Modeling NFL Football Outcomes

Rhonda C Magel, Joseph Roith

1. Introduction

The purpose of this research is to develop models for NFL football games which help identify the important factors in determining the outcome of the game. A variety of different methods will be used in developing these models: ordinary least squares regression, logistic regression, and proportional odds. Models will be used to explain an outcome and they will also be used to help predict an outcome based on differences of certain in-game statistics.

Values from thirty-five in-game statistics for both teams playing in each game were collected for 752 NFL games over the three seasons between the years 2011 and 2014. Data from the first two seasons in this time period were used to develop the models and data from the last season was used to test the models. Data from these games was found on the websites, *NFL Scores and Pro Football* [1] and [2].

Much of the recent literature on the NFL is focused on forecasting professional football games with respect to the efficiency of the betting market [3 ,4 ,5, 6]. Boulier and Stekler [7] used rankings of teams developed by a power score printed weekly in the New York Times and looked at the probability of higher ranked teams winning each game. Their method predicted approximately 61% of the games correctly over seven NFL seasons which was worse than the 66% accuracy of the betting market over the same period of time. It was unknown how the rankings were developed [7].

Harville [8] looked at games from the 1970's to create a predictive model and found more success than most current research. Over the course of seven years, he was able to achieve a 70% accuracy rate in predicting games. However, compared to the betting market during that time, which was able to correctly predict 72% of games, his model still underperformed. It is possible that in earlier eras of the game, talent from team to team was less equally distributed, and therefore, the outcomes of the games could have been easier to model with a few very good teams playing against many that were mediocre.

Stefani [9], also looked at professional football during the 1970's and developed a point spread model using least squares regression. With the 1970-71 season as his training data set, he used his model to predict all games over the next nine seasons. The main covariates Stefani used were an adjusted team rank and an estimated home field advantage. The final accuracy for predicting games correctly was 64.2%, and the model produced a point spread with an average absolute deviation of 10.98 points from the observed game point spread.

Of course, these ideas and methods are not constrained to football at the professional level, or even solely to the sport of football. Long and Magel [10] considered football games at the collegiate level, analyzing games from the NCAA Division I Football Championship Subdivision. They used regression techniques to identify significant in-game statistics and to develop prediction models for the outcome of games. They concluded that six factors contribute to wins for collegiate teams: difference in turnovers; difference in the probability of pass completion; difference in the probability of a 3rd down conversion, difference in the number of sacks, difference in the number of punt returns; and difference in the number of offensive yards per play. Combining these with a computer ranking of the individual teams playing, they were able to predict the correct winner of games around 73% of the time.

In other sports, Roith and Magel [11] use discriminant analysis to determine the difference of offensive and defensive statistics in the National Hockey League (NHL). Claims exist throughout many sports that defense is more important than offense, and hockey is one of those sports. Again, the challenge is to quantify this difference somehow. Looking at teams that made the playoffs compared to those that did not, discriminant analysis was performed on goals allowed and goals scored. This is the most basic sense of offense and defense, and it was found that the magnitude effect on making the playoffs of allowing one less goal was 41% larger than scoring one more goal during the season.

Roith and Magel [11] also developed logistic regression models and point spread models that consistently showed areas of the game which recorded defensive aspects of hockey, had a greater impact on the outcomes of games. Their models were able to forecast the results of future games with 65% accuracy, significantly better when compared to a handicapping website with 55% accuracy.

Unruh and Magel [12] considered NCAA Division I basketball games and the important influencers that determine the winners of individual games, along with predicting the winners of the championship tournament at the end of each season. Ultimately, their models performed with around 67% accuracy.

There is not very much literature available today that deals with the underlying reasons for teams to have success in the NFL. Most of it deals with finding a way to outdo the current betting market. The problem with this approach is that it is hard to take anything away from these models to improve the performance of an individual team or organization. Our goal in this paper is to help provide a way to optimize decision making and formation of strategy for teams to improve the product they put on the field. We would like to propose a more efficient way to evaluate teams by knowing what factors, including which areas of the game, can affect the outcome. We would like to even expand upon that by providing the relative degree to which those factors affect the game as well.

2. Methods -Development of Models .

Point Spread Model

The first model was developed using Ordinary Least Squares Regression (Abraham & Ledolter [13]) to help explain the point spread of an NFL football game. The point spread being points scored by the home team minus points scored by the away team. The following variable differences, or margins, between the home team and the away team were considered for entry into the model: WinPerMargin; FirstDownMargin; FirstDownOnPassMargin; FirstDownOnRunMargin; FirstDownOnPenaltyMargin; TotalPlayMargin; TotalYdMargin; YardPerPlayMargin; PassingMargin; YardPerPassMargin; YPRushMargin; PenYardsMargin; TurnoverMargi;, 3DownPercentMargin; SackYardsMargin; AveKickReturnMargin; AvePuntReturnMargin.. Stepwise selection technique [14] was used with an entry level of $\alpha = 0.10$. and a stay significance level of $\alpha = 0.15$ to help derive the initial model for consideration using OLS.

Once the initial model was determined using the Stepwise selection technique, we considered the amount that adjusted R-square was increased when each variable was the last to be added to the model. We also considered the amount that the predicted R-square changed when each variable was the last to be added to the model. We tested for multicollinearity in the independent variables by examining the variance inflation factors (VIF) associated with each of the independent variables [15]. If multicollinearity exists, then the estimates of the parameters cannot be interpreted. We wanted to be able to interpret the parameter estimates to determine

which factors had the most effect on the outcome of the game. All the VIFs should be less than 10 [15].

Logistic model for estimating probability of home team winning

We next developed a model using the logistic regression modeling technique [13] which estimates the probability of the home team winning the game given the differences, or margins, of in-game statistics found to be significant. The same variables that were considered in OLS regression were considered for entry into the Logistic Regression model. The stepwise regression technique was used again with the same entry and exist levels as were used in the OLS regression model [16].

Once the initial model was developed using the stepwise selection technique, we did consider future model adjustments by taking out a variable at a time and considering two criteria: the max rescaled R-square value and the Receiver Operator Characteristic (ROC) curve. The max rescaled R-square value represents the change in the likelihood function between the current model, and the baseline "intercept only" model containing no independent variables. The ROC curve is a graphical plot that represents the rate of true positives classified and false positives classified [17]. We also tested the logistic model for goodness of fit using the method proposed by Hosmer and Lemeshow [18].

Proportional Odds Model in Estimating a Win by a large or small margin

The last type of model that was fit was the proportional odds model [19]. This model is similar to the logistic regression model except in this case, the score margin response variable was separated into four different ordered categories: strong victory for away team (away team won by 10 or more points); weak victory for away team (the away team won by fewer than 10 points); weak victory for home team (home team won by fewer than 10 points); and strong victory for home team (home team won by 10 or more points). The value 10 was used as a cutoff point to represent one team winning by more than a touchdown with the extra point and a field goal or winning by less than that. The same selection procedure used in logistic regression was also used in this case.

3. Results

Results- OLS Point Spread Model

The OLS regression model was used to predict the score margin at the end of NFL games if the in-game statistics were known. The stepwise selection procedure selected 11 differences of in-game statistics which were all significant at a level of $\alpha = 0.05$ and an R-square of 0.8475.

Upon further examination, it was found that if we removed the last five variables that were entered into the model (AvePRM, YPRushM, PenYardsM, WinPerM, and AveKRM), we only would lose a total of 1.58% in our overall R-square value, but simplify the model by having fewer variables.

The final model, given in Table 1, shows a summary of the coefficients. The intercept was included in this model since it was also found to be significant. In this case, the intercept can be interpreted as an underlying point advantage for the home team during an NFL game. Although it is not very large, we would expect that for any given game, the home team will have the benefit of about one extra point in the final score margin.

The selected model has an adjusted R-square value of 0.8279, and a predicted R-square value of 0.8245. This indicates that the model should do a good job of explaining the point spread if the in-game statistics are known. Each of the variance inflation factors associated with the

independent variables in the model is less than 4.36, which indicates there are no problems with multicollinearity or in interpreting the estimated coefficients.

Table 1. OLS Regression Model Summary

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T VALUE	PR > |T|
INTERCEPT	1.00306	0.29165	3.44	0.0006
FIRSTDOWNM	1.37997	0.07967	17.32	<.0001
TOTALPLAYM	-0.53459	0.04064	-13.15	<.0001
YPPASSM	1.00567	0.13400	7.50	<.0001
TURNOVERM	-3.88568	0.15333	-25.34	<.0001
3DPERM	0.17715	0.01802	9.83	<.0001
SACKYARDSM	-0.12464	0.01840	-6.77	<.0001

In-game statistics from the games in the testing data set were used to validate the model. There were 256 games in the testing data set. The model was able to correctly identify the winner for 220 of these games, for an accuracy of 85.9%. The mean absolute deviation of the predicted score differences from the observed score differences for the games in the testing data set was 4.83 points. On average, the model produced a score margin that was within five points of the actual margin, or less than one touchdown. The final point spread equation can be defined as:

Point Spread = 1.00306 + 1.37997*(First Down Margin) – 0.53459*(Total Play Margin) + 1.00567*(Yards per Pass Margin) – 3.88568*(Turnover Margin) + 0.17715*(3rd Down Conversion Percent Margin) – 0.12464*(Yards Lost to Sacks Margin) (Eq 1)

It is noted that the coefficient associated with FirstDownMargin is approximately 1.4. This indicates that with every extra first down the home team gains over their opponent, we would expect them to score about 1.4 more points, with everything else held constant. The coefficient with the largest magnitude is turnover margin. For every extra turnover the home team commits over their opponent, this will cost them an estimated 3.9 points in the final score margin.

Results- Logistic Regression Model

The logistic regression model was coded in terms of the home team winning, "1", or losing, "0", the game. The initial logistic model obtained through the stepwise procedure contained 9 variables, but the last 2 variables that entered into the model had little effect on the results and were taken out. The final model is given in Table 2 along with the odds ratios and associated confidence intervals. The intercept is not significant, but it shows the advantage of the home team. If all margins are zero, the home team has a 55% chance, calculated by (exp (.2128)/(1+exp(.2128)), of winning the game. This is close to the winning percentage that we observed in our data set for the home team of 56%.

Table 2 Parameter Estimates and Odds Ratios for Game Model.

With the intercept included along with the seven most significant variables, we can check the goodness of fit for the model. Using the Hosmer-Lemeshow test, we found a p-value of 0.8758. This indicates that the logistic model can be used to estimate the probability of a home team winning the game. The max rescaled R-squared value for this model was 0.7867, and the area under the ROC curve is 0.9621, indicating that this model is doing well at estimating the probability of the home team winning the football game when the in-game statistics are known for the initial data set.

Table 2. Parameter Estimates and Odds Ratios for Game Model

Parameter	Estimate	Standard Error	Pr > ChiSq	Odds Ratio	95% Confidence Limits	
Intercept	0.2128	0.1688	0.2074			
FirstDownM	0.2865	0.0562	<.0001	1.332	1.193	1.487
TotalPlayM	-0.1074	0.0267	<.0001	0.898	0.852	0.947
YPPassM	0.4646	0.0878	<.0001	1.591	1.340	1.890
PenYardsM	-0.0191	0.0057	0.0008	0.981	0.970	0.992
TurnoverM	-1.3074	0.1524	<.0001	0.271	0.201	0.365
3DPerM	0.0704	0.0117	<.0001	1.073	1.049	1.098
SackYardsM	-0.0459	0.0112	<.0001	0.955	0.934	0.976

The interpretation of the model parameters is best done through the odds ratios. The odds ratio for the first down margin is 1.332 indicating that the odds of the home team winning the game get multiplied by 1.332 for each extra first down the home team has over their opponent. The odds ratio for turnover margin is .271 which indicates that the odds of the home team winning the game get multiplied by 0.271 (decrease) for each extra turnover the home team commits over their opponent.

This model was used with the in-game statistics for each of the games in the testing data set. The model correctly predicted 220 of the 256 games, or 85.9% of the games correctly, thus validating this model.

Results – Proportional Odds Model

The final model that we developed was an extension of the logistic regression model. Using the individual game data set and the marginal variables once again, we took advantage of the fact that score margin could be separated into different ordered groups. Four categories were created that indicated the winner of the game, along with a classification of a close scoring game or a game in which the score margin differed by at least 10 points. The four categories will be referred to as; "Strong Away Win", "Weak Away Win", "Weak Home Win", and "Strong Home Win". It is important to keep in mind that these categories are ordered, since the interpretation of the proportional odds model employs the odds of moving from one lower category to the next higher one.

After developing the model, we estimated the probability that a given observation will fall into each of the four categories. The probabilities are as follows:

$$p_{sa} = \frac{e^{\theta_{sa}}}{1+e^{\theta_{sa}}} \quad \text{(Eq. 2)}$$

$$p_{wa} = \frac{e^{\theta_{wa}}}{1+e^{\theta_{wa}}} - p_{sa} \quad \text{(Eq. 3)}$$

$$p_{wh} = \frac{e^{\theta_{wh}}}{1+e^{\theta_{wh}}} - (p_{wa} + p_{sa}) \quad \text{(Eq. 1)}$$

$$p_{sh} = 1 - (p_{sa} + p_{wa} + p_{wh}) \quad \text{(Eq. 5)}$$

where

$$\theta_{sa} = log\ \frac{P_{sa}}{1- P_{sa}} = -3.62 + \boldsymbol{X}\beta \quad \text{(Eq. 6)}$$

$$\theta_{wa} = log \frac{P_{wa}}{1- P_{wa}} = -0.2148 + \boldsymbol{X}\beta \quad \text{(Eq. 7)}$$

$$\theta_{wh} = log \frac{P_{wh}}{1- P_{wh}} = 3.4949 + \boldsymbol{X}\beta \quad \text{(Eq.8)}$$

and, $\boldsymbol{X}\beta$ = 0.278*(First Down Margin) - 0.038*(Yards lost to Sack Margin) - 1.229*(Turnover Margin) + 0.066*(3rd Down Conversion Percent Margin) + 0.015*(Rush Yards Margin) + 0.011*(Pass Yards Margin) - 0.013*(Penalty Yards Margin) - 0.171*(Total Play Margin).

The model was used to classify the categories in the training data set by calculating the estimated probability that an observation would fall into that category and then placing the observation in the category having the largest associated probability. This had a 70.6% accuracy rate. This means not only did the model have the correct outcome of the game, but it also gave the correct group for the score margin. If we just consider whether or not the model got the winner correct, not considering the margin of victory, it was 88.9% accurate. Therefore, our proportional odds model is just as accurate as the logistic model in fitting the game winner, and it can tell us something about the probability of the final score margin.

Using our results from the proportional odds model on the testing data set, we found that 68.8% of the games were correctly categorized, and 86.3% were classified correctly as a home win or home loss.

Comparing the effect of each variable in the proportional odds model, turnover margin is once again the most influential, along with first down margin. For every increase in the turnover margin, the odds of the home team moving from one category to the next, say from weak away win to weak home win, is reduced by a factor of 0.293. As first down margin increases one unit, the odds of winning by a larger margin (or losing by a smaller margin) increases by a factor of 1.320.

One advantage of the proportional odds model is that hypothetical scenarios could be considered for a game. For example, if we wanted to know the winner and final score of a game where every significant marginal variable had a value of zero, we can get an estimate from the point spread model and also a probability of the home team winning from the logistic model, but that does not tell us any more information. We can find the expected probabilities when both teams have the exact same performances in the significant variables of the home team falling into each of the categories. The percentage of times that the home team will win by more than ten points is 2.94%, and will win by less than ten points is 52.4%. Overall, with everything equal, the home team has a 55.3% chance of winning the game. This is similar to the logistic model, and reflects the actual historical home winning percentage in the NFL over the last ten years.

With this model we can give the theoretical winning probability and score margin for any combination of marginal variable values. For example, if the away team has five more first downs, fifty more rushing yards, seventy-five more passing yards, a 3rd down conversion percentage ten points higher than the home team, and close values for the other marginal statistics, we can see that the away overall percentage of winning is around 78.2%, with a 10.7% chance of the game being a blowout by the away team.

The proportional odds analysis provides a better way to visualize the possible outcomes of a game using our models. It also reminds us that there is a lot of variation and uncertainty when applying these models to new data sets that should be accounted for. The fitted value should not be taken directly as a prediction of how a game will end, but rather as a point estimate that should be used along with a range of possible outcomes and their probabilities.

4. Conclusion

The goal of this paper was to determine the most significant variables, collected over the course of games and seasons in the NFL, that contribute to success. On a short term basis, success can be defined as winning individual games. Three different models were developed that all

worked well on the testing data set. Models identified similar sets of variables that were significant. From these models, one can see the effect of turnovers, third down conversion percentages, sack yardage lost, among others, on the point spread and probability of winning the game.

This research was different than other recent sport research studies which focus on forecasting results. This research found significant variables explaining the point spread of a football game. It also developed models to estimate the probability of a team winning based on given values of the significant variables. These models will give teams an idea as to the most important factors of winning a football game and what they should concentrate their efforts on.

REFERENCES

[1] *NFL Scores*. [2009-2013]. Retrieved July 2014, from ESPN.com: http://scores.espn.go.com/nfl/scoreboard

[2] *Pro Football Reference* [2016]. Retrieved November 2014 from https://www.pro-football-reference.com/years/2016/#all_team_stats

[3] Zuber, R.A. [1985]. "Beating the Spread: Testing the Efficiency of the Gambling Market for National Football League". *Journal of Political Economy*, Vol. 93, No. 4, pp. 800-806.

[4] Glickman, M.E. & Stern, H.S. [1998]. "A State-Space Model for National Football League Scores". *Journal of the American Statistical Association*, Vol. 93, No. 441, pp. 25-35.

[5] Stern, H.S. [1991]. "On the Probability of Winning a Football Game". *The American Statistician*, Vol. 45, No. 3, pp. 179-183.

[6] Baker, R.D. & McHale, I.G. [2013]. "Forecasting exact scores in National Football League games". *International Journal of Forecasting,* Vol. 29, pp. 122–130.

[7] Boulier, B.L. & Stekler, H.O. [2003]. "Predicting the outcomes of National Football League games". *International Journal of Forecasting,* Vol. 19, pp. 257–270.

[8] Harville, D. [1980]. "Predictions for National Football League Games Via Linear-Model Methodology". *Journal of the American Statistical Association*, Vol. 75, No. 371, pp. 516-524.

[9] Stefani, R.T. [1980]. "Improved Least Squares Football, Basketball, and Soccer Predictions". *IEEE Transactions on Systems, Man and Cybernetics*. Vol. 10, No. 2, pp. 116-123.

10] Long, J. & Magel, R. [2013]. "Identifying Significant In-Game Statistics and Developing Prediction Models for Outcomes of NCAA Division 1 Football Championship Subdivision (FCS) Games". *Journal of Statistical Science and Application.* Vol. 1, No. 1, pp. 51-62.

[11] Roith, J. & Magel, R. [2014]. "An Analysis of Factors Contributing to Wins in the National Hockey League". *International Journal of Sports Science*, Vol. 4, No. 3, pp. 84-90.

[12] Unruh, S. & Magel, R. [2013]. "Determining Factors Influencing the Outcome of College Basketball Games". *Open Journal of Statistics.* Vol. 3, pp. 225-230.

[13] Abraham, B & Ledolter J. [2006]. *Introduction to Regression Modeling* (1st ed..) Belmont, CA: Thomson Brooks/Cole.

[14] Derksen, S. & Kesselman, H.J. [1992]. "Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables". *British Journal of Mathematical and Statistical Psychology,* Vol. 45, No. 2,

pp. 265-282.
[15] Liao, D. & Valliant, R. [2012]. "Variance Inflation Factors in the Analysis of Complex Survey Data". *Survey Methodology*. Vol. 38, No. 1, pp. 53-62.
[16] Agresti, A. [2002]. *Categorical Data Analysis* (2nd ed..) New York: Wiley.
[17] Hanley, J.A. & McNeil, B.J. [1982]. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve". *Radiology*, Vol. 143, No. 1, pp. 29-36.
[18] Hosmer, D.W. & Lemeshow, S. [2000]. *Applied Logistic Regression* (2nd ed.). New York: John Wiley & Sons, Inc.
[19] Faraway, J.J. [2006]. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (1st ed.) Boca Raton, FL: Chapman & Hall/CRC.