Asymmetric Incentives in the National Football League

Joseph Whitman[*]

Last Updated: November 18, 2019

Abstract

I use data on National Football League (NFL) games from the 1970-2017 regular seasons and newspaper sources to identify end-of-season matches where one team advances to the playoffs with a win while the opposing team's playoff status remains unchanged no matter the outcome. The asymmetric incentives to win these matches provide a way to identify and measure the effect of particularly strong incentives on group performance. I control for relative team strength by constructing and calibrating Elo ratings – a method of ranking competitors used in a wide variety of other sports – for NFL football, and compare this to results using point spread betting data. I find that both point spreads and Elo ratings are an accurate and highly significant predictor of the outcome of NFL games. I also find evidence that teams generally perform better when faced with stronger incentives to win. These effects are not present when using point spreads as a control, suggesting that the betting market uses this information when forming spreads.

1. Review of Literature

Economists generally agree that stronger incentives lead to increased effort. Prendergast (1999) provides an excellent review of the literature linking incentives to performance in firms. As shown by Lazear and Rosen (1981), tournament theory predicts that effort supplied should vary with the structure of incentives. Empirical studies such as Ehrenberg and Bognanno (1990) and Lazear (2000) provide further support for this assertion. Indeed, the literature on sports forecasting explicitly takes these incentive effects into account. For example, Goddard (2005) develops a model to forecast results of association

[*] I would like to thank Jonathan Hamilton, Steven Slutsky, Roger Blair, Dennis Coates, J.C. Bradbury, and Brian Mills for their helpful comments. I am grateful to Brad Humphreys for his comments and assistance with obtaining historical point spread data, as well as Rich Zuber, John Gandar, Andy Weinbach and Rodney Paul for their work compiling point spread data. All errors are my own.

football (soccer). In doing so, they develop an algorithm for determining the game's significance for each team. They ultimately find that teams for whom the game is significant have a higher probability of winning.

Athletes are skilled professionals who are highly compensated for their performance. In team sports, they individually benefit when their team advances to the playoffs. Depending on the terms of their contracts, players may be eligible for bonuses when their team makes a playoff appearance. However, players may benefit most from an increase in their expected future earnings. A memorable playoff performance or a championship victory can lead to lucrative sponsorships and more bargaining power when the player's contract is set to expire. Given such strong incentives to succeed, one might expect athletic performance to be at its best when the stakes are at their highest.

However, particularly strong incentives could reduce performance due to stress or psychological pressure. This phenomenon is often referred to as "choking under pressure".[1] In recent years, a growing number of empirical studies have attempted to find evidence of this "choking" effect using experiments or sports performance. For example, Dohmen (2008) finds evidence of this phenomenon in a study of penalty kicks in professional soccer. Hickman and Metz (2015) use professional golf tournament data from the PGA Tour to show that players are less likely to make a shot as the prize money riding on that shot increases. In an experimental study, Ariely et al. (2009) find that higher rewards generally lead to worse performance. Cao, Price, and Stone (2011) find evidence of choking during free throws in NBA basketball games. Apesteguia and Palacios-Huerta (2011) find that the team which takes the first kick in a penalty shoot-out is significantly more likely to win. They attribute this to psychological pressure on the players kicking second.

This paper primarily contributes to the literature by using NFL playoff seeding to identify and measure the effects of strong incentives on team performance. There has been relatively little research on this topic as applied to team performance, and – to my knowledge – none as applied to NFL football. All of the aforementioned studies focus on individual performance in high pressure situations with the exception of Goddard (2005) and, to some extent, Apesteguia and Palacios-Huerta (2011).

Aside from the general question of the existence of an incentive effect, this may be applicable to managers, executives, politicians, and anyone who wants to optimize group performance. An *individual* may "choke" under pressure, but that does not necessarily imply that a *team* will do the same. This choking effect may be not be present in teams for a number of reasons. Perhaps the psychological pressure is diffused among the team members or there exists some complementarity between team members that offsets this

[1] See Hill et al. (2010) for a review of the psychological literature on "choking".

negative effect.

In the NFL, there are often end of season games where one team advances to the playoffs if they win while their opponent's playoff status is unchanged no matter the outcome.[2] Individual players have a clear incentive to advance to the playoffs as this will result in greater bargaining power when renewing their contracts. Furthermore, teams have an incentive to win in order to increase fan interest, thereby generating additional revenue in the future. This incentive structure provides a natural experiment where I can identify the incentive effect by focusing on these cases of asymmetric incentives to win.

This paper also contributes to the literature by using Elo ratings to control for the relative strength of NFL teams. Betting market odds initially seem ideal for this purpose as they are a market-based forecast of game outcomes (Sauer, 1998). However, if NFL betting markets are efficient, incentive effects will also be taken into account. I expect this to be the case as NFL betting markets appear to be efficient (Boulier, Stekler, and Amundson, 2006). Therefore, I need to use a different measure that explicitly does not depend on any additional incentives to win. I follow an approach first proposed by physicist Arpad Elo (1978), a physicist who developed a ranking system for competitive chess players. This system has been modified for use in a wide variety of individual and team sports. These rankings explicitly do not depend on any additional incentives to win. There are many variations on the Elo rating system, but the version I use was first implemented by Silver (2014). I describe this rating system in greater detail in section 4.

2. History of NFL Franchises

In 1970 the NFL merged with its rival, the American Football League (AFL), to form a unified league consisting of 26 teams split into two conferences: the American Football Conference (AFC) and the National Football Conference (NFC). These conferences were further divided into three divisions per conference: East, Central, and West. Each season consisted of 14 games. The winners of these divisions – teams with the highest win percentage in their respective divisions – qualified for the playoffs in addition to one "wild card" team[3] from each conference for a total of eight playoff teams. This is widely regarded as the beginning of the modern era of American football, making this the ideal point to begin the sample period.

[2] It is possible to identify games with these characteristics earlier than the last week of the season. However, I ignore these cases because the outcome of other unrelated games will also determine the team's playoff odds in these earlier weeks. It is unclear that these games should be identified as particularly high pressure when the outcome only indirectly determines playoff seeding.

[3] A "wild card" team is one with the highest win percentage that did not win its division.

The number of teams, length of the regular playing season, and number of playoff qualifying slots have changed several times since the NFL-AFL merger. In 1976, the NFL added two additional teams, the Tampa Bay Buccaneers and the Seattle Seahawks. In 1978, the NFL added a second wild card for each conference which brought the total number of playoff teams to ten. The regular season was also extended to 16 games. This was followed by an era of expansion and relocation. The Oakland Raiders moved to Los Angeles in 1982, followed by the relocation of the Baltimore Colts to Indianapolis in 1984. In addition, the St. Louis Cardinals moved to Arizona in 1988. In 1990, The NFL expanded the playoffs to twelve teams by adding a third wild card slot for each conference. This was followed by the addition of two more teams – the Carolina Panthers and the Jacksonville Jaguars – in 1995. The Cleveland Browns relocated and became the Baltimore Ravens in the following year. The new Cleveland Browns were added to the league in 1999. The 32nd and final additional team was the Houston Texans, added in 2002. The NFL realigned its conferences into four divisions with eight teams each: North, South, East, and West. The NFL kept the playoffs limited to 12 teams by eliminating the third wild card slot in each conference.

There are two seasons in the sample which were shortened due to labor strikes. In 1982, the season was shortened to 9 games. This resulted in a unique 16-team playoff tournament in which division standings were ignored. Instead, the 8 highest ranked teams in each conference qualified. In 1987, the strike was shorter so the season was only limited to 15 games. However, games 4-6 of the season were played using replacement players who were of much lower quality than the professional roster. I ignore these seasons in my analysis.[4]

Table 1 summarizes these changes to the structure of the NFL over the sample period. All teams are tracked through their various relocations and are treated as the same franchise in the data. For example, the Cleveland Browns prior to 1996 and the Baltimore Ravens thereafter are treated as a single “franchise”. This is reasonable because teams retain most players and personnel during relocation.

[4] Including these seasons does not substantively change my results.

Table 1: NFL Timeline

1970	NFL-AFL Merger
1976	Tampa Bay Buccaneers and Seattle Seahawks added
1978	Second wild card added to each conference; regular season extended to 16 games
1982	Labor strike shortens season to 9 games; Oakland Raiders relocate to Los Angeles
1984	Baltimore Colts relocate to Indianapolis
1987	Second labor strike shortens season to 15 games; 3 games played using replacement players
1988	St. Louis Cardinals relocate to Arizona
1990	Third wild card added to each conference
1995	Carolina Panthers and Jacksonville Jaguars added; Los Angeles Raiders move back to Oakland; Los Angeles Rams move to St. Louis
1996	Cleveland Browns relocate, become Baltimore Ravens
1997	Houston Oilers move to Tennessee
1999	New Cleveland Browns added; Oilers renamed to Titans
2002	Houston Texans added; third wild card eliminated; conferences realigned into four divisions
2016	St. Louis Rams return to Los Angeles
2017	San Diego Chargers move to Los Angeles

3. Data

I use NFL football box score data collected from Pro-Football-Reference.[5] Data from the 1960-1969 seasons is used to generate the initial Elo scores for each team. I then use data from the 1970-2017 NFL regular seasons for the empirical analysis. Observations are at the game level. That is, one observation is one game coded as the home team versus the away team. Table 2 reports the summary statistics for the data set.

Table 2: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Points	10,644	22.159	10.442	0	14	29	62
Opp.Points	10,644	19.438	10.097	0	13	27	62
YardsGained	10,644	328.834	84.683	−7	271.8	385	653
Opp.YardsGained	10,644	313.772	85.844	26	255	372	676
Year	10,644	1,995.297	13.735	1,970	1,984	2,007	2,017
Points.Diff	10,644	11.361	9.811	−38	4	17	59
Points.Sum	10,644	41.597	14.285	3	31	51	106
YardsGained.Sum	10,644	642.607	120.572	234	559	720	1,125
Elo	10,644	1,497.895	86.735	1,240	1,436.2	1,558.0	1,787
Elo.Opp	10,644	1,501.667	86.586	1,233	1,440.3	1,563.5	1,790
in.	10,644	0.017	0.131	0	0	0	1
out	10,644	0.031	0.174	0	0	0	1
Elo Diff.	10,644	−3.700	123.186	−404	−86.0	77.7	452
homeline	8,904	−2.479	5.867	−27.000	−6.500	2.500	19.000

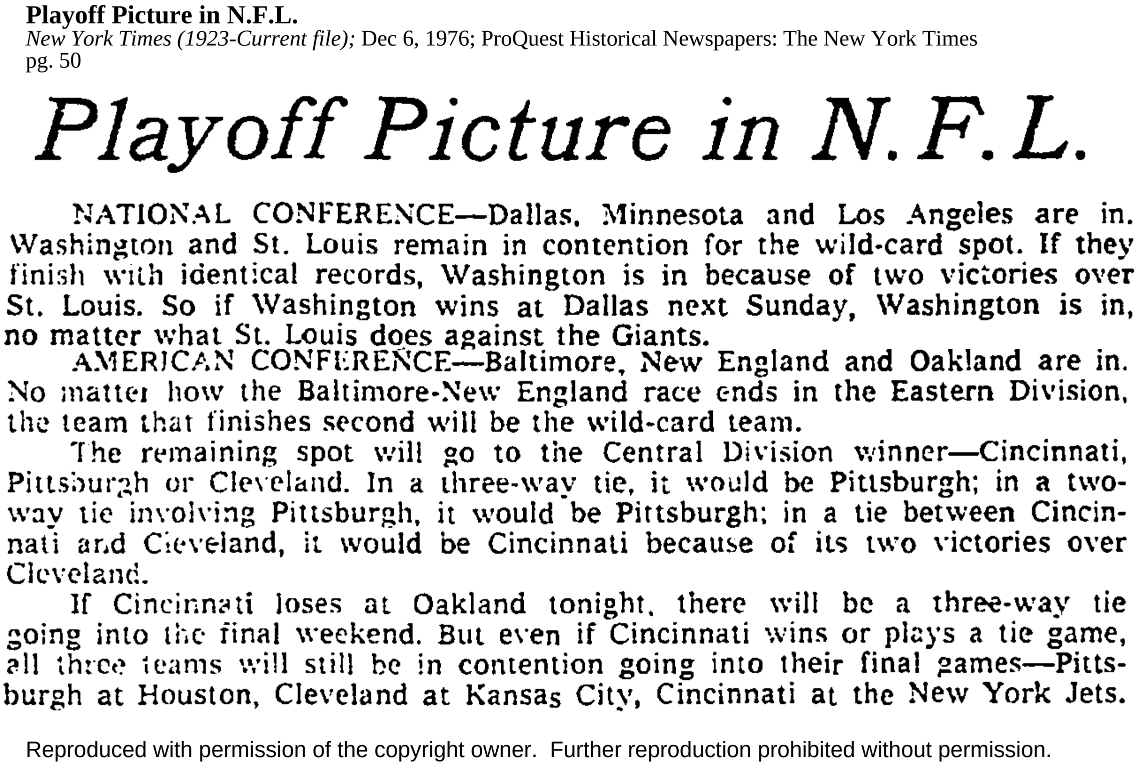
As I show in Section 4, Elo ratings are constructed using only box score data. No other information – such as injuries, individual player statistics, and so on – are used to generate the Elo ratings. Therefore, box score data is sufficient to create Elo ratings for the entire sample.

Data on NFL point spread betting odds are compiled from several sources. From 1978-2002, the data were compiled by hand from newspaper sources. I am grateful to Rich Zuber and John Gandar for collecting this data and to Brad Humphreys for providing it to me. Data from 2003-2017 were purchased from Sports Insights.[6] Betting data for 1970-1977 is unavailable. These years account for 1,484 regular season games, 106 of which occur in the final week of the season. This leaves 571 final week observations. In all cases, the point spread collected is the closing spread – that is, the last reported

[5] `http://www.pro-football-reference.com/`

[6] This data used to be available for purchase from SportsInsights.com, but as of January 2018 Sports Insights no longer sells access to their database of historical data.

spread before the game begins. Point spreads are recorded as positive numbers if the home team is favored to win, and negative if the visiting team is favored. To determine which end-of-season games exhibit asymmetric incentives, I compiled each team's playoff standings entering the final week from the sports sections of a variety of newspapers.[7] Figure 1 shows an example of the type of newspaper articles which are used to determine each team's playoff possibilities.



Playoff Picture in N.F.L.
New York Times (1923-Current file); Dec 6, 1976; ProQuest Historical Newspapers: The New York Times
pg. 50

Playoff Picture in N.F.L.

NATIONAL CONFERENCE—Dallas, Minnesota and Los Angeles are in. Washington and St. Louis remain in contention for the wild-card spot. If they finish with identical records, Washington is in because of two victories over St. Louis. So if Washington wins at Dallas next Sunday, Washington is in, no matter what St. Louis does against the Giants.

AMERICAN CONFERENCE—Baltimore, New England and Oakland are in. No matter how the Baltimore-New England race ends in the Eastern Division, the team that finishes second will be the wild-card team.

The remaining spot will go to the Central Division winner—Cincinnati, Pittsburgh or Cleveland. In a three-way tie, it would be Pittsburgh; in a two-way tie involving Pittsburgh, it would be Pittsburgh; in a tie between Cincinnati and Cleveland, it would be Cincinnati because of its two victories over Cleveland.

If Cincinnati loses at Oakland tonight, there will be a three-way tie going into the final weekend. But even if Cincinnati wins or plays a tie game, all three teams will still be in contention going into their final games—Pittsburgh at Houston, Cleveland at Kansas City, Cincinnati at the New York Jets.

Figure 1: Example Source for Playoff Scenarios

I will focus my analysis on games played in the final week of each season because playoff standings in weeks prior are difficult to pin down. The NFL has complex rules for breaking ties in playoff standing between teams and these rules have changed many times over the years. Furthermore, the structure of the NFL's playoff system often results in complex cases where the conditions for a team to enter the playoffs depend on the outcome of many other future games. As one journalist remarked, "...the National Football League playoff formula – a puzzle that, if it has not befuddled brains since the dawn of Time, certainly has done so since the dawn of Sports Illustrated. [...] the NFL play-

[7] These sources are listed individually in the reference list.

off system is impossible to understand."[8] Although going back to earlier weeks in the regular season would yield more data, the dependency of teams' playoff chances on the outcome of other teams' games would make it highly unlikely to observe any incentive effects.

The full sample contains 11,403 total observations. After removing the 1982 and 1987 strike seasons and all post-season games, there are 10,644 observations remaining. Of these, 677 observations are games that take place during the final week of the season. Teams can have four possible playoff standings as they are entering the final week:

1. Will advance to the playoffs with a win. These teams are commonly referred to as being "on the bubble".

2. Already qualified ("in").

3. Cannot qualify regardless of outcome ("out").

4. Need to win and one or more other teams to lose to advance to the playoffs ("in the hunt").[9]

Table 3 shows the number of observations for all possible permutations of standings between the home and away teams for each game. Notice that several of these combinations have few observations.

Table 3: Summary of Last Week Standings

		Away				
		Bubble	In	Out	Hunt	**Total**
Home	Bubble	22	17	30	9	**78**
	In	12	42	112	19	**185**
	Out	33	87	174	40	**334**
	Hunt	8	20	46	6	**80**
	Total	**75**	**166**	**362**	**74**	**677**

[8] Allen Abel, "Finding a Formula to the NFL Puzzle", The Globe and Mail (Dec. 12, 1978)

[9] There are often cases where a team needs a confluence of one or more events to occur in addition to a win. These scenarios can become quite complex. For example, the 1989 Pittsburgh Steelers needed to win and have four other teams win their games in order to advance. In fact, all four of those other teams won and Pittsburgh entered the playoffs.

4. The Elo Rating System

The Elo rating system was created by Arpad Elo to rank competitive chess players (Elo (1978)). It has since been adapted for other competitive sports such as tennis, basketball, soccer, and many more. Each competitor (or team) is assigned an Elo rating based on the results of past games. This rating is meant to be a measure of a competitor's current strength. Hvattum and Arntzen (2010) provide a generic version of the Elo rating system for team sports which I will summarize below.

Each game consists of two teams playing against each other, the home team and the away team. Suppose Elo_0^H is the initial Elo rating for the home team and Elo_0^A is the rating for the away team. Then define the home team's expected probability of winning[10] as:

$$x^H = \frac{1}{1 + c^{(Elo_0^A - Elo_0^H)/d}} \tag{1}$$

for the away team:

$$x^A = 1 - x^H = \frac{1}{1 + c^{(Elo_0^H - Elo_0^A)/d}} \tag{2}$$

The actual result of the game for the home team is simply:

$$r^H = \begin{cases} 1 & \text{if the home team won} \\ 0.5 & \text{if the game results in a tie} \\ 0 & \text{if the home team lost} \end{cases} \tag{3}$$

After the game, the Elo ratings for each team are updated as follows:

$$Elo_1^H = Elo_0^H + k(r^H - x^H) \tag{4}$$

Note that any points gained by one team are necessarily lost by the opposing team. This means all changes in the distribution of Elo ratings are mean-preserving. The parameters c and d determine the scale for the ratings and the parameter k determines the impact of new game results on the Elo rating. A low value of k means each individual game is worth relatively little so the rating will adjust to new information slowly. Conversely, a high value of k means the rating to be determined mostly by the latest games, causing the Elo rating to fluctuate more rapidly between games.

The initial Elo ratings are also important. Incorrect initial ratings will yield incorrect

[10] The expected probability of winning is derived from a logistic distribution of base c. I follow the literature and use a logistic distribution. However, it is possible that other distributions could provide a better fit.

predictions. Therefore, it is important to calibrate the initial ratings by using out of sample data. I initialize each team with the mean Elo rating in 1960, then use data on AFL and NFL teams during the 1960-1969 pre-merger seasons to generate initial ratings for the start of the sample period in 1970.

This basic formulation of Elo ratings does not account for margin of victory. This can be included by replacing the k parameter with another expression:

$$k = k_0(1+|PD|)^{\lambda} \tag{5}$$

where $k_0 > 0$, $\lambda > 0$, and $|PD|$ is the absolute value of the point difference.

Hvattum and Arntzen (2010) apply this system to professional soccer data and compare the predictive power of Elo ratings to several other alternative benchmarks. The scaling parameters are set as $c = 10$ and $d = 400$.[11] The adjustment parameters are then calibrated to $k_0 = 10$ and $\lambda = 1$. They conclude that Elo ratings are less predictive than betting odds, but better than the other methods tested. They speculate that betting odds are more predictive because they take into account other relevant information – such as player injuries – that Elo ratings lack. However, their analysis is focused on professional soccer so it is unclear if their results will generalize to the NFL.

4.1. FiveThirtyEight Elo Ratings

Silver (2014) adapts the Elo rating system to the NFL. He proposes the following modified margin of victory term which discards the λ parameter in favor of a multiplier that is a function of the difference in Elo ratings:

$$k = k_0 \ln(1+|PD|)\frac{s}{s+(Elo_W - Elo_L)*.001} \tag{6}$$

where s is a parameter and Elo_W and Elo_L are the Elo ratings for the winning and losing team, respectively. This new term replaces the k parameter in equation (4). This method discounts the margin of victory for strong teams and inflates it for weak teams. Silver argues that this is desirable because strong teams are more likely to win and often do so by a large margin. However, he does not elaborate on how he chose this particular expression. Silver uses the same scaling parameters ($c = 10$ and $d = 400$) as Hvattum and Arntzen (2010) and chooses parameter values $k_0 = 20$ and $s = 2.2$. He does not elaborate on how he selected these values.

[11] The scaling parameters are inconsequential. Only the difference in Elo ratings matters for predictive purposes.

NFL football teams undergo considerable personnel changes from season to season. Last season's championship team may be missing several of their star players at the start of the next season. Therefore, it is reasonable to discount the previous season's Elo ratings when moving onto the next season. Silver adjusts for this by reverting all team scores to the mean by one third. I will adopt this method for dealing with seasonal changes going forward.[12] I also need to account for expansion teams entering the league and the relocation of existing teams. I give new teams the mean Elo rating (1500) while relocated teams retain their existing rating. Although new teams may not truly be of average quality upon entering the league, I will use some simulations in section 5 to argue that they will converge to their "true" rating after several games.

4.2. Optimal Elo Ratings

It is unclear which version of the Elo rating system is a better predictor of the outcome of a game. We have two competing methods for capturing the margin of victory: equations 5 and 6. Furthermore, the parameters k_0, λ, and s need to be calibrated to maximize the predictive power of either system. The parameters c and d simply set the scale for the rating system so I use the same values ($c = 10$ and $d = 400$) as Hvattum and Arntzen (2010) and Silver (2014) for ease of comparison.

I use the mean squared error because it is perhaps the most commonly used statistical loss function for evaluating the accuracy of probabilistic predictions. However, different loss functions may yield different optimal parameters. For example, choosing parameters to minimize the mean absolute error yields different optimal parameter values. I use the mean squared error because it is most common, but comparing results across different loss functions may be of interest for future research.

The MSE is given by the following function

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(r^H - x^H)^2 \tag{7}$$

where n is the number of observations in the sample. To calibrate the rating system, I choose parameter values to minimize the mean squared error (MSE) between the expected probability of winning and the actual result of the game for every team-game observation in the sample period. I computed the Elo ratings over the entire sample for 3000 different combinations of parameters for each of the two competing rating systems. I then computed the MSE over the sample for each combination of parameters and found

[12] Whether this is the most optimal way to deal with season to season changes is debatable. I leave this question for future research.

the minimum. The results of my calibrations are reported below in Table 4.

Table 4: Benchmarks vs. Calibrated Parameters

	Silver	Silver (calibrated)	H-A	H-A (calibrated)
k_0	20	17.5	10	15.1
s	2.2	1.7	-	-
λ	-	-	1	1.1
MSE	0.22404	**0.22389**	0.22548	0.22403

The calibrated version of the Elo rating system with Silver's margin of victory term has the lowest mean squared error. Therefore, I will proceed using this implementation of Elo ratings for the remainder of the paper.

5. Robustness of Elo Ratings

In this section, I present some simulations to test the robustness of the Elo rating system. For example, what would happen if a bad team was incorrectly given a high Elo rating? How quickly would their rating adjust to a reasonable value? To answer this question, I simulate a season in which a team starts with a particularly high rating (1700) and loses by 10 points to opponents with Elo ratings drawn randomly between 1300 and 1700. I chose this range because nearly all the ratings in the sample fall within this interval, as shown in Figure 2.

I repeat this simulation 100 times and present the results in Figure 3. Note that the horizontal line corresponds to the mean Elo rating (1500). Although this team is overrated for the first half of the season, it has an appropriately low value by the end. I do the same procedure for a team with a particularly low rating (1300) continuously winning in Figure 4. Once again, the team's Elo rating more accurately reflects the team's true strength by the end of the season. This suggests erroneous Elo ratings converge to their "true" value as the season progresses. Therefore, Elo ratings are a suitable proxy for relative team strength – at least by the end of the season – even if the initial values are calibrated incorrectly. I am primarily interested in the outcome of end-of-season games, where Elo ratings are most accurate. This suggests Elo ratings are an acceptable control for relative team strength.

One might also be concerned that there is some underlying autocorrelation in Elo ratings. In Figure 5, I simulate 100 seasons of a team with the mean Elo rating playing against opponents with random Elo ratings. The outcome of each game is also drawn randomly. As expected, the figure is just noise.

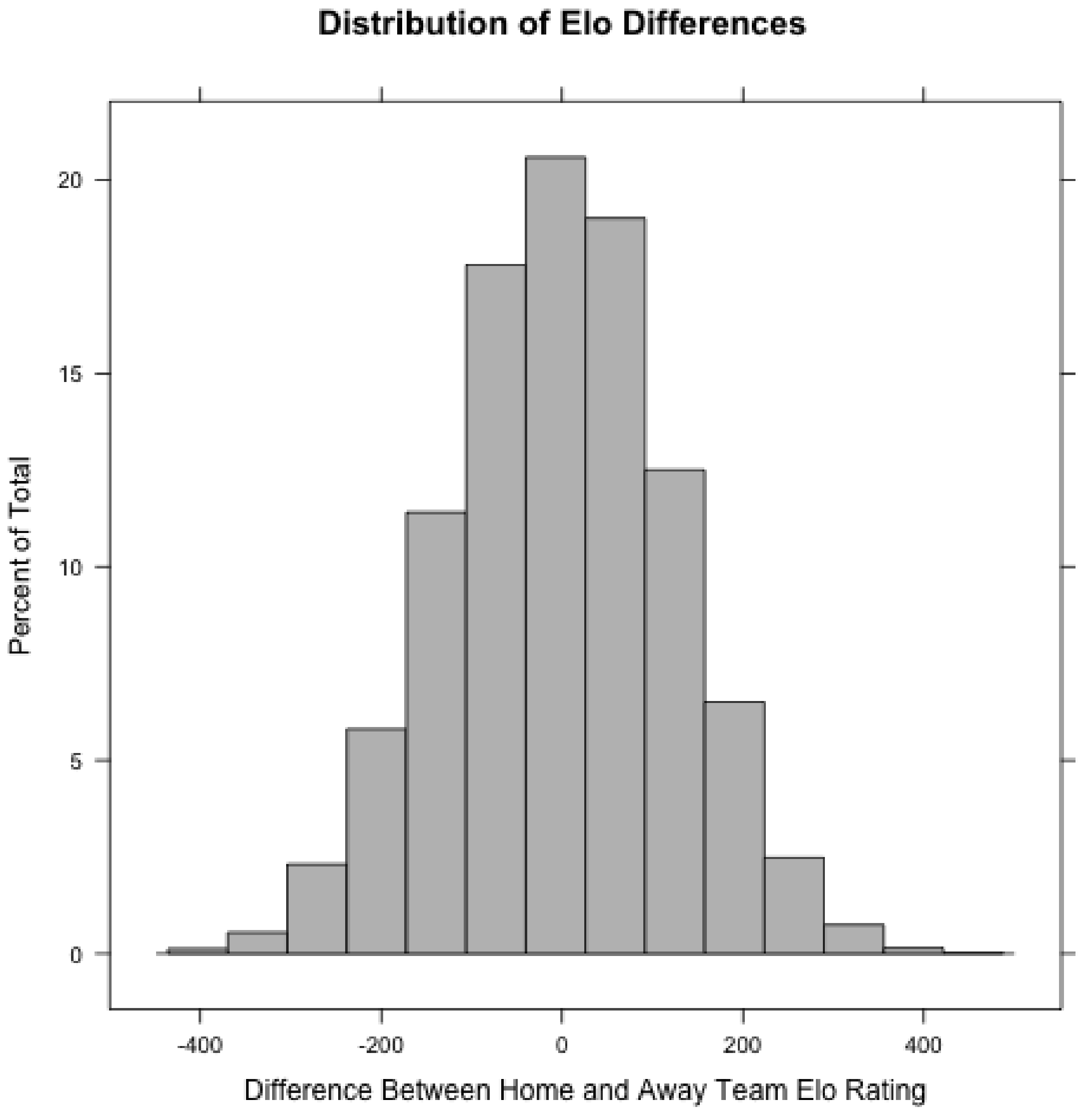


Figure 2: Distribution of Elo Ratings in the Sample

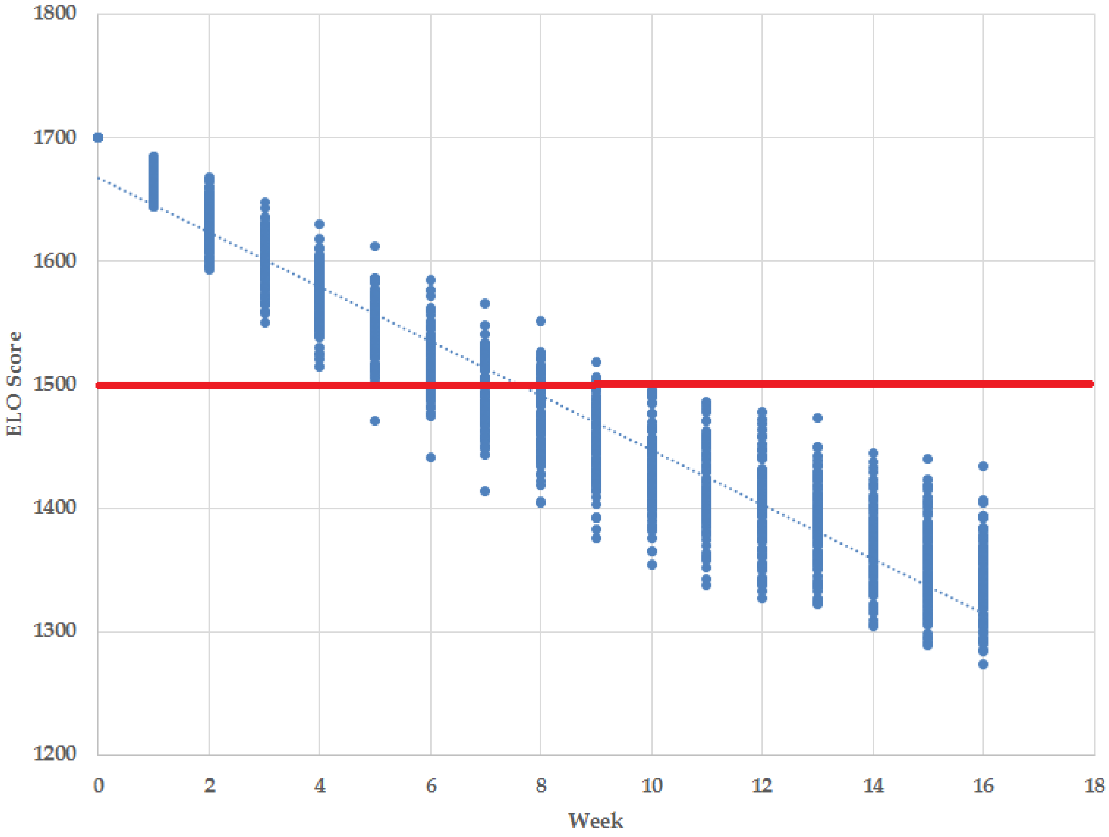


Figure 3: Simulation of a highly rated team continuously losing to random opponents

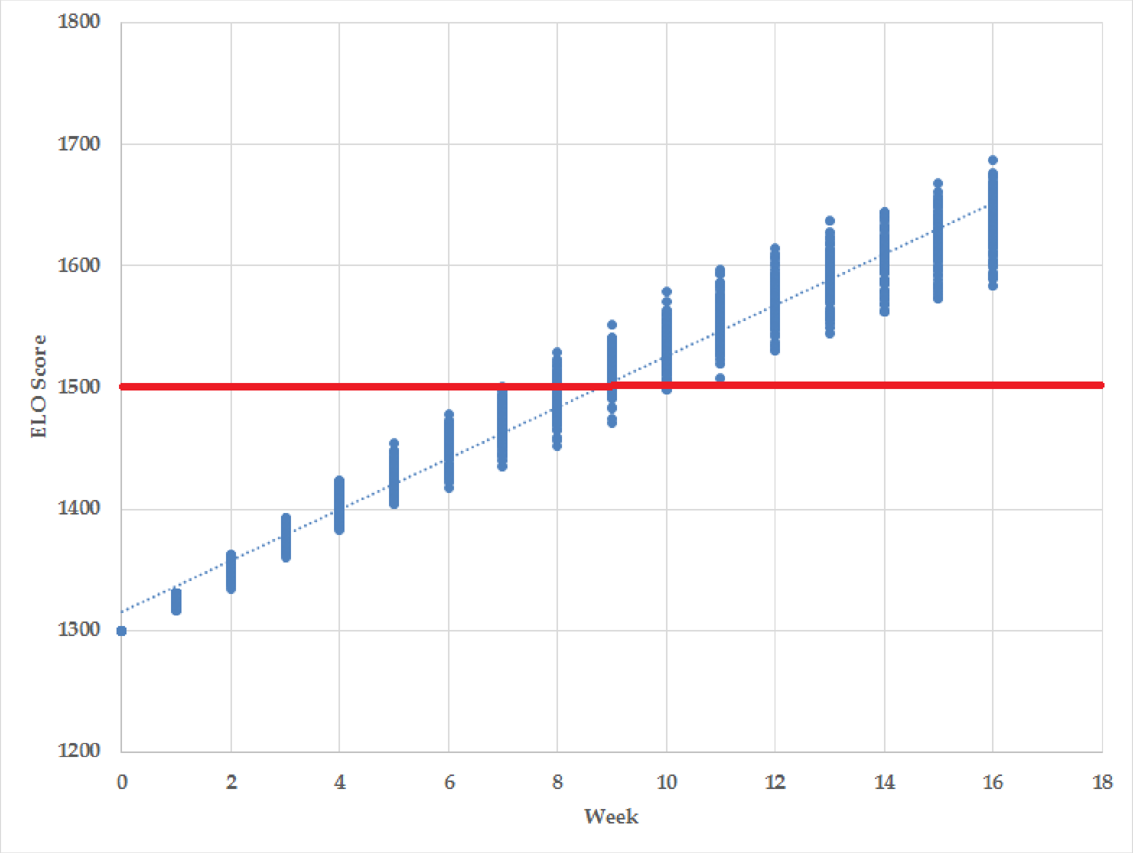


Figure 4: Simulation of a low rated team continuously winning against random opponents

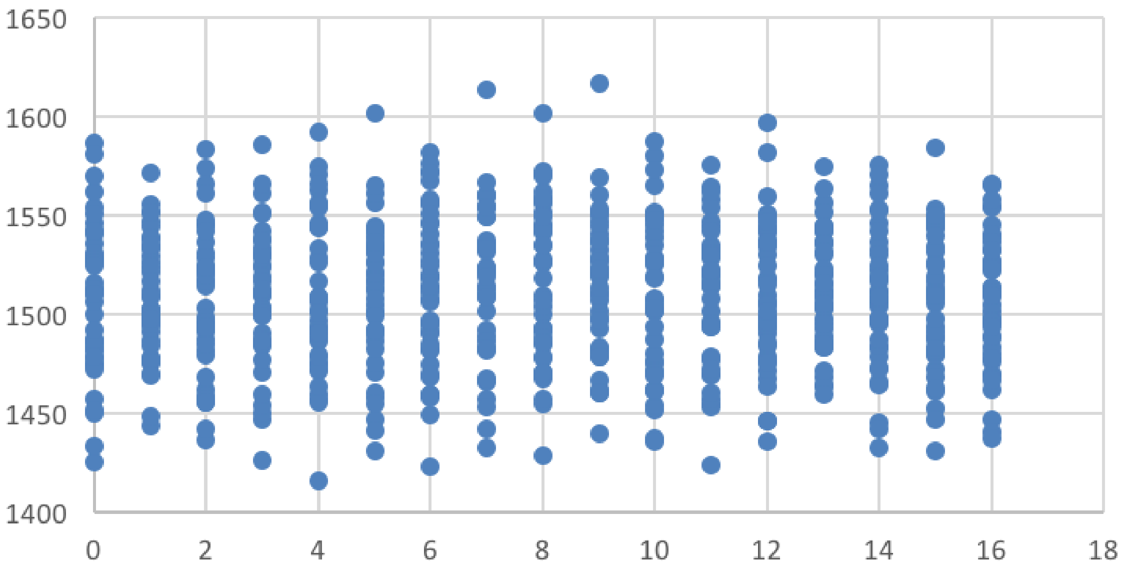


Figure 5: Simulation of an average team randomly winning against random opponents

I also simulate what would happen if a team was incorrectly rated at the beginning of the 2016 season. Figure 6 shows how the Elo rating for the Minnesota Vikings would develop over the season with four different starting values: 1700, 1500, 1300, and their actual Elo rating (1560). Although these Elo ratings converge closer to their true values as the season goes on, there remains a persistent gap. This occurs because the Vikings can only gain rating points with a win and can only lose rating points with a loss. Therefore, the Vikings' Elo rating cannot fall even when starting at an artificially high value (1700) until they start losing.

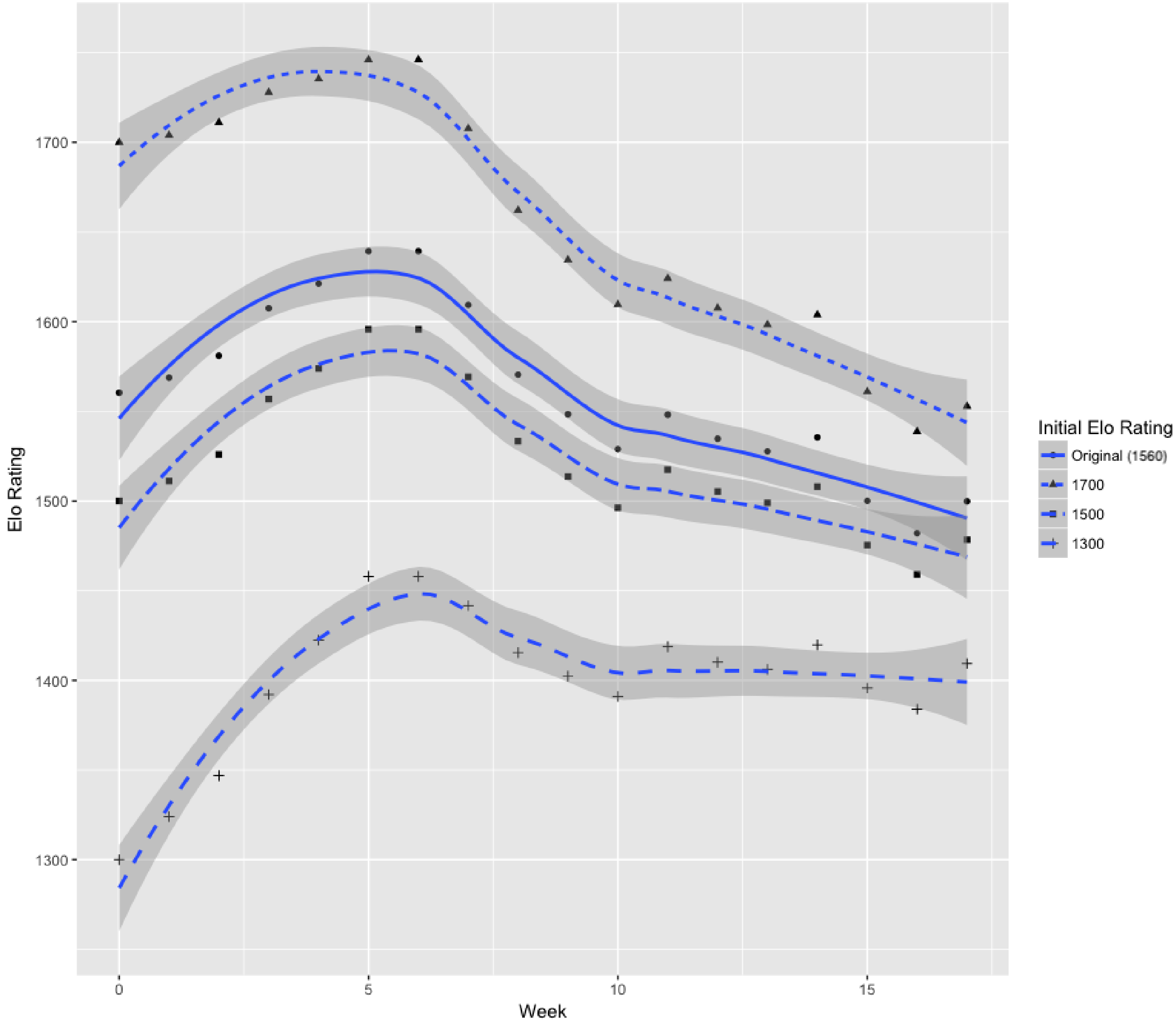


Figure 6: Ratings Development for the Minnesota Vikings (2016)

Figure 7 shows the same simulation for the Jacksonville Jaguars team. Because this team lost many games during the 2016 season, their Elo rating falls quickly even when they start with an incorrect rating of 1700. Conversely, they do not raise their rating when they start with an incorrectly low rating of 1300 except when they score wins.

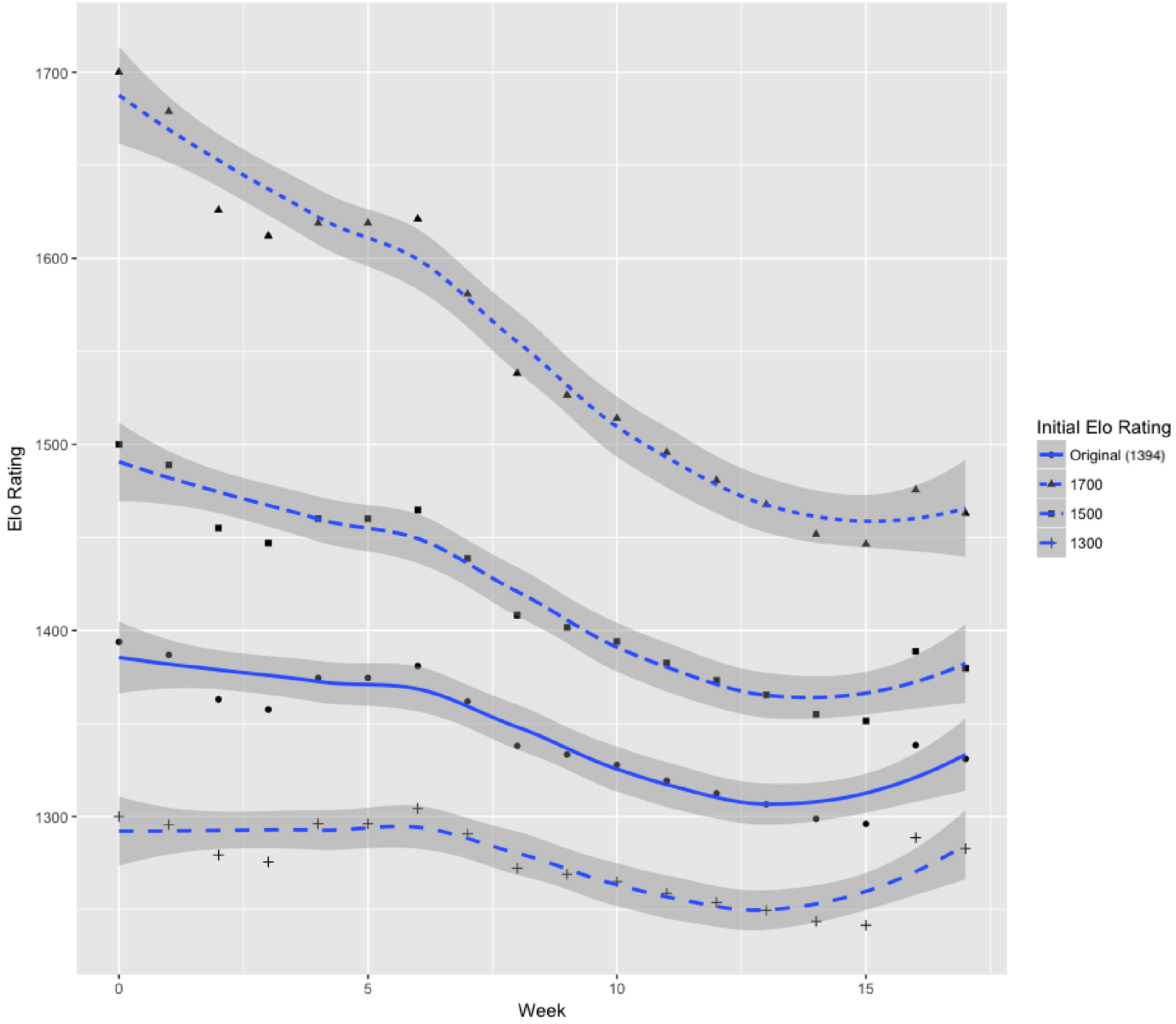


Figure 7: Ratings Development for the Jacksonville Jaguars (2016)

Figure 8 again repeats this simulation but with the New England Patriots team. This team won most of their games during this season, so their Elo rating quickly rises when incorrectly started at 1300.

Figure 9 is a conditional density plot showing the relationship between the game result for the home team and the difference in elo rating between the home team and the visiting team. The dark grey region shows the probability the home team will lose conditional on the difference in Elo ratings. As expected, this probability is very high when the home team is facing a much more highly rated team and approaches zero as the dif-

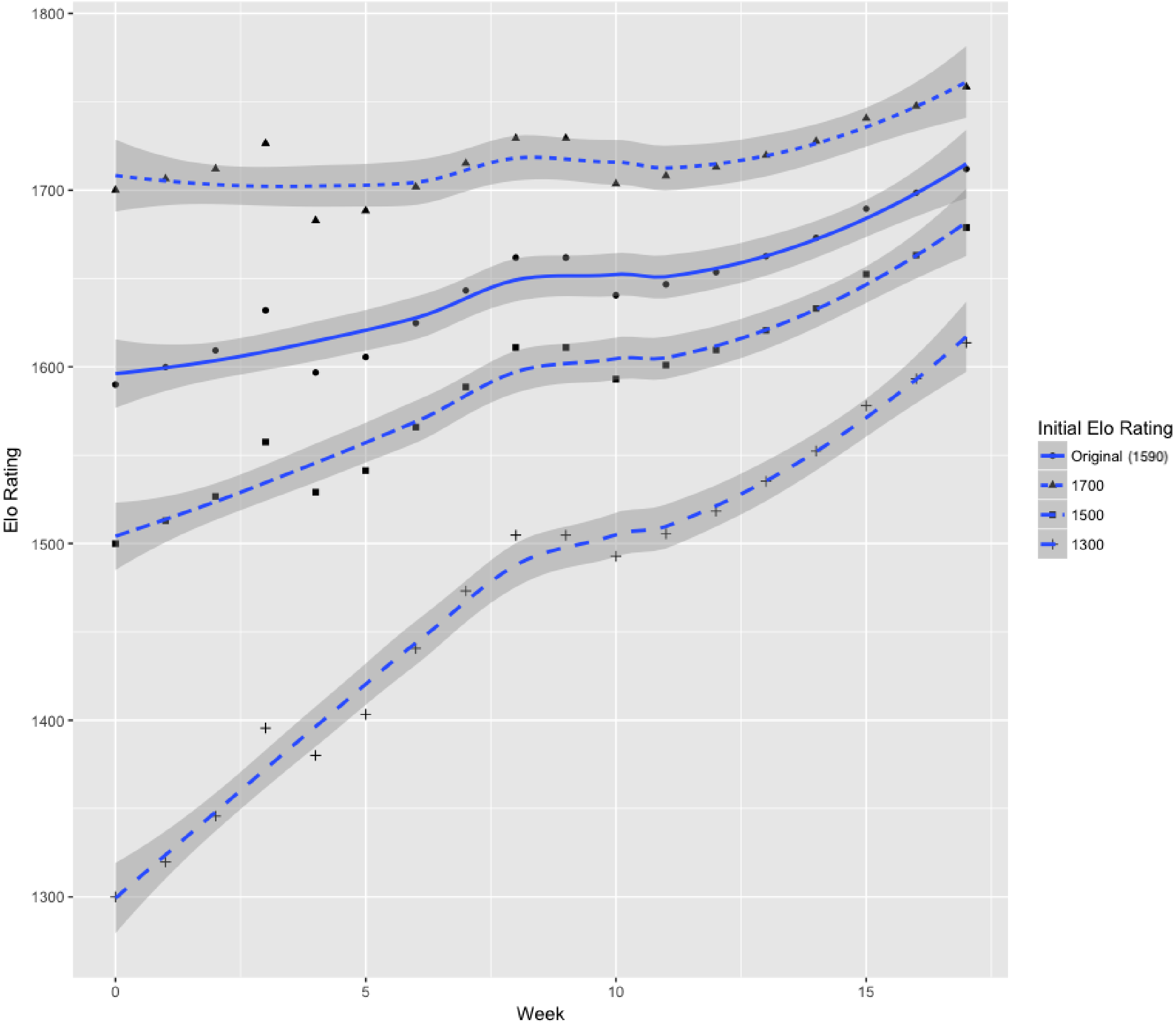


Figure 8: Ratings Development for the New England Patriots (2016)

ference in Elo ratings increases in favor of the home team. Note that this relationship appears to be approximately linear over the range of Elo ratings -200 to 200, but appears nonlinear near the extremes (-400 and 400). I observe that most games (approximately 89.3%) fall in the -200 to 200 range. This suggests the nonlinearity at the extremes occurs because there are relatively fewer games with such large differences in Elo ratings.

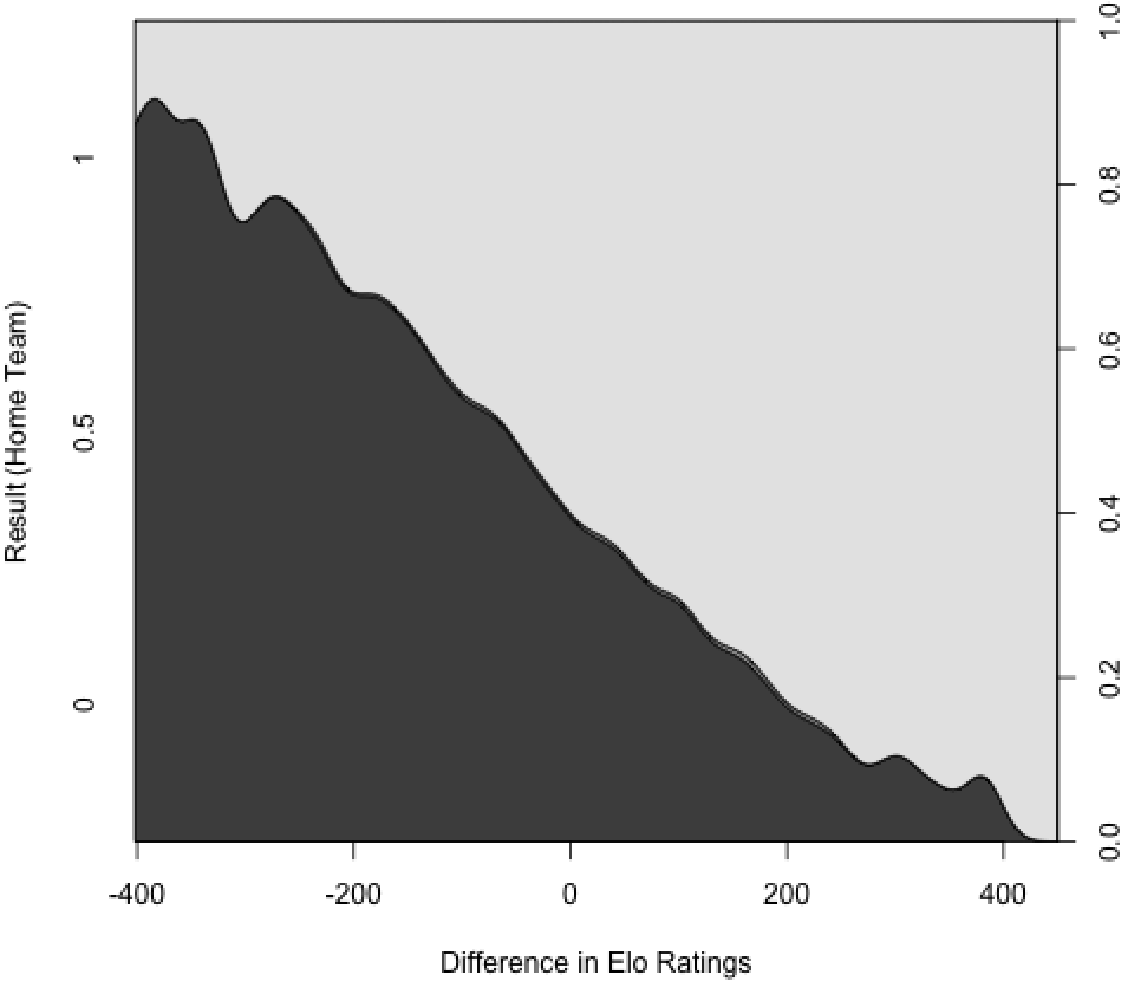


Figure 9: Conditional Density Plot (Result vs. Difference in Elo Rating)

This is also illustrated in Figure 10, which shows the results of a simple logistic regression with the difference in Elo ratings as the only independent variable. The vertical axis is the fitted values from the regression and the horizontal axis is the difference in Elo rating. As observed in the conditional density plot, the relationship between Elo rating difference and game outcomes is approximately linear.

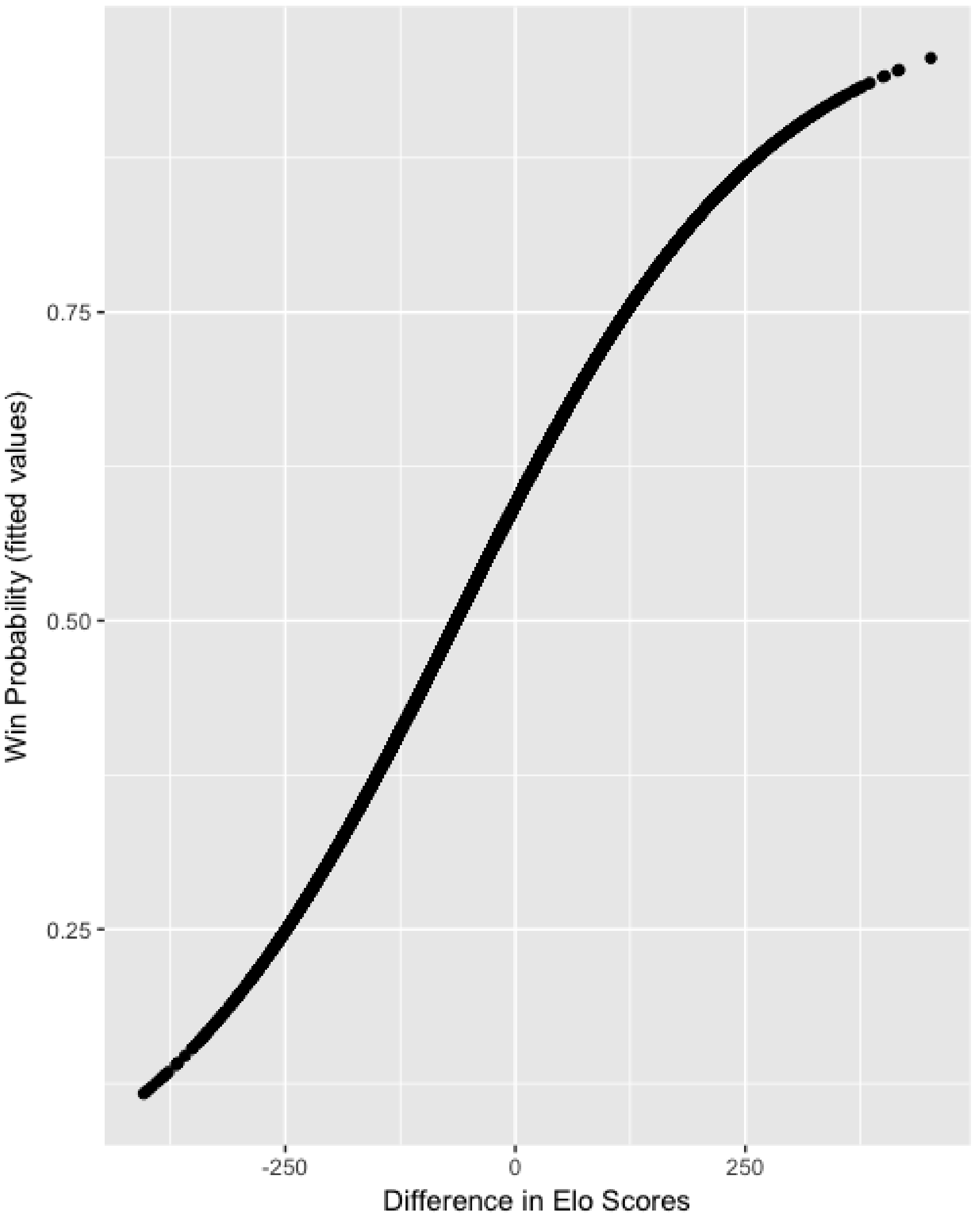


Figure 10: Win Probability versus Difference in Elo Rating (fitted values)

So far, I have argued that the difference in Elo ratings is predictive. However, it is possible the Elo rating *levels* also matter. In other words, is it possible that teams in the highest quantile of the rating distribution perform differently than those in the lowest quantile even with the same rating difference? To test this, I split the data by into quantiles by Elo rating and replicate Figure 10. The result is displayed in Figure 11. The lowest rated are in the first quantile while the highest are in fourth quantile. Interestingly, it appears that – for a given difference in Elo ratings, home teams in the highest quantile have a higher probibility of winning. However, this result is not statistically significant. When plotting the 95% confidence interval for each quantile as in Figure 12, we can see that the confidence intervals overlap.

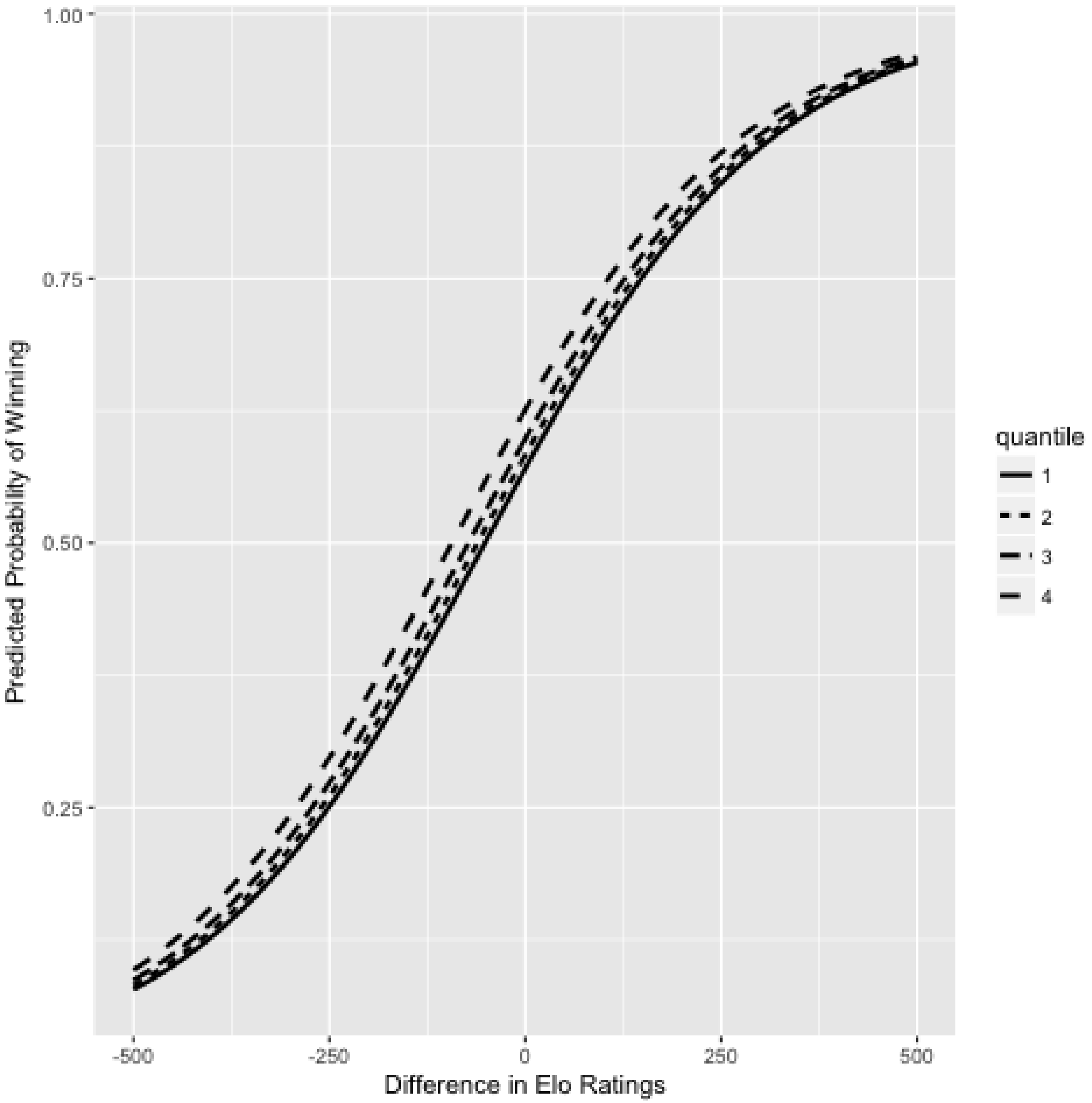


Figure 11: Win Probability versus Difference in Elo Rating, Grouped by Quantile

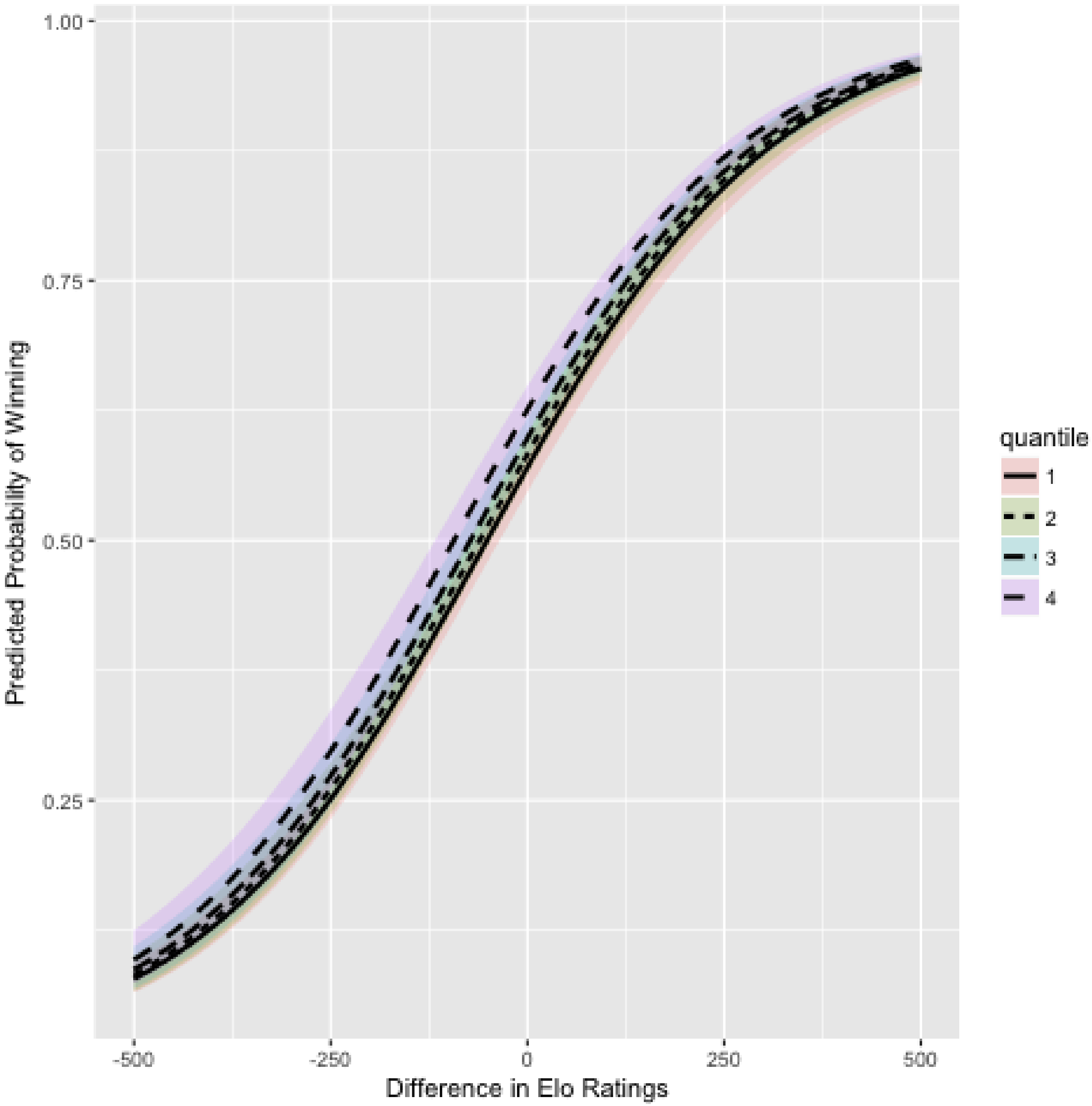


Figure 12: Win Probability versus Difference in Elo Rating, Grouped by Quantile (with error bars)

I also use the Ramsey RESET test to verify that a linear relationship between Elo rating difference and probability of winning is appropriate.[13] I find that no non-linear specifications were statistically significant.

6. Methodology

I use a binary logit regression model with the game result as a binary dependent variable equal to one if the home team wins and zero otherwise (tie or loss). The previously mentioned indicator variable ($Bubble$) is the independent variable of primary interest.

At first, it may seem that an ordered logit model is preferable to a binary logit model here because game outcomes are not truly binary due to the presence of tie games. However, ties are relatively rare in NFL football (there are only 51 tie games in the entire sample) and there are no tie games in the final week of the regular season.[14]

The following is the general form of the regression equation which will be used:

$$\text{Result}_{iy} = \alpha + \beta \text{Strength}_{iy} + \gamma \text{Bubble}_{iy} + \epsilon_{iy} \tag{8}$$

where i indexes each individual game and y indexes the season. Result_{iy} is the result of game i during season y. This is a binary variable which is equal to one if the home team wins and zero otherwise (home team loses or there is a tie). Strength_{iy} controls for the difference in "strength" between the teams. This will be proxied by the point spread or difference in Elo ratings depending on the specification used. Bubble_{iy} is the variable of interest. It is a vector of indicator variables for playoff standing of the teams playing game i in the final week of season y. α is a constant term which is estimating the home team advantage. Lastly, ϵ_{iy} is the error term.

If Elo ratings and point spreads are good controls, the estimators for both will be positive and statistically significant. The magnitude of these estimators should also be large enough to have practical significance.

If betting markets are efficient, point spreads should already take any additional incentive effects into account. When they are used to control for team strength, I expect to find the playoff standing indicators to be small in magnitude and/or statistically insignificant. The constant should also be insignificant as the betting market should be accounting for home advantage.

[13] See Ramsey, J. B. (1969). "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis". *Journal of the Royal Statistical Society Series* B. 31 (2): 350–371.

[14] I have also used an ordered logit model for all reported regressions but, unsurprisingly, the results do not substantively change.

However, the constant term and playoff standing indicators should have a positive and significant effect when Elo ratings are used in place of point spreads. Elo ratings are constructed such that they only include past performance and none of the forward-looking expectations an efficient market would.

A team's playoff standing entering the final week of the season is, of course, not random. It is determined by the team's record of wins and losses in prior games. Thus there make exist some endogeneity between the Bubble variable(s) and the control for team strength. One might be concerned that stronger teams are somehow more (or less) likely to end up on the bubble, which would bias the estimated coefficient upward (or downward). However, the schedule of games is determined by the league prior to the beginning of the season. Therefore, the matchup in the final week explicitly does not depend on the team's strength. Furthermore, as mentioned previously in Section 3, a team's playoff standing may also depend on the outcome of other unrelated games which is exogenous to the strength of the team in question.

It is unclear what to expect from teams who are not on the bubble. Teams that have already qualified may be more likely to win as they can potentially gain better seeding in the playoffs. They may also be less likely to win if they choose to rest their best players so to avoid injuries before the playoffs.

Teams that are already out may have an incentive to continue losing in order to gain better draft position for the next season. This is known as "tanking" or "sandbagging". Although this may be advantageous in theory, as shown in KrÃ€kel, 2014, it is unclear if NFL teams do this in practice.

7. Empirical Results: Point Spreads

Table 5 reports the logit regression results when the point spread is used to proxy for team strength. The first column is a specification with the point spread as the only independent variable. The second column adds a simplified set of playoff standings indicators: "Home Bubble" and "Away Bubble". The former is equal to one if the home team advances to the playoffs with a win and zero otherwise. The latter is the same but for the away team. The third and final column includes indicators for all permutations of playoff standings between the home and away team as shown in Table 3. For example, "Bubble vs. Bubble" is equal to one if both teams need to win to advance to the playoffs and zero otherwise. "Out vs. In" is equal to one if the home team cannot advance regardless of the outcome and the away team has already qualified for the playoffs, and zero otherwise. The excluded category is "Out vs. Out".

Table 5: Logit Regression Results: Point Spreads

	Dependent variable:		
	Result		
	(1)	(2)	(3)
Strength (Point Spread)	0.160^{***}	0.159^{***}	0.159^{***}
	(0.018)	(0.018)	(0.023)
Home Bubble		0.249	
		(0.322)	
Away Bubble		0.072	
		(0.299)	
Bubble vs. Bubble			−0.620
			(0.491)
Bubble vs. In			2.565^{**}
			(1.090)
Bubble vs. Out			−0.090
			(0.564)
Bubble vs. Hunt			1.044
			(1.101)
Out vs. Bubble			0.330
			(0.456)
Out vs. In			−0.287
			(0.336)
Hunt vs. Hunt			-2.170^{*}
			(1.127)
Out vs. Hunt			−0.421
			(0.448)
In vs. Bubble			−0.045
			(0.743)
In vs. In			0.182
			(0.404)
In vs. Out			0.572
			(0.366)
In vs. Hunt			−0.271
			(0.545)
Hunt vs. Bubble			1.926^{*}
			(1.095)
Hunt vs. In			0.447
			(0.581)
Hunt vs. Out			-0.753^{*}
			(0.409)
Constant	0.096	0.062	0.081
	(0.101)	(0.110)	(0.186)
Observations	571	571	571
Akaike Inf. Crit.	661.979	665.156	655.889
Note:		*p<0.1; **p<0.05; ***p<0.01	

As expected, the point spread is statistically significant in all three regressions. The marginal effect is approximately a 3.7% increase in probability of winning for a 1 point increase in the spread. The constant term is statistically insignificant, indicating that the betting market priced in the home advantage. The indicator variables are not statistically significant at the 5% level, except for “Bubble vs. In”. This is the case where the home team needs to win to advance and the away team has already qualified for the playoffs. Compared to a game where both teams are out of playoff contention, the marginal effect is a 32.5% increase in the probability the home team wins. Although this is very large and statistically significant, it is worth nothing that there are only 17 observations in this category so this result may be due to small sample size.

8. Empirical Results: Elo Ratings

Table 6 reports results of the same regressions as the previous section, but using difference in Elo ratings instead of the point spread to control for team strength.

The control for team strength is once again highly statistically significant, this time with Elo ratings. The marginal effect is positive as expected but seemingly small in magnitude. A difference of one point only raises the probability of winning by 0.14%. However, differences in Elo ratings are typically much larger. One standard deviation of Elo difference in the data is approximately 143 points, which corresponds to a 20% increase in the win probability. The constant term is now highly significant. The marginal effect is 10.8% in model (2) and 12.86% in model (3). Once again, the matchup category “Bubble vs. In” is significant and even larger than the specification with point spreads. There is also a significant and large increase in probability of winning (14.43%) when the home team is “on the bubble”. Interestingly, there is no corresponding advantage for the away team. Notice that this effect is even higher than the home advantage. Neither of these effects were present when using point spreads. This is suggestive evidence that point spread betting markets are operating efficiently. That is, they are incorporating these matchup dependent incentive effects and the home advantage into the construction of the point spread.

An additional concern when comparing these models is the issue of goodness of fit. Could the model be improved by removing some of the matchup categories? Would it be more predictive in the simple case (only consider teams “on the bubble” versus those who are not)? To answer this question, we can use goodness of fit tests.

I used two approaches to answer this question: the likelihood ratio test and the Wald test. Using the likelihood ratio test to compare the simple model with the full model (all

Table 6: Logit Regression Results: Elo Ratings

	Dependent variable:		
		Result	
	(1)	(2)	(3)
Strength (Diff. in Elo Ratings)	0.006***	0.006***	0.005***
	(0.001)	(0.001)	(0.001)
Home Bubble		0.672**	
		(0.296)	
Away Bubble		−0.351	
		(0.269)	
Bubble vs. Bubble			−0.590
			(0.461)
Bubble vs. In			2.677**
			(1.052)
Bubble vs. Out			0.477
			(0.537)
Bubble vs. Hunt			0.658
			(0.820)
Out vs. Bubble			−0.276
			(0.403)
Out vs. In			−0.521
			(0.318)
Hunt vs. Hunt			−2.082*
			(1.112)
Out vs. Hunt			−0.756*
			(0.386)
In vs. Bubble			−0.642
			(0.624)
In vs. In			0.026
			(0.364)
In vs. Out			0.288
			(0.360)
In vs. Hunt			−0.421
			(0.521)
Hunt vs. Bubble			1.585
			(1.083)
Hunt vs. In			0.960*
			(0.549)
Hunt vs. Out			−0.398
			(0.372)
Constant	0.494***	0.461***	0.554***
	(0.086)	(0.094)	(0.163)
Observations	677	677	677
Akaike Inf. Crit.	812.292	810.124	799.394
Note:		*p<0.1; **p<0.05; ***p<0.01	

matchup categories), I reject the null hypothesis that the reduced model provides a better fit than the full model at the 1% significance level. This is true for both the specification including point spreads and Elo ratings.

I also use the Wald test to make pairwise comparisons between matchup categories. When comparing the most empirically relevant categories such as “Bubble vs. Bubble”, “Bubble vs. In”, and so on, all pairwise comparisons are jointly significant. Only categories with low explanatory power, such as “In vs. Hunt” can be rejected from the model. In all cases, the sign and magnitude of the coefficients does not significantly differ from the results presented above.

9. Concluding Remarks

I find that the difference in both point spreads and Elo ratings is a highly significant predictor of game outcome in the NFL. Home advantage and most playoff standings results are not significant when point spreads are used as a control for team strength. When using Elo ratings in place of point spreads, these effects become significant. This indicates the point spread is taking all available information into account, which suggests the sports betting market is operating efficiently. This supports previous findings in the literature such as Boulier, Stekler, and Amundson (2006). When replacing point spreads with Elo ratings, I also find evidence that the additional incentive from being on the bubble is significant and large, but only for the home team. This increased performance in high pressure situations with stronger than normal incentive to win contradicts the findings of the choking under pressure literature and affirms prior findings showing larger incentives leads to higher effort.