Monte Carlo simulation of the 2021-2022 Premier League season (and beyond)

Conor Smith (20209919)

Abstract

In order to simulate the entire season of the Premier League (PL) with Monte Carlo methods, we model the outcome of each match using the basic Poisson model whereby the number of goals scored by each team is assumed to be Poisson distributed. In implementing this model, we apply judgement in: selecting the relevant data and period for our model, estimating the strengths for promoted teams, and investigating the effect of home advantage over a season. Our model is back-tested by simulating last season's results (2020-21), while we also perform a forward simulation of the 2021-2022 season. The final simulation is run over the next ten seasons as an exercise to see how the model behaves over time.

1 Introduction

Prior to a ball being kicked, we see many football pundits and fans alike sharing their predictions as to how the upcoming Premier League will unfold. The event of most interest is informally known as the 'Title Race', that is who will be crowned champions come the end of the season. While those with expert football knowledge can make very credible forecasts on this event, our aim is to create a statistical model that could potentially rival those forecasts.

20 teams compete in the Premier League with each team playing 38 games over the course of a season - that equates to 380 matches in total with the final league positions of each team determined by the outcome of these matches. We know that each team's final position in the league is a function of the results that team obtains over the season. Therefore to construct predictions for each team's position in the final league table, we need a method that will determine the outcome of each team's fixtures (i.e determine the outcome of all 380 fixtures in the Premier League). A popular approach is to assume a Poisson process for the number of goals scored by each team in a game. This will give us a probability distribution for the number of goals scored by each team during the game from which we can then derive the probability distribution for the match outcome (i.e probability distribution for home win, draw or away win)

Creating a probability distribution for each individual fixture isn't enough, however, to make a forecast for the final league table. Calculating the probability of any team becoming champions deterministically would be a tedious task given the number of permutations involved. Consequently, we apply a simulation based method known as Monte Carlo. Monte Carlo allows us to estimate the probability distribution of an event by repeated sampling. To do this, we simulate each individual fixture by sampling from the Poisson distribution to get the number of goals scored by each team (which we can do easily), determine the match outcome and then update the table based on the match outcome. By repeating this process a large number of times, we can infer probability distributions for events come the end of season (i.e probability of relegation or probability of being champions) based on the results from the simulations.

2 The Basic Poisson Model

The most popular approach used to model the outcome of football matches in the existing literature is based on modelling the number of goals per a game using the Poisson distribution. The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

We can say a discrete random variable X is said to have a Poisson distribution, with parameter $\lambda > 0$, if it has a probability mass function given by:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

One of the most cited publications in this field is Maher (1982)[2] where he proposed independent Poisson distributions for the number of goals scored by each of the home and away teams to model the outcome of matches, with means that are specific to each team's past performance. Maher (1982) also proposed that the home team's scoring rate is a function of an additional parameter that represents 'home advantage'. This is related to the empirical finding that teams win more at home than away on average. This parameter is the same for all teams and can be interpreted as the additional increase in the home team's scoring rate after controlling for the attacking and defensive strengths of both teams.

In a match between teams indexed i and j, let $X_{i,j}$ and $Y_{i,j}$ be the number of goals scored by the home and away sides respectively. Then

$$X_{i,j} \sim Poisson(\alpha_i \beta_j \gamma),$$

$$Y_{i,j} \sim Poisson(\alpha_j \beta_i),$$

where $X_{i,j}$ and $Y_{i,j}$ are independent, $\alpha_i, \beta_i > 0 \ \forall i$, the α_i measures the attack rate of two teams, the β_i measure the defence rates and $\gamma > 0$ is a parameter that allows for home advantage.

Using the Poisson distribution to model goals scored during a football match can be considered over-simplistic as we can think of counter-examples that would violate the assumptions of this model. For example, a team may employ more defensive tactics upon scoring a goal. This would potentially shift the mean rate of goals scored during a game (which is assumed to be constant) and introduce some dependence on the events (e.g if team A scores, then team A is now less likely to score again etc).

To judge the suitability of the Poisson distribution for our task, we produce a histogram of the number of goals scored per game for champions Manchester City using data from the 2020-21 Premier League season. Figure 1 shows the frequency of the goals Man City scored away-from-home last season while Figure 2 shows the frequency of the goals Man City scored at home. We overlay the best-fit Poisson distribution for both sets of data to compare.

We can already see the potential limitations of using the basic Poisson model from these plots. While the Poisson distribution appears to capture Manchester City's away goals per game reasonably well, it performs poorly at capturing Manchester City's home goals per game. From inspection, it appears Manchester City's home goal data is relatively over-dispersed (i.e variance is greater than the mean) which the Poisson distribution cannot allow for given its assumptions.

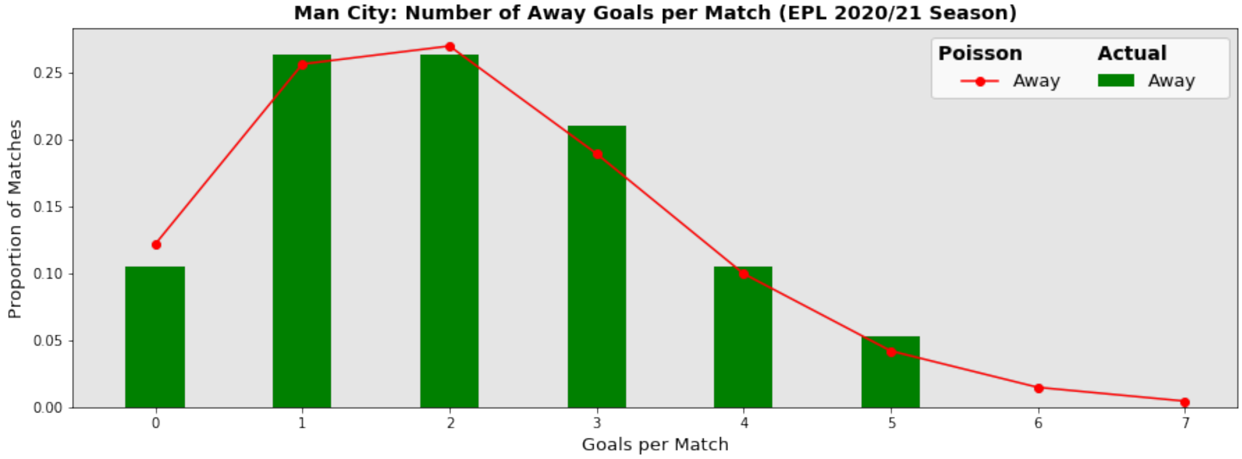


Figure 1: Best fit Poisson distribution for Manchester City goals per away game during the 2020-21 Premier League Season.

Despite our concerns, the Poisson distribution provides a reasonable approximation of the data-generating process on average, so we will apply the model set out by Maher(1982) for our Monte Carlo simulation. The parameters which define this model can be estimated efficiently using what's known as a Poisson regression model which we discuss further in Section 3.

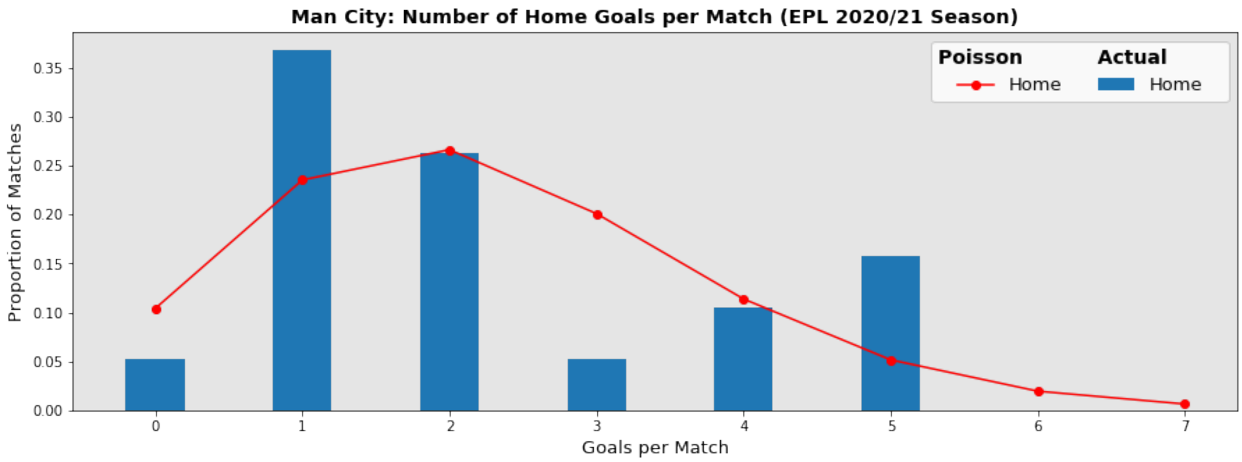


Figure 2: Best fit Poisson distribution for Manchester City goals per home game during the 2020-21 Premier League Season.

3 Model Choice and Model Inference

Our task is to simulate the 2021-22 Premier League and subsequent seasons using an appropriate model. In simulating subsequent seasons (i.e 2022-23 season onwards), we need to account for teams who exit the league (due to relegation) and enter the league from the division below (promoted teams). To determine which teams get promoted, we ultimately need to simulate The Championship in addition to the Premier League. Therefore, the following model inference is applied to all teams in the Premier League and The Championship. For our purposes, we assume promotion to The Championship is closed so we only account for movement of teams between the Premier League and The Championship.

3.1 Data

To fit the model put forward by Maher(1982), we must decide 1) what set of match results to include to estimate the parameters and 2) over what time period should we select our historic match results. Both of these choices are somewhat subjective so we will use footballing judgement to guide our choice for these settings.

Besides the league, teams compete in two other domestic cup competitions; the League Cup and FA Cup while the top Premier League teams also play in European competitions, so there are multiple sources of football results to choose from. In recent years, winning the cup competitions has become less prestigious with teams often fielding weakened teams so as to prioritise their performance in the Premier League[1]. Therefore, a team's performance in the

[1] Source: https://bleacherreport.com/articles/114125-the-glory-of-the-fa-cup-is-dead-and-buried-as-teams-prioritise-whats-important

cup competitions is not necessarily a reflection of their performance in the Premier League (for example, Manchester United won 5 Premier League titles during the period 2006-2014 while they didn't win a single FA Cup). Consequently, it seems prudent to only use league match results (Premier League and Championship) in our data set if our goal is to simulate the Premier League/Championship only.

The period from which we should choose the set of league results is a trickier decision. Ideally, we want to select a period which would adequately capture the strength parameters reflecting each team's ability going into the new season. So, if our time period is too short (e.g each team's last 10 games) we risk basing attacking/defensive strengths on temporary form rather than underlying quality while if our time period is too long (e.g 5 years) then we risk basing attack/defensive strengths on results that are no longer relevant for a team (e.g different players, different manager).

Dixon and Coles (1997)[1] who modified the original model by Maher(1982), arbitrarily chose the three most recent seasons to estimate each team's strengths (pp.267). Alternatively, we could treat the number of seasons as a hyper-parameter which we could tune with cross-validation. However, we take a pragmatic approach by only including the most recent season's match results in the data set. Using some football judgement, it seems intuitive that each team's strength for the upcoming season should be based exclusively on its most recent season's performance. Including one season of data also allows us to incorporate an additional model for promoted teams (Section 3.4) which would be more difficult to implement with more than one season's worth of data.

3.2 The Poisson Regression Model

Having decided on our data set and that the number of goals scored during a game follows a Poisson process, we now need to estimate each team's mean rate of goals scored for each match. A popular approach is to fit a Poisson regression model to our data where we estimate the means based on some covariates. If our response variable *lambda* is assumed to be Poisson distributed, then we can express $log(\lambda)$ as a linear function of a set of independent variables (i.e covariates which we think affect λ).

Applying this to our problem, consider a match between team i and team j. The rate at which team i scores goals in this match will depend on: 1) the (attacking) strength of team i, 2) the (defensive) strength of team j and 3) whether team i is at home or not. The Poisson regression model is then:

$$log(\lambda_i) = \alpha_0 + \alpha_1 Team_i + \alpha_2 Opponent_j + \alpha_3 Home_i + \epsilon_i$$

$Team_i$ is the team whose mean we wish to estimate (categorical variable), $Opponent_j$ is the opposition for $Team_i$ (categorical variable) while $Home_i$ is a binary variable for whether $Team_i$ plays at home or not. λ_i is the rate at which $Team_i$ scores during the game against $Opponent_j$. α_1 is a parameter which captures the attacking strength of $Team_i$ while α_2 captures the defensive strength of the opponent of $Team_i$ ($Opponent_j$). Note that while we say that α_2 represents the defensive strength of the opposition, a larger positive value for α_2 corresponds to a lower defensive strength for the opposition (and vice-versa).

We fit the model to the data set described in Section 3.1 using the 2020-21 season's results as our data set (it was not possible to include this output in the Appendix, however the model summary can be viewed from executing the script on Github). Using the coefficients, we can compute the mean rates for two teams in a given match. For example, we can estimate the means when the top two teams in the Premier League play each other: Liverpool (LIV) vs Man City (MC) with Liverpool playing at home. To estimate Liverpool's mean, our 'team' variable is Liverpool, our 'opponent' variable is Man City, while 'Home' is equal to 1. Taking the exponential of both sides of the fitted Poisson regression model yields:

$$\lambda_{LIV} = e^{0.0406+0.2158-0.1694+0.0976} = 1.20$$

To get Manchester City's mean rate for this match, we just reverse the values of the team and opponent variables and set home equal to 0.

$$\lambda_{MC} = e^{0.0406+0.4053+0.0877} = 1.71$$

So, for this particular match Man City will have a higher mean goal rate per game than Liverpool. It's worth noting that the mean rate at which each team scores will vary according to the opposition and whether the team plays at home or not. So in a sense, the rate at which a team scores is dynamic to the specific match the team is playing and is not constant over matches. If Man City plays at home against a weaker team (say Crystal Palace) then the mean rate for Man City for this particular match would be:

$$\lambda_{MC} = e^{0.0406+0.4053+0.5130+0.0976} = 2.88$$

3.3 Home Advantage

In addition to the attacking and defensive strength parameters, we must also fit the parameter representing the 'home effect' γ. To see how the 'home effect' behaves over time, we fit our Poisson regression model for every Premier League season since its inception and record the parameter estimate for each season. We produce a scatterplot of the 'home effect' over time in Figure 3.

From Figure 3, we generally see the 'home effect' ranges between 0.2 and 0.4 with no trend over time, however there is one noticeable outlier, the 2020-21 season, where there is practically no home advantage at all. We can attribute this outlier to the fact that games were

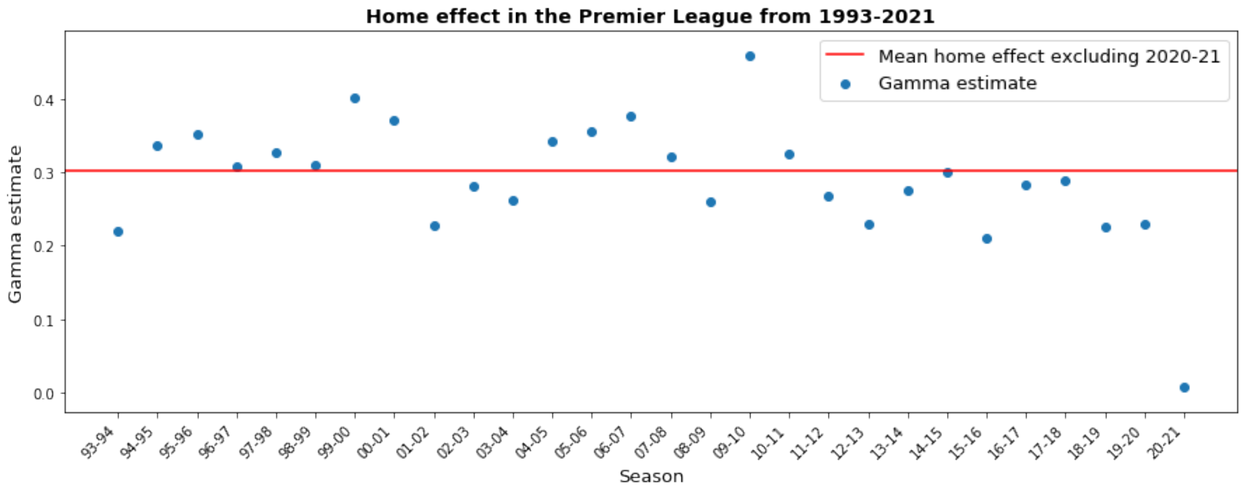


Figure 3: Gamma (home effect) parameter estimates in the Premier League from 1993-2021

played in empty stadiums over the 2020-21 season due the Covid-19 pandemic so home teams likely struggled to perform without the backing of their fans. With the UK now exiting lockdown and no capacity restrictions on attendance for the upcoming 2021-22 Premier League season, we would expect the ’home effect’ to return for home teams. While we use last season’s data to estimate the team strengths, we should not use last season’s ’home effect’ estimate in our model for simulating this season’s matches.

As the data appears stationary with no obvious trend, a suitable approach here would be to take the mean ’home effect’ estimate over every season, excluding the estimate for the 2020-21 season, and use this estimate to fit our Poisson regression model when simulating the 2021-22 season. This value corresponds to the red line in Figure 3 and equals approximately 0.3. As each team plays an equal number of home and away games over the season, the choice of gamma is likely to not have much effect on a team’s standing in the final table, however the choice of gamma will have a noticeable effect on each individual football fixture that we simulate.

3.4 Treatment of Promoted and Relegated teams

Every year, the three teams who finish in the bottom three positions of the final Premier League table are relegated to the division below while three new teams are promoted into the Premier League. The Poisson regression model that we’ve specified will estimate attack/strength coefficients for each team in the Premier League and The Championship, however, the attacking/defensive strengths of the promoted teams will be biased upwards as their strengths have been estimated using teams in the Championship and not in the Premier League. So we need a method to downweight the original attacking/defensive estimates for the promoted teams so as to capture the jump in difficulty from the Championship to the Premier League.

We investigated how promoted teams fared in their first season of the Premier League historically by comparing their average goals scored/conceded per game during their promotion winning season in the Championship with their corresponding goal statistics the subsequent year in the Premier League. We produce a scatterplot of the goal scoring rates for every promoted team during their promotion winning season versus their goal scoring rate the subsequent season in Premier League in Figure 4.

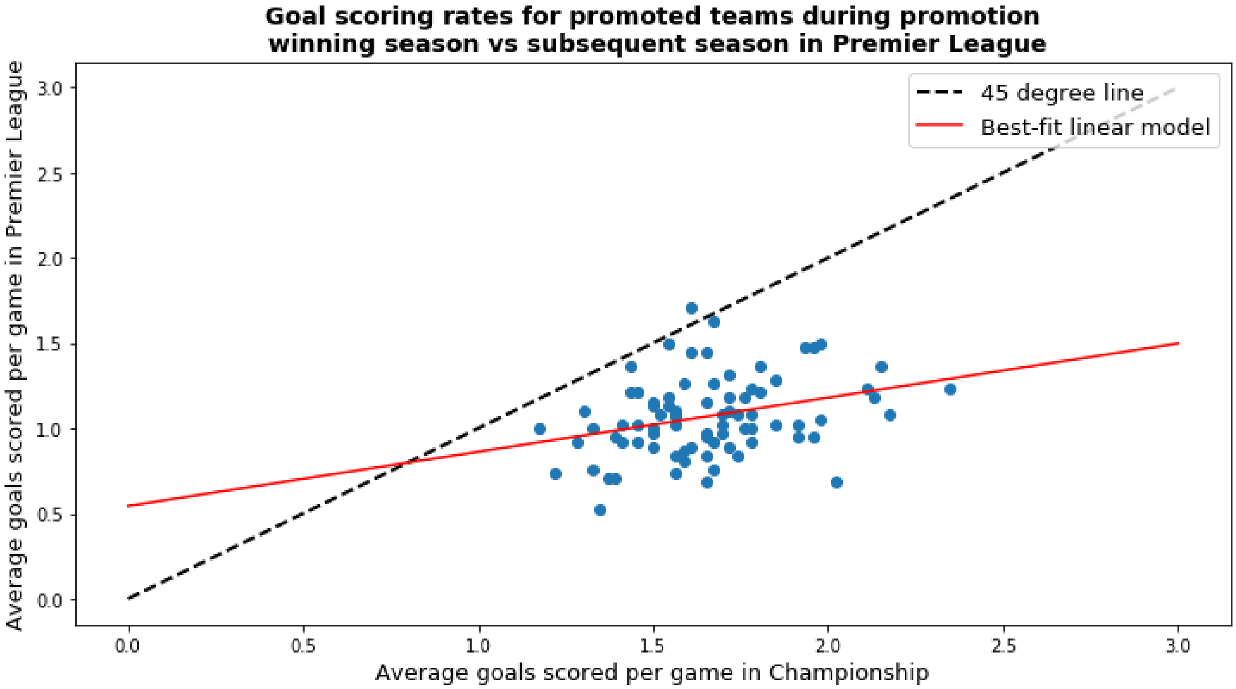


Figure 4: Goal scoring rates for every promoted team during their promotion winning season versus their goal scoring rate the subsequent season in Premier League

As expected, we see that each promoted team fared worse in the Premier League versus their prior season in The Championship as each team's average number of goals scored per game is lower in the Premier League versus the Championship (i.e practically all data points are below the 45 degree line). Based on the observed data, we fit a linear model using Least Squares to see if the relationship is statistically significant:

$$\lambda_{i_{t+1}PL} = \alpha_0 + \alpha_1 \lambda_{i_tC} + \epsilon_i$$

where $\lambda_{i_{t+1}PL}$ corresponds to the average number of goals scored per game by the promoted team in the Premier League and λ_{i_tC} corresponds to the average number of goals scored per game by the promoted team during their promotion winning season in the Championship. This model will reduce the rate at which a promoted team scores against a non-promoted team in the Premier League. We produce the output from this fit in Appendix (Figure 9) where we see both the slope and intercept coefficient are statistically significant from 0 at the 5% significance level.

We fit a similar model for the average number of goals that each promoted team **conceded** in their promotion winning season against the same statistic in their subsequent season in the Premier League, details of which are available in the Appendix (Figure 10). This model will increase the rate at which the (non-promoted) opposition scores against a promoted team in the Premier League.

We simply invert these models to adjust the rates for the teams who were relegated to The Championship (i.e need to upweight these teams strengths when they compete in The Championship).

4 Monte Carlo Simulation and Results

Our Monte Carlo approach will involve simulating the number of goals scored by the home team and away team for each individual fixture over the season. We use this data in determining the overall match result (i.e home win, draw or away win). For each fixture, the simulation works as follows:

1. Calculate the means (i.e rate parameter in the Poisson model) for both teams using the fitted Poisson regression model, applying adjustments to promoted/relegated teams for fixtures involving these teams.
2. Simulate the number of home and away goals scored by the home and away teams using two independent Poisson distributions corresponding to the means estimated from step 1. We use a function from the numpy package in Python to do this rather than derive a technique from first principles.
3. Determine the overall match result using the simulated goals from step 2 (e.g if home team scores more goals than away team, assign a win for the home team). Update the league table based on the match outcome.
4. Repeat for every fixture in the fixture list.

The procedure outlined above corresponds to one iteration so in order to perform Monte Carlo simulation, we perform the above procedure a large number of times. Given the computational burden of this simulation, we used only 1,000 iterations in our Monte Carlo simulation, though we would have increased the number of iterations in ideal circumstances (e.g 100,000).

4.1 Back-test of 2020-21 PL season

Before using our model to simulate forward in time, we can measure the accuracy of the fitted model by simulating a previous footballing season and then comparing our results with the true historical results. This approach can be seen as a form of cross-validation: our model will be fitted using data from season t (training set), we use this data to simulate season $t+1$, which we then compare with the true results for season $t+1$ (test set). The idea here is that by applying cross-validation, we may encounter issues which we may need to address before running the simulation forward in time.

We decide to simulate the previous Premier League season (2020-21 season), fitting our model using data from the 2019-20 season and then comparing our simulated league table with the historic final league table from 2020-21. We present our comparison in the form of a boxplot in Figure 5, where the distribution of each team's final points tally is compared with each team's actual points tally for that season. A final league table based on the average of all simulations is also available in the Appendix (Figure 12)

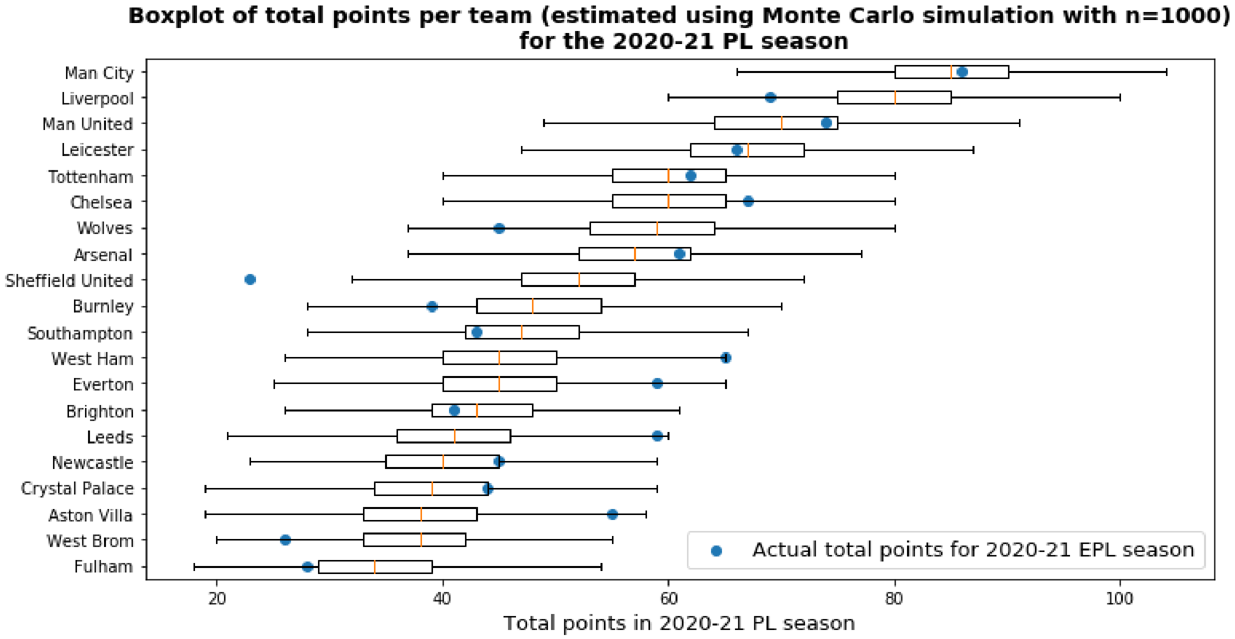


Figure 5: Total points per team (estimated using Monte Carlo simulation with n=1000) for the 2020-21 PL season

Despite Liverpool accumulating a massive 99 points during their title winning season in 2019-20 and Manchester City finishing a considerable 18 points behind Liverpool in second position [2], our model predicted Manchester City to be champions and Liverpool in second position. While Liverpool managed to only finish third in reality during 2020-21, we would have expected Liverpool to finish top in our simulation given the model was fit exclusively to data in the season which Liverpool won the league. The reason for this behaviour comes down to the fact that we have modelled the strength of teams based on their goal statistics rather than their points or position in the table.

To see this, consider the 2019-20 season where Man City scored an impressive 102 goals con-

[2] All final Premier League tables can be viewed at https://www.premierleague.com/tables

ceding 35 goals while Liverpool managed to score 85 goals and concede 33 goals. Therefore, our model would favour Man City over Liverpool given their superior goal scoring statistics. So, it appears that the goal difference (i.e the difference between the total goals scored and total goals conceded) of each team plays a big role on where that team ultimately finishes in the table in our model.

The specific model used to adjust the strengths of the promoted teams appears to have done quite well in our case. Two of the promoted sides, West Brom and Fulham, finished in the bottom three during the 2020-21 season while the average positions for these teams in our Monte Carlo simulation were 19th and 18th respectively (i.e bottom three). Furthermore, our model also predicted that Leeds United, the third promoted team, would not get relegated (Leeds finished in 9^{th} position during the 2020-21 season while Leeds' average position was 14^{th} in our Monte Carlo simulation). All in all, this piece of the overall model performed reasonably well.

There is one team in our boxplot who is a noticeable outlier - Sheffield United. Sheffield United enjoyed an impressive season in the top-flight during 2019-20 where they finished in ninth position but endured a disastrous following season, finishing rock bottom with 23 points. Injuries to key players, the loss of their goalkeeper to Manchester United and off-the-pitch matters all played their part in Sheffield United's torrid season. Despite these factors, Sheffield United were still expected to have a comfortable season in the Premier League with an expected finishing position of 14th before the Premier League started as forecasted by famed statisticians FiveThirtyEight[3]. So our model wasn't unique in failing to correctly predict Sheffield United's season.

4.2 Forward simulation of the 2021-22 PL season

We now put our model into action to simulate the 2021-22 Premier League season. This time around, however, we do not have any historical data to benchmark our results. So we adopt a different approach to measure success by comparing the probability of events implied by our Monte Carlo simulation to the bookmaker's implied probabilities for these same events. Using Monte Carlo methods, we can estimate the probability for events of interest by simply counting the number of iterations where the event occurred and dividing by the total number of iterations. For example, if our event is "Man City to win the Premier League", we can compute this probability by counting the number of times Man City finished top in our simulation and divide this by the total number of iterations.

Our best source for implied probabilities is the Betfair Exchange, rather than any traditional bookmaker. The Betfair Exchange behaves like the stock market where people can lay or back events at the market prevailing odds (much like how people can buy and sell stocks at the current price), where the odds are determined by market forces (i.e supply and demand).

[3] Historic predictions available at https://projects.fivethirtyeight.com/soccer-predictions/premier-league/

Just like the stock market with the efficient market hypothesis, the prices available at Betfair reflect the true price/odds of those events happening (in theory). There is no margin built into these odds (unlike a traditional bookmaker) which makes the implied probabilities from the Betfair Exchange comparable to the probabilities we estimate with Monte Carlo simulation.

The main event of interest in any Premier League campaign is who will be crowned champions come the end of the season. Figure 6 shows the probability of each team winning the Premier League estimated with our Monte Carlo model versus the corresponding probabilities available on Betfair.[4]Our model has Man City as heavy favourites with a probability of 0.70 of winning the title. The Betfair Exchange also has Man City as favourites but the implied probability is just 0.55. Betfair has given higher probabilities to Chelsea, Liverpool and Manchester United winning the Premier League relative to our estimated probabilities.

History tells us that successfully defending the Premier League title has proved to be much harder than claiming it. Only once in the last 12 years has a team successfully retained the title having won it the season before - co-incidentally this was also Man City who won the league in 2017-18 and then in 2018-19.[5] We can hypothesize that the odds quoted on Betfair reflect this peculiarity based on the differences we see between our Monte Carlo estimated probabilities and those quoted on the Betfair Exchange.

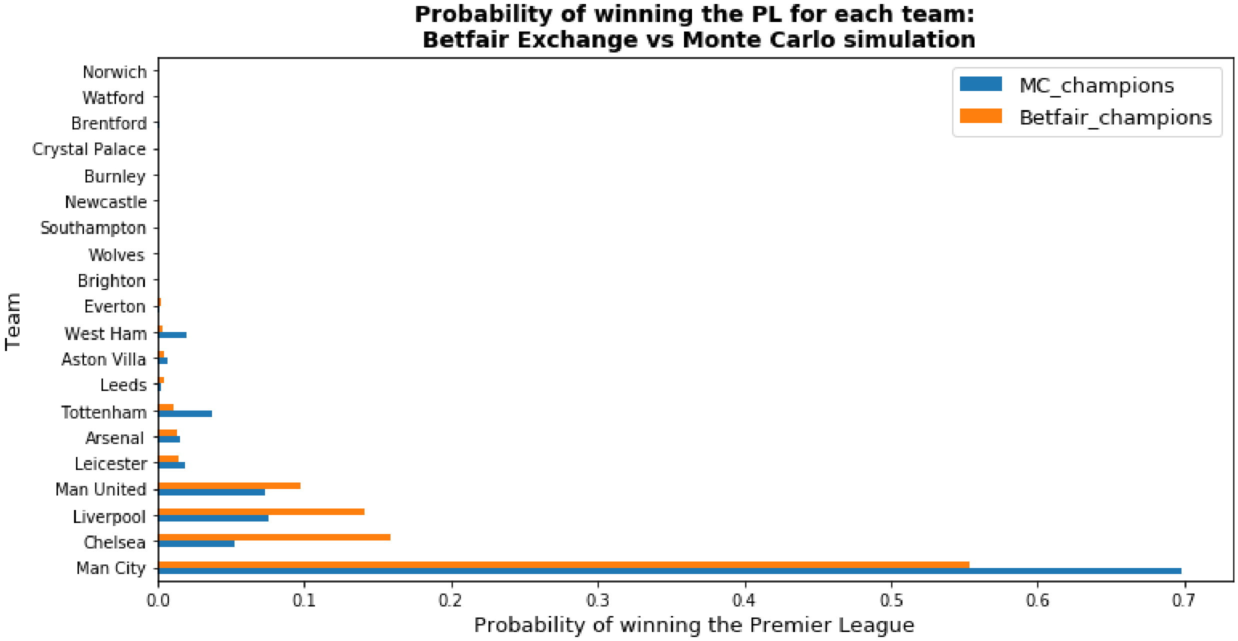


Figure 6: Probability of winning the PL: Monte Carlo vs Betfair

We turn our attention to the bottom of the table where we estimate the probability of each team being relegated from the Premier League (i.e finishing in the last three positions in the

[4] All odds were retrieved on 11/08/21 from the Betfair Exchange website

[5] Source: https://www.myfootballfacts.com/premier-league-winners/

table 18^{th} - 20^{th}), with our estimates in Figure 7. The first thing to note is that our model has underestimated the probability of relegation for the three promoted teams (Norwich, Watford and Brentford) relative to the probabilities at Betfair. Furthermore, all three teams do not finish in a relegation place when looking at the average table for this simulation which can be seen in the Appendix (Figure 13).

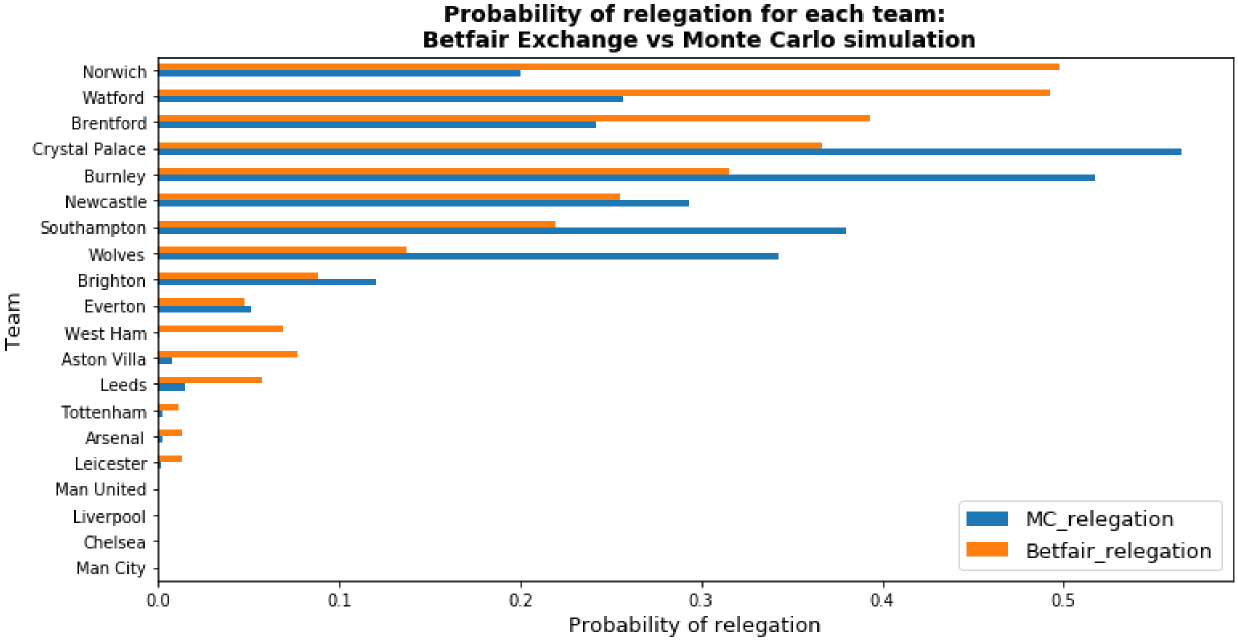


Figure 7: Probability of relegation from PL: Monte Carlo vs Betfair

We can use some footballing knowledge to explain these differences. Norwich City won the Championship last season with 97 points which is an impressive total points tally for this particular division. Given how we adjust the strengths of the promoted teams using the goal statistics in their promotion winning season, it's not unreasonable to expect Norwich to have an average finishing position of 13^{th} (i.e not in a relegation position) given their impressive goal statistics during the 2020-21 season in which they won promotion.

However, we can see why the betting public (i.e Betfair) would not be so optimistic. Norwich have sold their best player during the summer who heavily contributed to the amount of goals they scored during 2020-21 (so these goal statistics which we calibrate Norwich's strength in our model are biased upwards for the 2021-22 season). Similarly, Norwich finished bottom of the Premier League the last time they competed in this division only two years ago (which we do not capture in our model), so this historical event would likely reduce peoples expectations of Norwich performing well this season.

A similar narrative can be applied to Brentford and Watford who had impressive seasons in The Championship (relative to previous promotion winning sides) but are favourites with Norwich to go down amongst the betting public. The reasons are likely related to the fact that neither of these teams have spent money to improve the quality of their playing squad (rather than on their previous season's performance). Despite the bleak outlook in store for these teams by the public, two of the promoted teams have **stayed up** on average over the last ten years so it will be interesting to see how these sides perform over the season.

4.3 Forward simulation of the PL for the next ten seasons

Our last task is to simulate the Premier League for the next ten seasons and to see how the league evolves over time using our model. For this particular task, we don't have much options available to us in the form of a benchmark (you won't find a bookmaker offering odds for the Premier League champion in the 2031-32 season!) so the focus of our analysis will be on the patterns and trends we observe from one season to the next.

For this simulation, we make additional modifications from the single season simulation scenario:

1. We allow for promotion and relegation to and from the Premier League from one season to the next. The three teams that finish bottom of the PL in one season are relegated to The Championship for the next season while the three teams that finish top of The Championship are promoted to the PL as well. As a result, we implicitly simulate The Championship in order to determine the promoted teams each season.
2. We model each team's strengths for a given season by fitting a Poisson regression model to data within the previous season only. For example, when simulating season $t + 4$, we estimate each team's strength for season $t + 4$ using match results from season $t + 3$ only. Our approach can be interpreted as a Markov Chain with the next state depending on the previous state only.

Given the fact we allow for promotion and relegation, it doesn't make too much sense to compute an average table for each of the seasons. However, we can look at the estimated probability distribution for the league winners for each of the ten seasons. For each season, Figure 8 shows the "expected" league winners where "expected" is defined to be the mode of the league winner probability distribution for that season (i.e the team that finished top more than any other team that season). We plot the expected league winners probability of winning the league for each season as well.

There are two observations to make based on Figure 8. Firstly, Man City are the most likely team to win the Premier League in each of the next ten seasons. This probably isn't too surprising given that Man City were the clear favourites to win the league in 2021-22 as seen in Section 4.2 while our model bases each team's strength exclusively using the previous season's results. Secondly, despite Man City being the most likely team to win each season, the probability of Man City winning the league decreases over time. Again, this observation is intuitive; our best guess to win the league in ten seasons time is still Man City, however,

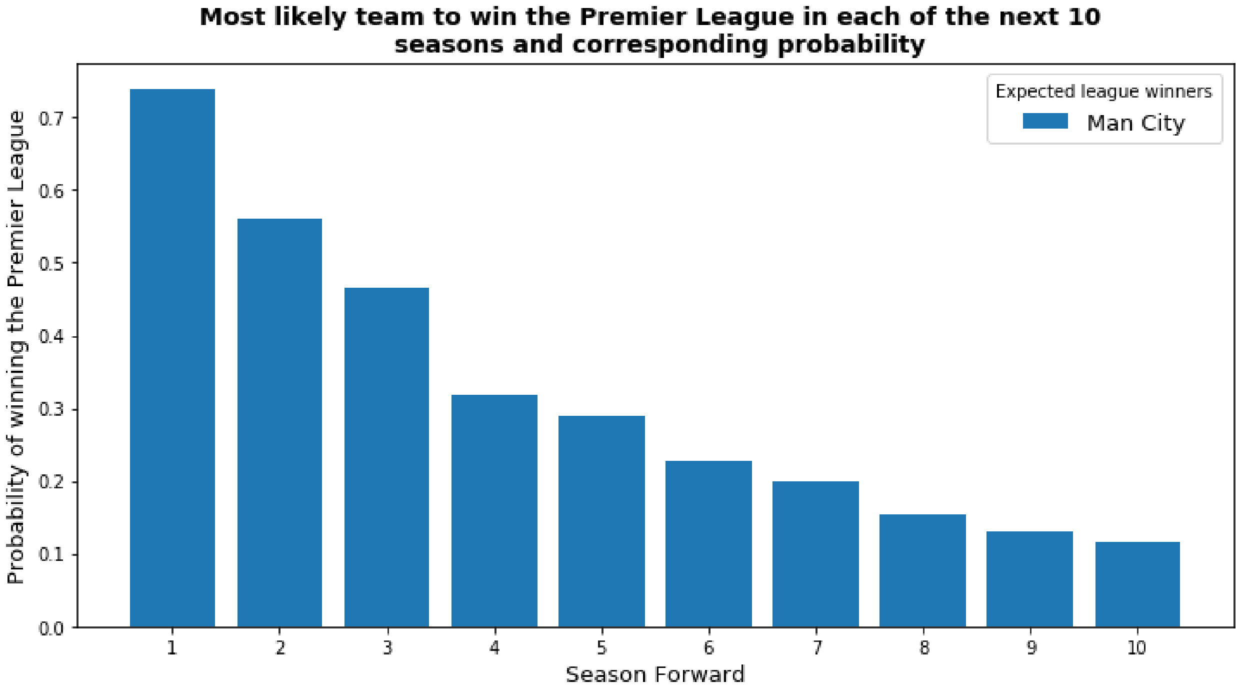


Figure 8: Expected league winners for each of the next ten seasons and corresponding probability

we are much more uncertain of the outcome. Contributing to this uncertainty is the fact we allow promotion and relegation, where in a small few number of our iterations a promoted team finished top of the league in season 10. To use some statistical terminology, we can say that the entropy of this probability distribution increases over time.

5 Conclusion

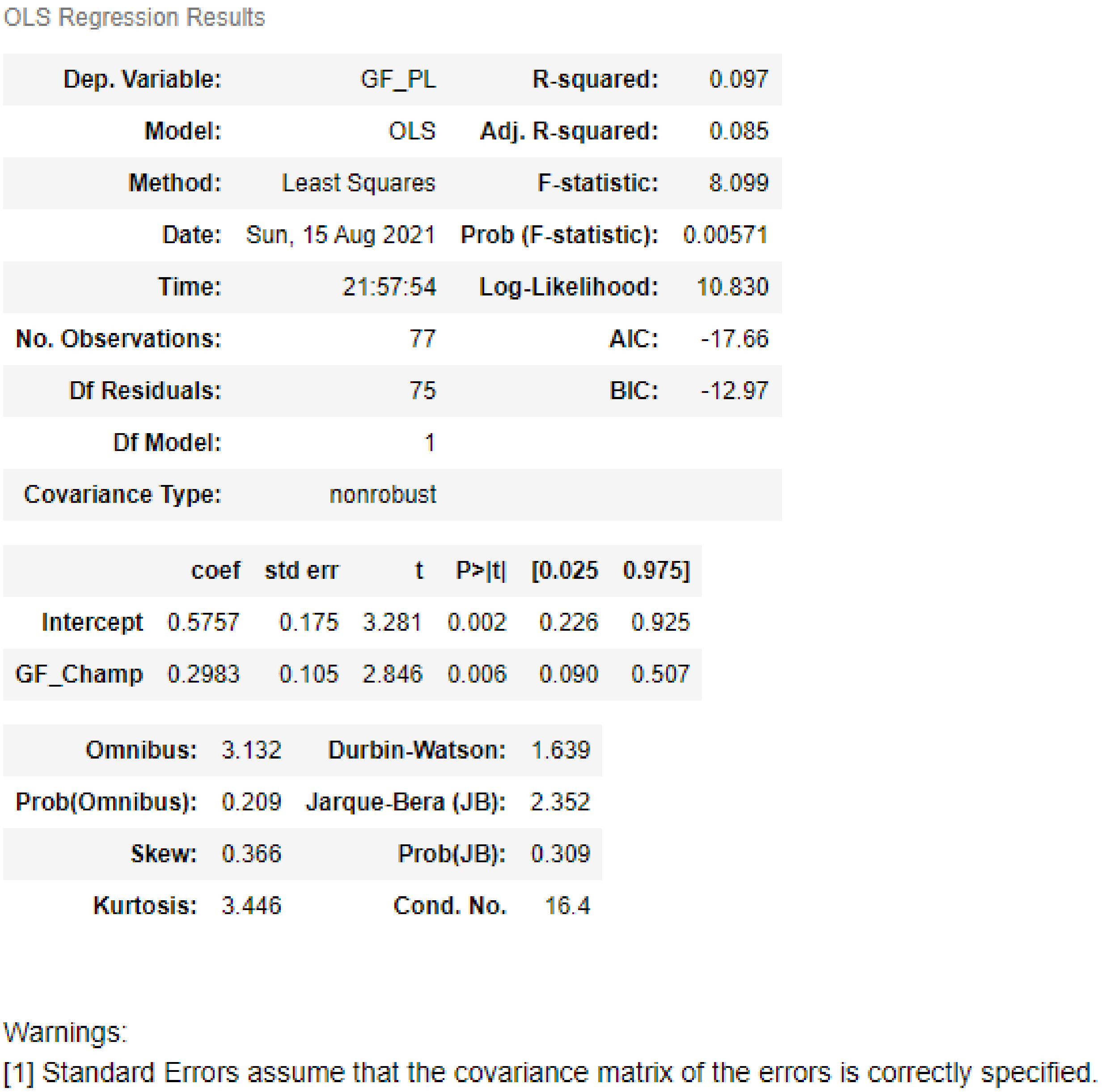
We have presented a method to model the outcome of individual games which formed the basis for our end of season forecasts with Monte Carlo simulation. The Poisson model used to simulate the match outcomes is relatively basic with potential improvements possible by considering additional covariates for each team. For example, we could incorporate the total transfer value of each team as an additional covariate that affects each team's strength. While basic, our model has provided reasonable end-of-season forecasts in both the one season and ten season simulations.

References

[1] Mark J Dixon and Stuart G Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280, 1997.

[2] Michael J Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

A Appendix



OLS Regression Results

Dep. Variable:	GF_PL	**R-squared:**	0.097
Model:	OLS	**Adj. R-squared:**	0.085
Method:	Least Squares	**F-statistic:**	8.099
Date:	Sun, 15 Aug 2021	**Prob (F-statistic):**	0.00571
Time:	21:57:54	**Log-Likelihood:**	10.830
No. Observations:	77	**AIC:**	-17.66
Df Residuals:	75	**BIC:**	-12.97
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.5757	0.175	3.281	0.002	0.226	0.925
GF_Champ	0.2983	0.105	2.846	0.006	0.090	0.507

Omnibus:	3.132	**Durbin-Watson:**	1.639
Prob(Omnibus):	0.209	**Jarque-Bera (JB):**	2.352
Skew:	0.366	**Prob(JB):**	0.309
Kurtosis:	3.446	**Cond. No.**	16.4

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 9: Fitted model summary for goal scoring adjustment model for promoted teams

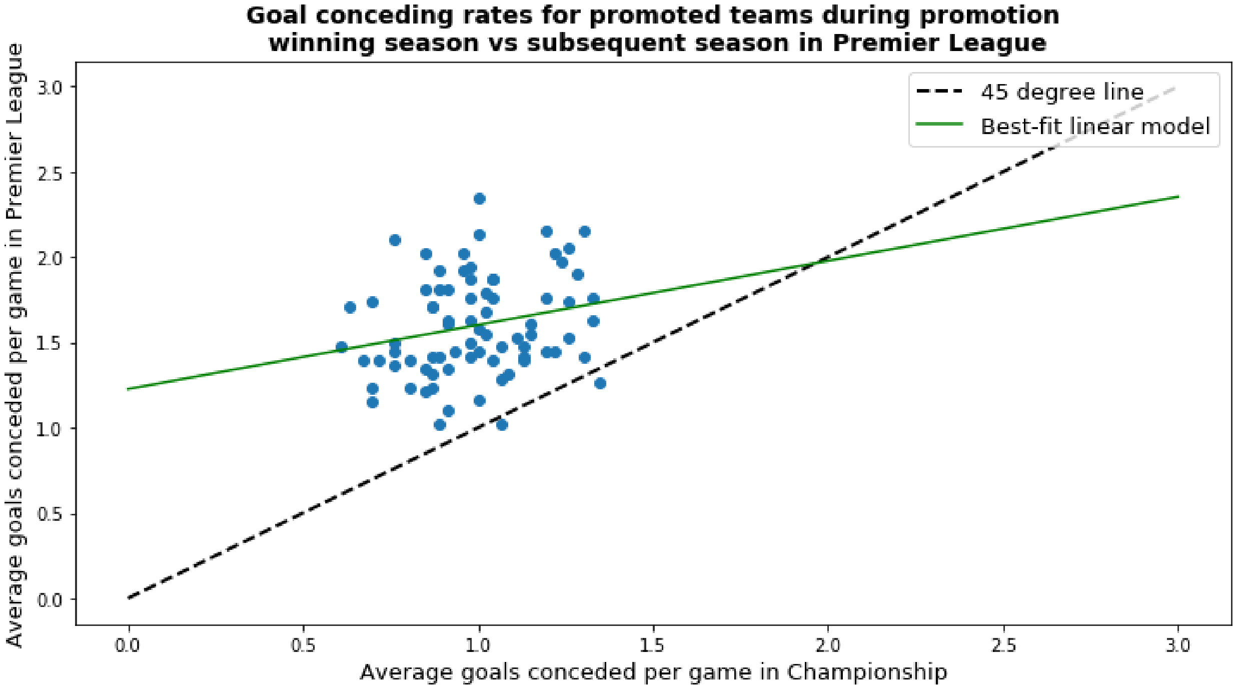
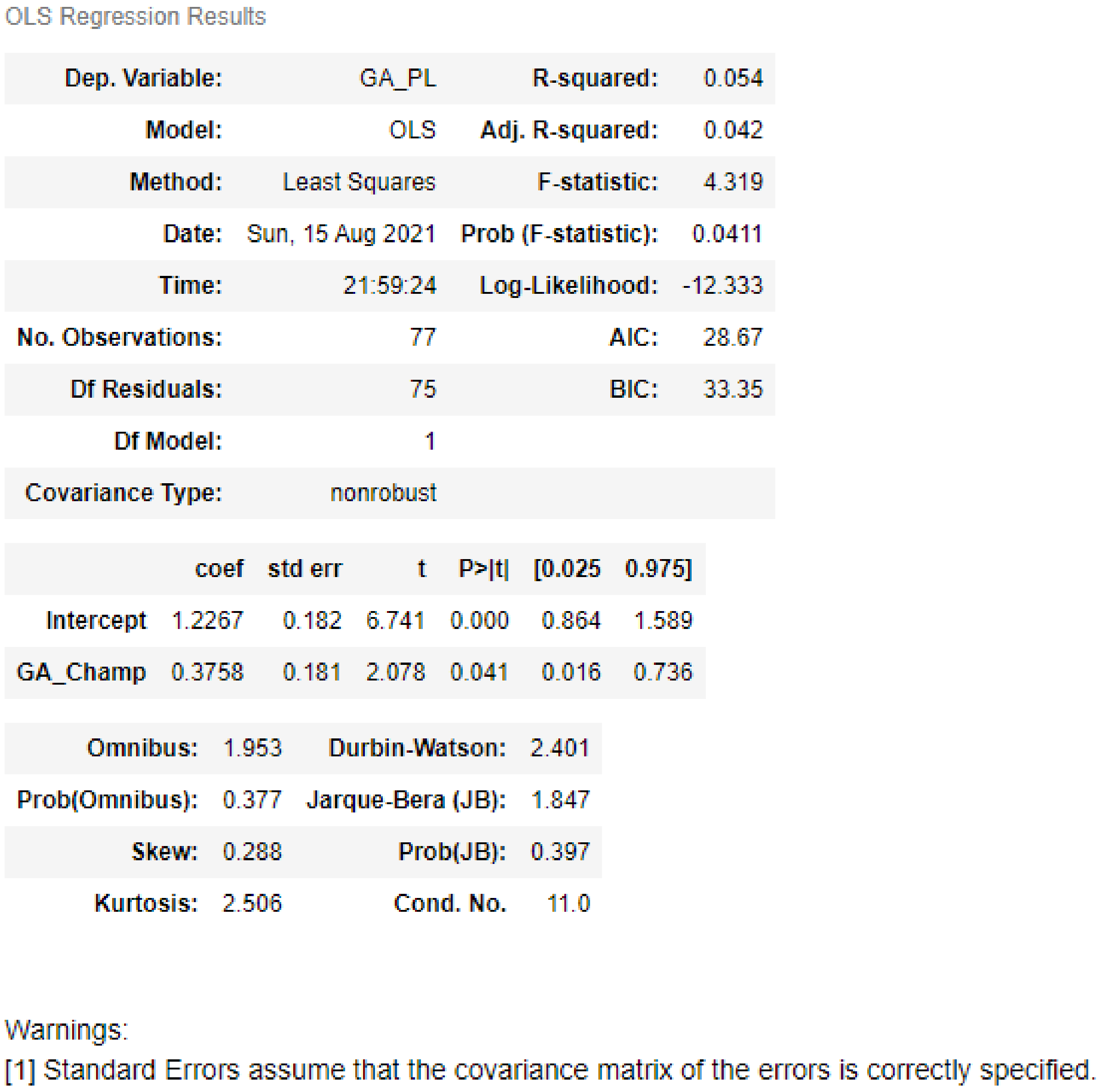


Figure 10: Goal conceding rates for every promoted team during their promotion winning season versus their goal conceding rate the subsequent season in Premier League



OLS Regression Results

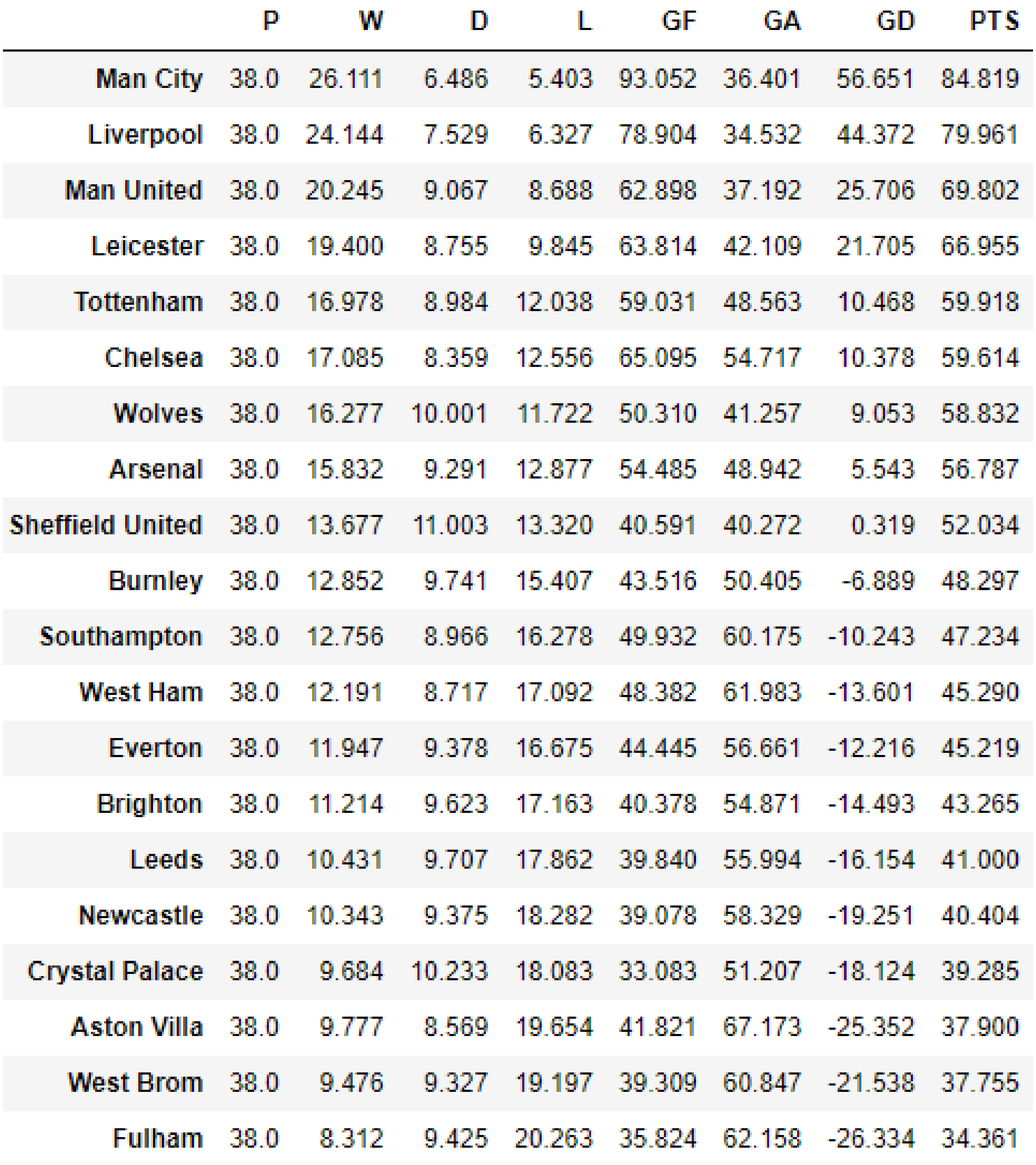
Dep. Variable:	GA_PL	**R-squared:**	0.054
Model:	OLS	**Adj. R-squared:**	0.042
Method:	Least Squares	**F-statistic:**	4.319
Date:	Sun, 15 Aug 2021	**Prob (F-statistic):**	0.0411
Time:	21:59:24	**Log-Likelihood:**	-12.333
No. Observations:	77	**AIC:**	28.67
Df Residuals:	75	**BIC:**	33.35
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1.2267	0.182	6.741	0.000	0.864	1.589
GA_Champ	0.3758	0.181	2.078	0.041	0.016	0.736

Omnibus:	1.953	**Durbin-Watson:**	2.401
Prob(Omnibus):	0.377	**Jarque-Bera (JB):**	1.847
Skew:	0.288	**Prob(JB):**	0.397
Kurtosis:	2.506	**Cond. No.**	11.0

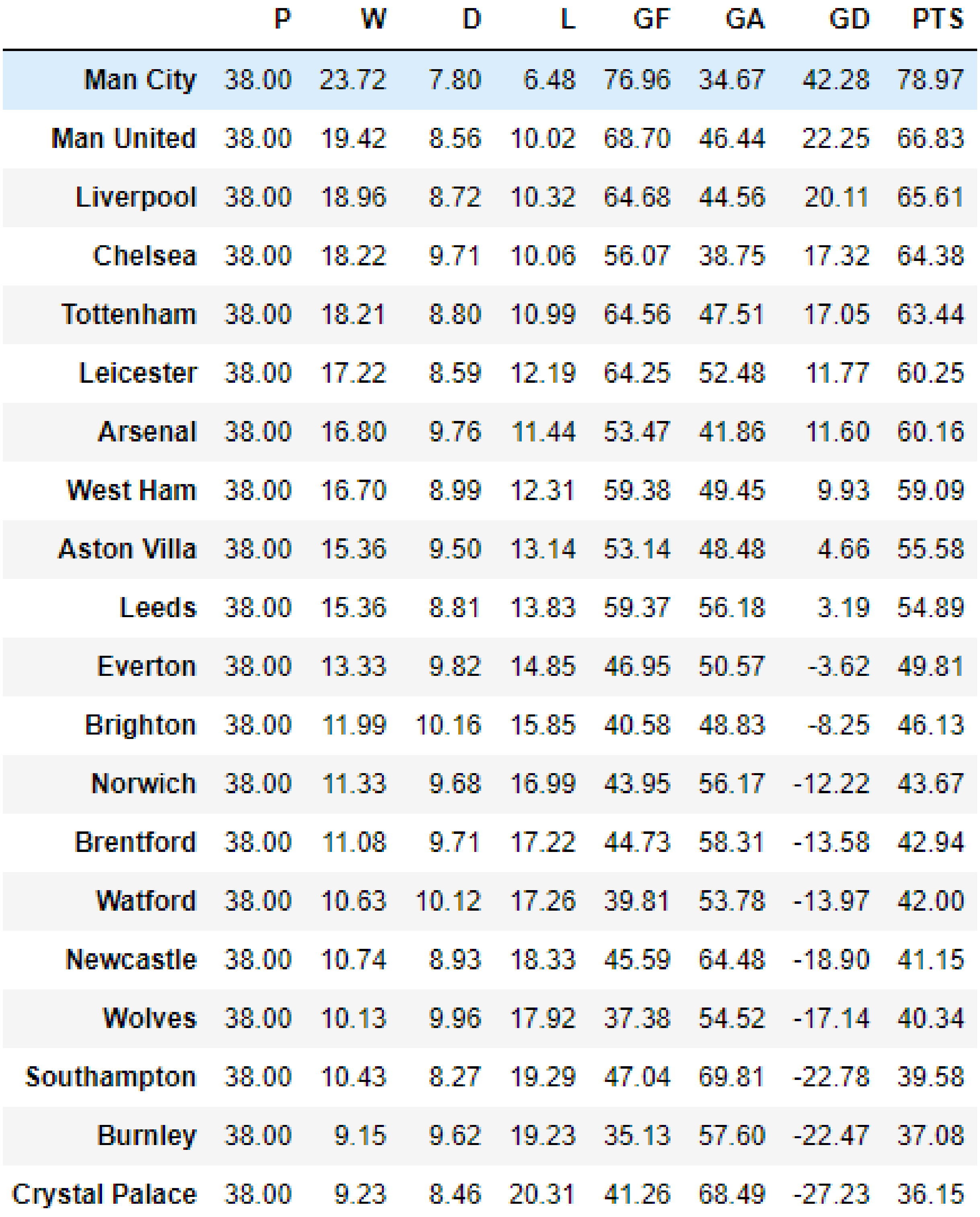
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 11: Fitted model summary for goal conceding adjustment model for promoted teams



	P	W	D	L	GF	GA	GD	PTS
Man City	38.0	26.111	6.486	5.403	93.052	36.401	56.651	84.819
Liverpool	38.0	24.144	7.529	6.327	78.904	34.532	44.372	79.961
Man United	38.0	20.245	9.067	8.688	62.898	37.192	25.706	69.802
Leicester	38.0	19.400	8.755	9.845	63.814	42.109	21.705	66.955
Tottenham	38.0	16.978	8.984	12.038	59.031	48.563	10.468	59.918
Chelsea	38.0	17.085	8.359	12.556	65.095	54.717	10.378	59.614
Wolves	38.0	16.277	10.001	11.722	50.310	41.257	9.053	58.832
Arsenal	38.0	15.832	9.291	12.877	54.485	48.942	5.543	56.787
Sheffield United	38.0	13.677	11.003	13.320	40.591	40.272	0.319	52.034
Burnley	38.0	12.852	9.741	15.407	43.516	50.405	-6.889	48.297
Southampton	38.0	12.756	8.966	16.278	49.932	60.175	-10.243	47.234
West Ham	38.0	12.191	8.717	17.092	48.382	61.983	-13.601	45.290
Everton	38.0	11.947	9.378	16.675	44.445	56.661	-12.216	45.219
Brighton	38.0	11.214	9.623	17.163	40.378	54.871	-14.493	43.265
Leeds	38.0	10.431	9.707	17.862	39.840	55.994	-16.154	41.000
Newcastle	38.0	10.343	9.375	18.282	39.078	58.329	-19.251	40.404
Crystal Palace	38.0	9.684	10.233	18.083	33.083	51.207	-18.124	39.285
Aston Villa	38.0	9.777	8.569	19.654	41.821	67.173	-25.352	37.900
West Brom	38.0	9.476	9.327	19.197	39.309	60.847	-21.538	37.755
Fulham	38.0	8.312	9.425	20.263	35.824	62.158	-26.334	34.361

Figure 12: Average final Premier League table from 2020-21 (back-test) Monte Carlo simulation



	P	W	D	L	GF	GA	GD	PTS
Man City	38.00	23.72	7.80	6.48	76.96	34.67	42.28	78.97
Man United	38.00	19.42	8.56	10.02	68.70	46.44	22.25	66.83
Liverpool	38.00	18.96	8.72	10.32	64.68	44.56	20.11	65.61
Chelsea	38.00	18.22	9.71	10.06	56.07	38.75	17.32	64.38
Tottenham	38.00	18.21	8.80	10.99	64.56	47.51	17.05	63.44
Leicester	38.00	17.22	8.59	12.19	64.25	52.48	11.77	60.25
Arsenal	38.00	16.80	9.76	11.44	53.47	41.86	11.60	60.16
West Ham	38.00	16.70	8.99	12.31	59.38	49.45	9.93	59.09
Aston Villa	38.00	15.36	9.50	13.14	53.14	48.48	4.66	55.58
Leeds	38.00	15.36	8.81	13.83	59.37	56.18	3.19	54.89
Everton	38.00	13.33	9.82	14.85	46.95	50.57	-3.62	49.81
Brighton	38.00	11.99	10.16	15.85	40.58	48.83	-8.25	46.13
Norwich	38.00	11.33	9.68	16.99	43.95	56.17	-12.22	43.67
Brentford	38.00	11.08	9.71	17.22	44.73	58.31	-13.58	42.94
Watford	38.00	10.63	10.12	17.26	39.81	53.78	-13.97	42.00
Newcastle	38.00	10.74	8.93	18.33	45.59	64.48	-18.90	41.15
Wolves	38.00	10.13	9.96	17.92	37.38	54.52	-17.14	40.34
Southampton	38.00	10.43	8.27	19.29	47.04	69.81	-22.78	39.58
Burnley	38.00	9.15	9.62	19.23	35.13	57.60	-22.47	37.08
Crystal Palace	38.00	9.23	8.46	20.31	41.26	68.49	-27.23	36.15

Figure 13: Average final Premier League table from 2021-22 Monte Carlo simulation