Econometrics Research Paper

What is the largest contributing factor in Premier League soccer players’ weekly salary?

Aedan Radvanyi

Table of Contents

[Abstract 3](#_Toc512541510)

[Introduction 4](#_Toc512541511)

[Literature Review 6](#_Toc512541512)

[Data 9](#_Toc512541513)

[Table 1: Variables 11](#_Toc512541514)

[Table 2: Summary Statistics 12](#_Toc512541515)

[Methodology 13](#_Toc512541516)

[Empirical Results 14](#_Toc512541517)

[Table 3: Model 1 Results 14](#_Toc512541518)

[Figure 1: Boxplot wage 15](#_Toc512541519)

[Table 4: Model 2 Results 16](#_Toc512541520)

[Table 5: Model 3 Results 17](#_Toc512541521)

[Figure 2: Scatterplot logwage – appsq 18](#_Toc512541522)

[Table 6: Model 4 Results 20](#_Toc512541523)

[Figure 3: Scatterplot att - logwage 21](#_Toc512541524)

[Figure 4: Scatterplot mid – logwage 21](#_Toc512541525)

[Conclusion 22](#_Toc512541526)

[Bibliography 23](#_Toc512541527)

[R Code 24](#_Toc512541528)

# Abstract

The English Premier League (EPL) is the top level of domestic English soccer, financially the largest soccer league in the world, with revenues competitive with that of Major League Baseball, and widely considered to be the most-followed and most notorious soccer league in the world. In the 2016-17 season, Premier League revenues reached a record high of $6.4 billion, and such colossal capital translates into significant payoffs for the players involved in the competition (CNN 2018). As of January 2018, EPL players make as much as £500,000 per week, and of course with such enormous fees comes prominent questioning of whether such amounts can be justified. A somewhat controversial issue with regard to whether players should earn as much as they do, one wonders if physical attributes are as important as performance statistics in calculating a players’ salary. This paper will look to uncover and identify the major factors that result in determination of weekly salaries in England’s leading competition. Variables considered include goals scored, goals assisted, wins, losses, appearances and physical characteristics for players competing in the Premier League season as of the 2016-17 season. Initial results, appeared to put greater emphasis on goals, assists and appearances.

# Introduction

The EPL is the highest-level of the English soccer league system, topping a pyramid of professional soccer leagues consisting of teams for England and Wales. The league has always been based on the injection of revenues, with the league as it is now breaking off from the football league as it was in 1992. As of 2013-14, that deal was worth £1 billion per year, with yearly domestic and international television rights averaging out to €2.2 billion and as of 2016-17, the 20 teams competing in the league were bringing in £2.4 billion each (Premier League 2017). While arguments for the EPL being the highest standard of league soccer in the world is still often debated, what is not debated is that the financial strength of the league dwarfs that of other European leagues. Individual players have been sold for as much as €120 million and brought in for €105 million in the past two years alone, highlighting the upward trend in financial acquisitions and movement of capital in the league. As such the contracts of the players within the league are becoming increasing lucrative, particularly at the top end of the league, leading to questions over the wage gap that is being created within the league. As of November 2017, the average weekly wage in the EPL stood at £50,817, generous in its own right but somewhat belittled in comparison to that of Alexis Sanchez’ salary of almost 10 times that figure, signed less than 2 months after said figure was published (BBC Sport 2017). With such colossal figures being touted, one can only wonder: what is the common characteristic that these top earners have that lead to earnings of £1 million over the course of a fortnight?

While salary caps are commonplace in the US (hard salary caps in the NFL and NHL and a soft cap in the NBA), no such system exists in the foremost soccer league in England. As such, salaries in the EPL have been growing year on year as broadcasting revenues grow and find their way to the professional clubs. This is leading to scenarios where clubs are having outlays in the double figure millions each week for their players in order to fend off competition from rivals, leading to the importance of understanding how a player can land themselves such a contract becoming increasingly significant. While there is no league wide salary cap, and despite the growing financial inflows that EPL teams are benefitting from, some teams still have their own wage structure systems that mean they are restricted in the way they allocate wages. Tottenham Hostpur, for example, are a team that competed in the Champions League (the pinnacle of European cup competition), and yet they cap their own player salaries at £120,000 per week in order to maintain reasonable parity within the team as well as team harmony. A team like that could benefit from understanding which factors are intrinsically linked with higher wages in order to help them make player purchasing decisions that fit their wage structure. Moreover, as players, knowing what can lead to increasing their weekly salary can allow them to focus their training efforts in order to attain greater value contracts in the future, and thus better themselves financially.

My analysis will look at a number of individual player characteristics that include physical characteristics and in game and overall EPL career statistics. Through my work I intend to find both the factors that positively affect salaries and also the factors that negatively affect them, in order to better understand how salaries are determined. In general, attacking players, the players who win games with important goals or assists for goals, tend to earn the most money. Of the top 10 earners in the league as of the 2017-18 season, 7 of the players involved are attack minded players, be that forwards or attacking midfield players. In addition, of the remaining 3 of the list, 2 are players who are known for their goal-scoring and goal-assisting exploits (Fox Sport 2018). For that reason, I predict that goals and assists will be strongly linked with higher wages.

# Literature Review

With the increasingly prominent trend of upward pressures on players’ financial demands in the EPL, recent years have seen an influx of publications with regard to the analysis of the financials involved in the world’s largest soccer league. The majority of publications have tried to link increased wage expenditures from teams to success on the field. Others have looked more specifically at what you can expect from players who are earning in the upper bracket of salaries, and therefore if said players are deserving of their salaries in reference to their output. Finally, there have also been similar studies to that of mine in looking at what factors relate most closely to players earning greater salaries, however few studies have looked at this topic specifically.

A study conducted by Chris Lonsdale looked into the use of the increasing revenues made by EPL clubs in recent years and how they haven’t necessarily translated into greater profits due to the increasing demand from players to see their share of the cash inflow (Lonsdale 2004). It was found that the players were receiving the majority share of the extra capital coming into the league. Lonsdale looked more so at overall team expenditure on wages than individual characteristics that affected wages, but his analysis looked at the growing trend with respect to wages and also at the top earning players in the league, a point that certainly affects the earning understanding for top earners in my model.

Dr. Adam Cox examined, in his analysis of club performance and revenue sharing in the EPL, the effect of wage expenditure from teams on the chances that a team will win a match and also how it increases the quality of the team. While the overall report looked in greater detail at effects of variables on spectator demand and overall club performance, Cox looked into the effects of changing wage rates in his holistic analysis of the EPL. This included a similar comparison to Lonsdale in that of revenues of clubs compared with wage expenditure and profit, but he also found through his work that higher wage expenditure correlated to higher overall wins (broken down into home wins and away wins) (Cox 2016). Cox’s findings that from 2008 to 2013 wages increased at a higher rate than revenues, falls directly in line with the work of Lonsdale. His work relates to my analysis as I am exploring the possibility that wins (and therefore losses) could be one of the major factors in wage determination, and while his work does not directly line up with mine it is useful in establishing that wins and losses are likely important factors in the wage calculation.

An ESPN article by Michael Caley investigated the link between teams’ wage expenditure and their top-four finishes in the EPL. A top-four finish for a team in said division comes with the lucrative benefit of qualification for the top level of European competition in the following season, and so while topping the table is the target for teams, a top-four finish is no mean feat. Based on data from 2000-01 seasons up until 2013-14, over 80% of top-four placements have been filled by teams with the top four wage bills in the league (Caley 2015). Again, the data draws a link between top performers and higher wage imbursement, and thus this publication, again not directly related to my analysis, allows me to deduce that wins and losses (or more broadly the level to which players are successful with their teams) are key variables in finding the factors that influence EPL salaries.

Fiona Carmichael, Ian McHale and Dennis Thomas explored the relationship between on field exploits of players and the commercial success of EPL teams. They used data to link on field success, unsurprisingly, to players skills and abilities, but further than wage expenditure steadily linked higher wages to players who were performing at a higher level, through quantitative measures such as goals and assists (Carmichael, Thomas and McHale 2010). The paper stated the interpretation that their statistical evidence suggested that investment in players (by way of wage packages) essentially bought on field success, which leads to the idea that higher wages could result in more wins for individual players, rather than the somewhat more expected winning more as individual players results in higher wages. It is interesting, with regard to my analysis, that this relationship could be positively correlated in a way that appears to make one factor cause the other in both instances, however above all it reiterates my need to include success of the team in the calculation of wages for EPL players.

With regard to individual player salaries, rather than as a component of a team’s wage expenditure, computer scientists Lara Yaldo and Lior Shamir used machine learning and data science to compute player salaries and compare predicted salaries with real world salary data, publishing their work in the International Journal of Computer Science in Sport (Yaldo and Shamir 2017). The study, which included attribute measurements relating to performance, behavior and abilities, combined salary data with 55 different attributes of players to determine whether residuals of the salary data were significant or not. While the findings of their study are not symbiotic with my analysis, their conclusion, that the top six highest paid players would be Lionel Messi (forward), Cristiano Ronaldo (forward), Luis Suarez (forward), Neymar (forward), David De Gea (goalkeeper) and Mesut Ozil (attacking midfielder), does further the theory that attack minded players (5 out of 6) usually attract, or are at least linked to, the highest wage offerings. This additionally supports the inclusion of attacking output-based metrics such as goals and assists as a key factor in determining what leads to high wages for EPL players.

The literature I found, relative in any way to my topic, suggested that there has not been extensive research specifically into what causes players to earn higher wage. I intend to use more recent data than that of the literature I have examined, largely because the changes in wages that have led me to research the topic have resulted from increased revenue from broadcasting deals in the past few years, thus requiring that data I use should be as up-to-date as possible in order to see the true results. Moreover, I am placing the focus of the variables in my analysis on individual characteristics and statistics. Whereas the literature I have examined tended to focus more on teams and their successes and failures in reference to wage expenditure, I will only use the team statistics in considering a players’ wins and losses, considering a player could not win or lose a game without the involvement of a team. Finally, I am taking into consideration the different positions on the field that players fall into, allocating each into either a defensive position, midfield position, or attacking position, by way of dummy variables, to further distinguish between the wages of the players in my dataset. As far as the literature I have examined explained, I have not seen such a distinction in looking at player wages thus far.

# Data

In order to investigate the link between player characteristics and EPL player salaries, I first sought out a dataset with information including individual statistics for EPL players. I found a dataset on github, a public sharing website for sharing projects, often including datasets and analyses, that had been compiled of data for the 2016-17 EPL season. The creator of the data had compiled information from a number of sources for the purpose of attempting to predict betting tips based on past player data.

The data began with 568 observations, with 57 different variables, however I found with many of the observations, with respect to the variables that I wanted to test, the information was missing. As such, to clean the data I went through and took out any players who had incomplete information for the 7 initial variables I wanted to examine. They were *height, weight, wins, losses, goals, assists and appearances*. I also wanted to include *age* as a variable in my analysis and while I was not given that directly by the dataset I was given a *date of birth* variable, from which I was able to compute player ages to be added to the dataset.

I realized that I would need to compile salary data separately and input them for the player that I had full information for. Using spotrac, an online database of major sports league contract breakdowns, I entered data for all players in the data set for whom data was available, removing players who did not have information available. This created the *wage* variable, with absolute values for weekly wage of 358 EPL players. From this new set of data, I created a *logwage* variable, in order to allow me to view changes in percentage wage with respect to variables in my dataset.

Next, I wanted to add something I felt was lacking from other analyses of a similar nature. That was the introduction of dummy variables with respect to player position. There were four major groups players could fall into: goalkeepers, defenders (*def*), midfielders (*mid*) and attackers (*att*). I introduced the latter four as dummy variables, with goalkeepers acting as the control variable due to the fact that they appeared less common in comparison to outfield players. This required using the Premier League official website database of squads in order to characterize players into each group.

Finally, I had been given the *appearances* variable as a start point, however I felt that it would be of greater use to my analysis by creating a quadratic term, *appsq*, in order to better understand how appearances affected wage. With the quadratic I was given the ability to find a turnaround point, through the use of derivative calculation, after which increases in appearances do not equate to increases in wages.

Table 1 below shows the final variables used in my four regression models with a short description of each.

## Table 1: Variables

|  |  |
| --- | --- |
| Variable | Explanation |
| Wage | Player weekly wage as of 2016-17 EPL season |
| Logwage | Natural log of *wage* variable |
| Age | Player age as of April 17 2018 |
| Height | Player height measured in centimeters |
| Def | Dummy variable that reads 1 if player regularly plays in defense for his team or 0 if not |
| Mid | Dummy variable that reads 1 if player regularly plays in midfield for his team or 0 if not |
| Att | Dummy variable that reads 1 if player regularly plays in attack for his team or 0 if not |
| Weight | Player weight measured in kilograms |
| Wins | The number of wins a player has been involved in as of the 2016-17 EPL season |
| Losses | The number of losses a player has been involved in as of the 2016-17 EPL season |
| Goals | The number of goals scored a player has achieved as of the 2016-17 EPL season |
| Assists | The number of goal assists a player has achieved as of the 2016-17 EPL season |
| Appearances | The number of appearances in games, either substitute or full game, that a player has been involved in as of the 2016-17 EPL season |
| Appsq | Squared function of the *appearances* variable |

Table 2 (below) shows the summary statistics for the variables I used in my analysis, namely the mean values, standard deviation, maximum value and minimum value of each. The summary statistics indicate that with my data, as was the case with the EPL in November 2017, the average wage was above the £50,000 level, showing just how high wages have spiked in recent times. Also, we can observe that there is a wide range in wage data available, with a minimum of £1,000 per week and a maximum of £300,000.

## Table 2: Summary Statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ***wage*** | ***logwage*** | ***age*** | ***height*** |
| Mean | 54743.855 | 4.601 | 28.573 | 182.804 |
| Standard Deviation | 43865.765 | 0.382 | 4.426 | 6.989 |
| Minimum | 1000.000 | 3.000 | 19.000 | 163.000 |
| Maximum | 300000.000 | 5.477 | 42.000 | 203.000 |
|  | ***def*** | ***mid*** | ***att*** | ***weight*** |
| Mean | 0.355 | 0.349 | 0.199 | 76.415 |
| Standard Deviation | 0.479 | 0.477 | 0.400 | 7.135 |
| Minimum | 0.000 | 0.000 | 0.000 | 60.000 |
| Maximum | 1.000 | 1.000 | 1.000 | 96.000 |
|  | ***wins*** | ***losses*** | ***goals*** | ***assists*** |
| Mean | 32.461 | 25.735 | 2.082 | 5.936 |
| Standard Deviation | 42.060 | 30.950 | 6.154 | 11.398 |
| Minimum | 0.000 | 0.000 | 0.000 | 0.000 |
| Maximum | 254.000 | 176.000 | 74.000 | 96.000 |
|  | ***appearances*** | ***appsq*** |  |  |
| Mean | 78.419 | 14803.095 |  |  |
| Standard Deviation | 93.155 | 34614.865 |  |  |
| Minimum | 0.000 | 0.000 |  |  |
| Maximum | 602.000 | 362404.000 |  |  |

# Methodology

I ran four different regressions using R, in an attempt to link contributing characteristics to salary. The first model included all variables mentioned above as independents, with the exception of the dummy variables for player position and using absolute values for weekly salary as the independent variable, rather than the logged variable. Also excluded from the independent variables was the quadratic term *appsq*, which would be used in later regressions. The first model for EPL salary was:

1. Wage = β0 + β1height + β2weight + β3wins + β4goals + β5losses + β6appearances + β7assists + β8age + u

The second model saw the introduction of the dummy variables into the regression. In the first model I wanted to see if there was an outstanding variable that could be linked with wages, but also wanted to consider how each variable would work in relation to the absolute numbers for the salary figures. With the introduction of the dummy variables, and therefore the inclusion of information about player position, due to the literature I had read and the presumptions I already had, I expected there to be stronger correlation with the *mid* and certainly the *att* variables. The second model for EPL salary was:

1. Wage = β0 + β1height + β2weight + β3wins + β4goals + β5losses + β6appearances + β7assists + β8age + β9def + β10mid + β11att + u

For the third and fourth models, I changed the absolute variable of wage for the logged version (*logwage*) in order to view results as a percentage change. Again, I did two regressions, one with the dummy variables and one without. However, now I included the quadratic variable in my model (*appsq*) so as to examine the turnaround point in the variable. That was accomplished by taking the derivative of the quadratic with respect to the original *appearances* term and then setting the derivative equal to 0. The third model I ran on EPL salary was:

1. Logwage = β0 + β1height + β2weight + β3wins + β4goals + β5losses + β6appearances + β7appsq + β8assists + β9age + u

The final model saw the third model as the basis and the introduction once again of the dummy variables for player position. Similarly, to with model 2 I expected *mid* and *att* to be positively correlated to salary, and with the logged independent variable I expected a clear link with between the variables. The fourth and final model I ran on EPL salary was:

1. Logwage = β0 + β1height + β2weight + β3wins + β4goals + β5losses + β6appearances + β7appsq + β8assists + β9age + β10def + β11mid + β12att + u

# Empirical Results

## Table 3: Model 1 Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 48475.08 | 71633.89 | 0.677 | 0.499 |
| Height | 124.86 | 505.66 | 0.247 | 0.805 |
| Weight | -37.18 | 507.44 | -0.073 | 0.942 |
| Wins | -74.58 | 523.63 | -0.142 | 0.887 |
| Goals | -222.82 | 386.99 | -0.576 | 0.565 |
| Losses | -125.37 | 632.67 | -0.198 | 0.843 |
| Appearances | 54.35 | 415.02 | 0.131 | 0.896 |
| Assists | -102.24 | 310.32 | -0.329 | 0.742 |
| Age | -394.21 | 683.22 | -0.577 | 0.564 |

The first model examined the effect of a number of variables on the salary of EPL players as an absolute number. Looking firstly at the intercept, the starting salary falls around the £50,000 mark that was touted earlier, before the effect of the different variables, which suggests that the model appreciates that as the basis, with any additional figures based on the characteristics tested. It was notable to me that none of the variables tested were statistically significant at any level, given that the presumption would be that these are major statistical points that would affect the salary decision for a player. The lowest p value recorded in this model (ignoring the intercept) was 0.564 on *age*, which shows that despite having the greatest significance of these variables in this model, is still not very significant.

### Figure 1: Boxplot wage

Upon reflection of the coefficients, the positive values on *height* and *appearances* are standalone in that all other factors appear to act detrimentally to calculation of a player’s salary. Particularly surprising, from this model, was that higher goal and assist output detrimentally affected wage. This led to later models involving *logwage* as an independent variable, so as to better understand the dependent variables in relation to a percentage change in wage. As shown by figure 1 above, the majority of wages fall in the lower half of the dataset, explaining the lower intercept given the range in the wage data.

## Table 4: Model 2 Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) | Significance |
| (Intercept) | 68774.42 | 71902.15 | 0.957 | 0.3395 |  |
| Height | -105.36 | 506.35 | -0.208 | 0.8353 |  |
| Weight | 125.95 | 504.55 | 0.250 | 0.8030 |  |
| Wins | 49.97 | 523.82 | 0.095 | 0.9240 |  |
| Goals | -270.26 | 383.44 | -0.705 | 0.4814 |  |
| Losses | -65.67 | 628.09 | -0.105 | 0.9168 |  |
| Appearances | -11.05 | 412.82 | -0.027 | 0.9787 |  |
| Assists | -143.20 | 309.41 | -0.463 | 0.6438 |  |
| Age | -440.67 | 681.69 | -0.646 | 0.5184 |  |
| Def | 3010.49 | 8270.24 | 0.364 | 0.7161 |  |
| Mid | 16310.80 | 8253.68 | 1.976 | 0.0489 | \* |
| Att | 19695.85 | 8944.81 | 2.202 | 0.0283 | \* |
| Signif. codes: | 0 ‘\*\*\*’ | 0.001 ‘\*\*’ | 0.01 ‘\*’ | 0.05 ‘.’ | 0.1 ‘ ’ |

The second model was based off the first, with the addition of dummy variables for player position, *def*, *mid*, and *att*. This model raised the intercept by £20,000, suggesting that the inclusion of player position is important because the player position has a large effect on salary. As we can see from the significance on two of the three dummy variables, *mid* and att at the 0.01 level, the inclusion of these variables is important to the model. The inclusion of those variables also changed the coefficients of some variables from the first model. *Height* and *appearances* flipped from a positive coefficient to a negative, whereas *weight* and *wins* have flipped from negative coefficients to positive ones. The coefficients on those dummy variables are also notable. A defensive player tends to make only £3,000 more than the £68,000 intercept, whereas a midfield player makes £16,000 more and an attacker makes almost £20,000 more. The effect of being further forward on the field appears to be very important.

## Table 5: Model 3 Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 4.7160000 | 0.6268000 | 7.524 | 4.61e-13 |
| Height | -0.0009215 | 0.0044230 | -0.208 | 0.835 |
| Weight | 0.0015670 | 0.0044300 | 0.354 | 0.724 |
| Wins | -0.0012040 | 0.0046130 | -0.261 | 0.794 |
| Goals | 0.0014280 | 0.0033840 | 0.422 | 0.673 |
| Losses | -0.0003364 | 0.0055780 | -0.060 | 0.952 |
| Appearances | 0.0005091 | 0.0037570 | 0.136 | 0.892 |
| Appsq | -0.0000002 | 0.0000014 | -0.167 | 0.867 |
| Assists | 0.0012510 | 0.0027090 | 0.462 | 0.645 |
| Age | -0.0023080 | 0.0060120 | -0.384 | 0.701 |

The third model saw the introduction of *logwage* as the independent variable, removal of the dummy variables, and introduction of *appsq* the appearances variable quadratic. The result of a change from absolute values to a percentage perspective was a coefficient outlook that looked more like what I had been expecting of the variables. The variables in this mode, like in the first model, lacked any significance at a substantial level. However, contrasting the first model, and as earlier predicted, I began to see some correlation between the *goals* and *assists* variables and wage growth. The model continued the trend of *appearances* being correlated with wage growth as well. Below I examined the effect of the quadratic variable *appsq*.

### Figure 2: Scatterplot logwage – appsq

There isn’t a clear trend from the scatterplot of the quadratic variable, however it does depict that as appearances increase there are occurrences of higher wages. Using differentiation to find a derivative and thus a turnaround point below, I attempted to find the point at which increases in appearances no longer increased wages. The calculations are as follows:

Appearances + 2(Appsq) = 0

0.0005091 + 2(-0.0000002) (Appearances) = 0

0.0005091 = 0. 0000004(Appearances)

1272.75 = Appearances

Of course, this figure of 1273 appearances, as a turnaround point, doesn’t seem to make sense, with an EPL season lasting 38 games and therefore our turnaround point being the equivalent of 33.5 seasons, however when you consider the manner of a professional soccer players career some sense can be made of the data. Players tend to be at their highest level when competing in the EPL, and therefore when they are no longer at that highest level, because of the level of competition, players who cannot compete tend to leave the league. In general, players tend to leave to leagues that are considered easier than the EPL in the latter stages of their career, Zlatan Ibrahimović and David Beckham’s moves to Major League Soccer (MLS) are just two high profile examples, and so it makes sense that players wages increase with added appearances and there is no feasible end point because players leave the league as they approach their twilight. As such the somewhat unfeasible turnaround point of 1273 does shed some light on the link between appearances and wages in the EPL.

## Table 6: Model 4 Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) | Significance |
| (Intercept) | 4.8170000 | 0.6270000 | 7.682 | 1.65e-13 | \*\*\* |
| Height | -0.0027210 | 0.0044160 | -0.616 | 0.53812 |  |
| Weight | 0.0030590 | 0.0043930 | 0.696 | 0.48674 |  |
| Wins | -0.0001192 | 0.0045970 | -0.026 | 0.97933 |  |
| Goals | 0.0008933 | 0.0033450 | 0.267 | 0.78956 |  |
| Losses | 0.0000458 | 0.0055180 | 0.008 | 0.99338 |  |
| Appearances | -0.0000251 | 0.0037200 | -0.007 | 0.99462 |  |
| Appsq | -0.0000002 | 0.0000014 | -0.152 | 0.87905 |  |
| Assists | 0.0010710 | 0.0026940 | 0.398 | 0.69121 |  |
| Age | -0.0033510 | 0.0059830 | -0.560 | 0.57578 |  |
| Def | 0.0967900 | 0.0721000 | 1.342 | 0.18033 |  |
| Mid | 0.1910000 | 0.0718700 | 2.657 | 0.00824 | \*\* |
| Att | 0.2162000 | 0.0780100 | 2.771 | 0.00589 | \*\* |
| Signif. codes: | 0 ‘\*\*\*’ | 0.001 ‘\*\*’ | 0.01 ‘\*’ | 0.05 ‘.’ | 0.1 ‘ ’ |

The final model saw the introduction of the dummy variables to the third model I ran. Above we can see that both the *mid* and *att* were significant at the 0.001 level, which relates to the results of the second model where the dummy variables had a large effect on the results of the regression. I observed that all positions are correlated with percentage changes in salary but that midfielders earned more than defenders and attackers earned more than midfielders, as supported by the scatterplots below with higher distributions for attackers than all other players.

### Figure 3: Scatterplot att - logwage



### Figure 4: Scatterplot mid – logwage

e

# Conclusion

I created the models in order to find a player characteristic that strongly affected wage growth in the EPL. After running four different regressions I found that many of the variables I thought would impact wages were not significant and therefore did not have major bearing on wage changes. Two such variables, goals and assists, proved to be somewhat correlated but not significant. I feel that the most likely reason for this is that there are teams who value defensive players, be that defenders or goalkeepers as much as they value their attacking stars, despite the tendency for clubs to value attack-minded players above all else. One such example is Manchester United goalkeeper David De Gea, who is one of the team’s highest paid players despite being a goalkeeper and thus not contributing goals or assists. While the output metrics for attacking players were not significant in their correlation with wages, the positions themselves proved to be the largest indicator of high wages. Being an attacker or midfielder, as opposed to goalkeeper or defender, yielded higher wage growth. As such if I were to run this model again I would likely make a further distinction between attack-minded midfielders and defensively-inclined midfielders, to further examine where the positional divide is that results in wage differences. Overall, I would say that for an aspiring player this information could be useful in that if they want to attain the highest salaries in the EPL, and benefit from the new revenues coming into the league, they would be best served seeking an attacking position within a potential team.

# Bibliography

BBC Sport. 2017. *Premier League average weekly wage passes £50,000, says new study.* November 27. Accessed April 23, 2018. https://www.bbc.co.uk/sport/football/42130297.

Caley, Michael. 2015. "Premier League wage bills mean top-four finishes, but not always." *ESPN Soccer.* September 10. Accessed April 24, 2018. http://www.espn.com/soccer/blog/tactics-and-analysis/67/post/2476622/premier-league-dominance-is-down-to-wages-but-can-be-broken.

Carmichael, Fiona, Dennis Thomas, and Ian McHale. 2010. "Maintaining Market Position: Team Performance, Revenue and Wage Expenditure in the English Premier League." *Bulletin of Economic Research.*

CNN. 2018. *Premier League revenues hit record $6.4 billion.* April 19. Accessed April 23, 2018. http://money.cnn.com/2018/04/19/news/companies/premier-league-record-revenue/index.html.

Cox, Dr Adam. 2016. "An economic analysis of spectator demand, club performance, and revenue sharing in English Premier League football." *Doctoral Thesis.*

Fox Sport. 2018. *Premier League Wages: Ridiculous wages of PL's tops earners.* February 1. Accessed April 24, 2018. https://www.foxsports.com.au/football/premier-league/premier-league-wages-ridiculous-wages-of-pls-top-earners/news-story/b25074aa386267158b13deed6056d15f.

Lonsdale, Chris. 2004. "Player power: capturing value in the English football supply network." *Supply Chain Management: An International Journal* 383-391.

Premier League. 2017. *Premier League value of central payments to Clubs.* June 1. Accessed April 23, 2018. https://www.premierleague.com/news/405400.

Yaldo, Lara, and Lior Shamir. 2017. "Are the world's highest paid football players overpaid? Big data says yes." *International Journal of Computer Science in Sport.*

# R Code

> getwd()

[1] "/Users/aedan"

> setwd("/Users/aedan/Documents/Econometrics/Paper")

> Players = read.csv("player\_data copy.csv")

> fix(Players)

> Playersreg1=lm(wage ~ height+weight+wins+goals+losses+appearances+assists+age, data=Players)

> summary(Playersreg1)

Call:

lm(formula = wage ~ height + weight + wins + goals + losses +

appearances + assists + age, data = Players)

Residuals:

Min 1Q Median 3Q Max

-59344 -28796 -10437 16727 241913

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 48475.08 71633.89 0.677 0.499

height 124.86 505.66 0.247 0.805

weight -37.18 507.44 -0.073 0.942

wins -74.58 523.63 -0.142 0.887

goals -222.82 386.99 -0.576 0.565

losses -125.37 632.67 -0.198 0.843

appearances 54.35 415.02 0.131 0.896

assists -102.24 310.32 -0.329 0.742

age -394.21 683.22 -0.577 0.564

Residual standard error: 44240 on 348 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.008339, Adjusted R-squared: -0.01446

F-statistic: 0.3658 on 8 and 348 DF, p-value: 0.9381

> Playersreg2=lm(wage ~ height+weight+wins+goals+losses+appearances+assists+age+def+mid+att, data=Players)

> summary(Playersreg2)

Call:

lm(formula = wage ~ height + weight + wins + goals + losses +

appearances + assists + age + def + mid + att, data = Players)

Residuals:

Min 1Q Median 3Q Max

-62409 -28289 -10969 16830 237548

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 68774.42 71902.15 0.957 0.3395

height -105.36 506.35 -0.208 0.8353

weight 125.95 504.55 0.250 0.8030

wins 49.97 523.82 0.095 0.9240

goals -270.26 383.44 -0.705 0.4814

losses -65.67 628.09 -0.105 0.9168

appearances -11.05 412.82 -0.027 0.9787

assists -143.20 309.41 -0.463 0.6438

age -440.67 681.69 -0.646 0.5184

def 3010.49 8270.24 0.364 0.7161

mid 16310.80 8253.68 1.976 0.0489 \*

att 19695.85 8944.81 2.202 0.0283 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43760 on 344 degrees of freedom

(2 observations deleted due to missingness)

Multiple R-squared: 0.03804, Adjusted R-squared: 0.007279

F-statistic: 1.237 on 11 and 344 DF, p-value: 0.261

> Playersreg3=lm(logwage ~ height+weight+wins+goals+losses+appearances+appsq+assists+age, data=Players)

> summary(Playersreg3)

Call:

lm(formula = logwage ~ height + weight + wins + goals + losses +

appearances + appsq + assists + age, data = Players)

Residuals:

Min 1Q Median 3Q Max

-1.60525 -0.19741 0.04321 0.25303 0.87336

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.716e+00 6.268e-01 7.524 4.61e-13 \*\*\*

height -9.215e-04 4.423e-03 -0.208 0.835

weight 1.567e-03 4.430e-03 0.354 0.724

wins -1.204e-03 4.613e-03 -0.261 0.794

goals 1.428e-03 3.384e-03 0.422 0.673

losses -3.364e-04 5.578e-03 -0.060 0.952

appearances 5.091e-04 3.757e-03 0.136 0.892

appsq -2.334e-07 1.396e-06 -0.167 0.867

assists 1.251e-03 2.709e-03 0.462 0.645

age -2.308e-03 6.012e-03 -0.384 0.701

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3862 on 347 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.003332, Adjusted R-squared: -0.02252

F-statistic: 0.1289 on 9 and 347 DF, p-value: 0.9989

> Playersreg4=lm(logwage ~ height+weight+wins+goals+losses+appearances+appsq+assists+age+def+mid+att, data=Players)

> summary(Playersreg4)

Call:

lm(formula = logwage ~ height + weight + wins + goals + losses +

appearances + appsq + assists + age + def + mid + att, data = Players)

Residuals:

Min 1Q Median 3Q Max

-1.54932 -0.19520 0.03045 0.25034 0.85866

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.817e+00 6.270e-01 7.682 1.65e-13 \*\*\*

height -2.721e-03 4.416e-03 -0.616 0.53812

weight 3.059e-03 4.393e-03 0.696 0.48674

wins -1.192e-04 4.597e-03 -0.026 0.97933

goals 8.933e-04 3.345e-03 0.267 0.78956

losses 4.583e-05 5.518e-03 0.008 0.99338

appearances -2.510e-05 3.720e-03 -0.007 0.99462

appsq -2.106e-07 1.383e-06 -0.152 0.87905

assists 1.071e-03 2.694e-03 0.398 0.69121

age -3.351e-03 5.983e-03 -0.560 0.57578

def 9.679e-02 7.210e-02 1.342 0.18033

mid 1.910e-01 7.187e-02 2.657 0.00824 \*\*

att 2.162e-01 7.801e-02 2.771 0.00589 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3811 on 343 degrees of freedom

(2 observations deleted due to missingness)

Multiple R-squared: 0.03424, Adjusted R-squared: 0.0004528

F-statistic: 1.013 on 12 and 343 DF, p-value: 0.4358