Word Count: 2412/2500

An Analysis of the determinants and impact of goals scored in the English Premier League

Introduction

The objective of football teams is to score more goals than their opponent. Thus, the analysis of goals scored is an interesting research area that has been extensively explored. The literature I have reviewed highlights several factors that influence goals scored including set plays, passes and the time period in a football match (e.g., the last 15 minutes of a match).

Although popular football tournaments produce a significant amount of data, there is a lack of comprehensive seasonal analysis of league competitions (Krzysztof and Paweł, 2014). This project aims to address this issue by assessing the impact of several variables influencing goals scored in the English Premier League based on six seasons worth of player performance statistics (2015/16 season – 2020/21 season).

From my research topic, I have formed the following two questions:

- What is the impact of uncorrelated explanatory variables on goals scored in the English premier league?
- Does goals scored impact the position of a Player?

Keywords: Set plays, Continental Wins, Union of European Football Associations (UEFA) Champions League, Goalkeeper, Defender, Midfielder, Forward, Transitions in Play, Counterattacks, Open Play, Penalty Box, Organised Offenses

Literature Review

This literature review is based on carefully selected sources that provide insight into some of the determinants of goals scored in football:

- The impact of set plays on goals scored
- The impact of passes on goals scored
- The impact of the time period in a football match on goals scored
- The impact of counterattacks or transitions in play on goals scored

The impact of set plays on goals scored

TotalFootballSchools.com (n.d.) states that set plays are ‘any passage of play that comes from a restart to the normal flowing game. This could be the kickoff, a free kick, a corner or a throw in’. Evidence from research that analysed different tournaments and levels of play in football (e.g., recreational, and competitive) found that 25% to 35.6% of goals were scored from set plays (Bangso and Peiterson, 2000 as cited by Wright et al, 2011). Additionally, in the 2006 World Cup football teams that performed well had a higher set-play to goal ratio; 1:7.5 versus 1:14 (Bell-Walker et al, 2006 as cited by Wright et al, 2011).

In addition to this, 30.1% of goals scored in the English Premier League in the 2008/09 season were as a result of set plays (Krzysztof and Paweł, 2014). This study highlights that set plays are a determinant of goals scored. However, set plays were found to have no impact on goals scored within a logistic regression which suggests that coaching staff should carefully consider

the time they spend on practicing set play tactics relative to open play tactics (Wright et al, 2011). Open play in the context of football means when teams are running and passing with the aim of scoring goals with their teammates.

Wright et al (2011) highlighted the impact of several variables on goals scored but also stated that 'Despite this the R^2 would still suggest there is plenty of variance which has not been accounted for, which might warrant further investigation'. This statement has highlighted the idea that set plays are one of the many factors affecting goals scored.

The impact of passes on goals scored

Passes are another variable that existing research has shown influences goals scored. For example, Wright et al (2011) found that '85% of goals were scored via short (0-4) passing sequences'. This supports the statement of Wright et al (2011) that 70% of goals would be scored via 'relatively short passing exchanges'. Additionally, Krzysztof and Paweł (2014) state that 'referring to research conducted by Charalampos et al. [8], much more goals are observed when the ball is passed short'.

However, research has found that longer passing resulted in a greater frequency of shots per 1000 possessions than short passing, which directly opposes findings from previous research (Hughes and Frank, 2005 as cited by Wright et al, 2011). This suggests that passes influence goals scored, but the type of passes (short or long passes) that have the most significant impact on goals scored is unknown to researchers.

The impact of the time period in a football match on goals scored

Wunderlich, F, Seck,A and Memmert, D (2021) Researchers have found that a significant amount of evidence indicating that the frequency of goals scored during a match is time-dependent and increases systematically throughout the match, 'with a climax in the last period from minute 75 on' (Wunderlich, F, Seck,A and Memmert, D, 2021). This could be due to fatigue, as well as the current result of the game in the final 15 minutes, and the reduced time for scoring encouraging the players of a team to be resourceful with time they have by finding a way to score a goal (Alberti et al, 2013). Additionally based on data from professional European football leagues research has found that the phenomenon of higher goals being scored in the final 15 minutes of a match is not season nor league dependent (Alberti et al, 2013). This should give readers confidence that this phenomenon is genuine. However, in future research, Alberti et al (2013) could identify whether this phenomenon is observable in non-European professional football leagues, given the global nature of professional football.

The impact of transitions in play or counterattacks on goals scored

Counterattacks or transitions in play are a type of attack in football initiated by one team that involves recovering the ball and immediately trying to reach their opponents penalty box (box directly in front of both goals on a football pitch) as quickly as possible and with the least number of touches. This allows a team to take advantage of an opponent's weak positioning (Fieldinsider.com, n.d.), with the aim of scoring a goal.

Yiannakos and Armatas (2006) as cited by Wright et al (2011) 'found that counter-attacks occurred less frequently (20.3%) in comparison to organised offences (44.1%) and set plays (35.6%)'. Nevertheless, a study found that regardless of the low frequency of counter attacks in modern football (4.9%), 16.9% of counterattacks resulted in a goal relative to only 11.1% of organised offenses resulting in a goal (Armatas et al, 2005 as cited by Wright et al, 2011). Organised offenses in this context refers to the movement of the ball by players in a systematic way in order to score goals. This suggests that it may be beneficial for football coaches to spend more time improving their team's ability to create more counterattacking situations in football matches.

Data and Method

The dataset I will be using in this project will be an unbalanced panel dataset consisting of six seasons worth of 51 player performance statistics from the English premier league. I have decided to use this dataset because the English premier league is undoubtedly the most exciting league in the world. There are a number of reasons for this, but I believe the main reason is the competitive balance that the league has. For example, according to BleacherReport.com (2014) 'Domestically, the English are helped by a league with an enormous amount of parity as the gap between first and last is only 31 points'. Additionally, based on continental wins (wins by teams in continental competitions e.g., the UEFA Champions league), the seven most highly ranked English premier league teams have the best wins per team average in Europe of 3.71 (BleacherReport.com, 2014).

Fixed Effects Model:

$$Goals = \beta0 + \beta1Forward + \beta2\ tacklesuccess + \beta3\ interceptions + \beta4\ recoveries + \beta5\ duelswon + \beta6\ passes$$

I will be using the dependent variable goals scored in the fixed effects model specified above as the focus of this research is to analyse factors that influence goals scored hence it the variable that is of most interest to me. Additionally, I will be using the explanatory variable Forward as I would like to determine whether the data proves that forwards typically score more goals than non-Forward players which is what would be expected as forwards are typically in more goal scoring positions than non-Forward players throughout the course of a football match.

Also, I will be using the explanatory variables tackle success, interceptions, recoveries, duels won and passes. Wright et al (2011) used similar variables in their analysis of one season's worth of English premier league goals data. They found that the variables they used had statistically insignificant effects on goals scored. However, the dataset I will be using is based on six seasons worth of English premier league player performance statistics. Therefore, I have the opportunity to use a greater number of observations of the explanatory variables of interest in my analysis.

Multinominal Logit Model:

$$PlayerPosition = \beta0 + \beta1goals + \beta1\ goalspermatch + \beta2\ goalswithrightfoot + \beta3\ goalswithlefttfoot + \beta4\ headedgoals$$

I will be using the categorical unordered dependent variable 'PlayerPosition' as I would like to determine whether the position of a player is explained by variables related to goals scored including goals per match and goals with a player's right foot. I believe it is appropriate to investigate the impact of goals on a player's position as goals as mentioned previously are the most important determinant of a football team's success in a football match. Therefore, it may be the case that players hold certain positions due to the number of goals they score.

Discussion of the suitability of the selected models

The basic advantage of using the linear fixed effects model specified above is that it will reveal the linear relationships between the dependent variable and the independent variables. Additionally, the model will account for individual unobserved heterogeneity, which a random

effects model would not. However, it might be the case that non-linear relationships exists between variables and therefore the linear model specified above will not reveal this.

The multinominal logistic regression model I have specified above will give me the opportunity to ascertain whether player's hold certain positions due to the number of goals they score. First by establishing a base group and then by making comparisons accordingly. Additionally, the second model I will be estimating has a categorical unordered dependent variable and a multinominal logistic regression model is more amenable than a standard logit model for running a regression with a categorical unordered dependent variable as it is capable of dealing with several categories of a dependent variable unlike a logit model.

Empirical Analysis

Table 1: The fixed effects regression of goals on Forward, tacklesuccess, interceptions, recoveries, duelswon and passes

Regression results

goals	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
Forward	-.017	.013	-1.32	.188	-.043	.008	
tacklesuccess	-.089	.075	-1.18	.238	-.236	.059	
interceptions	-.025	.006	-4.16	0	-.037	-.013	***
recoveries	.004	.004	1.11	.266	-.003	.011	
duelswon	.013	.003	4.13	0	.007	.019	***
passes	.001	0	1.27	.204	0	.001	
Constant	.03	.031	0.96	.337	-.031	.091	

Mean dependent var	0.389	SD dependent var	1.185
R-squared	0.210	Number of obs	2790
F-test	.	Prob > F	.
Akaike crit. (AIC)	5388.534	Bayesian crit. (BIC)	5418.203

**** p<.01, ** p<.05, * p<.1*

I have chosen the model shown above based on the logic specified in the decision tree for selecting different estimators. The dependent variable goals scored is a numerical variable, thus I ran poolability test. I found evidence for individual unobserved heterogeneity and therefore I ran a Hausman test. Then I found evidence to reject the null hypothesis that a random effects model is appropriate, hence why I have run a fixed effects regression model.

I expected to see a positive sign on the coefficients of the following variables: Forward, tacklesuccess, interceptions, recoveries, duelswon and passes as I believe these variables have a positive impact on goals scored.

Table 1 reveals that tacklesuccess, interceptions have negative impacts on goals scored contrary to my expectations. This may be because successful tackles and interceptions may be predominantly in the defensive third of a football pitch and players typically attack less from this position on the pitch. Therefore, instead of scoring goals players may instead prevent their team from conceding goals. This suggests there may be a negative relationship between tackle success and goals conceded and interceptions and goals conceded that may be worth exploring in future research.

Additionally, there is a negative coefficient (-.017) on Forward which indicates that Forwards between the 2015-16 to the 2020/21 English premier league seasons on average scored -0.017 fewer goals than non-Forward players. This is an unexpected result and I believe it could be due to the lack of goals scoring opportunities created by the teammates of forwards and potentially due to players in other positions e.g., Midfielders and Defenders finding or creating more goal scoring opportunities for themselves.

Table 2: Multinominal logit model of PlayerPosition on goals, goals per match, goals with right foot, goals with left foot, and headed goals

PlayerPosition	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
goals	1.602	1.088	1.47	.141	-.53	3.734	
goalspermatch	-2.461	1.703	-1.44	.149	-5.799	.878	
goalswithrightfoot	-.294	.614	-0.48	.632	-1.497	.909	
goalswithleftfoot	12.408	1.086	11.42	0	10.279	14.537	***
headedgoals	-1.613	1.392	-1.16	.247	-4.341	1.116	
Constant	4.577	.228	20.04	0	4.129	5.024	***

Mean dependent var	1.993	SD dependent var	0.083
Pseudo r-squared	0.059	Number of obs	2905
Chi-square	3373.195	Prob > chi2	0.000
Akaike crit. (AIC)	236.880	Bayesian crit. (BIC)	272.725

**** p<.01, ** p<.05, * p<.1*

The Player position variable is a discrete categorical unordered variable that has 3 levels (0 representing Goalkeepers, 1 representing Defenders and 2 representing Midfielders and Forwards). With 1 being the base group and 2 being the group I will compare with the base group. This led me to the 'Pool Waves?' branch of the decision tree for selecting estimators. The dataset that I have is an unbalanced panel data set consisting of English premier league player performance statistics for the last six seasons (2015/16 – 2020/21 season). Therefore, the dataset is indeed a pooled set of data in waves. Finally, I decided to run a multinominal logit model.

I expect to see a positive sign on all the coefficients of the explanatory variables in the model above as they are indicators of goals scored and the comparison group Midfielders and Forwards typically score more goals than the base group; Defenders. Again, contrary to my expectations, Table 2 reveals that goalspermatch, goalswithrightfoot and headedgoals have negative coefficients. This suggests to me that players in positions where they have a higher likelihood of scoring (Midfielders and Forwards) may not hold their positions based on the number of goals they score but rather due to other factors unaccounted for in the model above.

Conclusion

The multicollinearity detected by stata prior to me running my first regression resulted in a reduced number of regressors in the first model I estimated. Nevertheless, I was still able to identify the effect of several important variables on goals scored included passes, which is a variable well documented in literature as a factor influencing goals scored.

Unfortunately, I was not able to regress goals scored on the following variables that have been found to have an impact on goals scored in previous research: set plays, time period in the football match and transitions in play. This was due to the lack of data in the dataset I used for this project.

Finally, as mentioned previously, Wright et al (2011) who used one season's worth of English premier league goals data found that a set of determinants of goals scored including interceptions had statistically insignificant effects on goals scored. However, I have found that interceptions and duelswon have a statistically significant effect on goals scored based on six seasons worth of English premier league player performance statistics. This suggests that the determinants of goals scored are more likely to have a statistically significant effect on goals scored over an extended period.

Bibliography

1. Krzysztof, D and Paweł, B (2014). *Analysis of goals and assists diversity in English Premier League.* Online at: (PDF) Analysis of goals and assists diversity in English Premier League (researchgate.net) (accessed 16/12/2021).
2. Totalfootballschools.com (n.d.). *Set Plays*. Online at: https://totalfootballschools.com/start-learning/insight/set-plays (accessed 16/12/2021).
3. Wright et al (2011). *Factors Associated with Goals and Goal Scoring Opportunities in Professional Soccer*. Online at: Factors Associated with Goals and Goal Scoring Opportunities in Professional Soccer: International Journal of Performance Analysis in Sport: Vol 11, No 3 (tandfonline.com) (accessed 16/12/2021).
4. Wunderlich, F, Seck,A and Memmert, D (2021). *The influence of randomness on goals in football decreases over time. An empirical analysis of randomness involved in goal scoring in the English Premier League*. Online at: https://www.tandfonline.com/doi/abs/10.1080/02640414.2021.1930685?journalCode=rjsp20 (accessed 20/12/2021).
5. Alberti et al (2013). *Goal scoring patterns in major European soccer leagues.* Online at:https://www.researchgate.net/publication/262971485_Goal_scoring_patterns_in_major_European_soccer_leagues (accessed 20/12/2021).
6. BleacherReport.com (2014). *Statistically Ranking the World's Top 10 Football Leagues*. Online at: https://bleacherreport.com/articles/1922780-statistically-ranking-the-worlds-top-10-football-leagues (accessed 21/12/2021).
7. FieldInsider.com (n.d.). *How To Counter-Attack: A Complete Guide*. Online at: https://fieldinsider.com/how-to-counter-attack/ (accessed 13/01/2022).

Appendix

Stata do file commands

```
*tells stata which version is being used

version 17


*closing any do files open

clear all


*closes log files

capture log close


*tells stata where the data of interest is stored

global DataDir "C:\Users\uyier\OneDrive\Documents\University\Modules 2021_22\Applied Econometrics"


* stores all files where I say 'AnalysisDir' in do file

global AnalysisDir "C:\Users\uyier\OneDrive\Documents\University\Modules 2021_22\Applied Econometrics"


* creates and opens a log file

log using "$AnalysisDir\2_3_EPL_Player_Stats_Panel log file.log", replace


* opens the data

use "$DataDir\EPL_Player_Stats_2015_to_2021\pl_15_16",clear


* identifys the season that the observations come from

generate season = 1


* saves file

save "$AnalysisDir\Seasons", replace
```

```
* opens data
use "$DataDir\EPL_Player_Stats_2015_to_2021\pl_16_17", clear

* identifys the season that the observations come from
generate season = 2

* appends data to file
append using "$AnalysisDir\Seasons"

* saves file
save "$AnalysisDir\Seasons", replace

* opens data
use "$DataDir\EPL_Player_Stats_2015_to_2021\pl_17_18", clear



* identifys the season that the observations come from
generate season = 3

* appends data to file
append using "$AnalysisDir\Seasons"

* saves file
save "$AnalysisDir\Seasons",replace

* opens data
use "$DataDir\EPL_Player_Stats_2015_to_2021\pl_18_19", clear

* identifys the season that the observations come from
generate season = 4
```

```
* appends data to file
append using "$AnalysisDir\Seasons"

* saves file
save "$AnalysisDir\Seasons",replace

* opens data
use "$DataDir\EPL_Player_Stats_2015_to_2021\pl_19_20", clear

* identifies the season that the observations come from
generate season = 5

* appends data to file
append using "$AnalysisDir\Seasons"

* saves file
save "$AnalysisDir\Seasons", replace


* opens data
use "$DataDir\EPL_Player_Stats_2015_to_2021\pl_20_21", clear

* identifies the season that the observations come from
generate season = 6

* appends data to file
append using "$AnalysisDir\Seasons"

* saves file
```

```
save "$AnalysisDir\Seasons",replace

* creates id variable
encode name, generate(id)

* deletes observations that meet specified condition
drop if name == ""

* identifies panel var / identifies whether panel is balanced or not
xtset id season

* creates an id variable
bysort id: generate n = _N

* converts position var to numeric
encode (position), gen(position_converted)

* creates player position dummy variables and runs regression
xi:reg goals i.position_converted tacklesuccess interceptions recoveries duelswon passes

* renames player position dummy variables
rename (_Iposition__2 _Iposition__3 _Iposition__4) (Forward Goalkeeper Midfielder)

* identifies suspected multicollinearity
pwcorr goals tacklesuccess interceptions recoveries duelswon passes goals season
appearances cleansheets goalsconceded tackles lastmantackles blockedshots clearances
headedclearance clearancesoffline duelslost successful5050s aerialbattleswon aerialbattleslost
owngoals errorsleadingtogoal assists passespermatch bigchancescreated crosses
crossaccuracy throughballs accuratelongballs yellowcards redcards fouls offsides
headedgoals goalswithrightfoot goalswithleftfoot hitwoodwork goalspermatch
penaltiesscored freekicksscored shots shotsontarget shootingaccuracy bigchancesmissed
saves penaltiessaved punches highclaims catches sweeperclearances throwouts goalkicks
Forward Goalkeeper Midfielder, obs sig star(5)
```

* creates dummy variable for defender

generate Defender = 0

replace Defender = 1 if Goalkeeper == 0

* Technique 1 - Fixed Effects Model

*poolability test for individual unobserved heterogeneity

xtreg goals Forward tacklesuccess interceptions recoveries duelswon passes, fe

* Null hypothesis: ui = 0

* Alternative Hypothesis: ui is not equal to 0

* Prob > F = 0.00, therefore the null hypothesis that there is no individual unobserved heterogeneity is rejected

* Hausman test

* Null hypothesis: random effects model is appropriate

* Alternative Hypothesis: fixed model is appropriate

* runs fixed effects regression

xtreg goals Forward tacklesuccess interceptions recoveries duelswon passes, fe

* stores results of fixed effects regression

estimate store fe

* stores results of random effects regression

xtreg goals Forward tacklesuccess interceptions recoveries duelswon passes, re

* stores results of random effects regression

estimate store re

* executes hausman test in stata

hausman fe re

* Prob > chi2 = 0.0000 is < 0.05 therefore I reject the null hyopthesis that the random effects model should be used

* runs fixed effects model and stores results in word doc

asdoc xtreg goals Forward tacklesuccess interceptions recoveries duelswon passes, fe cformat(%9.3f) vce(robust) replace

* creates PlayerPosition column and populates it

generate PlayerPosition = -1

replace PlayerPosition = 0 if Goalkeeper == 1

replace PlayerPosition = 1 if Defender == 1

replace PlayerPosition = 2 if Midfielder == 1

replace PlayerPosition = 2 if Forward == 1

* runs multinominal logit model and stores results in word doc

asdoc mlogit PlayerPosition goals goalspermatch goalswithrightfoot goalswithleftfoot headedgoals, cformat(%9.3f) vce(robust) baseoutcome(1) replace