Asymmetric Incentives in the National Football League

Joseph Whitman

Last Updated: November 4th, 2017

Abstract

I use data on National Football League (NFL) games from 1970-2016 and newspaper sources to identify end-of-season games 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 games 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. I find that the difference in Elo ratings are an accurate and highly significant predictor of the outcome of NFL games. Controlling for Elo ratings, I find that groups with stronger incentives to win perform better.

1. Introduction

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. 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. Goddard (2005) develops a model to forecast association football (soccer) match results. In doing so, they develop an algorithm for determining match significance and find that teams for whom the match is significant have a higher probability of winning.

Athletes are skilled professionals who are highly compensated for their performance. 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

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

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 et al. (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 NFL team performance. Aside from Goddard (2005) and to some extent Apesteguia and Palacios-Huerta (2011), all of these aforementioned studies focus on individual performance in high pressure situations. There has been relatively little research on this topic as applied to team performance, and – to my knowledge – none as applied to NFL football.

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] This incentive structure provides a natural experiment where I can identify the incentive effect by focusing on these cases of asymmetric incentives to win.[3]

NFL players 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 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.

This paper also contributes to the literature by using a novel approach for measuring relative team strength. Betting market odds seem ideal at first unless bookmakers are taking incentive effects into account when determining the point spread. I expect this to be the case, so I need to use a different measure that explicitly does not depend on these additional incentives to win. Therefore, I follow an approach first proposed by the

[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] For a true experiment to occur, explicit randomization would be required to determine these situations.

physicist Arpad Elo (1978) for competitive chess and modified for use in a wide variety of individual and team sports that explicitly does not depend on these additional incentives to win. Specifically, I use a modified version of the Elo rating system implemented by Silver (2014), which I describe in section 3.

2. Background

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[4] 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.

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

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

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.[5]

Table 5 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 team in the data. For example, the Cleveland Browns prior to 1996 and the Baltimore Ravens thereafter are treated as a single "franchise". This is sensible because teams typically retain most players and personnel during relocation.

3. 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[6] 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:

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

[6] 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.

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

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 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 $d = 400$.[7] 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.

3.1. 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:

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

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

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.

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.[8] 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 4 to argue that they will converge to their "true" rating after several games.

3.2. 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

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

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 the minimum. The results of my calibrations are reported below in Table 6.

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.

4. Elo Rating Simulations

In this section, I present some simulations to test the robustness of the Elo rating system. 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 1.

I repeat this simulation 100 times and present the results in Figure 2. 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 3. 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 4, 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.

I also simulate what would happen if a team was incorrectly rated at the beginning of the 2016 season. Figure 10 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 9 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 11 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.

5. Data

I use NFL football box score data collected from Pro-Football-Reference[9]. Data from the 1960-1969 seasons is used to generate the initial Elo scores for each team. I then use data from the 1970-2016 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 7 reports summary statistics for the data.

To determine which end-of-season games exhibit asymmetric incentives, I use data from various newspaper sports sections. McKillop (2013) has compiled many of these sources but stops at 2012. Playoff scenarios for 2013 and 2014 are from Smith (2013) and Smith (2014) respectively. From 1970-2016 there are 11,146 total observations. After removing the 1982 and 1987 strike seasons and post-season games, there are 10,388 observations remaining. 158 of these are end-of-season games where one team needs a win to assure that it will advance to the playoffs. I will refer to these teams as being "on the bubble". The opponents of these teams can have four possible playoff standings:

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

1. Already qualified ("in").

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

3. Also "on the bubble".

4. Need to win and one or more other teams to lose to advance to the playoffs ("in the hunt").[10]

Figure 5 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 difference 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.

This is also illustrated in Figure 6, 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.

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 6. The result is displayed in Figure 7. 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 8, we can see that the confidence intervals overlap.

I also use the Ramsey RESET test to verify that a linear relationship between Elo rating

[10] There are often cases where a team needs a confluence of several events to occur in addition to a win. For example, the 1989 Pittsburgh Steelers needed four other teams to win their games in order to advance.

difference and probability of winning is appropriate.[11] I find that no non-linear specifications of the Elo difference were statistically significant.

6. Methodology

I use an ordered logit regression model with the game result as the dependent variable and the previously mentioned indicator variable ($Bubble$) as the independent variable of interest. The ordered logit model is preferable to a binary logit model here because game outcomes are not binary due to tie games. I have also run a binary logit regression but the results do not substantively change. Ties are relatively rare in NFL football (there are only 51 tie games in the whole sample) so this is unsurprising. I use the difference in Elo ratings and an indicator for home field advantage as controls. I also present alternative specifications with the number of turnovers[12] and yards gained as controls.

7. Empirical Results

Table 1 reports the logit regression results. The first column shows the regression results for the simplest model with no additional controls besides the difference in Elo ratings (Elo.Diff) and a binary indicator variable which is equal to 1 if the home team is “on the bubble” and equals 0 otherwise (bubble1). Both coefficients are greater than 1 and statistically significant as expected. This suggests that teams with an Elo rating advantage are more likely to win. A team with an Elo rating 100 points higher than their opponent has 1.6 times higher odds of winning than they would if they were evenly matched. Teams “on the bubble” are even more heavily favored. The odds of a win for the bubble team is nearly double that of an equivalent team that is not on the bubble.

As a robustness check, I also replicate the same regression results with a linear probabilty model in Table 2

Although the difference in Elo ratings remains significant in each specification, notice that the bubble coefficient loses statistical significance when turnovers and/or interaction terms are included in the model. I suspect this occurs because turnovers are endogenous to this model. The amount of turnovers committed by your team could be caused by superior defensive play by your opponent or errors by your own players. One

[11] 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.

[12] In NFL football, a “turnover” occurs when the team on offense loses the ball to the defense. This is a significant disadvantage as it allows the other team a chance to score. One would expect the number of turnovers committed by a team to be negatively correlated with said team’s probability of winning the game.

might expect a highly skilled team to force more turnovers from their opponent, commit fewer turnovers themselves, and have a high Elo rating. An omitted variable – the true measure of the team's skill – is the joint cause.

To test for the presence of this omitted variable bias, I run an OLS regression with turnovers (and a specification with net turnovers) as the dependent variable. The results are presented in Table 3. The coefficients for the difference in Elo ratings and the bubble indicator variable are both statistically significant and negative as expected. This provides some suggestive evidence that turnovers should be instrumented or excluded from the model due to endogeneity.

Using a simple indicator for whether or not the home team is on the bubble leaves out an important factor – the opposing team's status. A bubble team might be more likely to win only if their opponent is already out of playoff contention and has no incentive to win. Likewise, when the opposing team has already secured a playoff berth they may have a weaker incentive to win. For instance, teams are known to rest their best players to avoid the risk of injury. Table 8 shows the regression results with indicators for each of these instances.

Notice that although the effect of the difference in Elo ratings has not changed, the effect of being on the bubble has grown tremendously – but only for bubble teams that are playing against those that are already in. In the specification without turnovers, a team on the bubble has 17 times higher odds of victory against a team that is already in the playoffs. This grows to 23 times higher odds when turnovers are included in the model.

8. Concluding Remarks

I find that the difference in Elo ratings is a highly significant predictor of game outcome in the NFL. I also find some evidence that teams perform better in high pressure situations where they have a stronger than normal incentive to win. When comparing teams that are on the bubble to equivalent teams that are not, those on the bubble have nearly double the odds of winning. However, the latter result loses statistical significance as turnovers are added to the regression model.

Table 1: Logit Regression Results

	Dependent variable:					
	Result					
	(1)	(2)	(3)	(4)	(5)	(6)
Elo.Diff	0.006***	0.006***	0.007***	0.007***	0.006***	0.007***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Bubble	0.656**	−0.155	0.479	−0.165	0.765***	0.467
	(0.280)	(0.591)	(0.325)	(0.597)	(0.296)	(0.344)
TurnoversxBubble		−0.356		0.365		
		(0.233)		(0.236)		
OppTurnoversxBubble		0.790***		0.030		
		(0.282)		(0.283)		
Turnovers			−0.726***	−0.728***		−0.726***
			(0.020)	(0.020)		(0.020)
Opp.Turnovers			0.749***	0.748***		0.749***
			(0.021)	(0.021)		(0.021)
Bubble (opponent)					−0.326	0.032
					(0.262)	(0.327)
Constant	0.379***	0.379***	0.415***	0.420***	0.380***	0.415***
	(0.021)	(0.021)	(0.052)	(0.052)	(0.021)	(0.052)
Observations	10,388	10,388	10,388	10,388	10,388	10,388
Log Likelihood	−6,475.624	−6,469.560	−4,796.071	−4,794.896	−6,474.850	−4,796.066
Akaike Inf. Crit.	12,957.250	12,949.120	9,602.142	9,603.792	12,957.700	9,604.133
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 2: Linear Probability Models

	Dependent variable:					
	Result					
	(1)	(2)	(3)	(4)	(5)	(6)
Elo.Diff	0.003***	0.002***	0.003***	0.002***	0.003***	0.002***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Bubble	0.264**	0.138	−0.005	−0.013	0.300***	0.131
	(0.108)	(0.092)	(0.228)	(0.195)	(0.113)	(0.096)
Turnovers		−0.232***		−0.232***		−0.232***
		(0.005)		(0.005)		(0.005)
Opp.Turnovers		0.220***		0.220***		0.220***
		(0.005)		(0.005)		(0.005)
TurnoversxBubble			−0.098	0.134*		
			(0.083)	(0.071)		
OppTurnoversxBubble			0.189**	−0.021		
			(0.076)	(0.065)		
Bubble (Opponent)					−0.125	0.025
					(0.114)	(0.097)
Constant	2.162***	2.164***	2.162***	2.165***	2.163***	2.164***
	(0.009)	(0.017)	(0.009)	(0.017)	(0.009)	(0.017)
Observations	10,388	10,388	10,388	10,388	10,388	10,388
R^2	0.108	0.356	0.109	0.356	0.108	0.356
Adjusted R^2	0.108	0.355	0.108	0.355	0.108	0.355
Residual Std. Error	0.931	0.791	0.931	0.791	0.931	0.792
F Statistic	629.937***	1,432.528***	316.977***	955.785***	420.374***	1,145.934***
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 3: Turnovers

	Dependent variable:		
	Own Turnovers	Opp. Turnovers	Net Turnovers
	(1)	(2)	(3)
Elo.Diff	-0.001^{***}	0.001^{***}	-0.002^{***}
	(0.0001)	(0.0001)	(0.0002)
bubble1	-0.441^{**}	0.107	-0.548^{**}
	(0.176)	(0.176)	(0.252)
Constant	1.938^{***}	2.035^{***}	-0.097^{***}
	(0.015)	(0.015)	(0.021)
Observations	10,388	10,388	10,388
R^2	0.005	0.010	0.015
Adjusted R^2	0.005	0.010	0.014
Residual Std. Error (df = 10385)	1.519	1.518	2.173
F Statistic (df = 2; 10385)	27.802^{***}	54.205^{***}	76.413^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4: Comparing Performance of Elo Difference between Last Week and Full Sample

	Dependent variable:					
	Result					
	(1)	(2)	(3)	(4)	(5)	(6)
Elo.Diff	0.006***	0.006***	0.005***	0.006***	0.006***	0.007***
	(0.001)	(0.001)	(0.001)	(0.0002)	(0.0002)	(0.0002)
bubble1		0.598**	0.453		0.656**	0.479
		(0.294)	(0.327)		(0.280)	(0.325)
Turnovers			−0.641***			−0.726***
			(0.076)			(0.020)
Opp.Turnovers			0.604***			0.749***
			(0.075)			(0.021)
Constant	0.502***	0.438***	0.527***	0.383***	0.379***	0.415***
	(0.086)	(0.091)	(0.203)	(0.021)	(0.021)	(0.052)
Observations	675	675	675	10,388	10,388	10,388
Log Likelihood	-404.521	-402.309	-323.313	-6,478.611	-6,475.624	-4,796.071
Akaike Inf. Crit.	813.042	810.619	656.626	12,961.220	12,957.250	9,602.142
Note:						*p<0.1; **p<0.05; ***p<0.01

References

Apesteguia, J. and I. Palacios-Huerta (2011): “Corrigendum: Psychological Pressure in Competitive Environments: Evidence from a Randomized Natural Experiment,” *American Economic Review*, 101, 1636.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar (2009): “Large Stakes and Big Mistakes,” *The Review of Economic Studies*, 76, 451–469.

Cao, Z., J. Price, and D. F. Stone (2011): “Performance Under Pressure in the NBA,” *Journal of Sports Economics*, 12, 231–252.

Dohmen, T. J. (2008): “Do professionals choke under pressure?” *Journal of Economic Behavior and Organization*, 65, 636–653.

Ehrenberg, R. G. and M. L. Bognanno (1990): “Do Tournaments Have Incentive Effects?” *Journal of Political Economy*, 98, 1307.

Elo, A. E. (1978): *The rating of chessplayers, past and present*, Arco Pub.

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

Hickman, D. C. and N. E. Metz (2015): “The impact of pressure on performance: Evidence from the PGA TOUR,” *Journal of Economic Behavior & Organization*, 116, 319–330.

Hill, D. M., S. Hanton, N. Matthews, and S. Fleming (2010): “Choking in sport: a review,” *International Review of Sport and Exercise Psychology*, 3, 24–39.

Hvattum, L. M. and H. Arntzen (2010): “Using ELO ratings for match result prediction in association football,” *International Journal of Forecasting*, 26, 460–470.

Lazear, E. P. (2000): “American Economic Association,” *The American Economic Review*, 90, 1346–1361.

McKillop, A. (2013): “Teams That Controlled Their Own Destiny in the League’s Final Week,” *FootballGeography.com.*

Prendergast, C. (1999): “The Provision of Incentives in Firms,” *Journal of Economic Literature*, 37, 7–63.

Silver, N. (2014): “Introducing NFL Elo Ratings,” *FiveThirtyEight*.

Smith, M. D. (2013): “NFL Playoff Scenarios for Week 17,” *Pro Football Talk*.

——— (2014): “Week 17 Playoff Scenarios,” *Pro Football Talk*.

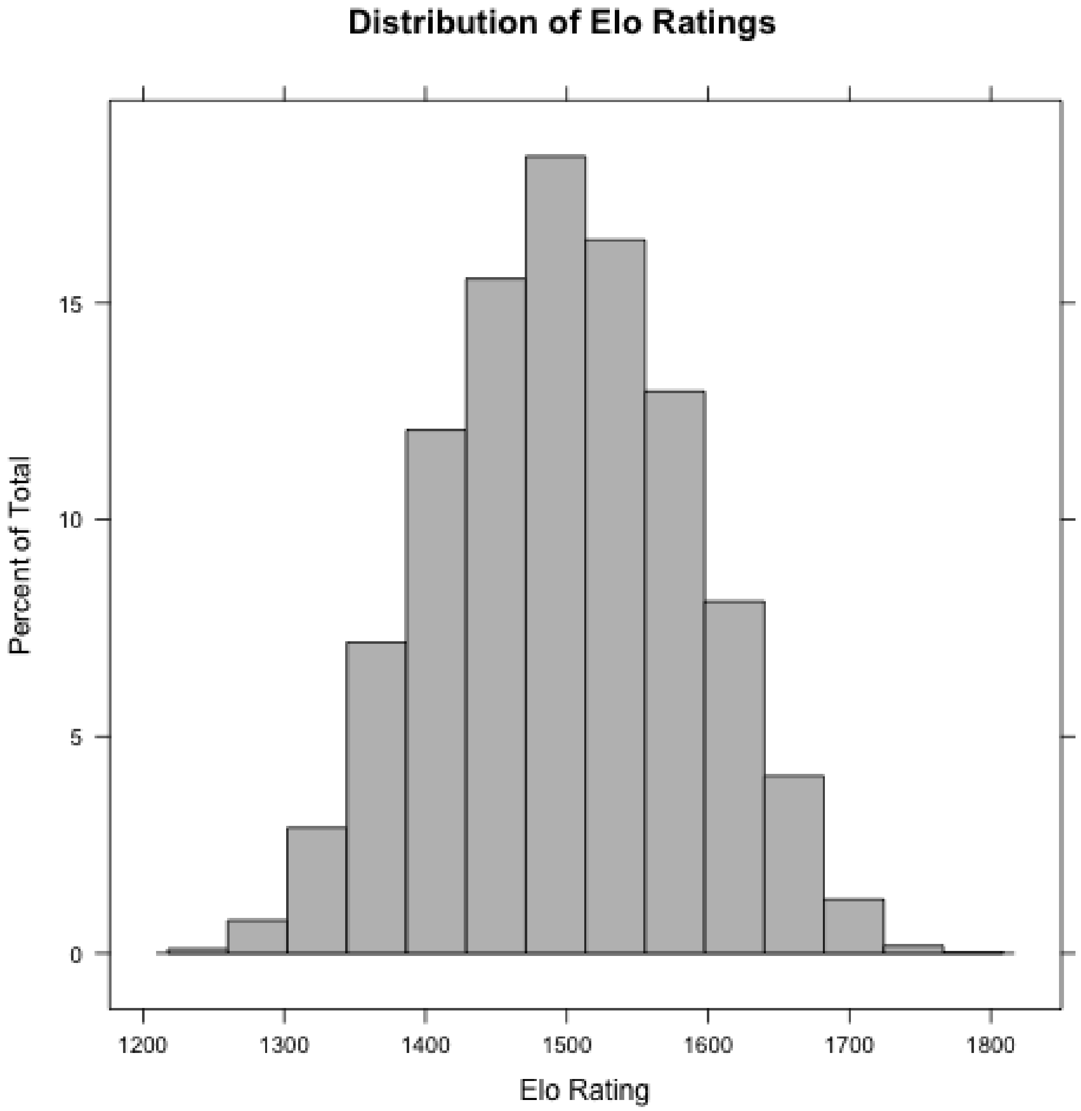

Figure 1: Distribution of Elo Ratings in the Sample

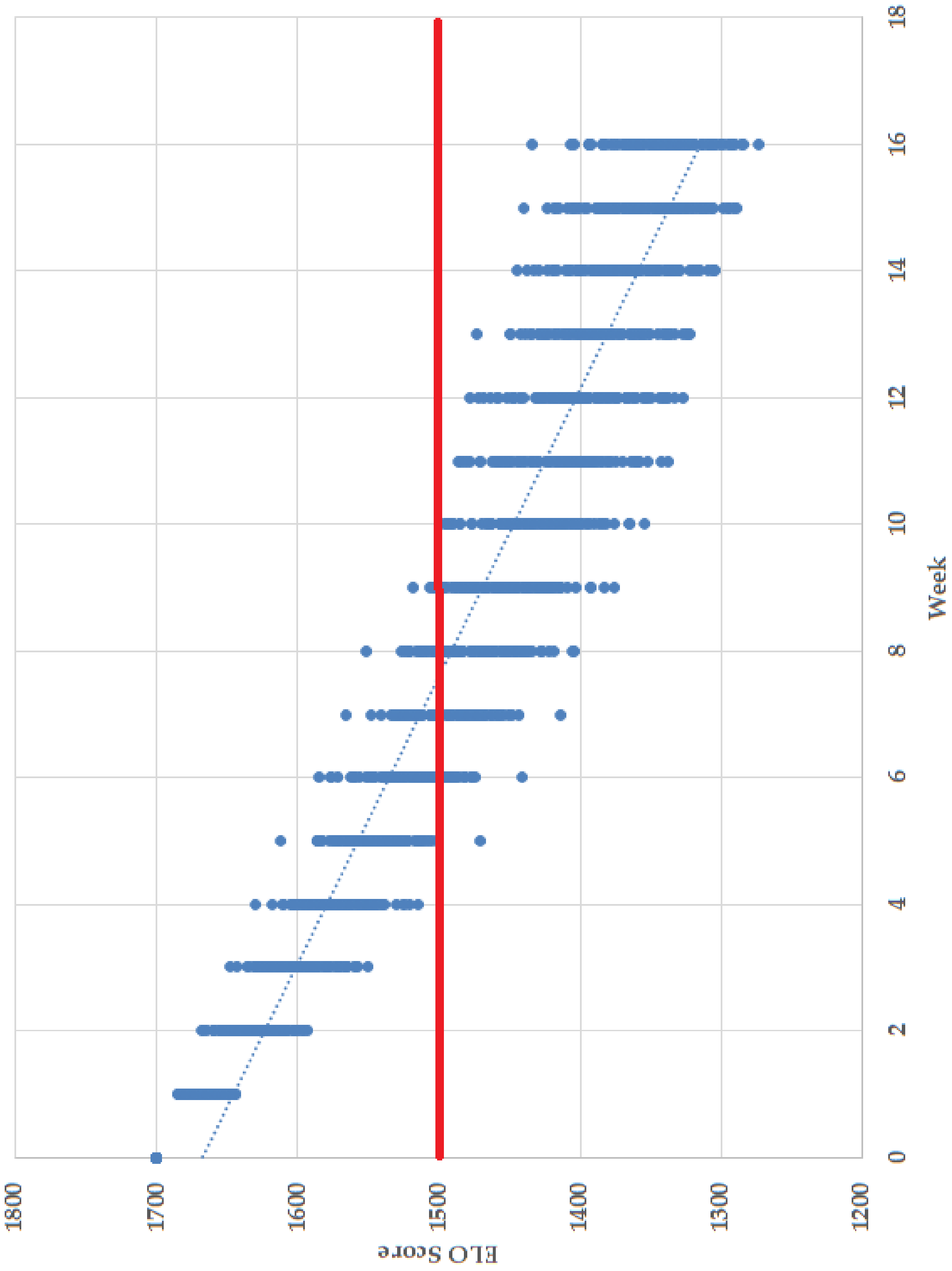

Figure 2: Simulation of a highly rated team continuously losing to random opponents

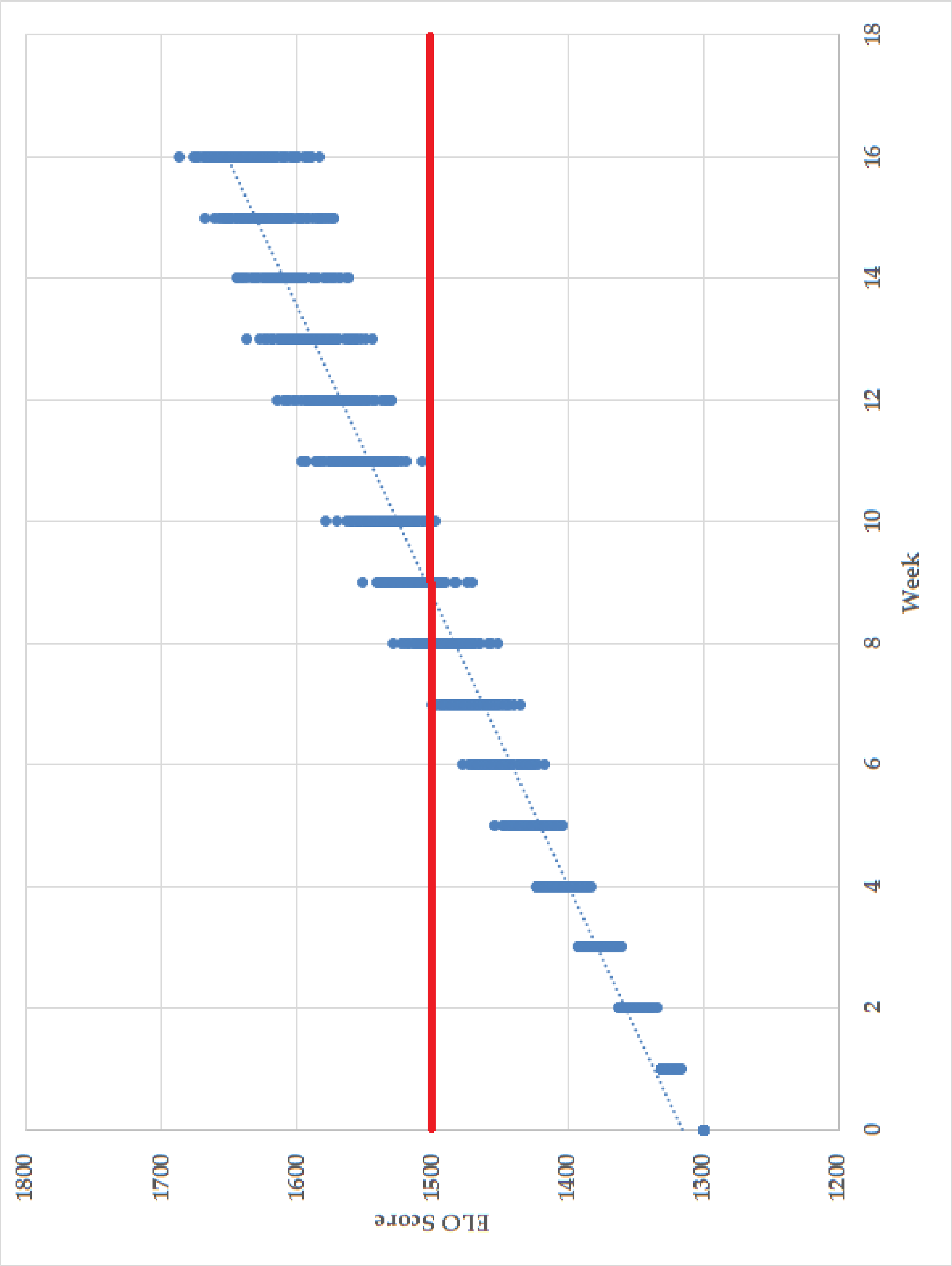

Figure 3: Simulation of a low rated team continuously winning against random opponents

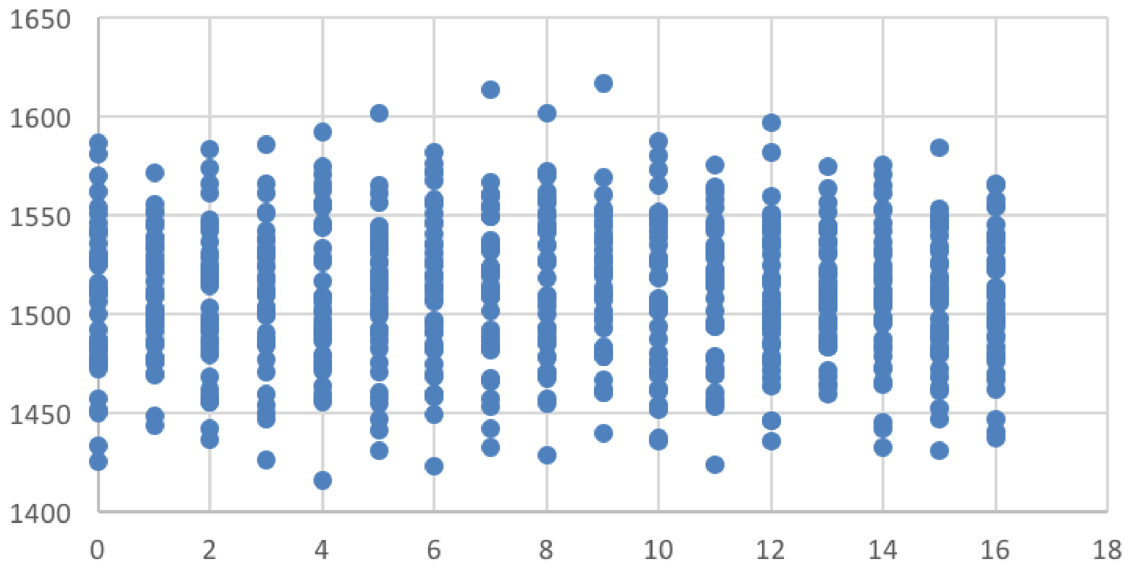

Figure 4: Simulation of an average team randomly winning against random opponents

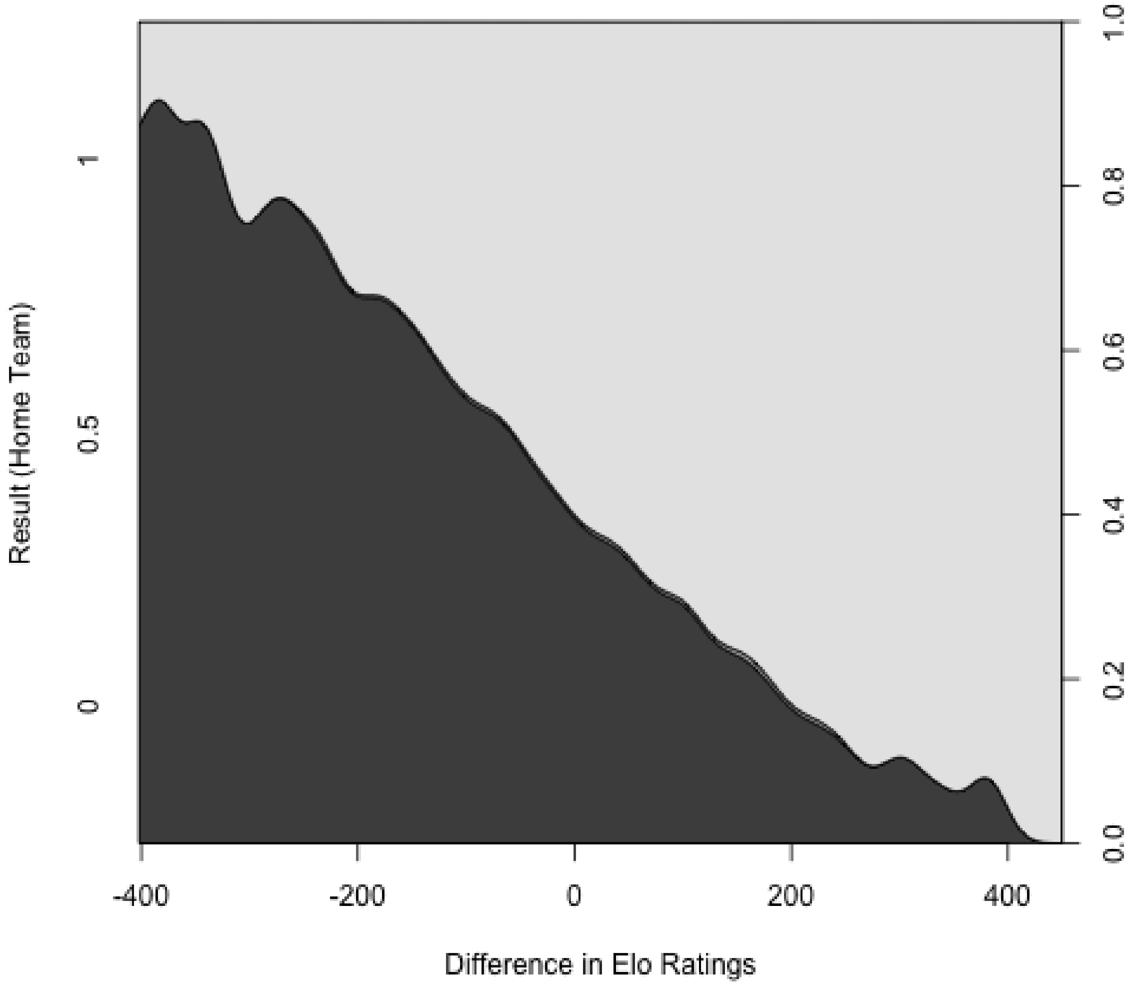

Figure 5: Conditional Density Plot (Result vs. Difference in Elo Rating)

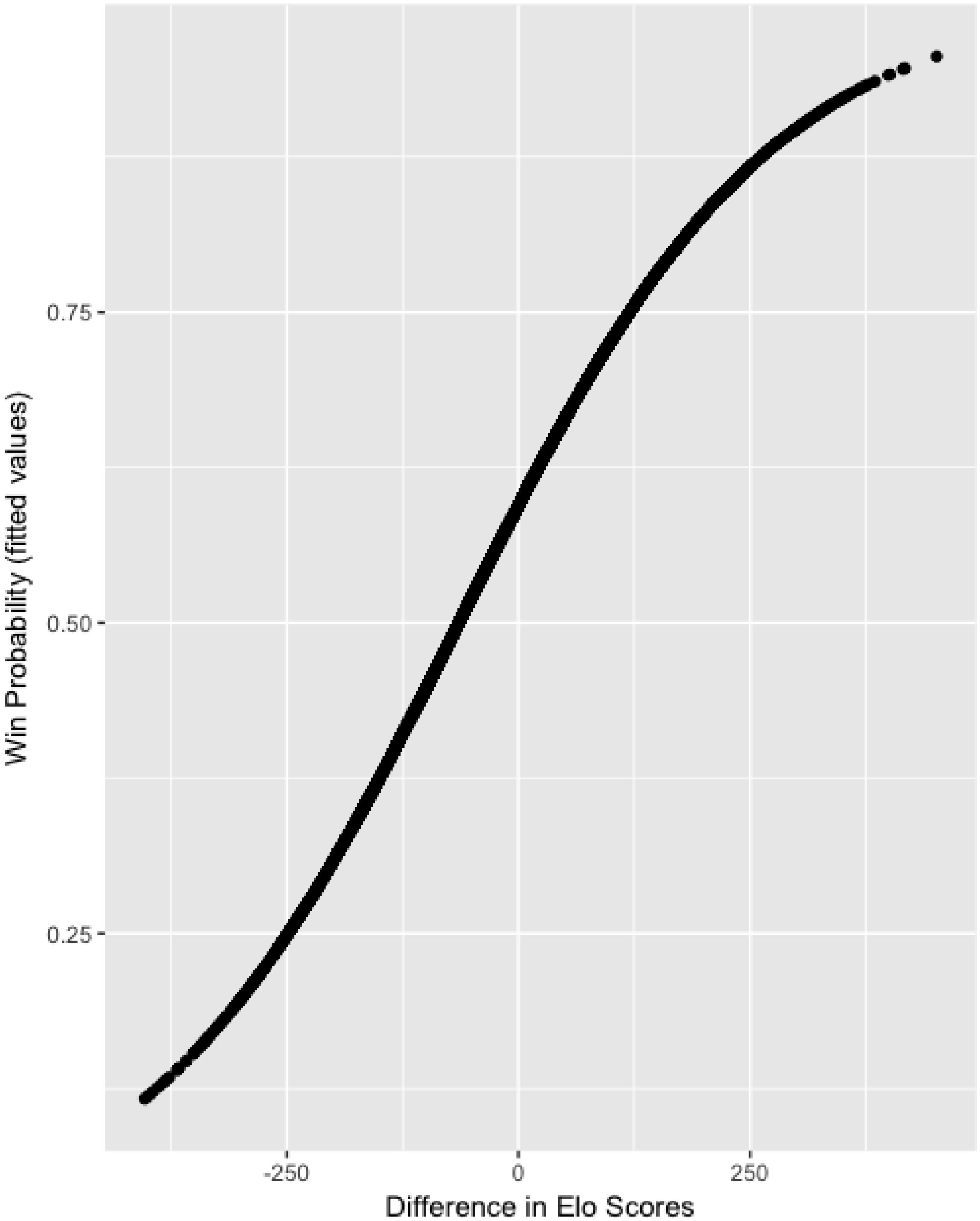

Figure 6: Win Probability versus Difference in Elo Rating (fitted values)

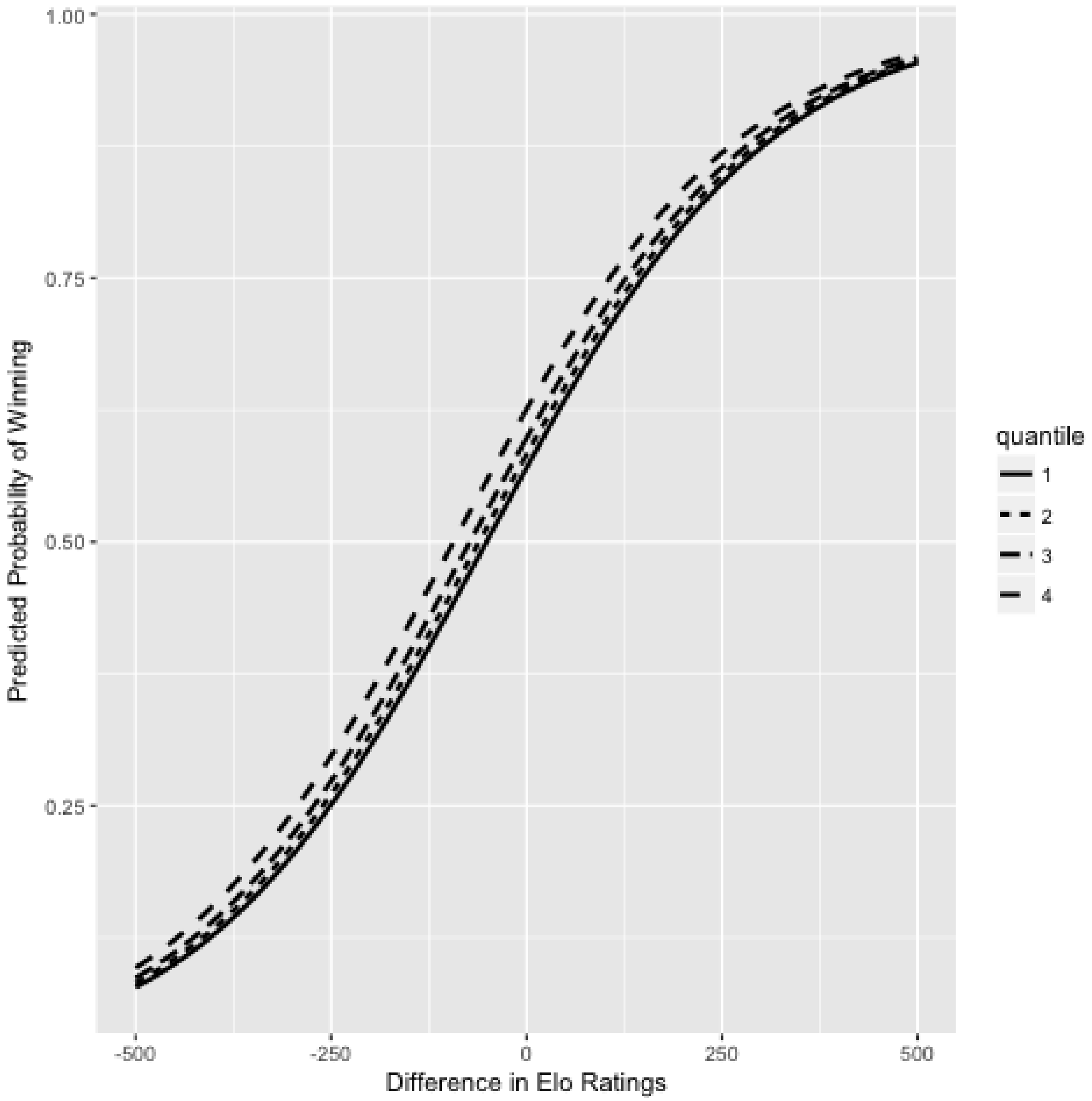

Figure 7: Win Probability versus Difference in Elo Rating, Grouped by Quantile

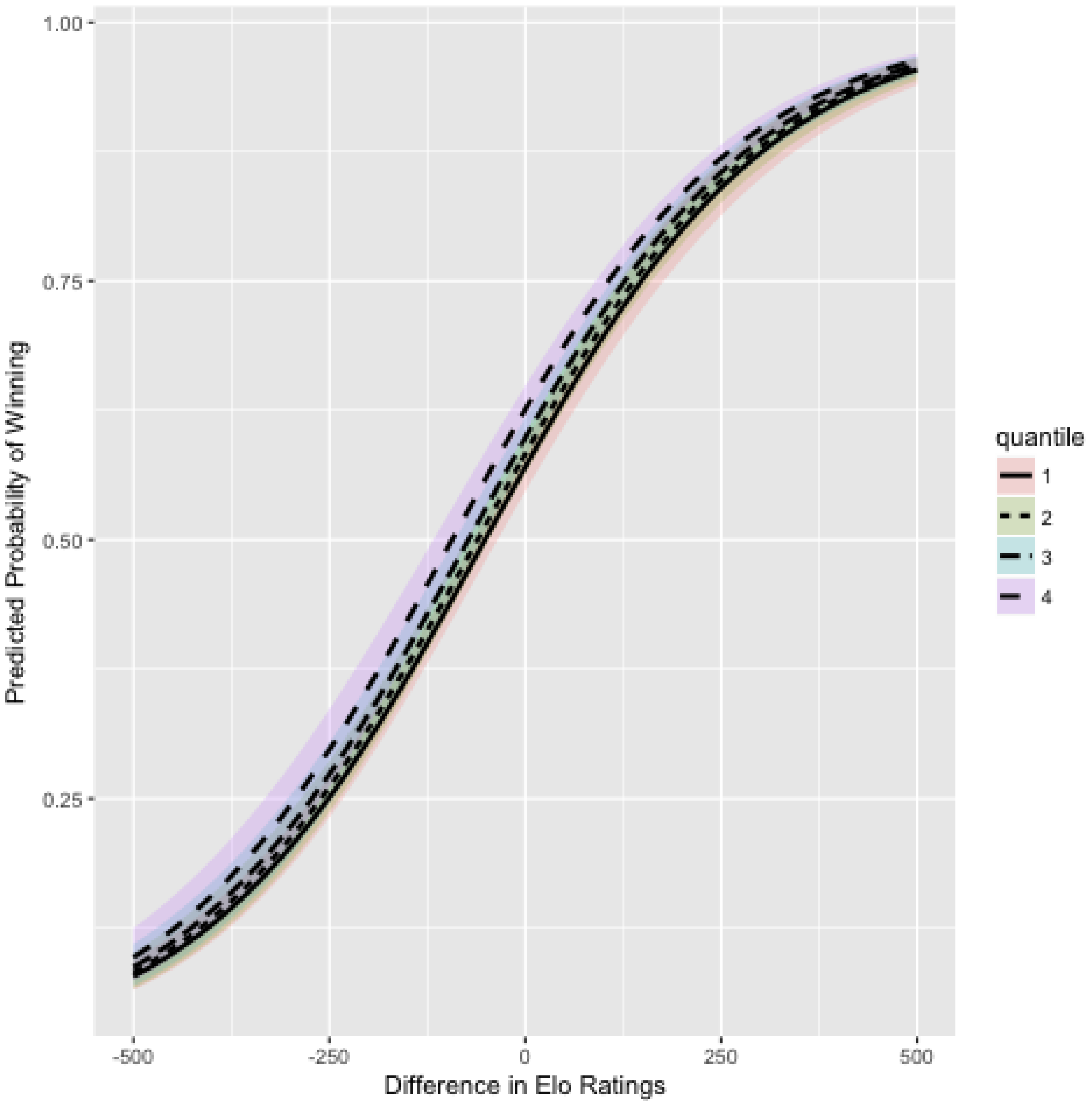

Figure 8: Win Probability versus Difference in Elo Rating, Grouped by Quantile (with error bars)

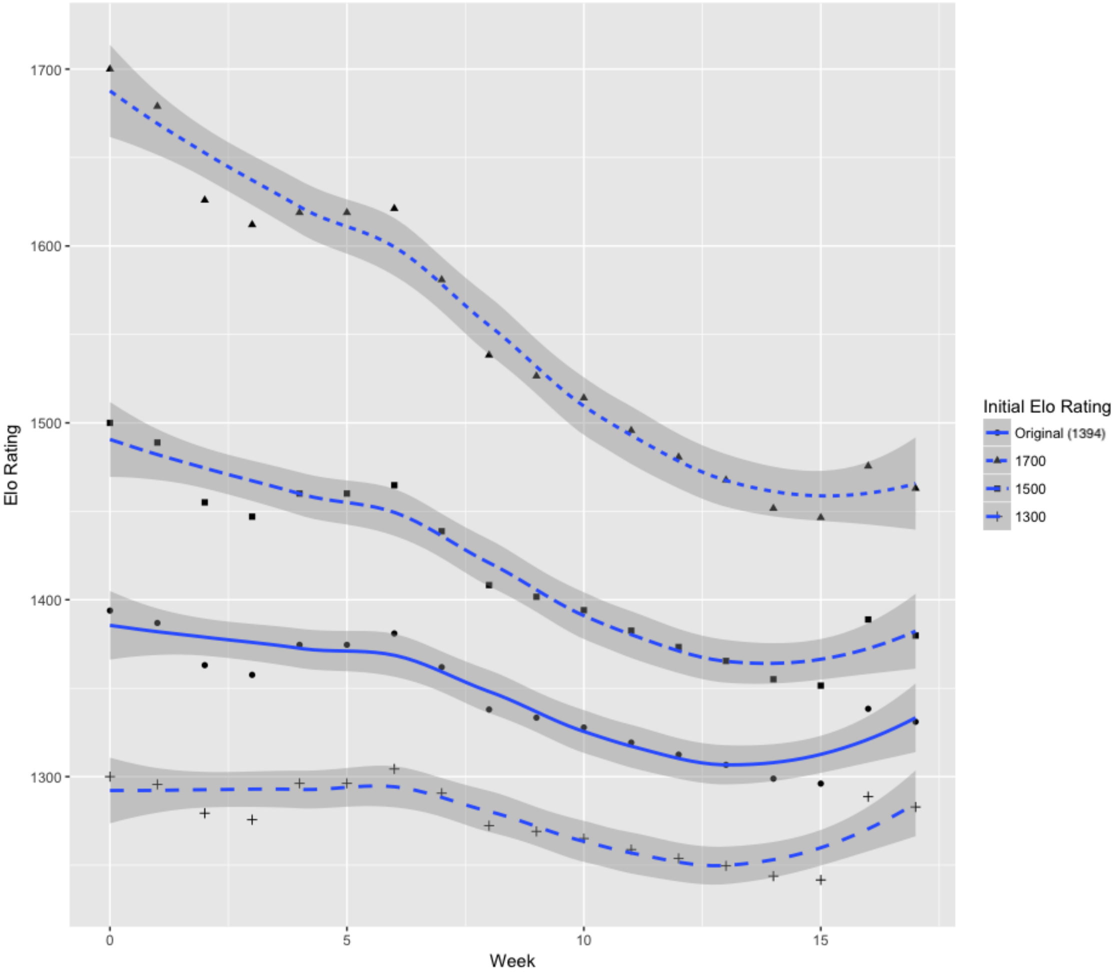

Figure 9: Ratings Development for the Jacksonville Jaguars During 2016

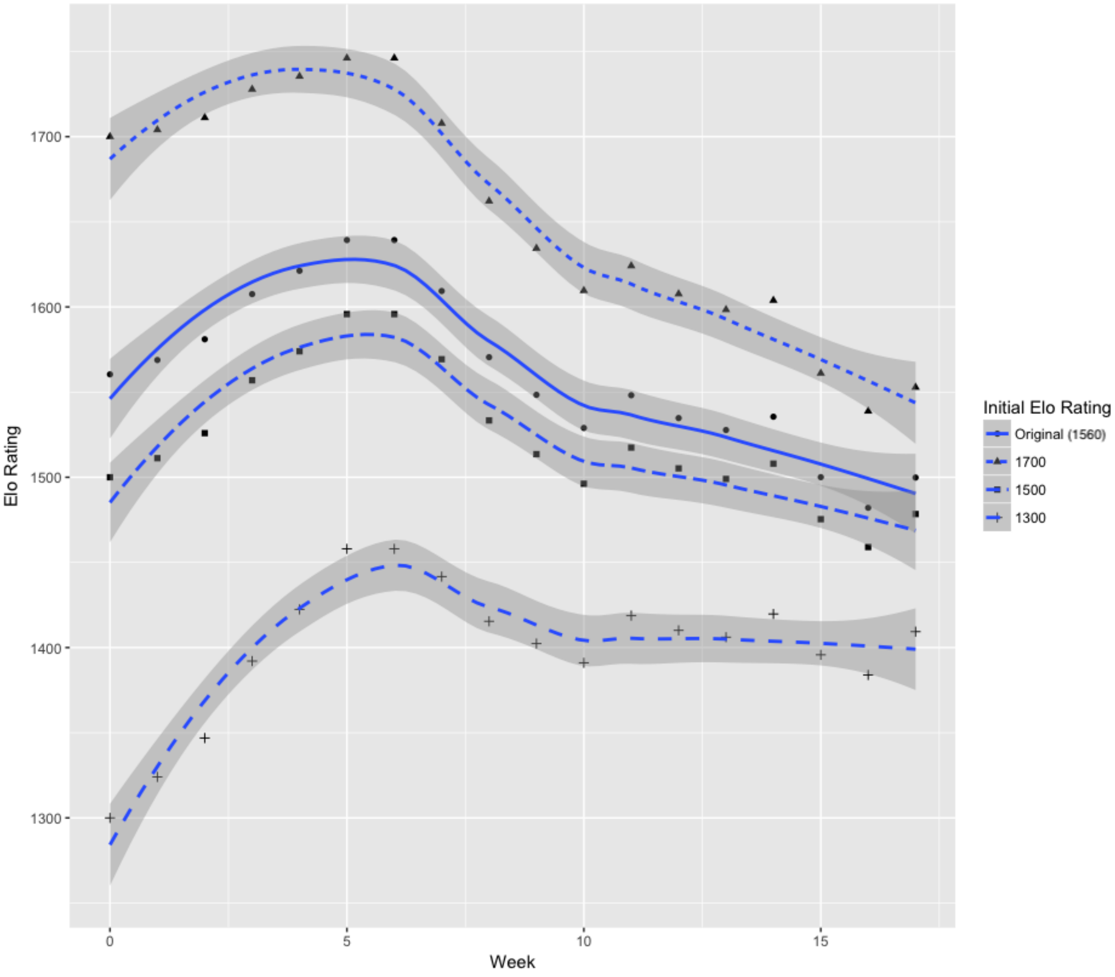

Figure 10: Ratings Development for the Minnesota Vikings During 2016

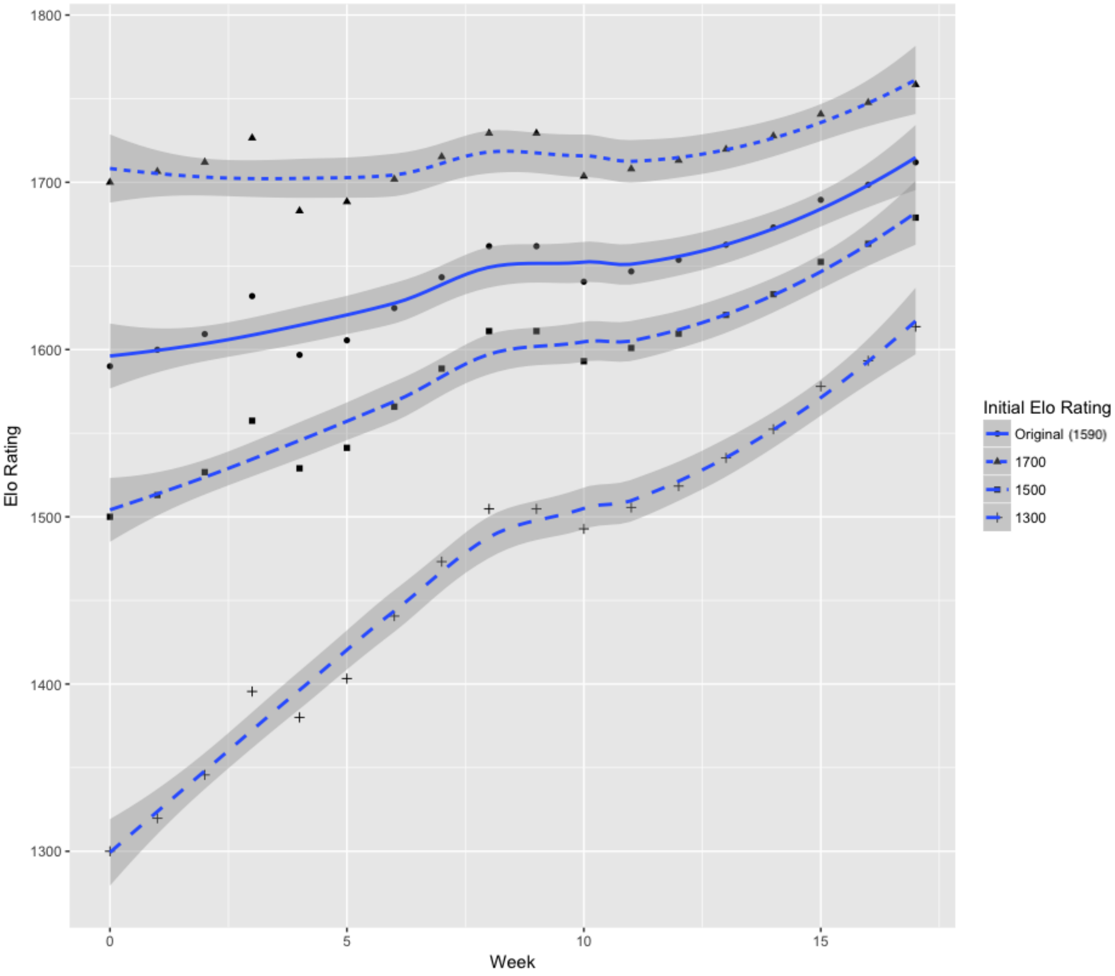

Figure 11: Ratings Development for the New England Patriots During 2016

Table 5 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

Table 6: 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

Table 7: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Points	10,388	22.139	10.436	0	62
Opp. Points	10,388	19.413	10.107	0	62
Yards Gained	10,388	328.449	84.785	−7	653
Own Turnovers	10,388	1.938	1.523	0	11
Opp. Yards Gained	10,388	313.529	85.888	26	676
Opp. Turnovers	10,388	2.031	1.526	0	10
Point Difference	10,388	11.350	9.831	−38	59
Net Yards Gained	10,388	52.210	109.863	−382	488
Elo	10,388	1,497.843	87.797	1,239.655	1,786.755
Opp. Elo	10,388	1,501.708	87.646	1,232.553	1,789.711
Elo Difference	10,388	−3.865	124.690	−403.834	451.609
Net Turnovers	10,388	−0.093	2.189	−10	9

Table 8: Logit Regression Results w/ All Matchup Categories

	Dependent variable:		
	Result		
	(1)	(2)	(3)
Elo.Diff	0.006***	0.007***	0.006***
	(0.0002)	(0.0002)	(0.0002)
Bubble vs. Bubble	−0.417	−0.711	−0.419
	(0.432)	(0.496)	(0.432)
Bubble vs. In	2.841***	3.160***	2.837***
	(1.040)	(1.120)	(1.040)
Bubble vs. Out	0.799	0.552	0.798
	(0.548)	(0.653)	(0.548)
Bubble vs. Hunt	0.832	0.477	
	(0.805)	(0.902)	
Out vs. Bubble	0.043	0.653	0.039
	(0.368)	(0.447)	(0.368)
Out vs. In	0.517	0.579	0.511
	(0.439)	(0.557)	(0.439)
Hunt vs. Hunt	0.078	−0.037	
	(0.111)	(0.134)	
Out vs. Hunt	0.104	0.537	
	(0.594)	(0.733)	
In vs. Bubble	−0.513	−0.866	−0.514
	(0.597)	(0.766)	(0.597)
In vs. In	−0.130	0.107	−0.133
	(0.666)	(0.769)	(0.666)
In vs. Out	0.190	0.070	0.191
	(0.454)	(0.534)	(0.454)
In vs. Hunt	−1.717*	−0.965	
	(0.878)	(0.937)	
Hunt vs. Bubble	0.933	2.343***	
	(0.819)	(0.895)	
Hunt vs. In	−0.401	0.683	
	(0.932)	(1.049)	
Hunt vs. Out	−0.694	−0.748	
	(0.607)	(0.654)	
Turnovers		−0.728***	
		(0.020)	
Opp.Turnovers		0.753***	
		(0.021)	
Constant	0.377***	0.412***	0.379***
	(0.022)	(0.052)	(0.021)
Observations	10,388	10,388	10,388
Log Likelihood	−6,463.734	−4,780.809	−6,468.121
Akaike Inf. Crit.	12,961.470	9,599.618	12,956.240
Note:			*p<0.1; **p<0.05; ***p<0.01