Using bookmaker match result odds to evaluate the bivariate Poisson distribution as a model of football goal scoring

John Thornley

Abstract

The bivariate Poisson distribution is a widely-applied model of association football (soccer) goal scoring. This paper uses bookmaker match result odds to investigate whether this simple three-parameter distribution (two goal-scoring rates and their correlation coefficient) is sufficiently powerful to represent the complexities of actual football match outcomes. Assuming a bivariate Poisson goal-scoring process with fixed correlation, we show that match result odds (win/draw/loss) uniquely determine the implied goal-scoring rates of the two teams. Analyzing over 40,000 matches from twenty years of English professional football we find that with a best fit correlation coefficient of 13% a bivariate Poisson distribution with odds-implied goal-scoring rates closely predicts the observed score distribution. This suggests that misprediction of draws and low/high scores reported in past work is more likely due to shortcomings of the specific regression factors and formulations than to inherent limitations in the applicability of the underlying bivariate Poisson distribution.

Keywords: Sports forecasting, Bivariate Poisson, Football, Betting, Bookmaker odds

1. Introduction

The bivariate Poisson distribution describes two discrete-valued random variables with positive correlation and each variable having a marginal Poisson distribution. It can be defined in terms of three parameters: the means of the two marginal distributions and their correlation coefficient. The distribution is a widely-accepted and physically plausible model for goal scoring in an association football (soccer) match, with the means corresponding to the goal-scoring rates of the two teams.

There is a large body of work going back many decades fitting predictive factors such as past match scores to bivariate Poisson regression formulations then evaluating the goodness of fit. The outcomes have been generally encouraging but have often exhibited misprediction of special cases such as draws and low/high scores, leading to proposals for ad hoc extensions to the bivariate Poisson model such as diagonal inflation (i.e., increased draw probabilities). Additionally, there has not been consensus on the need for or degree of correlation between the two goal-scoring processes.

It has not been clear whether the shortcomings of these fitted models result from inherent limitations of the bivariate Poisson distribution or insufficient information in the specific predictive factors and regression formulations. Published regression work has understandably tended toward using easily-quantified factors such as scores and shots on goal rather than more nuanced factors such as individual player fitness and form that are likely also important to accurate football predictions. We believe that it is premature to add complexity on top of the bivariate Poisson goal

generation process without first understanding the fundamental applicability of the model to football goal scoring, separately from any particular regression formulation.

To evaluate the bivariate Poisson distribution as a model of football goal scoring, we calibrate it using bookmaker match result (win/draw/loss) odds[1]. Betting market competition drives bookmaker odds to be accurate estimators of match result probabilities. We show that if we assume a bivariate Poisson goal-scoring process with a known correlation coefficient, the result probabilities for a match uniquely determine the implied goal-scoring rates of the two teams participating in the match. Therefore, by making the simplifying assumption that the correlation coefficient is fixed for all matches in a population, we can first fit the best correlation coefficient for the population then determine the odds-implied goal-scoring rates of each match.

We apply this analysis to twenty recent years of match results from the four fully-professional English football leagues (comprising over 40,000 matches) and find the correlation coefficient between the goal-scoring rates to be 0.13 using least-squares regression. With this correlation coefficient, the predicted distribution of scores closely matches the observed distribution and does not exhibit major misprediction of draws or low/high scores. This shows that the bivariate Poisson distribution with fixed correlation coefficient is a good model for football goal scoring when calibrated using sufficiently powerful predictive factors and implies that weaknesses in regressions are likely due to their factors having less predictive information than is used by the bookmakers.

Demonstrating that a well-calibrated bivariate Poisson model captures the dynamics of goal scoring without special-case adjustments is valuable in its own right and also provides a foundation for future work. Accurate ex-ante goal-scoring rates derived from match result odds could be used for better analysis of intra-match goal scoring by normalizing goal-scoring rates across matches. Quantified correlation coefficients and distributions of goal-scoring rates could be used to provide robust prior distributions for regressions to fit goal-scoring models to predictive factors.

2. Related work

The bivariate Poisson distribution was first defined by Campbell (1934) and Aitkin (1936). It is described by Holgate (1964) as "a useful model for investigating two-dimensional discrete-valued random variables with positive correlation where the marginal distributions are both Poisson".

There is a large body of past work going back over seventy years related to fitting bivariate Poisson models to football results. That work attempts to develop predictive models of future match results based on hypothetical variables such as team offensive and defensive strengths fitted using linear regression to factors from past match results. Success is often measured in terms of comparison with the predictive power of bookmakers, e.g., by simulated betting against bookmaker odds.

Among the formative work of this kind, Moroney (1951) finds that the goals scored in 240 football matches are well fitted by a population-wide Poisson distribution but better fitted if the means vary from match to match. Maher (1982) fits a dual independent Poisson model to goal-scoring from four seasons of English professional football, based on each team's attacking and defensive strength.

[1] Exact score odds typically do not cover all scores and have wider spreads and more statistical noise.

He finds an overestimation of the frequency of high scores, which leads him to propose a bivariate Poisson model with a small positive correlation coefficient.

Subsequent research extends Maher's work with different regression formulations and adjustments to handle poorly fitted cases. For example, Dixon and Coles (1997) describe a bivariate Poisson regression model that adds a specific adjustment to better fit low-scoring matches. Karlis and Ntzoufras (2003) investigate many adjustments including inflating the number of draws to improve the model fit. Goddard (2005) compares bivariate Poisson modeling of goals with ordered probit modeling of match results and finds little difference in predictive power.

This past work does not attempt to evaluate the bivariate Poisson distribution separately from the specific regression formulations. In contrast, we use the predictions implied by bookmaker match result odds to investigate the inherent applicability of the bivariate Poisson model to football goal scoring and do not attempt to demonstrate better predictive power than the bookmakers or any other model. We are able to definitively estimate the best bivariate Poisson correlation coefficient because we are not conflating the fitting of the correlation coefficient and two goal-scoring rates in the same regression.

The idea of using bookmaker odds as predictors of match results is a natural extension of the efficient market hypothesis in financial markets described by Fama (1970). In principle, bettors seek the best odds and bookmakers try to maximize their profits, so the prices at which they transact should represent the best market consensus on result probabilities. Empirical research supports this hypothesis for popular betting markets. Pope and Peel (1989) find odds to be efficient for the 1981-82 season of United Kingdom football and Cain et al. (2000) find very few profitable betting opportunities in the 1991-92 season despite identifying a slight favorite-longshot bias. More recently, both Angelini and De Angelis (2019) and Elaad et al. (2020) find the English football betting markets to be efficient over the 2006-17 and 2010-18 seasons respectively.

Regarding the conversion of bookmaker match result odds to match result probabilities, most of the work listed above prorates away the bookmakers' margin by normalizing the total probability to one. We find that this simple approach yields unbiased probabilities for our data drawn from large professional football leagues with competitive betting markets. Štrumbelj (2014) summarizes prior work and makes a case for more sophisticated methods of factoring out the margin using regression or modeling of the bookmakers' objective function in smaller markets.

3. Data description and sources

The data that we analyze consist of the 40,247 matches from the 2000-01 to 2019-20 seasons of the four fully-professional English football leagues: the Premier League (EPL), League Championship (ELC), League One (EL1), and League Two (EL2). All four leagues have the same structure: every team in a league plays every other team in the same league once at home and once away each season. Matches consist of two 45-minute halfs with stoppage time (usually no more than a few minutes) added to the end of each half at the discretion of the referee to make up for unusual delays during play. In all these leagues, there is no extra time or other tie-breaker if the scores are equal when the match finishes.

We only include matches from the 2019-20 season up to and including March 10, 2020. After this date, matches in all four leagues were suspended due to the Covid-19 pandemic. The remainders of

	EPL	ELC	EL1	EL2	Total
Teams	20	24	24	24	92
Matches per season (2000-01 to 2018-19)	380	552	552	552	2,036
Matches per season (2019-20)	288	444	400	440	1,572
Total matches	7,508	10,932	10,888	10,928	40,256
Matches with missing odds	0	3	2	4	9
Matches in our data set	7,508	10,929	10,886	10,924	40,247

Table 1: Matches included from 20 seasons of English professional football.

the EPL and ELC seasons were eventually played in June and July 2020 under unusual circumstances (no spectators and a compacted schedule) and the remainders of the EL1 and EL2 seasons were canceled. As a consequence, our candidate data set contains 40,256 matches, rather than the 40,720 matches that would have been played over 20 full seasons.

For each match, the data that concern us are the final score and the pre-match bookmaker match result odds. Bookmaker odds are published several days before a match and usually remain unchanged until the match begins. Match results and odds data were obtained from Football-Data.co.uk. Match results were verified against data and media reports from ESPN.co.uk/football, Soccerbase.com, and BBC.com/sport/football.

The bookmaker match result odds that we use consist of triples of home-win, draw, and away-win odds from up to ten different individual bookmakers per match. (FootballData.co.uk also provides the maximum odds from across a larger set of bookmakers but we do not use them as they are not available over our entire data set.) Of our 40,256 candidate matches, FootballData.co.uk is missing odds for 9 matches, leaving 40,247 matches included in our data set for analysis, as shown in Table 1.

Bettors have a choice of which bookmaker to bet with, so for each match we consider only the best (i.e., maximum) home-win, draw, and away-win odds from across all bookmakers. For approximately 3.5% of matches, these best odds allow a sure-win arbitrage bet across multiple bookmakers in a single match, which is likely an artifact of FootballData.co.uk loading odds from bookmakers at different times or promotions limited to small bets. In these cases, we automatically discard the outlier bookmakers (almost always just one is needed) that most efficiently remove the arbitrage. Throughout the rest of this paper, where we refer to odds, we mean the best pre-match bookmaker result odds with apparent arbitrage opportunities removed in this manner.

4. Match result odds and odds-implied result probabilities

We denote each match as being between two teams, A and B, with three possible results: win_A, $draw$, and win_B. There is no requirement that A and B correspond to the home and away teams consistently. The odds for a match are denoted as θ_{win_A}, θ_{draw}, and θ_{win_B}. Odds are given as the total payout to a successful bettor, e.g., a bet of \$10 at odds of 1.5 will pay \$15 if successful and \$0 if unsuccessful.

The *over-round* or bookmaker's *margin* for a match is defined as:

$$\hat{O} = \frac{1}{\theta_{win_A}} + \frac{1}{\theta_{draw}} + \frac{1}{\theta_{win_B}} \geq 1 \tag{1}$$

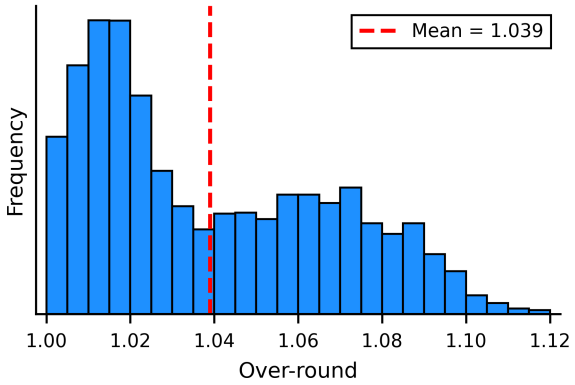

Figure 1: Observed over-round distribution across all seasons and leagues.

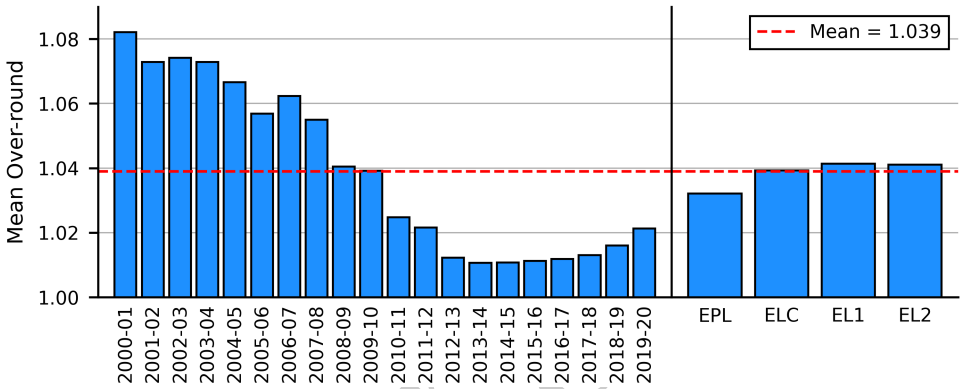

Figure 2: Observed mean over-round by season and league.

The over-round can be thought of as a fee that bettors pay the bookmakers to provide betting services. An over-round of less than one represents an arbitrage opportunity, i.e., a risk-free guaranteed-profitable bet. In practice, arbitrage opportunities are usually illusionary (e.g., outdated odds), short-lived, or limited to very small bets. As we discuss in Section 3, all matches in our data set have arbitrage opportunities removed and hence have over-rounds of at least one.

Figure 1 shows the distribution of over-rounds for our data set of twenty seasons of four football leagues. The average over-round is a little less than 104%. Figure 2 shows that the average over-round has generally decreased over time as betting markets have become more competitive and is lower for the English Premier League (EPL), which has high worldwide betting volumes, than for the three English Football League competitions (ELC, EL1, EL2), which have lower betting volumes.

We define the *odds-implied probabilities* for the results of a match as follows:

$$\pi_r = \frac{1}{\hat{O}.\theta_r}, r \in \{win_A, draw, win_B\} \tag{2}$$

This definition normalizes the total probability to exactly one by presuming that the bookmakers have approximately pro-rated the over-rounds across the three different results on top of their best predictions. Štrumbelj (2014) points out that bookmakers might not add their profit margin to the

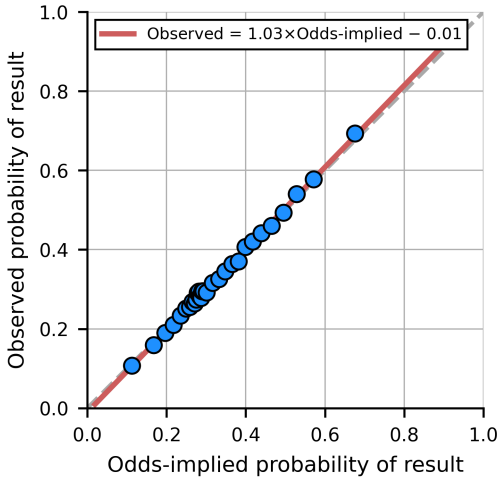

Figure 3: Observed versus odds-implied result probabilities, for result $\in \{win_A, draw, win_B\}$ with equal-sized bins and best linear fit.

fair odds in a pro-rated manner, but he also found that this basic normalization is more accurate than commonly-used regression-based approaches and is particularly accurate for football because of market efficiency resulting from large betting volumes. This simple approach is well-suited to our data set with an average over-round of less than 4% above fair odds.

Figure 3 shows that the odds-implied probabilities of results (win_A, $draw$, and win_B) almost exactly match the observed probabilities of those results. Visually, there appears to be a slight overestimation of low-probability results and underestimation of high-probability results, i.e., slightly lowered odds for long-shots and raised odds for favorites, which matches the bias direction reported by Goddard and Asimakopoulos (2004) and Daunhawer et al. (2017). However, in linear regression of observed results against odds-implied result probabilities, we cannot reject a null hypothesis of slope equal to one and intercept equal to zero with high confidence ($p = 0.08$ for the slope and $p = 0.07$ for the intercept), so we do not adjust for this low-significance effect.

5. Bivariate Poisson goal-scoring distribution

It is reasonable to approximate goal scoring in a football match as dual Poisson processes. Over short time horizons, two concurrent scoring processes is clearly an inaccurate model because there is only one ball and the opposing goals are approximately 100 meters apart, but this is not material in practice because the goal-scoring rate in football is low (2.6 goals per match for our data set). The simplest form is two independent Poisson goal-scoring processes. However, the bivariate Poisson distribution is a more useful model because positive correlation can model anecdotally-described behavior such as the team that is ahead playing defensively.

Two random variables, X_A and X_B, follow a bivariate Poisson distribution, $P(\mu_A, \mu_B, \rho_{A,B})$, if the marginal distributions of X_A and X_B follow Poisson distributions with means μ_A and μ_B respectively and correlation coefficient $\rho_{A,B}$. For our purposes, X_A and X_B are the numbers of goals scored by teams A and B in a specific match, and μ_A and μ_B are the means for the number

of goals scored (i.e., the goal-scoring rates) by the two teams over that specific match[2]. The usual formulation is in terms of three Poisson random variables, Y_A, Y_B, and Y_C, with means λ_A, λ_B, λ_C respectively. If we define $X_A = Y_A + Y_C$ and $X_B = Y_B + Y_C$, X_A and X_B will follow a bivariate Poisson distribution with probability function:

$$P(X_A = a, X_B = b \mid \mu_A, \mu_B, \rho_{A,B}) = \exp(-\lambda_A - \lambda_B - \lambda_C) \frac{\lambda_A^a}{a!} \frac{\lambda_B^b}{b!} \sum_{k=0}^{\min(a,b)} \binom{a}{k} \binom{b}{k} k! \left(\frac{\lambda_C}{\lambda_A \lambda_B}\right)^k$$

$$\begin{aligned} \text{where:} \quad \mu_A &= \lambda_A + \lambda_C, \\ \mu_B &= \lambda_B + \lambda_C, \\ \rho_{A,B} &= \frac{\lambda_C}{\sqrt{(}\mu_A \mu_B)}, 0 \leq \rho_{A,B} \leq \min\left\{\sqrt{\frac{\mu_A}{\mu_B}}, \sqrt{\frac{\mu_B}{\mu_A}}\right\} \end{aligned} \tag{3}$$

If λ_C is zero and hence $\rho_{A,B}$ is zero, X_A and X_B are independent Poisson random variables and the joint distribution is known as a double or dual Poisson distribution.

6. Bivariate Poisson match result distribution

From the bivariate Poisson distribution of goals scored (Equation 3), we can define the resulting probability distribution of the match result, R, as follows:

$$P(R = r \mid \mu_A, \mu_B, \rho_{A,B}) = \begin{cases} \sum_{a=1}^{\infty} \sum_{b=0}^{a-1} P(X_A = a, X_B = b \mid \mu_A, \mu_B, \rho_{A,B}), & r = win_A \\ \sum_{x=0}^{\infty} P(X_A = x, X_B = x \mid \mu_A, \mu_B, \rho_{A,B}), & r = draw \\ \sum_{b=1}^{\infty} \sum_{a=0}^{b-1} P(X_A = a, X_B = b \mid \mu_A, \mu_B, \rho_{A,B}), & r = win_B \end{cases} \tag{4}$$

Figure 4 shows bivariate Poisson match result probabilities versus goal-scoring rates, for correlation coefficients ranging from zero through to one. For $\rho_{A,B} = 0$, the two Poisson goal-scoring processes are independent and μ_A and μ_B are unconstrained. As $\rho_{A,B}$ increases, allowed values of μ_A and μ_B become more constrained and the probability of a draw increases for a given μ_A and μ_B. At the extreme, for $\rho_{A,B} = 1$, the goal-scoring processes are completely correlated, μ_A and μ_B are constrained to be equal, and all results are draws.

7. Inverse Bivariate Poisson match result distribution

It is not immediately obvious and to our knowledge has not been pointed out in past work that for a bivariate Poisson goal-scoring process with a known correlation coefficient, the result probabilities for a match uniquely determine the goal-scoring rates of the two teams. If the correlation

[2]There is no presumption that the goal-scoring rate of a team in one match bears any relationship to the goal-scoring rate of that team in a different match. Even the same two teams playing each other on two separate occasions may have very different expected goal-scoring rates on each occasion.

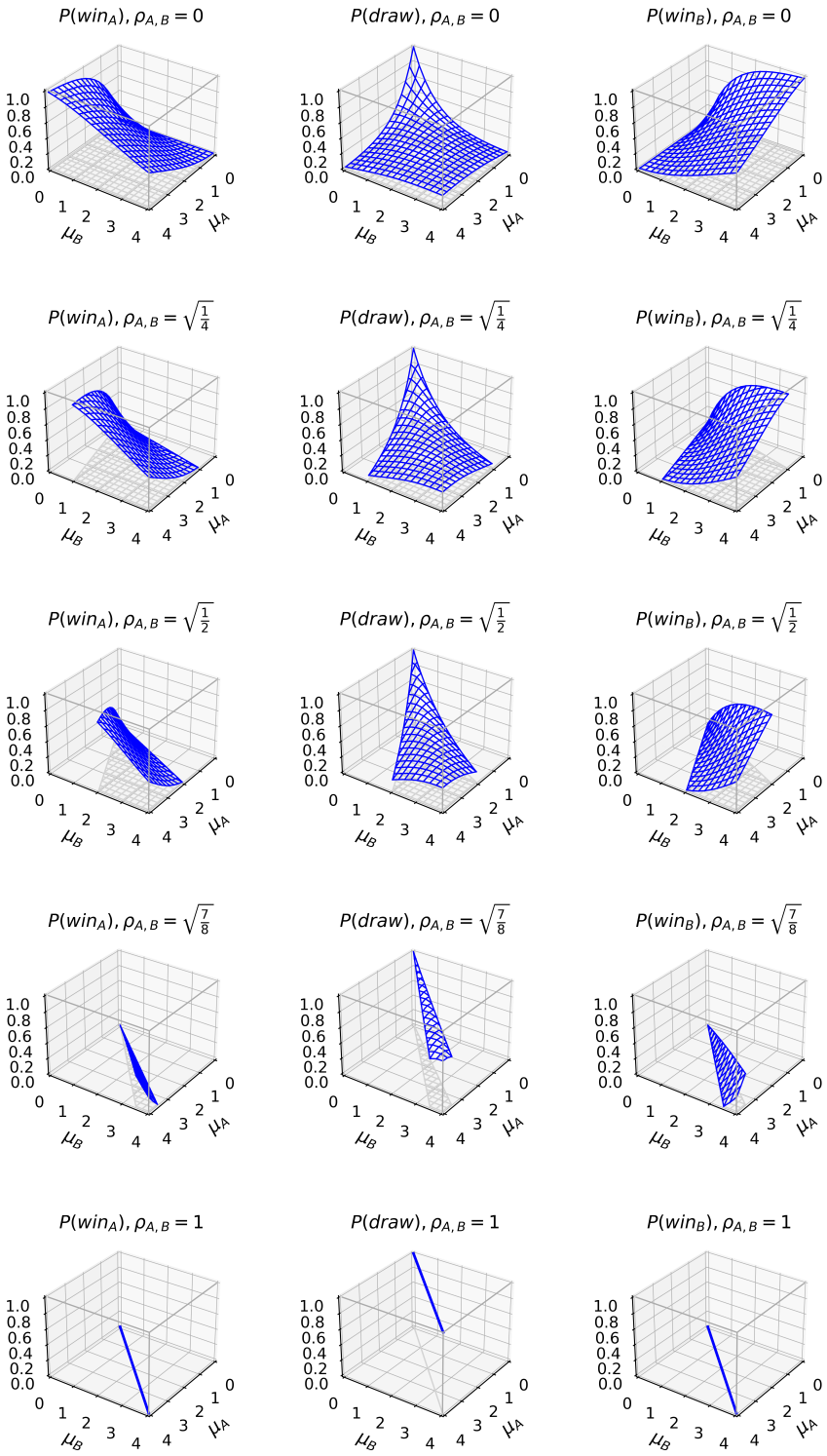

Figure 4: Bivariate Poisson match result probabilities, $P(win_A)$, $P(draw)$, and $P(win_B)$, versus goal-scoring rates, μ_A and μ_B, for a range of correlation coefficients, $\rho_{A,B}$.

coefficient between the two goal-scoring processes is known, the Bivariate Poisson match result distribution (Equation 4) becomes a function from two known variables (the goal-scoring rates of the two teams) to two unknown variables (the probabilities of the three possible match results - since any two determine the third). To determine unknown goal-scoring rates from known match result probabilities, we can define the inverse of the Bivariate Poisson match result distribution, as follows:

$$\begin{aligned} P^{-1}(P_{win_A}, P_{win_B} \mid \rho_{A,B}) = \mu_A, \mu_B \text{ such that} \\ P(R = win_A \mid \mu_A, \mu_B, \rho_{A,B}) = P_{win_A} \text{ and} \\ P(R = win_B \mid \mu_A, \mu_B, \rho_{A,B}) = P_{win_B} \\ \text{where: } 0 \leq P_{win_A}, 0 \leq P_{win_B} \\ 0 \leq P_{win_A} + P_{win_B} < 1 \\ 0 \leq \rho_{A,B} < 1 \end{aligned} \tag{5}$$

The inverse function is not defined for $\rho_{A,B} = 1$ (where a draw is certain and any $\mu_A = \mu_B$ is a solution) or $P_{win_A} + P_{win_B} = 1$ (where a draw is impossible and at least one of μ_A and μ_B is infinite). For other input values, the result of the inverse function can be determined using numerical multivariate root-finding algorithms. In our analysis, we use the Julia NLsolve package developed by Mogensen et al. (2020) with the default trust region method.

Figure 5 shows bivariate Poisson goal-scoring rates derived from match result probabilities for correlation coefficients ranging from zero through to one. As $\rho_{A,B}$ increases, the probability of a draw result increases, and therefore the μ_A and μ_B required to achieve given non-zero $P(win_A)$ and $P(win_B)$ results also increase.

8. Relationship between the correlation coefficient and goal-scoring rates

It is informative to understand the relationship between the correlation coefficient and goal-scoring rates in order to explain the characteristics of the bivariate Poisson model of football goal scoring. For any given match result probabilities, $P(win_A)$, $P(draw)$, and $P(win_B)$, the goal-scoring rates returned by the inverse bivariate Poisson distribution will by definition and construction predict exactly those match result probabilities for all correlation coefficients. However, a larger correlation coefficient increases the likelihood that at any level of goal scoring the two teams will be tied, therefore both goal-scoring rates need to be higher to achieve the given match result probabilities.

Figure 6 shows the odds-implied goal-scoring rate distributions and resulting predicted vs observed score likelihoods obtained by applying the inverse bivariate Poisson distribution to the odds-implied match result probabilities over the matches in our data set for three different correlation coefficients. For a correlation coefficient of zero, the odds-implied goal-scoring rates are lower than the observed rates, resulting in significant over-prediction of low scores and under-prediction of higher scores. For a correlation coefficient of 0.15, the odds-implied goal-scoring rates are close to the observed rates and there is only a slight (less than 1%) over or under-prediction of scores. For a correlation coefficient of 0.30, the odds-implied goal-scoring rates are higher than the observed rates, resulting in significant under-prediction of low scores and over-prediction of higher scores.

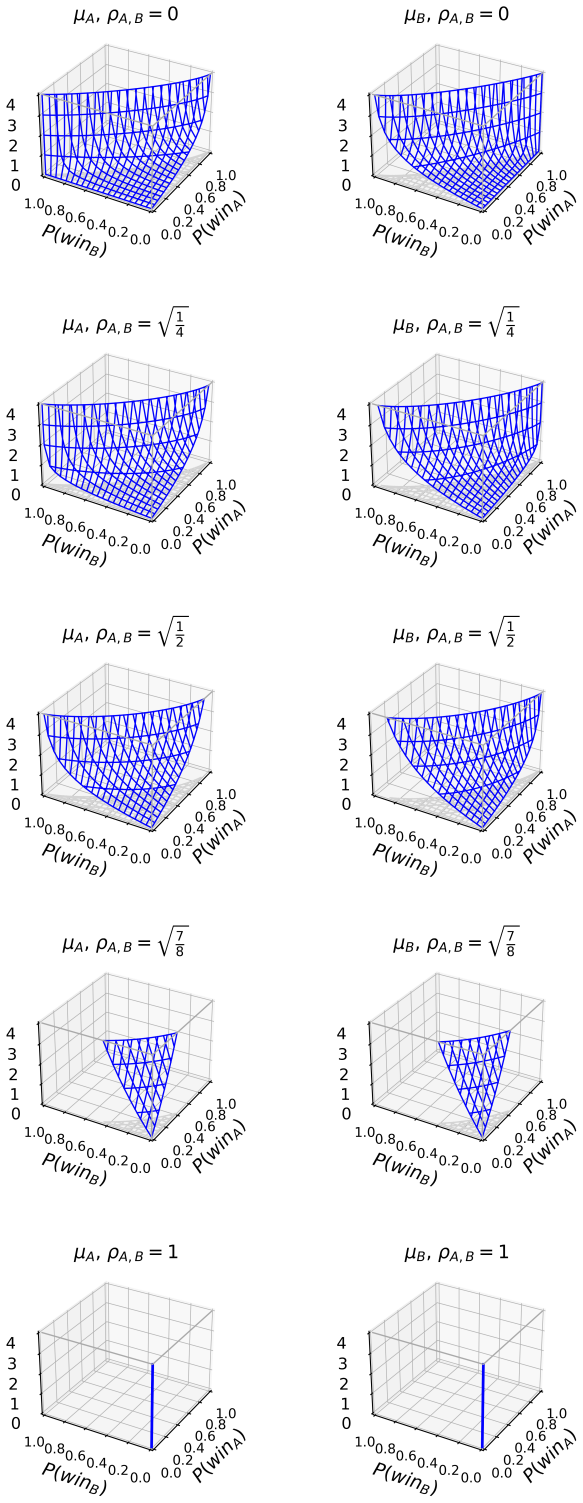

Figure 5: Bivariate Poisson goal-scoring rates, μ_A and μ_B, versus match result probabilities, $P(win_A)$ and $P(win_B)$, for a range of correlation coefficients, $\rho_{A,B}$.

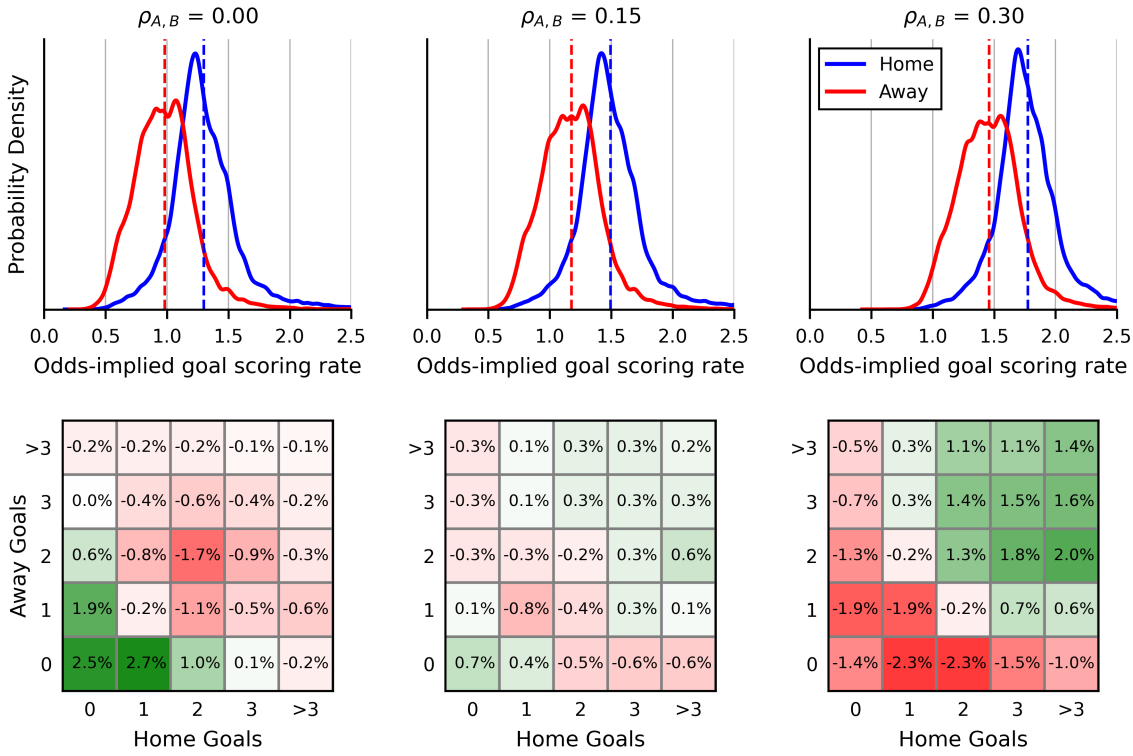

Figure 6: Bivariate Poisson odds-implied goal-scoring rate distributions (top) and predicted minus observed score likelihoods (bottom) for three different correlation coefficients. (For comparison, the average observed home and away scores are 1.46 and 1.14 goals respectively.)

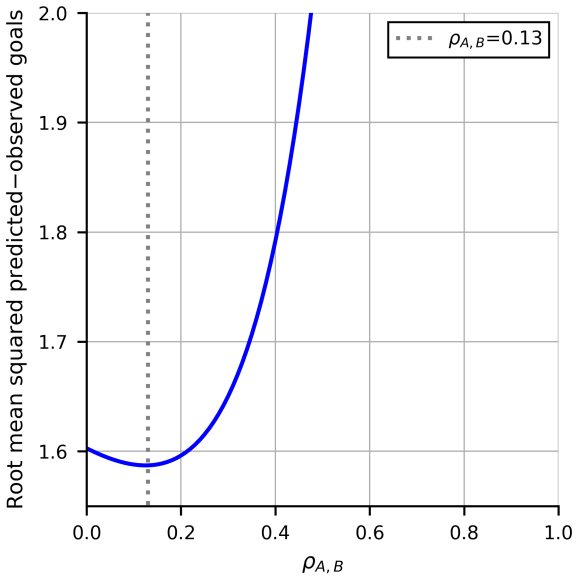

Figure 7: Root mean squared distance between bivariate Poisson predicted goal-scoring rates and observed match goals across all matches versus correlation coefficient.

9. Fitting the best correlation coefficient

For a given correlation coefficient, the inverse bivariate Poisson distribution applied to the odds-implied match result probabilities determines the odds-implied goal-scoring rates. Making the simplifying assumption that the correlation coefficient is fixed over a population of matches, we define the best fit correlation coefficient to be the value for which the odds-implied goal-scoring rates most accurately predict the observed match scores. More precisely, we take the least-squares approach of fitting the correlation coefficient, $\rho_{A,B}$, that minimizes the root mean squares of the distances between the predicted goal-scoring rates and observed match scores across all the matches, as follows:

$$min \left(\frac{1}{n} \sum_{i=1}^{n} ((\mu_A^i - x_A^i)^2 + (\mu_B^i - x_B^i)^2) \right)^{1/2}, \mu_A^i, \mu_B^i = P^{-1}(\pi_{win_A}^i, \pi_{win_B}^i \mid \rho_{A,B}) \tag{6}$$

where matches are numbered 1 to n, x_A^i and x_B^i are the observed number of goals scored in match i, $\pi_{win_A}^i$ and $\pi_{win_B}^i$ are the odds-implied win result probabilities for match i as defined by Equation 2, and P^{-1} is the inverse bivariate Poisson match result distribution function defined by Equation 5. We divide the sum by n and take the square root to normalize to a distance measure that is independent of the number of matches.

Figure 7 shows the root mean square of the distance between the predicted and observed goals, as defined by Equation 6, calculated over our data set for correlation coefficients from zero to one in increments of 0.01. This straightforward iterative computation is feasible even on a moderate desktop workstation, so more sophisticated numerical search techniques are not required. The distance smoothly decreases for correlations coefficients from zero through to best fit of 0.13 then exponentially increases for larger correlation coefficients. This least-squares best fit is consistent with the pattern shown by example in Figure 6 and confirms the benefit of including a positive

correlation coefficient term in a bivariate Poisson goal-scoring model.

10. Evaluating the best fit bivariate Poisson distribution

Finding the best fit correlation coefficient does not in itself indicate whether the fitted fixed correlation bivariate Poisson distribution is a good or bad model of the observed football match scores. Without an alternative goal-scoring model for comparison, it is difficult to assess whether a root mean squared distance of a little less than 1.6 goals achieved at the best fit correlation coefficient of 0.13 is a good fit. To evaluate the quality of the bivariate Poisson model of goal scoring in its own right we take the approach of comparing predicted to observed values for key statistics over the population of matches at the least-squares best fit correlation coefficient. We consider three population statistics:

1. Match result probabilities.
2. Home and away team goals.
3. Exact score probabilities.

These statistics provide measures of the quality of the fitted bivariate Poisson model in understandable aggregate terms and a basis to compare with mispredictions reported in past regression models.

Match result probabilities

By construction, the odds-calibrated model is guaranteed to predict exactly the same match result (win/draw/loss) probabilities as the bookmaker odds for each match, and we have shown that the odds-implied probabilities are accurate predictions in Section 4. It is an arithmetic identity that the fixed-correlation bivariate Poisson distribution (two free variables) can be fitted to odds-implied match result probabilities (two independent probabilities) without losing power with regard to predicting match results. It is not obvious whether this fitted model will accurately predict higher fidelity measures such as goals and exact scores.

Home and away team goals

Since the least-squares best fit minimizes the mean squared distance between the predicted scoring rate and observed goals, we expect the mean absolute predicted goals to be close to observed goals for the best fit correlation coefficient. However, that does not preclude systematic differences when broken down by home and away team, particularly given that the home team scores significantly more goals than the away team on average.

Figure 8 shows predicted and observed mean home and away goal-scoring rates versus correlation coefficient. As we discussed in Section 7, goal-scoring rates increase with increasing correlation coefficient. The predicted goal-scoring rates almost exactly match the observed mean goals scored for both home and away teams for the least-squares best correlation coefficient of 0.13. This demonstrates that a well-calibrated bivariate Poisson distribution with a fixed positive correlation coefficient can be an accurate predictor of mean goal-scoring rates for a population of professional football matches.

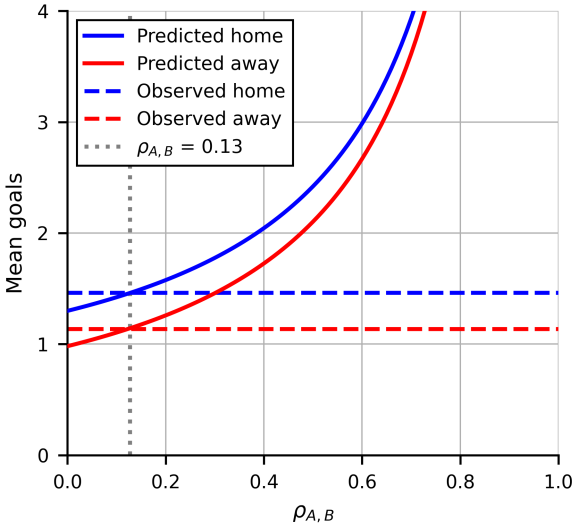

Figure 8: Predicted and observed mean home and away team goal-scoring versus correlation coefficient.

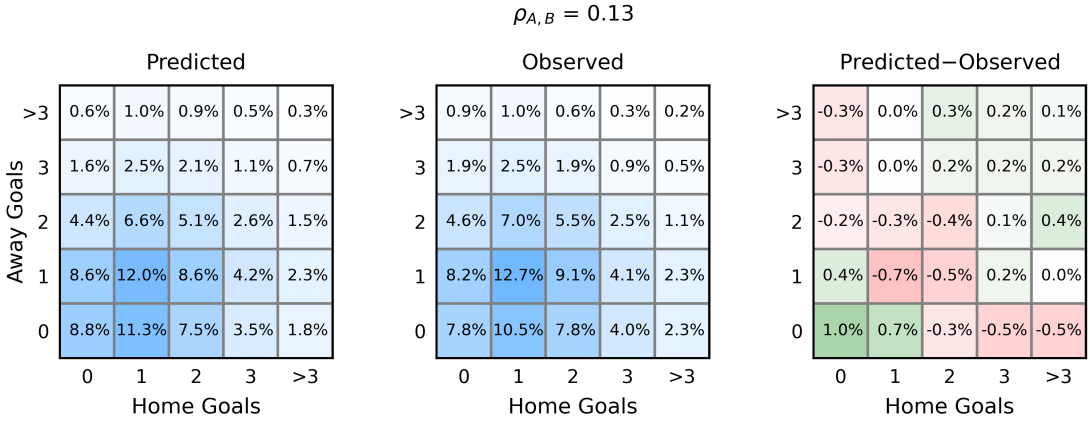

Figure 9: Predicted and observed score likelihoods for the least-squares best fit correlation coefficient of 0.13.

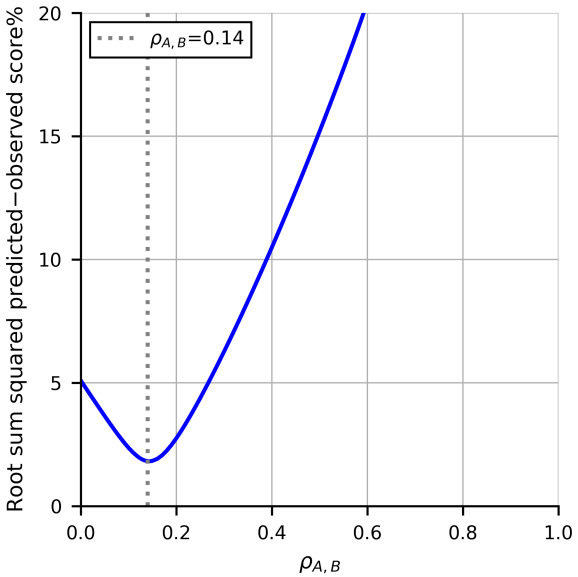

Figure 10: Root mean squared difference between predicted and observed score likelihood across all possible scores versus correlation coefficient.

Exact score probabilities

Predicting the mean goal-scoring rates of home and away teams accurately is one indication of a successful goal-scoring model. But to be useful, the model must also accurately predict the likelihoods of the distinct scores. Figure 9 compares the predicted probabilities of each possible score with the observed frequencies of those scores across all matches for the least-squares best fit correlation coefficient of 0.13. Although we deliberately highlight the differences in the left-most diagram, there is a close correspondence between predicted and observed score likelihoods with most differences less than 0.5% and none greater than 1%. This demonstrates that the bivariate Poisson distribution with a well-calibrated fixed positive correlation coefficient can be an accurate predictor of exact score likelihoods as well as overall goal-scoring rates.

However, there is a low-magnitude pattern of over-prediction of low and high scores and under-prediction of moderate scores highlighted by our color-coding. Figure 10 shows that the root mean squared difference between predicted score probabilities and observed score probabilities across all possible scores is minimized at a correlation coefficient of 0.14 with an almost identical value to that at the least-squares best fit correlation coefficient of 0.13, so a different correlation coefficient would not resolve the issue. Further investigation would be necessary to determine whether this slight systematic deviation from observed scores is attributable to the fixed-correlation bivariate Poisson model or shortcomings of the bookmaker odds used to calibrate the model.

11. Conclusion

The purpose of this paper is to better understand the fundamental applicability of the bivariate Poisson distribution to football match goal scoring and to add clarity to unresolved questions such as the need for and magnitude of the correlation coefficient between goal-scoring processes of the

two teams. Past work that regresses football results onto bivariate Poisson models with the goal of predicting future match results unavoidably conflates their specific regression formulation with the characteristics of the underlying bivariate Poisson distribution, making it unclear whether their conclusions are generalizable.

We take the novel approach of calibrating the bivariate Poisson goal-scoring model using a known source of accurate predictions - bookmaker match result odds. It is not immediately obvious, and we are not aware of it having been described in past work, that for a bivariate Poisson goal-scoring distribution with the simplifying assumption of a fixed correlation coefficient, match result probabilities uniquely determine the goal-scoring rates of the two teams. The primary results of our work are:

- A well-calibrated fixed correlation bivariate Poisson distribution models observed football goal scoring remarkably well. This result confirms the validity of regression from primary factors such as past results onto bivariate Poisson models.
- The bivariate Poisson distribution with positive coefficient is clearly superior to dual uncorrelated Poisson distributions for modeling football goal scoring. For our large data set of match results, the best fit correlation coefficient is 13%.
- We found only small deviations between predicted and observed score frequencies. This suggests that improving the regression formulation will likely be more effective than ad hoc changes to the underlying bivariate Poisson distribution for improving the fit of a regression model.

Validating the bivariate Poisson model of football goal scoring and giving a robust method of determining the best fit correlation coefficient, provides a prior distribution for the development of both traditional statistical and machine learning regression models. The success of the model also provides a basis for future work, in particular:

- Better analysis of intra-match goal-scoring rate changes in response to events such as goals scored and red cards, by normalizing to predicted goal-scoring rates. Traditional analysis based on averages tends to be swamped in statistical noise.
- Investigation of alternative goal-scoring models by also calibrating with bookmaker match result odds then comparing the fit with our bivariate Poisson model. For example, as we discussed in Section 5, the bivariate Poisson model is not accurate over short time periods.

Additionally, although our data set is large and covers a substantial time period, it may not be representative of other competitions, e.g., cup competitions with extra time and penalty deciders, international football, and competitions with inefficient betting markets. Understanding the variability of our results, e.g., the best fit correlation coefficient, would be valuable.

References

Aitkin, A.C., 1936. A further note on multivariate selection. Proceedings of the Edinburgh Mathematical Society 5, 37–40.

Angelini, G., De Angelis, L., 2019. Efficiency of online football betting markets. International Journal of Forecasting 35, 712–721.

Cain, M., Law, D., Peel, D., 2000. The favorite-longshot bias and market efficiency in UK football betting. Scottish Journal of Political Economy 47, 25–36.

Campbell, J.T., 1934. The Poisson correlation function. Proceedings of the Edinburgh Mathematical Society 4, 18–26.

Daunhawer, I., Schoch, D., Kosub, S., 2017. Biases in the football betting market. Available at SSRN: https://ssrn.com/abstract=2977118.

Dixon, M.J., Coles, S.G., 1997. Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics) 46, 265–280.

Elaad, G., Reade, J.J., Singleton, C., 2020. Information, prices and efficiency in an online betting market. Finance Research Letters 35, 607–711.

Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. The Journal of Finance 25, 383–417.

Goddard, J., 2005. Regression models for forecasting goals and match results in association football. International Journal of Forecasting 21, 331–340.

Goddard, J., Asimakopoulos, I., 2004. Forecasting football results and the efficiency of fixed-odds betting. Journal of Forecasting 23, 51–66.

Holgate, P., 1964. Estimation for the bivariate poisson distribution. Biometrika 51, 241–245.

Karlis, D., Ntzoufras, I., 2003. Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistican) 52, 381–393.

Maher, M.J., 1982. Modelling association football scores. Statistica Neerlandia 36, 109–118.

Mogensen, P.K., et al., 2020. NLsolve.jl. URL: `https://github.com/JuliaNLSolvers/NLsolve.jl`.

Moroney, M.J., 1951. Facts From Figures. Pelican Books, London. chapter 8: Goals, Floods, and Horse-kicks - The Poisson Distribution.

Pope, P.F., Peel, D.A., 1989. Information, prices and efficiency in a fixed-odds betting market. Economica 56, 323–341.

Štrumbelj, E., 2014. On determining probability forecasts from betting odds. International Journal of Forecasting 30, 934–943.