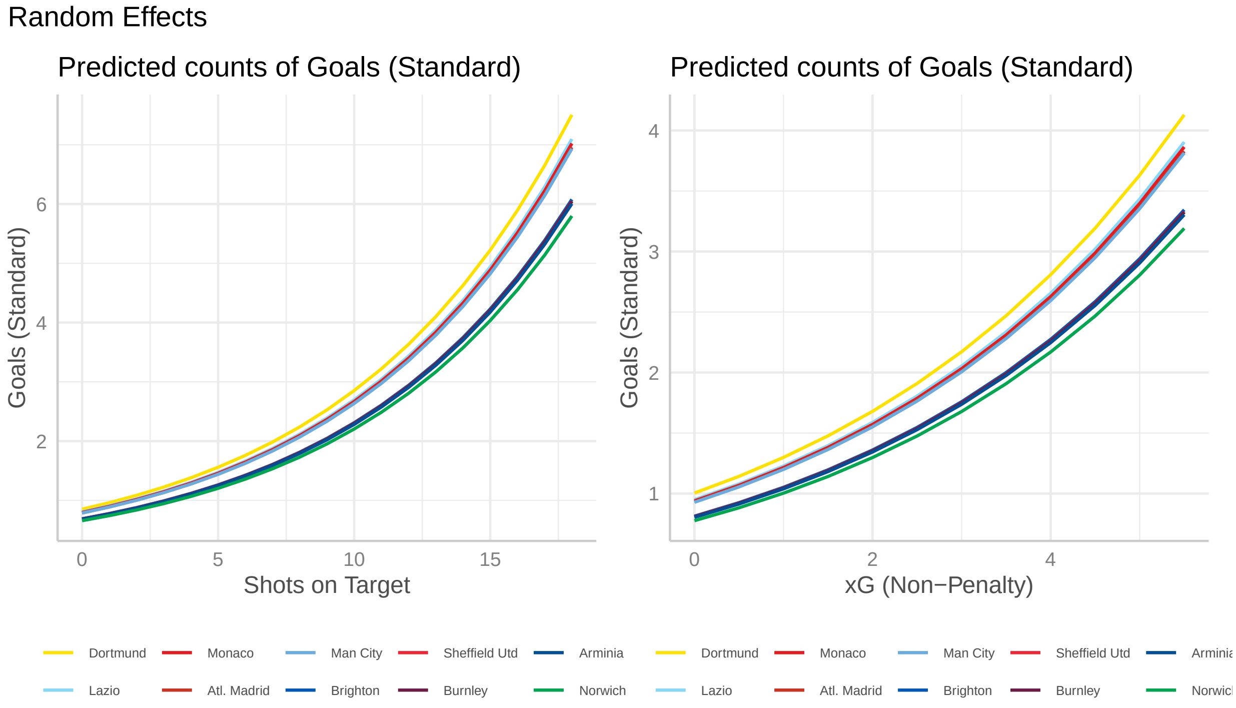

Figure 7

Figure 8

Figure 7 provides a visual representation of the estimated number of standard goals scored in a match with respect to the number of shots on target taken in a match for the teams highlighted in Table 5 and Table 6. Figure 8 provides a visual representation of the estimated number of standard goals scored in a match with respect to the amount of non-penalty xG generated in a match for the teams highlighted in Table 5 and Table 6.

It can be seen that for the teams with the highest random team intercepts, they were estimated to score more standard goals in a match than the teams with the lowest random team intercepts with respect to the number of shots on target taaken in a match. Each team has a similar slope and a unique intercept. These same patterns follow for the non-penalty xG plot as well. Overall, these results are expected and in line with the estimated random team intercepts from the model.

Based on these results for the random effects, it appears that the model effectively accomplishes the second goal of this analysis as it can estimate the number of standard goals scored in a match based on the number of shots on target taken and the amount of non-penalty xG generated while individually accounting for the variation in standard goals scored in a match due to a specific team playing after accounting for the number of shots on target taken and the amount of non-penalty xG generated.

Overall, the model accounted for the team-specific variation in standard goals scored in a match after accounting for the number of shots on target taken and the amount of non-penalty xG generated by estimating higher random team intercepts for higher quality teams which resulted in higher estimates of standard goals scored in a match and estimating lower random team intercepts for lower quality teams which resulted in lower estimates of standard goals scored in a match.

Conclusions

The two main goals of this analysis were to find shot-specific variables that can better explain the number of goals scored in a match and to fit a model that was able to account for the team-specific variation in goal scoring after accounting for the previously identified shot-specific variables. The previous section showed that accounting for the quality of the shots taken throughout a match by using the number of shots on target and the amount of non-penalty xG generated as fixed effects in the model helped to better explain the number of standard goals scored in a match. Furthermore, the model estimated these shot-specific variables as having positive relationships with respect to the number of standard goals scored in a match. Previous sections also showed that there is significant variability in the rate at which teams score standard goals in a match and the model accounted for that team-specific variation with random team intercepts.

There are other studies that could be done to further enhance this analysis. One aspect that could be further studied is fitting other models to better understand penalty goal scoring and/or own goal scoring and synthesizing the results of those models to be able to model all goal scoring activity in a match and not just standard goal scoring. Another aspect that could be studied is whether or not there is variability due to the opposing team in a match. This study would most likely require the collection of additional data so that there would be enough data for each match up being considered.

In conclusion, it can be said that the model proposed in this analysis which examined the quality of a shot while accounting for team-specific variation was a simple and effective model for understanding the number of goals scored in a match. Even with this simple model model proving effective, it is quite probable that further improvements to this model could be made to continue to improve the understanding of the number of goals scored in a match.

References

Hartig F (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.6, https://CRAN.R-project.org/package=DHARMa.