A STUDY OF INFLUENTIAL STATISTICS ASSOCIATED WITH SUCCESS IN THE NATIONAL FOOTBALL LEAGUE

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Joseph Michael Roith

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Statistics

April 2015

Fargo, North Dakota

North Dakota State University
Graduate School

Title

A Study of Influential Statistics Associated With

Success in the National Football League

By

Joseph Michael Roith

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Rhonda Magel
Chair

Ronald Degges

Gang Shen

Brian Slator

Approved:

4/13/2015
Date

Rhonda Magel
Department Chair

ABSTRACT

This dissertation considers the most important aspects of success in the National Football League (NFL). Success is defined, for this paper, as winning individual games in the short term, and making the playoffs over the course of a season in the long term. Data was collected for 750 different regular season games over the course of five seasons in the NFL, and used to create models that identify those factors which are most significant towards winning at both the short term and long term levels.

A point spread model was developed using an ordinary least squares regression method, and stepwise selection technique to reduce the number of variables included. Logistic regression models were also created to state the probability a team will win an individual game, and also the probability a team will make the playoffs at the end of the season. Discriminant analysis was performed to compare the significant variables in our models, and determine which had the largest influence. We considered the relationship between offense and defense in the NFL to conclude whether or not one area had a significant advantage over the other. We also fit a proportional odds model on the data set to categorize blowout games, and those that are close at the end.

The overwhelming presence of turnover margin, passing efficiency, first down margin, and sack yardage in all of our models is clear evidence that there are a handful of statistics that can explain success in the NFL. Using the statistics from games, we were able to correctly identify the winner around 88% of the time. Finally, we used simulations and historical team performances to forecast future game outcomes, our models classified the actual winner with a 71% accuracy rate.

Analytics are slowly gaining momentum in football, and the advantages are clear. Quantifying success in the NFL can benefit both individual teams, and the league as a whole, to present the best possible product to their audiences.

ACKNOWLEDGMENTS

I would like to thank Dr. Rhonda Magel for advising me throughout the process of writing and reviewing this dissertation. And also for encouraging and supporting a subject matter that so deeply interests me. Finally, I want to thank Dr. Magel for all of the opportunities offered to me during my time at NDSU.

I also want to thank the rest of my committee members, Dr. Ronald Degges, Dr. Gang Shen, and Dr. Brian Slator for the time they took and the input they contributed towards my dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Sports in the United States elicit a passionate and loyal following. In recent years, football has taken the place as the most popular sport, both on a professional, and amateur level (Luker [2011]). In particular, the National Football League (NFL) is the pinnacle of football talent and competition. As a business, the NFL is the most lucrative sports league in the world, with revenue of over $9 billion in 2013 (Burke [2013]). As the popularity and income increase, teams look for better ways to maximize their investments, and to put the best product on the playing field. This has raised the question of how statistics can relate and contribute to the arena of sports. In this paper, we will look at different ways to analyze NFL games, find factors that are important to winning those games, and create models that can best predict the outcome of future contests based on historical performances.

The National Football League was founded on August 20, 1920 in Canton, Ohio. The league originally consisted of fourteen teams, now almost 100 years later and after many mergers with rival associations, there are thirty-two teams across the United States. Currently, every season these teams play each other in three types of games; exhibition, regular season, and playoffs. Exhibition games start the season and do not count towards a team's spot in the standings. There are sixteen regular season games played by each team over the course of seventeen weeks that determine the relative position in the standings based on the win-loss record of the team. After the regular season, teams are ranked by record, and participate in a four round, single elimination playoff tournament to decide a league champion in the Super Bowl.

The league is divided into two conferences, the National Football Conference (NFC), and the American Football Conference (AFC). Each conference consists of sixteen teams separated into four divisions of four teams. At the end of the regular season schedule, the top team in every division secures a playoff berth. Then, from the remaining teams, the two in each conference with the best records, called wildcards, also make the playoffs. Thus, there are six playoff teams from both the NFC and AFC, and the seeding are as follows in each league:

Seed 1: Division winner with best record

Seed 2: Division winner with second best record

Seed 3: Division winner with third best record

Seed 4: Division winner with fourth best record

Seed 5: Wildcard team with best record

Seed 6: Wildcard team with second best record

For the purposes of this paper, we are focusing on regular season games, and the teams that make the playoffs at the end of each season. From 2011 to 2014, all NFL games were recorded for our data set. A multitude of commonly collected in-game statistics were kept, along with some more obscure measurements. Included in these were home and away team, points scored, total yards gained, number of turnovers committed, along with over fifty others. The current format and rules for the National Football League have been fairly consistent over this period of time, with only minor changes to address the safety of players, not the outcome of games. Therefore, we feel that this is a good representation of the current state of play in the league and a reasonable data set from which to develop our models.

With these data sets, we will set out to examine games in five different ways. First we will consider the score margin at the end of games. This tells us who won the game and by how many points. We will use ordinary least squares regression to create a model with score margin as the response. The model will be created using the first two years of our data set, and tested using the third year. Accuracy in not only identifying the winners of each game, but also the margin they won by is the objective with this analysis.

Next, we will develop models that can accurately derive the outcomes of games using in game statistics. The first two years of NFL games in our data set will again constitute our model training data set. Using a logistic regression model, we will report the probability that each team will win a matchup by means of the significant covariates chosen by our model selection procedure. We will also consider long term success, by the measure of making the playoffs. A logistic regression equation will be created to assess the probability of each team making the playoffs at the end of the season. Using statistic totals at the end of seasons, we hope to determine those that are the best indicators of successfully making the playoffs.

Then, we will use a discriminant analysis approach to determine the major factors that go into winning individual games, and also making the playoffs at the end of the season. Our goal here, is to see if there are certain combinations of in game statistical measurements that aid or prevent a team from experiencing success in the league. Once again, we will create models for both long term success over the course of a season, and short term success in individual games. One hypothesis is that defense is a greater contributor to victory than offense. With our discriminant analysis, we will be able to quantify these effects and make relative comparisons to see the overall influences from both offensive and defensive statistics.

Additionally, we will consider a proportional odds model to forecast the outcome of games, along with a general range of the margin of victory. By considering not only which team wins the contest, but also whether it was within ten points or decided by more than ten points, we will create different categories games can be classified into. These unique categories will represent any possible outcome at the end of the game, a large win by the away team, a small win by the away team, a small win by the home team, and a large win by the home team. The proportional odds model will assign a probability of the final score margin falling into each of the four categories. Whichever one has the largest probability will be cross validated with the actual outcome to check accuracy. This method of analysis provides a more or less combination of the logistic regression approach with the point spread model.

The last part of our analysis will be to validate and forecast the outcome of future games. With the models developed from the training data sets, we will fit the outcomes of contests from the third year of the collected games. Here, we hope to find that the models will work just as well when used on games that are outside of the data set they were developed with to validate our findings. In the case of the point spread model, we will also attempt to accurately predict the difference in scores, along with the winner.

Simulation will provide a useful tool as well, since there are many sources of underlying variation that go into any football game, we will simulate the future performances of teams based on the games that they have already played during that season. Each of the significant variables in a model will have its own distribution, and together with the data from the opposing team, one game can be simulated and the outcome recorded. With enough simulations of each game, we hope to get an accurate picture of how we can expect the teams to perform during their upcoming games, and check it against the actual results.

The goal of this paper is to develop models which emphasize the important areas that contribute to winning in the NFL. We would like to know if there are consistent areas that can be pointed to as indicators of highly performing teams. If there are consistent factors, we would like to quantify them, so as to eliminate any ambiguity a coach or casual fan may have while watching a game. We also want to offer evidence for, or against, the long held notion that “defense wins championships”. Finally, we would like to provide models that can forecast the outcome of games based solely on recent performances. With all of this analysis, we hope to prove that sports analytics has a viable place in the NFL, and that it continues to grow with support and implementation.

CHAPTER 2. LITERATURE REVIEW

Much of the recent literature on the NFL is focused on forecasting professional football games with respect to the efficiency of the betting market (see Glickman & Stern [1998] and Stern [1991]). Whether or not one can consistently outperform the bookmakers, and the economic impact that could have in the market, is a common theme. One example of such research was performed by Baker and McHale [2013]. In their paper, they propose to model the overall combined score of both teams at the end of the game. This is then compared to a previously determined over-under line of total points scored set by bookmakers to test whether they are more accurate in predicting scores in NFL games.

They model scoring as a birth process with a continuous-time Markov chain. Two hazard functions are offered, representing the probability of each team, home and away, scoring throughout the game. Both of these hazard functions are based on certain variables that are "in practice predictors of attacking and defensive strengths derived from previous games." They also rely on previous work by Zuber [1985], who develops his own model to predict games, to select their significant in game statistics. In the end, the models Baker and McHale propose can be broken down into three parts, the in game statistics, the current points scored for each team, and a recalibrated point spread. The best model was able to accurately predict the winner in about 67% of the contests and also had a mean absolute deviation of 7.7 from the actual point spread, which is a measure of how close their predicted score was to the actual score.

One of the main issues with the models introduced by Baker and McHale is that they are not trying to explain team success in football or even predict the winner of games (although it does a good job). They merely are trying to “beat” the over-under, in other words, predict the total number of points scored by both teams. Also, to do this, they use information from the book makers as a weight for the amount each team will win. Specifically, the coefficient for home points is (Point Spread + Over-Under)/2. Both Point Spread and Over-Under are taken from an online betting website, and while these numbers are most likely determined by some formula, ultimately they are subject to adjustments by the bookmaker based on the betting volume for each team, so as to limit their own financial loss. This leaves many unknown quantities and subjective elements which we would like to avoid in our research.

One nice aspect of the models, is that they do incorporate the current score. This allows flexibility for predictions and can produce a more complete picture with the added information, such as halftime scores, or quarter by quarter scores. But the authors seem to be more concerned with the distribution of scoring as a total measure from both teams, and also in their ability to prove or disprove the efficiency of the current betting market. Nowhere is there any mention about what leads to scoring during the football games, or what drives the total score values higher or lower in any given game.

In a paper by Boulier and Stekler [2003] a new approach to predicting NFL outcomes was presented. Taking the rankings of teams developed by a power score printed weekly in the New York Times, the authors looked at the probability of higher ranked teams winning each game. The power scores are meant to summarize the relative performance of each teams' past games. Since these ranks were taken from an independent source, no methodology was given as to how they are exactly computed. Boulier and Stekler mainly want to find out if choosing the higher ranked team can outperform the betting market, and also a "sports expert". Over the course of seven NFL seasons, this method was able to forecast the correct winner around 61% of the time, slightly better than the sports editor at the New York Times, at 60%, but still significantly worse than the 66% accuracy of the betting market over that same period of time.

From there, the authors looked a little closer at the individual matchups between the different ranks. They present a distribution of all the possible combinations of ranks playing in a game and recorded the winning percentages of the higher rank. As one may expect, when the difference in rank is larger, for instance twenty-four, the higher ranked team won more often, 75% of the time. And when the difference in rank was low, say two, the outcome of the game was less certain, the higher ranked team winning about 57% of the time.

Now, using the difference in ranks as a covariate, along with a dummy variable for the home team, they perform a series of recursive probit regressions to predict the outcomes for all remaining weeks each season in their data set. As may be expected, the models lose their predictive ability the further they get into future games, however, the point is made in the ability to use power scores and a ranking system to forecast.

It is interesting to observe the differences in the sport of football from its early days and the evolution to its current form. Harville [1980] looked at games from the 1970's to create a predictive model and found much 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, he still underperformed. It is possible that in earlier eras of the game, talent from team to team was less equally distributed. This could possibly create more discrepancy in the ability for some teams to perform at a high level, and therefore make prediction easier to model with a few very good teams playing against many that are mediocre.

Stefani [1980], also looked at professional football 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, developed earlier, and an estimated home field advantage. The final accuracy for predicting correct games 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. Our goal is to provide a similar model, with improved accuracy and a smaller deviation of final scores from the actual games.

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 [2013] considered football games at the collegiate level, analyzing games from the NCAA Division I Football Championship Subdivision. Here they used regression techniques to identify significant in game statistics and 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.

One interesting comparison will be the results of their model with the resulting models from our research. We may be able to point out the difference in effect of some in game statistics from an amateur level to the professional level. When there are fewer teams with more equal talent, as is the case in the NFL, are the effects of these statistics changed in any way?

In other sports, Roith and Magel [2014] use discriminant analysis to determine the difference of offensive and defensive statistics in the National Hockey League (NHL). Similar 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 [2014] 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 [2013] 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. So we can see that modeling individual games for most sports seems to have some limitations in predictive power regardless of the sport. In each case, the most accurate models can correctly predict the winner of the contest around 65-70% of the time. We will use this basis as a benchmark for our models, to determine if we can stay consistent with current methods of forecasting games, or improve on them.

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. We will also look at the predictive power of our models to see if it can be comparable to current research or possibly even surpass current forecasting abilities.

CHAPTER 3. DESIGN OF STUDY

The main focus of this paper is to identify the key contributors in NFL games that lead to both short term and long term success, namely winning individual games and making the playoffs at the end of the regular season. As mentioned earlier, we would like to not only identify these aspects, but also to quantify them, and will consider five analytical methods for several different NFL data sets; ordinary least squares regression, logistic regression, discriminant analysis, a proportional odds model, and simulation to forecast games. The data sets are divided into two groups, individual game statistics and seasonal total statistics.

The individual in game statistics were collected for nearly every regular season game, including ties, over three NFL seasons from 2011-2014, a total of 752 games. The values for thirty-five game statistics for both the home and away team were recorded from the box scores found on ESPN.com, creating seventy variables overall. Some of these measurements include total yards gained, total passing yards, turnovers, and sacks. A full list of all the initial variables collected can be found in Appendix A.

The game statistics were further separated into two groups, a model training group, and a model testing group. The model training group is comprised of all games played during the first two seasons of the collected data, the 2011-12 season and the 2012-13 season. The model testing group contains the remaining data from the 2013-14 NFL season. As it implies, the training data set will be used for all single game analysis. Each of the different models will be determined from these games. The testing data set will be used to forecast games that have not been included in the development of the models, therefore providing a more accurate check on validity. All games during this timeframe are played under the same rules, with similar players and no major changes to overall style of play in the league, so we are confident the samples should be homogenous.

The second type of data are seasonal statistic totals for teams each year. The data set covers each of the thirty-two teams for five seasons, from 2009 to 2013, a sum of 160 observations. The totals of twenty-seven different statistics were collected for each team, along with whether or not they made the playoffs that year. Some examples of the variables collected are; total points, total yards, total turnover margin, and total penalties yards. A full list of the initial variables collected and considered is available in Appendix B.

It should be noted that not every variable collected was initially considered for every model. Some of the statistics recorded are a result of the sum of other variables. For instance, turnover margin is the sum of interception margin and fumble margin, and first down margin is the sum of passing first down, rushing first down, and first down by penalty margins. To avoid redundant information, we will consider two sets of variables initially to fit models, one with the broader totals, and one that is more specific and includes the breakdown of each category. The total measurements and partial measurements of a particular variable will never be considered for the same model. Rather, two separate models will be developed and the best one chosen by the methods appropriate for each type of analysis. Appendices A and B also indicate which variables were considered total measurements and which were considered individual parts of the total.

To verify all of these approaches, we will use a combination of forecasting techniques and simulation to compare with the real world results and some of the previously researched methods mentioned earlier. Hopefully, through our analyses, we will be able to point out and quantify those facets of NFL games that players, coaches, and decision makers can look at and isolate to try and improve individual, along with team performances.

3.1. Ordinary Least Squares Regression

The first type of statistical analysis that we will perform on the data set will be Ordinary Least Squares (OLS) Regression (Abraham & Ledolter [2006]). The purpose of this technique is to develop models that exhibit some sort of underlying linear relationship between our response variables and the independent variables. The general form of an OLS regression model is:

$$y = \boldsymbol{X}\beta + \varepsilon \qquad \text{(Eq. 1)}$$

where $\boldsymbol{X}$ is a matrix of the observed values for each of the independent variables (including an intercept when appropriate). The vector β consists of the coefficients for each of the predictor variables included in $\boldsymbol{X}$. These coefficients represent the net effect that a single unit change in each variable has on the expected value of the response, y. Any inherent variation in the data that cannot be explained by the included variables is included with the error term, ε. In order to use the OLS model, certain assumptions about the error need to be met. Specifically, the error terms must be independent from each other, they must follow a Normal distribution, with an expected value of zero, and have the same constant variance. These assumptions will need to be checked and met for our models to show validity.

For our research, we will be using the OLS method to model the final point spread at the end of each regular season NFL game. In this case, the response variable is calculated by taking the score of the home team and subtracting the score of the away team. The result is our dependent variable, score margin. The score margin represents the winner of the game, as determined by the sign of the value, and also the number of points the team has won by. A negative value for score margin indicates the away team has won, and similarly a positive value that the home team has won.

The independent variables will be created in a similar fashion to the score margin. For all statistics collected, there is a value for the home team and the away team. Our covariate will be the difference between the home team and the away team. For example, a variable included in a model may be turnover margin, this is calculated by taking the number of turnovers for the home team and subtracting the number of turnovers for the away team.

Initially, we will consider all of the collected in game statistics in our regressor matrix, $\boldsymbol{X}$. However, since not all of these predictor variables contribute significantly to the determination of score margin, and because some even express multicollinearity between each other, we will use a selection process to simplify the model. Appendix A lists all of the variables along with those that will not be considered together due to close associations. We will not consider any interactions or higher order terms in our OLS model since interpretation is often difficult and the purpose of this paper is to clarify and quantify the effects of different statistical measures in the NFL.

As mentioned above, when selecting which covariates to include, we first need to eliminate those that are redundant. Since some of the variables represent a combination of others, we will develop several models for consideration but never include both the total value and the individual parts in any one model. If each is significant in its respective model, we can compare the efficiency and simplicity of the models to determine which is more appropriate.

Once we have pared down the initial variables we would like to include, a stepwise selection process will be performed to identify those variables that are most significant (Derksen & Kesselman [1992]). As each variable is considered in the model, a simple t-test is performed to assess its significance. To be incorporated in the model, a variable will have to be significant at an entry level of $\alpha = 0.10$. To stay in the model, a variable cannot exceed an exit significance level of $\alpha = 0.15$. This procedure will eliminate most of the superfluous predictors.

When we have all the variables deemed significant by the stepwise selection, any further elimination to simplify the model will be performed based on the value of R-squared, the coefficient of determination, and an adjusted coefficient of determination. R-squared is a measure, between 0 and 1, of the proportion of total variation in the response variable that can be explained by the current model. Every added variable will increase the R-squared value by some amount, so adjusted R-squared measures the same variation, but gives a penalty for each extra variable included. This prevents variables that contribute a small or negligible amount of extra explanation of variance from being included in the model.

We will also consider the predicted R-squared value, which is a measurement of the ability to create a model and fit the response for each individual observation in the training data set one at a time using the remaining observations. The purpose of the predicted R-squared value is to ensure there is not an over fit of the data with too many variables, in other words, we do not want to have an abundance of variables that can explain the variation in the model but perform poorly fitting the response values.

Multicollinearity is also a concern when selecting variables for an OLS regression model. This occurs when two or more variables in the model are highly correlated with each other, meaning one variable can be linearly predicted with the others. The variance inflation factor (VIF) is a measure of multicollinearity for each variable (Liao & Valliant [2012]). As a rule of thumb, any VIF value over ten indicates a variable that is highly correlated with other variables, and this will be the criteria for our analyses to indicate this problem if it occurs.

Once we have established the preferred model for the score margin, we will be ready to interpret the coefficients of the covariates, and compare the relative effect each one has towards gaining, or losing, points throughout the game. We will also be able to apply this model to our testing data set, trying to fit the score margins of those games not used in model development. If the model is able to accurately produce score margins from games outside of the training data set, it will be ready for further forecasting of future games using simulations.

3.2. Logistic Regression

The second form of regression analysis we will conduct is logistic regression (Abraham & Ledolter [2006]). Logistic regression is a specific technique that is used when the response variable consists of two dichotomous outcomes. With only two distinct possibilities that occur, the logistic model can provide a probability of each event occurring based on the values of the predictor variables. This results in a way to measure the effects independent variables have on the probability of an event occurring. Here is the general form of the logistic model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{X}\beta \qquad \text{(Eq. 2)}$$

where π is the probability of the response being classified into the category of interest. The right side of the equation can be defined similarly to the OLS model described in Section 3.1. Since we are interested in the estimation of π, Equation 2 can be solved to get an estimate:

$$\pi = \frac{e^{\boldsymbol{X}\beta}}{1+e^{\boldsymbol{X}\beta}} \qquad \text{(Eq. 3)}$$

Now the estimate, π, is bounded by 0 and 1, giving a value of the probability that the response will fall into our category of interest. And it can easily be seen that the probability of the response belonging to the other category is estimated as 1- π.

The interpretation of the vector of estimated coefficients, β, changes as well. Now, with one unit increase of the *i*th individual predictor variable, the odds of observing a response belonging to the category of interest changes by a factor of e^{β_i}, β_i being the coefficient for the *i*th variable. This value is referred to as the odds ratio.

With the NFL data, we will create two logistic regression models. The first will consider only the individual games played throughout the regular season. The two categories for the response variable will be determined by whether or not the home team won the game, "1" represents a home team win, while "0" represents a home team loss. We will not consider ties as a possible outcome since, while they do occur, they are not frequent enough to be of interest. Thus our logistic model will provide a measure of probability that the home team will win the game in question.

Similarly to the OLS model, we will use the marginal statistics between the home and away team. We also want to simplify the number of independent variables included in the logistic model. Again, a stepwise selection procedure is applied to narrow the options. The model entry level of significance is set at $\alpha = 0.10$, and the model exit level of significance is set at $\alpha = 0.15$. A Wald test calculated for each variable at every step provides the mechanism to measure significance (Agresti [2002]).

After the significant variables are determined, any further model adjustments will be performed based on the criterion of 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 (Hanley & McNeil [1982]). A model performs well when the true positive rate is high while the false positive rate is low. Finally, the model fit will be tested using the method proposed by Hosmer and Lemeshow [2000] for the goodness of fit for logistic regression models.

In addition to observing individual games, we will create a logistic regression model to provide probabilities of making the playoffs at the end of the season for each NFL team. In this case, the dichotomous response will be “1” when a team has made the playoffs, and “0” when a team has missed the playoffs. The independent variables for this logistic model will simply be the totals of each statistic collected at the end of one season. All of the same selection and diagnostic techniques used for the model of single games will be applied to the model of the entire season here. We hope to see whether those factors that are significant for winning one game can be extrapolated throughout the entire season and relate to long term success in the form of making the postseason. Also we would like to know if the quantitative effect of those factors are similar in both cases.

3.3. Discriminant Analysis

Next we will approach the data with a multivariate technique called discriminant analysis (Rencher [2002]). Discriminant analysis focuses on considering multiple different classes separately, and creating a linear function that transforms the original variables so as to best classify an observation into one of several groups. The transformation is determined by the linear combination that maximizes the difference, or separation between the groups. This simply means the technique tries to find some combination of the variables that provides the best differentiation between classes, essentially rotating the axes. The general form of the transformed observations is called the linear discriminant function and is represented by:

$$z_i = a_i'\boldsymbol{X} \qquad \text{(Eq. 4)}$$

where a is the linear combination of the variables for classification into the ith class. The result is I different linear combinations, one for each available class, and a new set of transformed data, z_i.

This has the added benefit of reducing the dimensionality of the data, particularly useful when considering a large number of covariates.

The next step is to test the linear discriminant function by classifying the observations based on the transformation of the independent variables. This process is called classification analysis, and the specific method of classification we will use is referred to as cross validation (Rencher [2002]). When performing cross validation, each observation is held out of the data set one at a time, then the linear discriminant functions are computed with the remaining observations. The held out data point is subsequently plugged into each of the linear discriminant functions and the values are ranked from largest to smallest. The classification of the observation is then made based on the class associated with the linear discriminant function that produced the largest value. This classification can then be checked against the actual class the observation belongs to and the accuracy recorded.

One of the biggest advantages of discriminant analysis is that when comparing the standardized coefficients for each linear combination, the magnitude of the coefficient is directly related to the effect that variable has on the final classification. Therefore, the magnitude of every standardized coefficient for each function can be ranked and we can directly compare which factors are more significant towards classification into that group. To standardize the linear discriminant functions, we will multiply them by the square roots of the diagonal elements of the pooled covariance matrix. This will provide a means to not only find influential statistics contributing to success in the NFL, but also to compare and contrast them.

For our NFL data, we will be applying this analysis in several different ways. First, we will consider the individual games. In this case, the number of classes we will need a linear discriminant function for is two, once again, to represent the home team winning the game, and the other to represent the home team losing the game. We will initially consider all collected variables in our model, however, we would still like to simplify it to the point where we can identify only a handful of the most important covariates. Consequently, another stepwise selection procedure will be performed to narrow down the number of variables considered in the model. A partial F-test is conducted on every variable at each step and the model is reevaluated to find the optimal addition. The threshold for a variable to enter the model is a significance level of $\alpha = 0.20$, and the criteria to stay in the model is a significance level of at least $\alpha = 0.25$.

The selected model will then be cross validated to see how many home wins and home losses were correctly classified and how many were incorrect. This will provide an overall error rate that we hope is relatively low. With a successful model, we can compare the standardized, or normalized, magnitudes of the significant variables to see which are most important. We will then relate the effects of these variables for both the linear function associated with the "Home Win" class and the linear function associated with the "Home Loss" class, noting any obvious differences.

We will also perform this analysis on the seasonal data. The two classes will be determined by whether a team makes the playoffs versus misses the playoffs. The same type of selection process will be performed on this model as was used for individual games. The resulting linear discriminant functions will be cross validated, counting the number of correctly classified playoff or non-playoff teams and the misclassified teams. In this case, the error rate is calculated using a weight of prior probabilities. Since we know that twelve out of thirty-two teams make the playoffs and twenty out of thirty-two teams miss them, the prior probabilities will be set at 37.5% and 62.5% respectively.

Once we have validated the model, the coefficients can be compared for importance in long term success in the NFL. Again, this is possible despite the difference in the scale of measurement for each variable because the coefficients can be standardized. Since the same data sets for both games and seasons are used here as they are with the logistic regression models, we can start to note if there are any consistencies in the variables being selected for the models. If we start seeing the same covariates regularly in each of these models, we can begin building an idea of reliable areas significant to winning in the NFL.

A third discriminant analysis will also be conducted, focusing on individual games once again. However, this time we will not consider the marginal variables, but the original variable measurements of the value gained, and the value allowed by each team. Home and away teams will be mixed as our team of interest, instead of simply using the home team as a reference in the previous marginal data sets. The classes for the analysis will be "Win" and "Lose", and the interpretation of the variables now has a different meaning as well, namely a change in each variable represents a change in the total amount gained or allowed, not a difference over the opponent.

A stepwise selection procedure will be performed, not necessarily to choose our significant variables, but to see if they agree with our models already created. We will directly compare the effects of offense and defense by considering the variables that were found to be significant for our earlier models. With the standardized discriminant functions, we will relate the effect sizes of both, but the connection will be between pairs of variables, with a greater concern for providing evidence of offensive or defensive preference. For example, if yards gained has a smaller coefficient than yards allowed, we can conclude that defense is a stronger indicator of winning in that particular area since giving up more yards has the larger effect. The same cross validation performed earlier will be implemented to ensure the model accurately classifies winning and losing.

3.4. Proportional Odds Model

The final model that we will fit to the data is a proportional odds model (Faraway [2006]). This type of model is similar to the logistic regression model in the fact that the response is qualitative. The difference is that there are more than two categories the response can be classified into, and they are ordinal. The general form for a proportional odds model with J ordered categories, and with a logit link function is:

$$\log \frac{P_j}{1-P_j} = \boldsymbol{X}\beta \qquad \text{(Eq. 5)}$$

for j = 1, 2, …, (J-1). This implies that there are J-1 different log odds ratios to be calculated. The interpretation of each of these is the change in odds of moving from one lower class to the next higher one.

Using the game data from the NFL, we will fit a proportional odds model. The score margin response variable, y_i, will be separated into four different ordered categories based on the winner of the game, and also the point difference at the end, to create a new response variable that has a multinomial distribution, w_i.

$$w_i = \begin{cases} Strong\ Away\ Victory & y_i = [-\infty, -10) \\ Weak\ Away\ Victory & y_i = [-10, 0) \\ Weak\ Home\ Victory & y_i = [0, 10) \\ Strong\ Home\ Victory & y_i = [10, \infty) \end{cases}$$

The goal here is to create a more specific interpretation of modeling for the winner of each game. Now, we can also include information about whether the game was close, within ten points for either side, or the game was a blowout, more than ten points for either side. The cutoff of ten points was selected to represent one touchdown and one field goal, in this scenario, the team behind cannot tie the game in just one possession, nor do they need to score two consecutive touchdowns to catch their opponent. In theory, the cutoff point of the categories can be set at any level, but for our purposes, we will set it at ten points. Hopefully this proportional odds model can provide an alternate way to view the analysis of both the logistic and OLS regression models described earlier.

Now that we know we will have four distinct ordinal categories for our response, the log odds ratios implied in Equation 5 can be defined:

$$log\frac{P_{sa}}{1-P_{sa}} = \mu_1 + \alpha x_1 + \cdots + \gamma x_k \tag{Eq. 6}$$

$$log\frac{P_{wa}}{1-P_{wa}} = \mu_2 + \alpha x_1 + \cdots + \gamma x_k \tag{Eq. 7}$$

$$log\frac{P_{wh}}{1-P_{wh}} = \mu_3 + \alpha x_1 + \cdots + \gamma x_k \tag{Eq. 8}$$

where k is the number of covariates included in the model. Each P_j is a cumulative probability of an observation belonging to class j, or any other previous class. In addition, let p_J be the probability that an observation belongs only to the *J*th class. Therefore, P_{sa} = p_{sa}, $P_{wa} = p_{sa} + p_{wa}$, $P_{wh} = p_{sa} + p_{wa} + p_{wh}$, and $P_{sh} = p_{sa} + p_{wa} + p_{wh} + p_{sh} = 1$. Once the model is fit, we can then solve for the probability of a response being predicted in any of the classes.

One advantage to the proportional odds model is the ability to interpret the parameter estimates. In Equations 6, 7, and 8, the coefficients of the variables are constant for each class, and can be interpreted in the same fashion as the logistic model. For example, one unit increase in the variable x_1, with everything else staying constant, would correspond to a change in the odds of moving from one category, say a weak away victory, to the next, a weak home victory, by a factor of e^{α}. The only difference in the three log odds ratio equations are the intercepts, μ_j, these represent the thresholds of moving into the next class.

When we perform this analysis, to control the number of variables included in the model, we will employ a stepwise selection procedure based on the Akaike Information Criterion (AIC) of each model, which is a measure of the relative quality of a model for a given data set (Hansen [2007]). Once we have a model with a minimum AIC value, we can then apply it to the testing data set. For every observation, when the values for the selected variables are entered into the model, it will provide probabilities that the game will result in the away team winning by more than ten points, less than ten points, and the home team winning by more than ten or less than ten points. These classes will at that point be ranked by the most likely outcome to occur, and checked with the observed outcome of the game. We can then see how accurate the model is at predicting the winning team and the margin of victory. Furthermore, we will still be able to tell if we correctly classified the winning team, even if not the correct class for margin of victory.

3.5. Simulation

Each of the models introduced above will be created using the training data set as mentioned. This involves fitting or classifying the response based on the actual values of in game statistics or season statistic totals for games and seasons already played out. Then, the verification of these models is a process that will use the testing data set, as mentioned earlier. Again, this is a process that looks at games that have already been played and, using the statistics from that game, determining if our models are accurate in correctly classifying the response or, providing a fitted score margin close to the actual one observed. If the models are able to perform adequately, then we can say they are generalizable to all NFL games and seasons played under the same conditions and rules.

It is also of interest to test the ability to forecast NFL games without using information from the games themselves. One of the most difficult things to do in any sport is to try to predict how teams and players will perform with and against each other. There is so much natural variability involved with player and team performances, it is fundamentally challenging to do. In addition to that variability, there is also the impact of coaches, weather, crowds, and a myriad of other factors that are inherently difficult to include in quantitative models like the ones we propose. Much of the literature mentioned in Chapter 2 is concerned with trying to predict the outcome of games as best they can, with our research, we are more interested in the effects of measureable performances in the game. One fascinating type of analysis we can perform is simulation of games without using the observed statistics.

For the models that are concerned with individual games in the NFL, the point spread, logistic, discriminant analysis, and proportional odds, we would like to see if the historical performance of the teams playing can be used to forecast the outcome with some degree of accuracy. Using simulation and the models we develop, this is possible to carry out. For most of the in game statistics that were collected, the values follow a normal distribution. So for each team, throughout the season, we can compute the mean and variance of every statistic with the games they have already played. Using these as parameters, we can then simulate a future performance against their opponent, based on the normal distribution. We expect that the teams will perform within a similar range of their previous games.

For example, say we find that total yards is a significant factor, and we want to simulate the outcome of a game between two teams during Week 9 of the NFL season. Then, for each team we can find their mean and variance for total yards from the previous eight weeks and, based on those statistics, simulate 10,000 total yard values they might have for the upcoming game. Likewise, we can look at how many total yards each team allowed their opponent to have through the previous eight weeks, and simulate 10,000 different values of total yards allowed. With all of these numbers, a new marginal statistic can be computed and entered into our models which is a difference of the average number of yards gained by one team and allowed by their opponent. Here is an example of how we will calculate the new marginal measure using the total yards gained and allowed by the home and away teams:

$$\left[\frac{(Home\ TotYds + Away\ TotYds\ Allowed)}{2}\right] - \left[\frac{(Away\ TotYds + Home\ TotYds\ Allowed)}{2}\right]$$

Then the fitted results of 10,000 simulated games can be viewed as a whole and we will be able to provide an expected result based on these simulations to compare with the observed results from the actual game.

We will need at least four games played prior to create our values of mean and variance for the significant variables, so only games played after Week 4 will be considered. That means we will study the 193 games from the testing data set after that point in the season. We will evaluate the OLS regression, logistic, discriminant analysis, and proportional odds models for individual games only, using the simulations. The offensive statistics simulated for each team will be used first to create the marginal variables, for instance, the simulated number of passing yards for the home team minus the simulated number of passing yards for the away team. Next, we will use the simulated values for each significant variable gained by each team, and allowed by each team to create a marginal value in the manner introduced above. This will represent the presence of defensive capabilities along with offensive prowess for both teams.

The models for making the playoffs using seasonal data will not be considered for this simulation analysis since teams change personnel so much from season to season. Not to mention, that there are even more chances for teams to have players change and get injured during a single season, which would greatly affect the forecasted values. A reliable way to calculate the mean and variance of statistics for season totals for each team across at least four seasons is simply not available.

We will then compare our simulation results to some of the previous research mentioned earlier, and also some different naïve approaches, such as selecting the home team to win, to see if it can perform similarly. We do not expect to be able to predict games with any specific amount of certainty, but any accuracy we do find should help indicate an idea of the level of consistency of performances from week to week. Simulation also provides a way to look at our fitted responses for models such as the point spread model, and determine the probability of a range of different outcomes. Just because the model provides a fitted value, we also want to know the distribution of the outcome, this can tell us if there is a wide variation in the simulated outcome or if we can be fairly confident in our predicted value.

CHAPTER 4. RESULTS

4.1. Ordinary Least Squares Regression

The goal of the OLS regression model is to develop a way to predict the score margin at the end of NFL games. Table 4.1 shows a summary of the stepwise selection procedure on the group of covariates for the model that was ultimately chosen. In total, eleven variables were selected to enter the model, and all eleven are significant at a level of $\alpha = 0.05$.

Table 4.1. OLS Regression Stepwise Selection Summary

STEP	VARIABLE ENTERED	NUMBER VARS IN	PARTIAL R-SQUARE	MODEL R-SQUARE	F VALUE	PR > F
1	YPPassM	1	0.4309	0.4309	374.82	<.0001
2	TurnoverM	2	0.2121	0.6430	293.55	<.0001
3	FirstDownM	3	0.0874	0.7304	159.78	<.0001
4	TotalPlayM	4	0.0452	0.7756	98.99	<.0001
5	3DPerM	5	0.0385	0.8140	101.52	<.0001
6	SackYardsM	6	0.0159	0.8299	45.89	<.0001
7	AvePRM	7	0.0041	0.8340	12.01	0.0006
8	YPRushM	8	0.0035	0.8375	10.57	0.0012
9	PenYardsM	9	0.0046	0.8422	14.23	0.0002
10	WinPerM	10	0.0019	0.8441	6.01	0.0146
11	AveKRM	11	0.0016	0.8457	5.17	0.0234

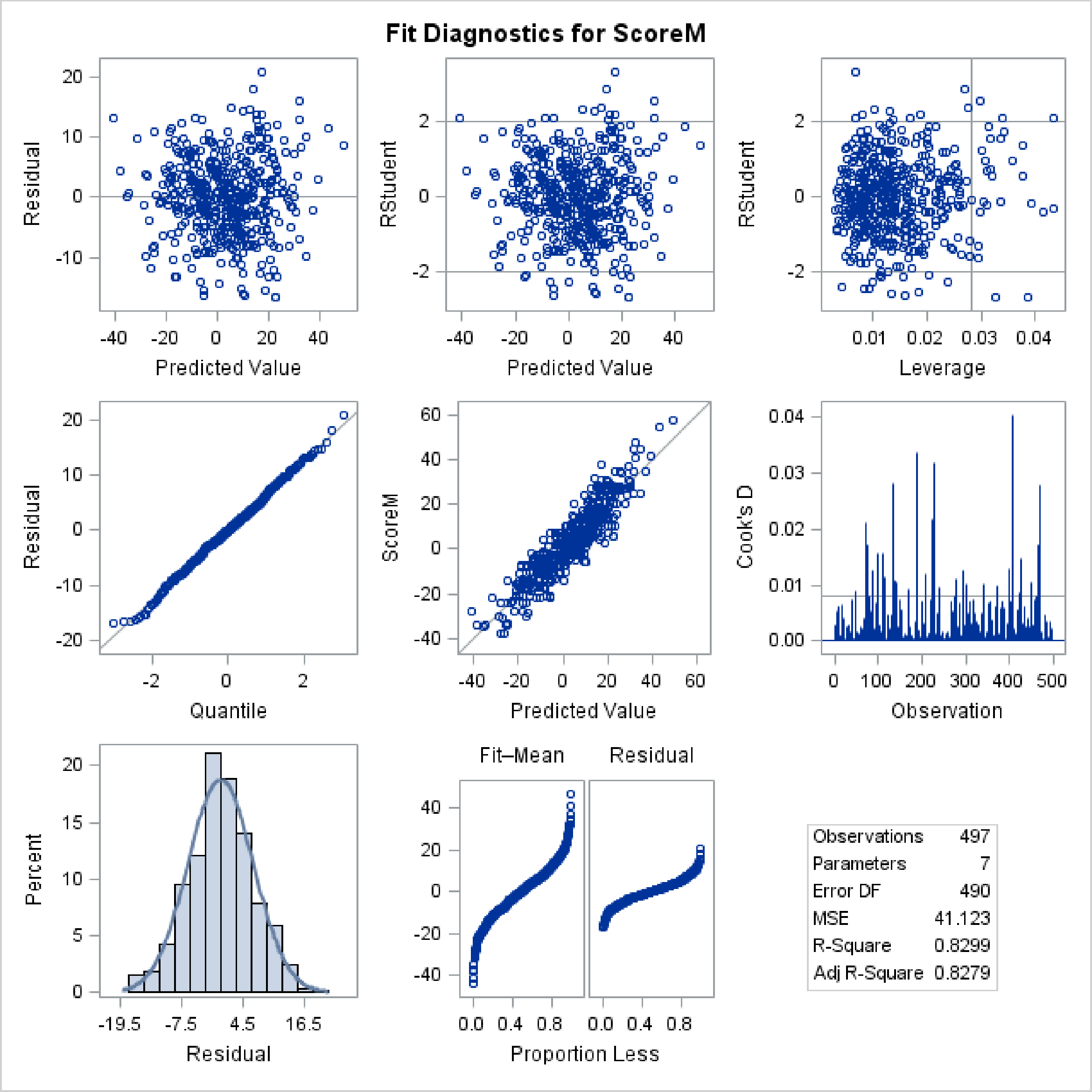
The R-squared value for the model is 0.8457, meaning 84.57% of the variation in score margin can be explained by the linear combination of these eleven variables. Typically in most real world modeling, any R-squared value over 0.80 is considered a very good percentage for the model. However, we would still like to reduce the number of variables we are using to fit the data. Considering the last few variables that were entered into the model, you can see that the partial R-squared values are all under 0.01. This mean that their inclusion in the model increases the explanation of the variance of score margin by less than 1%. If we remove all the variables with partial R-squared values under 0.01, we only lose a total of 1.58% in our overall R-squared value, but simplify the model by five variables, so they will be excluded.

The final model is then recalculated, and Table 4.2 shows a summary of the coefficients. It should be noted that the intercept was included in this model since it was also found to be significant. The way the variables were derived, marginal values of home team minus away team, 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.

Table 4.2. 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

The final model has an adjusted R-squared value of 0.8279, and a predicted R-squared value of 0.8245, meaning there is no issue with over fitting the data and it can be used to predict other games. The variance inflation factors for each of the variables is no higher than 4.36, which indicates there are no problems with multicollinearity. The rest of the diagnostics of the model fit show no violations in the initial model assumptions. Figure 4.1 provides a summary of these diagnostics. The residuals follow a normal distribution, with a mean of zero and a constant variance. There are very few observations that might be considered outliers, and even those do not affect the model significantly enough to warrant any additional action. Therefore, we will call this our final model and proceed to validation by submitting it to our testing data set.



Observations	497
Parameters	7
Error DF	490
MSE	41.123
R-Square	0.8299
Adj R-Square	0.8279

Figure 4.1. OLS Regression Model Diagnostics

The in game statistics for the testing model set will be entered into our model and a fitted score margin created. If the fitted margin has a positive value, we would expect the home team to win, if it is negative, the away team. Using the model on the testing data set, we were able to correctly identify the winner for 220 out of 256 games, an accuracy of 85.9%. Also, the mean absolute deviation of the predicted scores from the observed, 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. So, our model works just as well for games independent from the ones with which it was created. 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)

We can interpret this by saying, for example, 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. This is one of the larger coefficients, so we anticipate first downs to have a more significant effect on the game. The coefficient with the largest magnitude is turnover margin. For every extra turnover the home team commits over their opponent, they will give up around 3.9 points in the final score margin. Turnovers are the most costly area where a team can lose points, they are more rare, but lead to points for the opposing team more often.

We will look at how this model performs without using the in game statistics in Section 4.5, with simulation. But, here we can get an initial idea of what is important when trying to analyze NFL games. Turnovers seem to play a very critical role in the measure of the final score. As we perform further analyses, we will note any consistencies between variables chosen as significant for the subsequent methods.

4.2. Logistic Regression

4.2.1. Game Analysis

The first of our logistic regression models will focus on the outcome of individual games as the dependent variable. It will be coded in terms of the home team winning, “1”, or losing, “0”, the game. A summary of the initial stepwise selection procedure is presented in Table 4.3. Nine variable are included in the model. Upon further inspection, we will remove the last two variables entered into the model. The entry and exit levels of significance were set at a higher than usual level to ensure at least some variables are allowed into the model. With a p-value of 0.1294, we can feel comfortable removing yards per rush margin from the model. Also, average punt return margin will be removed to simplify the model, with little affect to the model fit.

Table 4.3. Game Logistic Regression Stepwise Selection Summary

	Effect		**Score**	
Step	**Entered**	**Removed**	**Chi-Square**	**Pr > ChiSq**
1	**YPPassM**		146.8567	<.0001
2	**TurnoverM**		102.1342	<.0001
3	**3DPerM**		63.4548	<.0001
4	**PenYardsM**		16.1683	<.0001
5	**SackYardsM**		16.3217	<.0001
6	**FirstDownM**		14.1141	0.0002
7	**TotalPlayM**		17.2579	<.0001
8	**AvePRM**		3.8950	0.0484
9	**YPRushM**		2.3002	0.1294

The analysis of our selected variables is performed again, and Table 4.4 shows the parameter estimates along with their odds ratios. It should be noted that we have included the intercept in the model despite the fact that it was tested to be insignificant. The interpretation of the intercept is similar to that of the OLS model, it is any inherent advantage the home team exhibits over the away team. Without the intercept, and with both teams at equal levels for each of the marginal statistics, the logistic model will give a 50% chance of either team winning. With the intercept included, and all other variables at zero, the model gives the home team a 55% chance of winning. This is very close to the actual home winning percentage we observed in our data set, around 56%. It is also right around the historical value of home winning percentage over the last ten years, 57.3%.

Table 4.4. Parameter Estimates and Odds Ratios for Game Model

		Standard			**95%**	
Parameter	**Estimate**	**Error**	**Pr > ChiSq**	**Odds Ratio**	**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

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, indicating that our model is indeed a good fit for the data. The max rescaled R-squared value for this model was 0.7867, indicating a major improvement over the baseline model. Also, checking the ROC curve in Figure 4.2, the area under the curve is 0.9621. The highest value for the area under an ROC curve is 1.0, meaning the model perfectly classified all observations, so the area for our model indicates a high rate of correct classification.

The interpretation of the model parameters is best done through the odds ratios. These indicate how the odds of the home team winning changes with an adjustment in each individual covariate. For example, with every extra first down over their opponent, the odds of the home team winning the game increase by a factor of 1.332, with everything else being constant. Figure 4.3 illustrates the change in the probability of winning for the home team as first down margin increases and everything else is held constant. Clearly, any positive marginal value for the home team in first downs will correspond to a high probability of winning the game. The largest negative effect for the logistic model is turnovers, for every extra turnover committed the odds of winning the game are decreased by a factor of 0.271.

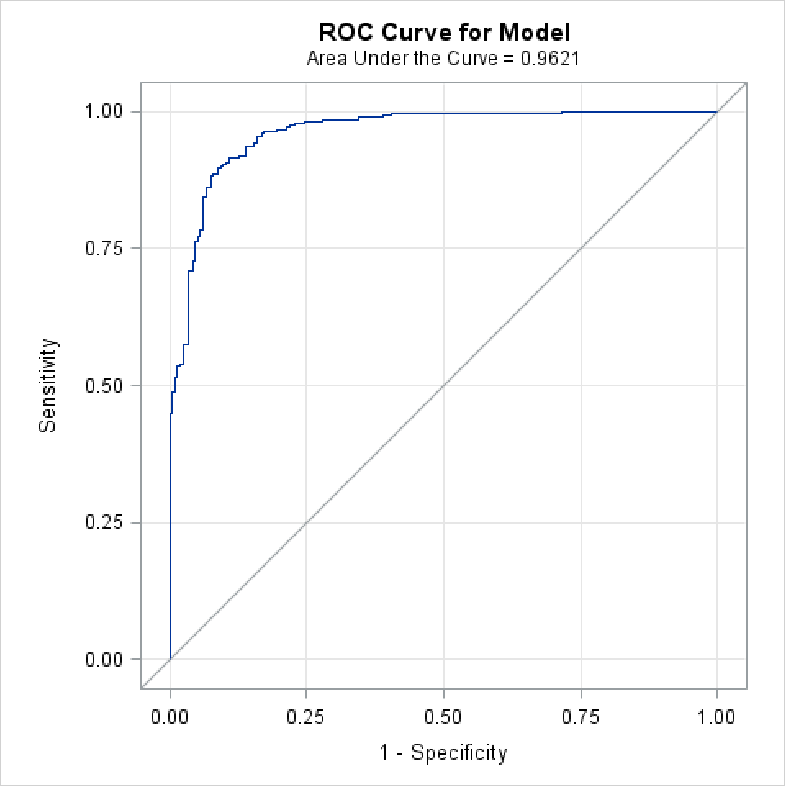


Figure 4.2. ROC Curve for Game Model

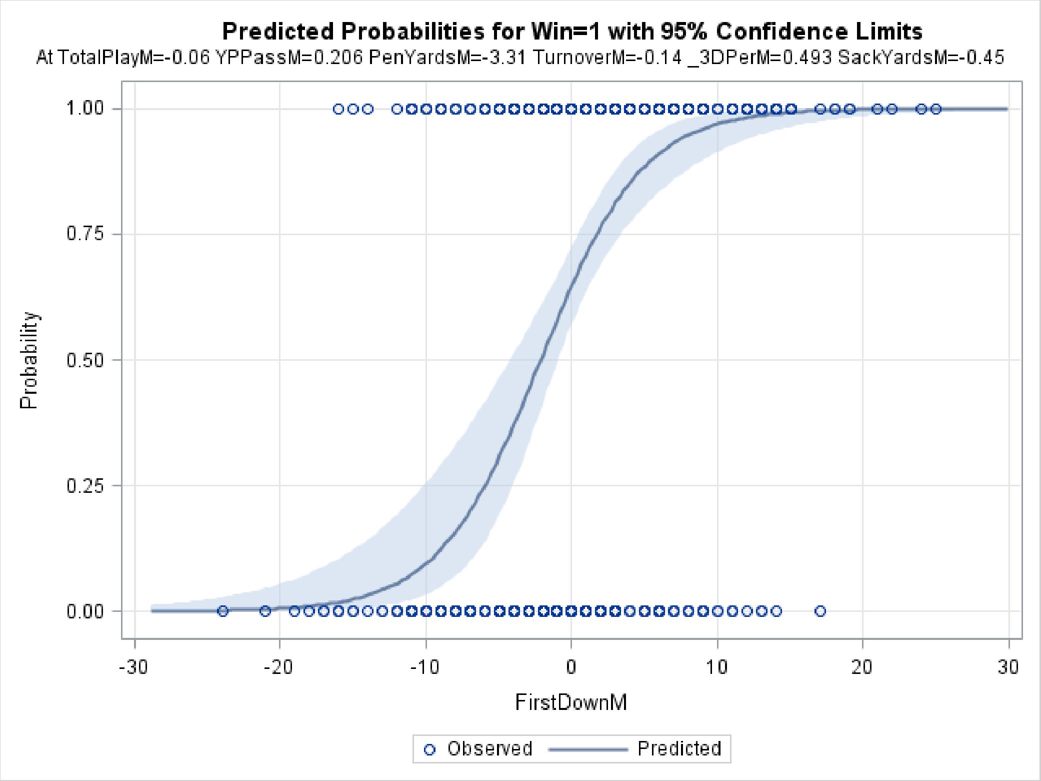


Figure 4.3. Effect of First Down Margin on Probability of Winning

Performing the model validation on the testing data set using in game statistics, we again saw 220 out of 256 games (although not the exact same games) correctly classified as home win or home loss. This means our model is valid for all games played under the current format. The simulation section will determine if we can use this model with historical performances to forecast games without the in game statistics.

The seven variables for this logistic model are the almost identical to the variables selected for the point spread model, with the only addition being penalty yards margin. Furthermore, it appears that turnover and yards per pass margin have the largest effect in both cases. This is the type of consistency that we would like to see throughout all of our different models created for the same data set.

4.2.2. Season Analysis

The second logistic regression model we will create uses the total season statistics for each team. The response variable in this model will be coded by whether the team of interest made the playoffs at the end of the season, “1”, or missed the playoffs, “0”. It is also important to note that the covariates are no longer marginal statistics, but represent simply the totals from sixteen games for the team. Table 4.5 shows a summary of the initial stepwise selection procedure for the logistic regression model.

Five variables are selected to stay in the model, and they are all significant at a level of at least α = 0.1. Table 4.6 provides the parameter estimates along with the odds ratio values and confidence intervals. For this seasonal model, the intercept was not significant, and unlike the previous case for in game marginal statistics, its inclusion does not make any sense in helping to explain the data better, so it was not considered necessary.

Table 4.5. Season Logistic Regression Stepwise Selection Summary

	Effect		**Score**	**Wald**	
Step	**Entered**	**Removed**	**Chi-Square**	**Chi-Square**	**Pr > ChiSq**
1	**Sc%**		53.6463		<.0001
2	**TO%**		17.1587		<.0001
3	**PenaltyYards**		3.9641		0.0465
4	**PassYds/TD**		2.7534		0.0970
5	**RushYds/TD**		4.2557		0.0391
6	**Plays**		4.3590		0.0368
7		**Sc%**		0.3633	0.5467

Table 4.6. Parameter Estimates and Odds Ratios for Season Model

		Standard			95%	
Parameter	**Estimate**	**Error**	**Pr > ChiSq**	**Odds Ratio**	**Confidence**	**Limits**
TO%	-0.3904	0.0956	<.0001	0.677	0.561	0.816
Plays	0.00993	0.00220	<.0001	1.010	1.006	1.014
PenaltyYards	0.00301	0.00152	0.0477	1.003	1.000	1.006
PassYds/TD	-0.0360	0.0103	0.0005	0.965	0.945	0.984
RushYds/TD	-0.0184	0.00572	0.0013	0.982	0.971	0.993

The Hosmer-Lemeshow test for goodness of fit provides a p-value of 0.4186, indicating that our model is a good fit for the data. The max rescaled R-square value was 0.6444, demonstrating the improvement of the model over the baseline intercept only model, this was the highest value of all other models considered. Figure 4.4 shows the ROC curve, and also an area of 0.9103 under the curve. All of these diagnostics signify that this is a good model for our seasonal data set.

There was no actual testing data set for this analysis, but it is interesting to use the model to check the playoff teams from the most recent NFL season, 2014-15. Inputting the seasonal values for the significant variables in our model for the thirty-two teams, we ranked the twelve teams with the highest probability of making the playoffs. The rest were classified as missing the playoffs. Our model correctly categorized twenty-six out of thirty-two teams for this past season, an 81.25% accuracy. Considering the different ways a team can make the playoffs, as mentioned earlier, this is a fairly good mark. A team with a good record and respectable statistics in one conference can miss the playoffs, while a worse team makes it in the other conference. For instance, in 2014, the Carolina Panthers made the playoffs with a record of 7-8-1, while the Philadelphia Eagles missed the playoffs with a record of 10-6.

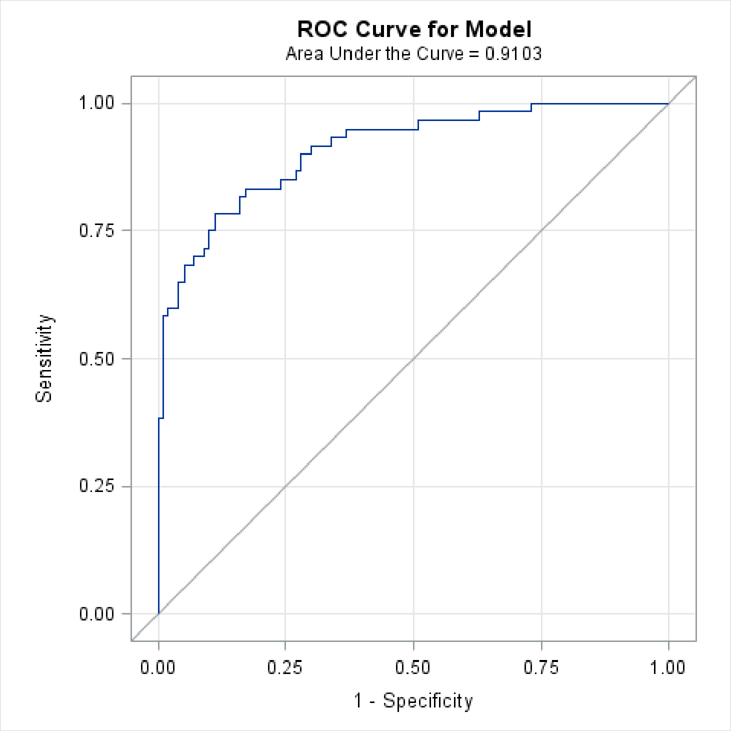


Figure 4.4. ROC Curve for Season Model

The most striking aspect of the season logistic model is the type of variables found to be significant. For the most part, statistics that represent efficiency were included, such as turnover percentage, and passing yards per passing touchdown. Turnover percentage is a measurement of the number of total turnovers committed by a team divided by the number of offensive drives they had throughout the season. For every one percent increase in turnover percentage, the odds of a team making the playoffs that year decrease by a factor of 0.677. Figure 4.5 illustrates the change in the probability of a team making the playoffs as turnover percentage increases while everything else is held constant. If a team can keep the turnover percentage below 10%, there is a good chance they will make the playoffs. On the other hand, few teams that have a turnover percentage over 15% will make the playoffs.

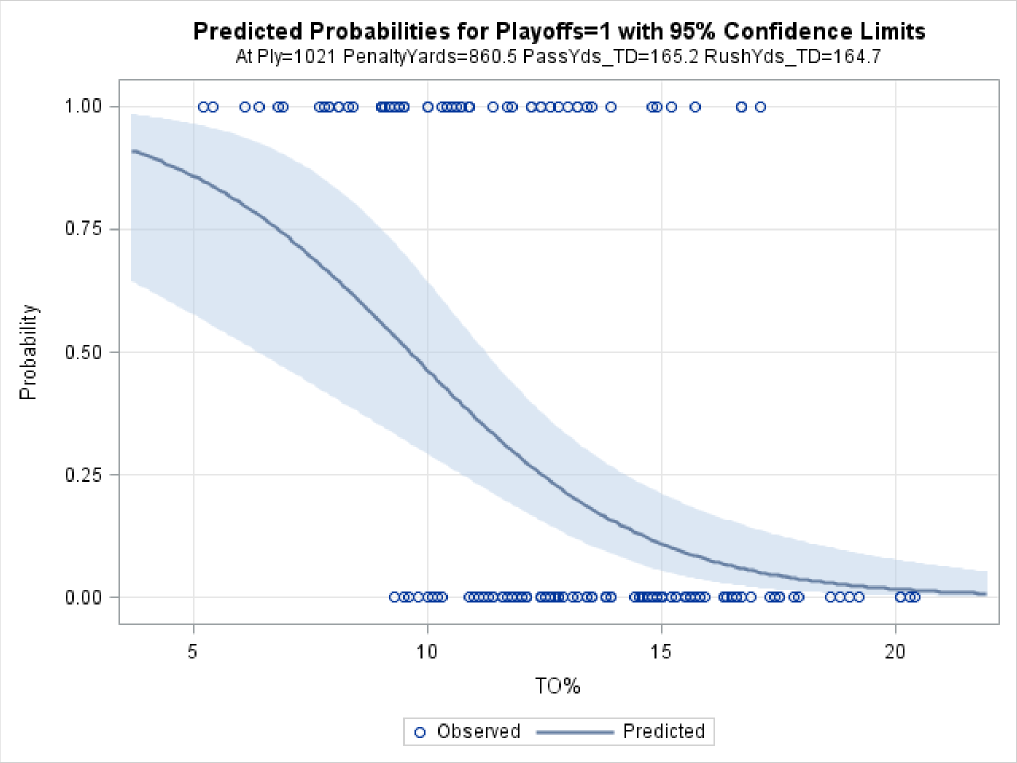


Figure 4.5. Effect of Turnover Percentage on Probability of Making Playoffs

Efficiency seems to play a vital role in long term team success over the course of a season. Once again, turnovers shows up as a crucial indicator, and the ability to quantify the effects they have is very beneficial. We will be comparing the results from our logistic regression model for NFL seasons with the analysis in the next section, using linear discriminant functions to classify playoff teams. Any agreement in significant variables found, is further evidence of our findings here.

4.3. Discriminant Analysis

4.3.1. Game Analysis

The discriminant analysis technique is one that is very useful, especially when relating and ranking the degree to which each variable is contributing to the end results. With the variables selected for the linear discriminant function, we will be able to compare the magnitudes of the standardized coefficients directly to each other to determine which has the largest effect. First, we will consider the game data, using the classes "Home Win" and "Home Loss" as our categories. Table 4.7 shows the stepwise selection process.

Table 4.7. Game Discriminant Analysis Stepwise Selection Summary

Step	Entered	Removed	Partial R-Square	F Value	Pr > F
1	**YPPassM**		0.2955	207.61	<.0001
2	**TurnoverM**		0.2127	133.47	<.0001
3	**3DPerM**		0.1361	77.64	<.0001
4	**SackYardsM**		0.0350	17.83	<.0001
5	**PenYardsM**		0.0336	17.08	<.0001
6	**YPRushM**		0.0133	6.58	0.0106
7	**FirstDownM**		0.0083	4.12	0.0430
8	**TotalPlayM**		0.0161	8.01	0.0048
9		**YPRushM**	0.0020	0.98	0.3220

All of the included variables are significant in the linear discriminant model at a level of at least α = 0.05, so we will proceed with these seven covariates. Table 4.8 shows the standardized linear discriminant functions for our two groups. In both cases, turnovers contribute the most towards an observation being classified into that group. As described in Section 3.3, the classification of each observation uses the discriminant function and inserts the value for each of the variables associated with that observation. The function that produces the largest value determines the group the data point is classified into. For every increase in the turnover margin on behalf of the home team, the discriminant function will penalize classification as a home win more than for any other category. The conversion percentage for 3rd down is another variable that has a large magnitude for both groups.

The sign of the coefficient is also informative, and makes natural sense in most cases. An increase in yards per pass margin is beneficial to the home team and increases the classification value for a home win, while at the same time, decreasing the classification value for a home loss. Total play margin is the only variable that seems to be counter intuitive, but as we will consider later on, this may be due to the efficiency of a team throughout the game.

Table 4.8. Standardized Linear Discriminant Functions for Game Model

Variable	**Home Loss**	**Home Win**
TurnoverM	0.79468	-0.82641
FirstDownM	-0.34551	0.78191
TotalPlayM	0.27263	-0.64415
3DPerM	-0.59911	0.52289
YPPassM	-0.45613	0.40363
SackedM	0.33322	-0.28138
PenaltyM	0.18074	-0.27553

With the linear discriminant functions, we can perform cross validation of the training data set as a way to assess the ability to correctly classify the two groups. Table 4.9 displays the results of the performance of the discriminant functions. With the hold out method, 195 out of 215 home losses were correctly classified, and 247 out of 282 home wins were correctly classified. That corresponds to an 89% accuracy rate for classifying a home win using these linear discriminant functions. So we can be confident in the variables that were chosen to be included.

When the discriminant functions are applied to the testing data set, 219 out of the 256 games were correctly grouped, an 85.5% accuracy. This is almost identical to the other models when the testing data set was considered. Therefore, the discriminant functions can be generalized to include any game played outside of the training data set, while still played under the same set of rules and conditions.

Table 4.9. Cross Validation Summary of Discriminant Functions for Game Model

Home Wins	***Classified Loss***	***Classified Win***	**Total**
Observed Loss	195	20	215
Observed Win	35	247	282
Total	230	267	497
Error Rate	0.0930	0.1241	**0.1086**

Comparing the variables in the discriminant functions for individual games to those found using the previous techniques of OLS and logistic regression, we see a lot of similarities. For the most part, the variables are identical, with the exception of the discriminant functions favoring the total number of penalties and sacks opposed to the yards lost to both. However, if we compare the values of the coefficients for each methods, some of the differences begin to appear. Turnovers has the largest coefficient in every case, but the next largest varies from model to model. Yards per pass margin and first down margin have large effects in the point spread and win probability models, but this is mainly due to the scales of measurements. There are only small deviations in the values of yards per pass margin, so an increase has a superficially larger effect in the score margin and probability of winning. But, we know from the standardized discriminant coefficients, that these do not have as much influence as they may seem to have.

One area that may be overlooked in the first two models is 3rd down conversion percentage margin. It may have a smaller coefficient in the point spread and logistic models, but it is clear there is a significant impact on whether or not a team will win the game, based on the ability to convert and prevent an opponent from converting 3rd downs opportunities.

4.3.2. Season Analysis

For the seasonal data, the discriminant analysis approach will classify the response similarly to the logistic regression methods. The two different groups considered are “Made Playoffs”, and “Missed Playoffs”. Table 4.10 summarizes the results of the variables chosen during the stepwise selection.

Table 4.10. Season Discriminant Analysis Stepwise Selection Summary

Step	Entered	Removed	Partial R-Square	F Value	Pr > F
1	**TO%**		0.2957	66.33	<.0001
2	**Yards**		0.1171	20.83	<.0001
3	**RushYds/TD**		0.0245	3.93	0.0493
4	**PassYds/TD**		0.0255	4.05	0.0459
5	**PenaltyYards**		0.0189	2.97	0.0871
6	**Yards/Play**		0.0092	1.42	0.2360
7		**Yards/Play**	0.0092	1.42	0.2360

These five variables are almost the exact same as the ones that made up the logistic model for this data set. Looking at the standardized discriminant functions, we can compare the coefficients and their effect sizes to determine if they also coincide to the previous model. Table 4.11 shows that indeed, the turnover percentage has the largest magnitude, and therefore contributes the most to the separation of the classes. For a team in the NFL, the best way to ensure a successful season and making the playoffs is to limit the number of giveaways. The other four variables, while significant in their own right, each contribute about the same, and not nearly as much as turnover percentage.

Table 4.11. Standardized Linear Discriminant Functions for Season Model

Variable	Miss Playoffs	Make Playoffs
TO%	0.50055	-0.83425
Yards	-0.22007	0.36678
RushYds/TD	0.21499	-0.35832
PassYds/TD	0.20977	-0.34961
PenaltyYards	-0.14281	0.23801

To validate these discriminant functions, we considered each of the four-team divisions in the NFL separately. First, we found the one team most likely to make the playoffs in each division as classified them as a playoff team. Next, the two wild card playoff teams from each conference were selected as the most likely to make the playoffs from the remaining teams in each conference. All remaining teams were classified as missing the playoffs.

The cross validation summary in Table 4.12 shows a slightly worse ability to categorize teams into the correct groups than the discriminant functions for individual games. Overall, 136 out of 160 team were correctly classified. The error rate is calculated with a slight adjustment in this case, since we know each season that twelve teams out of thirty-two total will make the playoffs, the prior probability of each class is known. Therefore we set these probabilities to 62.5% of the teams miss the playoffs, and 32.5% of the teams make them. Each individual error rate is multiplied by these proportions before the total rate of 15% is determined. For the most recent NFL season, we were able to classify twenty-six teams out of thirty-two correctly, exactly the same number when using the logistic model.

Table 4.12. Cross Validation Summary of Discriminant Functions for Game Model

Playoffs	***Classified Miss Playoffs***	***Classified Make Playoffs***	**Total**
Observed Miss Playoffs	88	12	100
Observed Make Playoffs	12	48	60
Total	100	60	160
Error Rate	0.1200	0.2000	**0.1500**
Prior Probability	0.6250	0.3750	

One interesting item of note is the positive effect total penalty yards has on season success. The logistic model in Section 4.2.2 also shows more penalties as a slight advantage. This could possibly be due to the fact that good teams take more chances on defense to try and increase their turnover margin. This aggressive type of play may result in more penalties and penalty yardage being called.

Overall, the variables selected show a consistent preference for measurements that quantify efficiency. This would seem to imply that simply racking up numbers during the season, such as passing yards and rushing yards, will have little impact if a team is not making use of those yards by scoring regularly. Also, most of the recorded season statistics are related to offensive values, with no indication of how preventing an opponent from scoring, passing, and so on, impacts success in the NFL. In the next section we will address this issue and observe the difference between offensive and defensive variables for winning an individual game.

4.3.3. Offensive and Defensive Comparison for Games

The final type of discriminant analysis we will perform is to compare offensive and defensive abilities to determine if either is consistently contributing more to winning individual games. Instead of the marginal data used in earlier models, we will consider just the individual team totals for both the home and away team. The two classes will now simply be, "Win" and "Lose". However, now we will have a designation of offensive amount gained, or defensive amount given up for each covariate. Table 4.13 gives the stepwise selection summary.

The first six variables selected to enter the model represent three different types of in game statistics; turnovers, yards per pass, and 3rd down conversion percentage. This is consistent with the variables that we have found significant in our other models. The next group of covariates entered into the model are different types of measurements for first downs, sacks, and penalties. Again, this is consistent with our findings thus far using alternative methods.

Table 4.13. Offensive vs. Defensive Discriminant Analysis Stepwise Selection Summary

Step	Entered	Removed	Partial R-Square	F Value	Pr > F
1	**YPPassGain**		0.2162	139.86	<.0001
2	**TurnoverForce**		0.1838	113.95	<.0001
3	**Per3DAllow**		0.0950	53.03	<.0001
4	**TurnoverCommit**		0.1003	56.17	<.0001
5	**Per3D**		0.0706	38.24	<.0001
6	**YPPassAllow**		0.0406	21.27	<.0001
7	**FDGain**		0.0289	14.91	0.0001
8	**TotalPlays**		0.0401	20.89	<.0001
9	**SackGain**		0.0253	12.93	0.0004
10	**SackYdsLost**		0.0209	10.65	0.0012
11	**Penalty**		0.0153	7.71	0.0057
12	**OppPenaltyYard**		0.0073	3.63	0.0573
13	**YPRushAllow**		0.0087	4.33	0.0379
14	**TotalYardAllow**		0.0087	4.33	0.0379
15	**FDAllow**		0.0040	1.99	0.1593
16	**TotalYard**		0.0027	1.33	0.2490
17		**TotalYard**	0.0027	1.33	0.2490

For our analysis, we would like to reduce the number of variables included, so we will only use the offensive and defensive pairs for turnovers, yards per pass, 3rd down conversion percentage, sack yards, and first downs. Penalties will be omitted since they don't reveal much information about offensive and defensive abilities. Table 4.14 displays the linear discriminant functions for the two classes.

Table 4.14. Standardized Linear Discriminant Functions for Offense vs. Defense Model

Variable	**Lose**	**Win**
TurnoverCommit	0.51450	-0.50449
TurnoverForce	-0.72730	0.71315
YPPassGain	-0.47315	0.46394
YPPassAllow	0.31672	-0.31056
Per3D	-0.34140	0.33476
Per3DAllow	0.46584	-0.45678
SackYdsLost	0.25987	-0.25481
SackYdsGain	-0.22023	0.21594
FDGain	-0.36894	0.36176
FDAllow	-0.03998	0.03920

To analyze these functions, we will need to compare the sets of variables. For instance, committing a turnover is a measurement of offense, and for every turnover committed, a team will move away from a "Win" classification by about 0.504 standardized units. However, if a team forces a turnover, which is a defensive measurement, they will move towards a Win classification by 0.713 standardized units. Clearly, for turnovers, it is better to force them than commit them, but all things being equal, it would be more beneficial to have one more turnover forced than one less turnover committed. Defense has the clear advantage for turnovers.

For the yards per pass variable, we see the opposite effect, the coefficient for the "Win" class has a magnitude of 0.464 for the yards per pass gained, while the yards per pass against measure has a magnitude of 0.311. When considering yards per pass, it is advantageous to be more efficient on offense than it is to prevent your opponent from having a lot of yards per pass. Similarly, for first downs, it seems as though offense is more important, making sure your team is moving down the field.

The remaining variables all favor the defensive side. A team should focus more on stopping their opponent on 3rd downs than converting their own, if they want to maximize their probability of winning the game. Sacks are also more significant from the defensive side, giving up sack yardage is not as critical as is sacking the opposing quarterback. To be clear, all of these variables will affect the game significantly, but the discriminant functions let us see that those effects are not necessarily equal.

The discriminant functions were able to correctly classify 88.2% of the observations using the cross validation method. This shows that the analysis performs well in relation to describing the data, but our primary concern in this section was the idea of offense versus defense. Of the variables that we are consistently finding significant in our models, it appears that there are different ways to maximize the benefits. There is no clear advantage given to the broad idea of offense or defense, instead it should be considered on a case by case basis. For areas such as turnovers, and 3rd down percentage, a team would be better off with a defensive mindset. As for yards per pass, first downs, and preventing sacks, an offensive strategy will help a team take full advantage of their chance to win the game.

4.4. Proportional Odds Model

The final model that we will create is an extension of the logistic regression model. Using the individual game data set and the marginal variables once again, we can take advantage of the fact that score margin can be separated into different ordered groups. We will create four categories that indicate the winner of the game, along with a classification of a close scoring game or a blowout, Section 3.4 provides the details of how the new ordinal response variable was created. 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 we have a model, we can calculate the probability that a given observation will fall into each of the four categories. Once those probabilities are ranked from most likely to least likely, the most likely group will be the fitted value and compared to the actual score margin observed at the end of the game. The winner of the game and the score margin category can then be checked against the actual outcome of the games. Table 4.15 gives a summary of the proportional odds model after a stepwise selection procedure, and the same variables are showing up again in the proportional odds model, along with a few others that we did not see earlier.

Table 4.15. Parameter Estimates and Odds Ratios for Proportional Odds Model

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T VALUE	PR > |T|	ODDS RATIOS
INT-SA|WA	-3.6261	0.2529	-14.339	<.0001	-
INT-WA|WH	-0.2148	0.1538	-1.396	0.0813	-
INT-WH|SH	3.4949	0.2399	14.5689	<.0001	-
FIRSTDOWNM	0.27753	0.038963	7.123	<.0001	1.320
SACKYARDSM	-0.03809	0.007115	-5.353	<.0001	0.963
TURNOVERM	-1.22858	0.089493	-13.728	<.0001	0.293
3DPERM	0.06562	0.007310	8.977	<.0001	1.068
RUSHM	0.01497	0.002238	6.689	<.0001	1.015
PASSINGM	0.01073	0.002028	5.294	<.0001	1.011
PENYARDSM	-0.01301	0.003692	-3.525	0.0004	0.987
TOTALPLAYM	-0.17052	0.017664	-9.654	<.0001	0.843

As described earlier, the proportional odds model actually creates three separate log odds models for our data, one for each of the first three classes. The variable coefficients are the same in each case and only the intercept changes, marking the boundary of the classes. If we plug in the values from the summary table to Equations 6, 7, and 8 on pg. 24 we obtain the cumulative probability that an observation will fall into each category or lower. Here are the equations with the current values:

$$\theta_{sa} = log\,\frac{P_{sa}}{1-P_{sa}} = -3.62 + \boldsymbol{X}\beta \qquad \text{(Eq. 9)}$$

$$\theta_{wa} = log\,\frac{P_{wa}}{1-P_{wa}} = -0.2148 + \boldsymbol{X}\beta \qquad \text{(Eq. 10)}$$

$$\theta_{wh} = log\,\frac{P_{wh}}{1-P_{wh}} = 3.4949 + \boldsymbol{X}\beta \qquad \text{(Eq.11)}$$

where, $\boldsymbol{X}\beta$ = 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).

Now, we can solve for these cumulative probabilities to find the probability that each observation will fall into each individual category. The general form of these probabilities is given here:

$$p_{sa} = \frac{e^{\theta_{sa}}}{1+e^{\theta_{sa}}} \qquad \text{(Eq. 12)}$$

$$p_{wa} = \frac{e^{\theta_{wa}}}{1+e^{\theta_{wa}}} - p_{sa} \qquad \text{(Eq. 13)}$$

$$p_{wh} = \frac{e^{\theta_{wh}}}{1+e^{\theta_{wh}}} - (p_{wa} + p_{sa}) \qquad \text{(Eq. 14)}$$

$$p_{sh} = 1 - (p_{sa} + p_{wa} + p_{wh}) \qquad \text{(Eq. 15)}$$

When the model was used to classify the categories in the training data set, it 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 the testing data set with the proportional odds model, we found that 68.8% of the games were correctly categorized, and 86.3% were classified correctly as a home win or home loss. So, our models are valid for games outside of the training data set. Considering the variables that were included with this method, again we can see a lot of similarity to all of the previous models. Some extra variables are included, such as total passing yard margin, and for the first time we have a rushing statistic found significant.

Comparing the effect of each variable, 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 this model is to consider hypothetical scenarios during 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 doesn't tell us anything more about what else could possibly occur. Figure 4.6 illustrates the expected probabilities when both teams have the exact same performances in the significant variables. The probability that the away team will win by more than ten points is 2.59%, and for the home team to win by less than ten points is a probability of 52.4%. Overall, with everything equal, the home team has a 55.3%

probability 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.

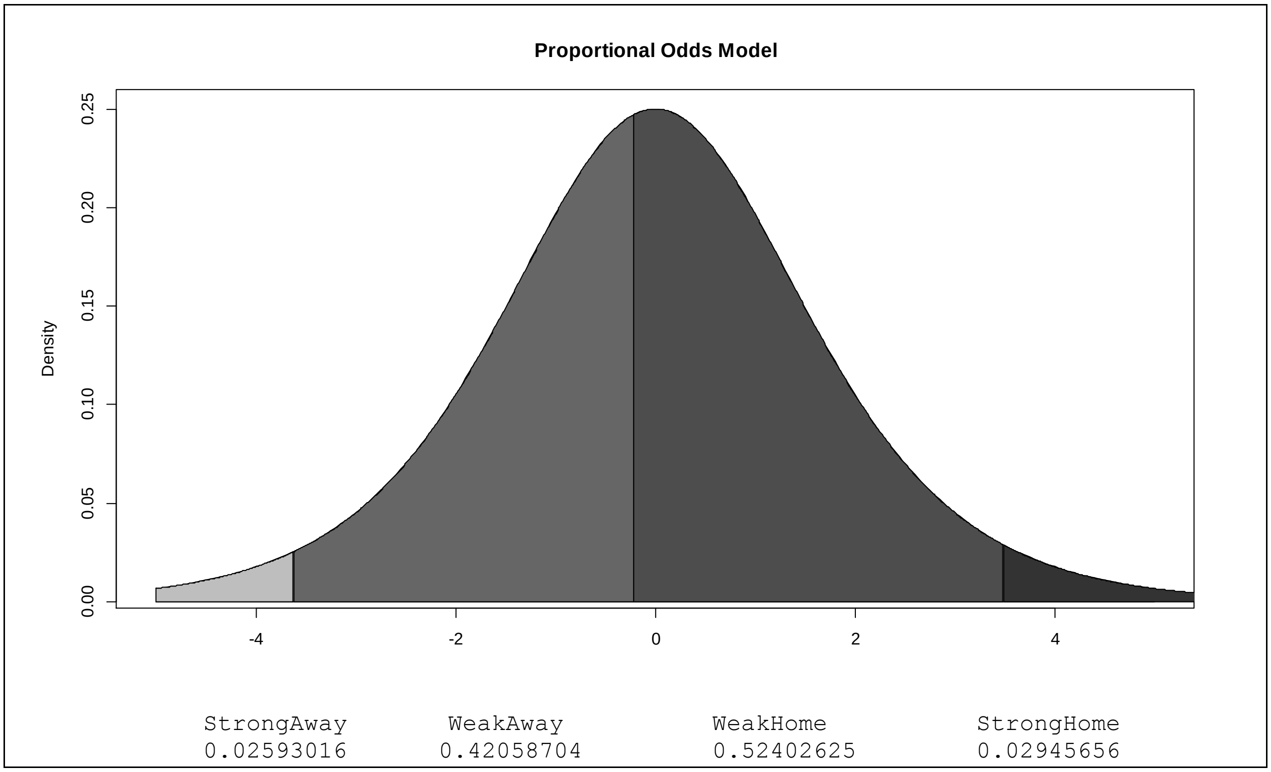
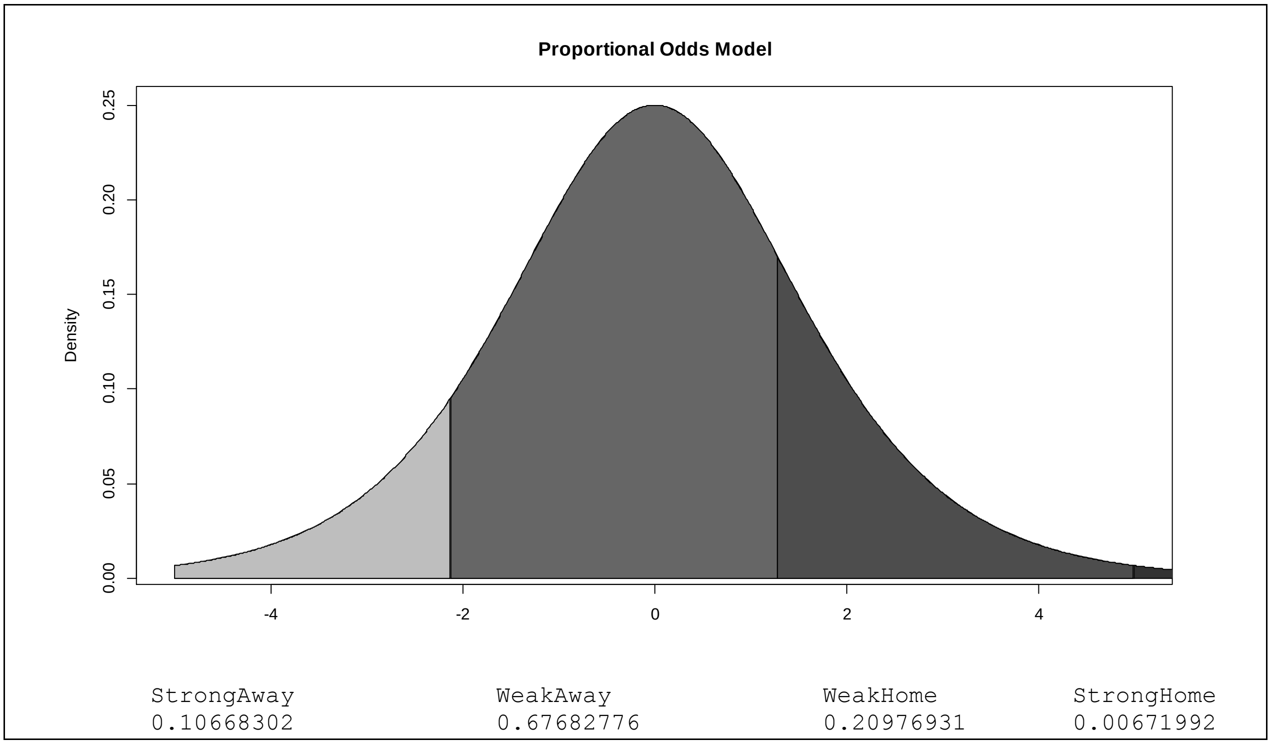


Figure 4.6. Proportional Odds Probabilities for No Advantage

With this model we can give the theoretical winning probability and score margin for any combination of marginal variable values. Another example is illustrated in Figure 4.7, a moderate advantage for the away team. 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 probability of winning is around 78.2%, with a 10.7% chance of the game being a blowout by the away team.



```
StrongAway          WeakAway            WeakHome            StrongHome
0.10668302          0.67682776          0.20976931          0.00671992
```

Figure 4.7. Proportional Odds Probabilities for Away Team Advantage

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. This leads us to our next section, where we will use simulation to forecast games.

4.5. Simulation

Up until now, all of the model validation that we have performed has been comprised of considering games outside of our testing data set and using the statistics collected from those games to fit the final score margin or classify the winner. This provides an extra level of confidence that the models we have chosen can work outside of the observations used to create them. While we have presented some quantitative analysis of the effects of those areas, such as turnovers and passing efficiency, the casual fan may respond with remarks which point to the fact that it is common knowledge that anytime a team commits three more turnovers than their opponent, they will find it difficult to win the game. What they would really like to know is, what is the probability a team will actually commit those turnovers in their next game? Is there any way to forecast the outcome of games without using the statistics from that contest?

This leads us to our final model evaluations, using simulations from historical performances. In our testing data set, all games after Week 4 were selected, 193 in total. For each of these games, the two teams involved had the means and standard deviations calculated for each of the variables in the models based on all of the games they played leading up to the game of interest. Then, from these statistics, 10,000 game simulations were created based on the marginal statistics of offensive performance, and also a combination of offense and defense which was introduced in Section 3.5. The results for each model were compared to those of the actual game results.

One example of simulations used to forecast a game outcome is presented here. On November 3rd 2013, during Week 9 of the NFL season, the Kansas City Chiefs played at the Buffalo Bills. We will illustrate using the point spread regression model on the simulated data. Table 4.16 shows the average values and standard deviations for all of the variables used, including the average values allowed by each team over the previous eight weeks. These are the numbers that were used to simulate 10,000 games being played, and the marginal variables were calculated for each one. Next, the marginal values are entered into the model and the average point spread for all the simulations is calculated. For this case, the average point spread was -3.7 using only the offensive numbers for each team, and -6.1 using the combination of both offensive and defensive statistics. This represents that we would expect the away team, the Chiefs, to win the game by somewhere around four to six points. The actual outcome of the game saw the Chiefs win by a margin of ten points. So we were correct in selecting the winner using both marginal statistics, and it appears that using the combination of offense and defense provided a fitted point spread that was a little closer to the actual point spread.

Table 4.16. Averages and Standard Deviations Through Week 8

Variable	Chiefs	Chiefs - Allowed	Bills	Bills - Allowed
First Downs	Mean = 19.0	16.0	18.88	21.38
	SD = 1.69	3.93	2.85	3.96
Total Plays	Mean = 67.75	62.25	70.75	71.88
	SD = 5.036	6.84	5.83	10.08
Yards per Pass	Mean = 5.78	5.9	5.8	7.10
	SD = 1.24	1.94	0.89	2.33
Turnovers	Mean = 1.0	2.5	1.63	1.88
	SD = 1.19	1.31	1.06	1.73
3rd Down Percentage	Mean = 36.02	25.54	35.97	37.88
	SD = 15.47	7.87	9.35	11.99
Yards Lost to Sack	Mean = 15.63	31.75	19.5	21.63
	SD = 11.38	20.38	10.60	12.74

Figure 4.8 shows the histogram of the point spreads for all of the simulated games, using the offensive and defensive marginal statistics. The overall percentage of simulated games won by the Chiefs was 76%, and while the average point spread value was -6.1, clearly there were many simulated games that resulted in a point spread close to -10, that of the actual game.

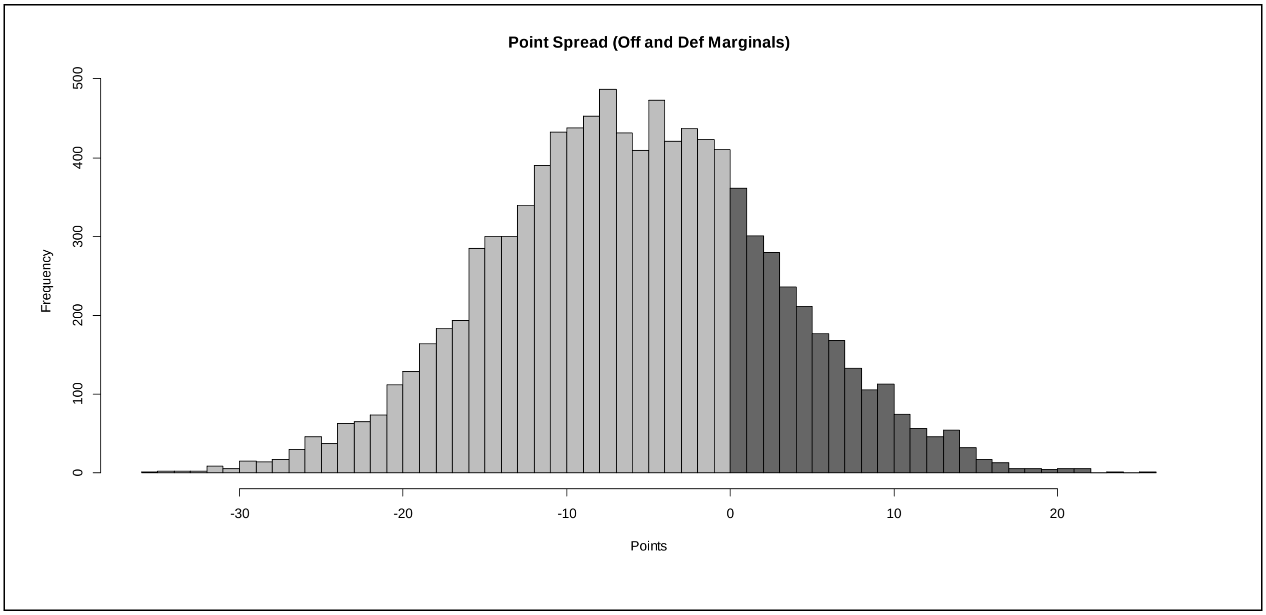


Figure 4.8. Histogram of Simulated Point Spreads

This process was repeated for 193 games, and for each of the individual game models. Table 4.17 summarizes the results and accuracy of each model when forecasting the outcome of games using historical team performance. Included in the table are the percentages of each model correctly choosing the winner based on the actual in game statistics observed, which was our previous validation method.

For projecting games without any knowledge from the event itself, we start with the naïve method of choosing a winner of the game by selecting the home team to win. With this technique, 57% of the games were correctly selected. For predicting the point spread, using a combination of both offensive and defensive statistics for each team resulted in finding the actual winner 67% of the time. The mean absolute deviation of the fitted point spread to the observed point spread was just under ten points, meaning that is how close our fitted value was to the observed value on average.

Table 4.17. Summary of Model Forecasting Accuracy

Model	**Using In Game Statistics**	**Offense Only Forecast**	**Offense and Defense Forecast**
Naïve (Home Team Wins)	57%	57%	57%
Point Spread	85.9%	65%	67%
Logistic	85.9%	63%	66%
Discriminant Functions	85.5%	63%	67%
Proportional Odds (Correct Category)	68.8%	37%	35%
Proportional Odds (Correct Winner)	86.3%	64%	**71%**

For the models we developed that only try to classify the game as a Win or a Loss, the best performing models were the proportional odds models. While they were only around 35% in predicting the actual category the final point spread fell into, the correct winner was selected for 71% of the games. Overall, for each of the models, it seems that simulating games using both offensive and defensive statistics provides a more accurate forecast. This would seem to make intuitive sense, since when we consider offensive values, we are only looking at half of the picture. Simulation seemed to create a lot of conservative average point spreads, with few games being classified as blowouts. But those models were just as successful as the others when we considered only if they produced the correct winner.

Overall, every one of the models we developed had an accuracy of around 63% or higher, much better than just choosing the home team to win. The best models we can provide are actually able to beat the performance of the others presented in Chapter 2, although by just a slight margin. This is further evidence that those variables we have identified as significant in explaining success in the NFL are indeed important. Also, this illustrates the capabilities that simulation has when looking forward to games that have not yet occurred.

CHAPTER 5. CONCLUSIONS

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, and over the long term, success can be considered as making the playoffs at the end of a season. In both cases, our analyses produced similar results. Turnovers, passing efficiency, first downs, and 3rd down conversion percentage consistently showed up in each of the models that we formulated. There are certainly other areas that can lead to success in the NFL, but these seem to be the ones that offer the most influence.

Indicating the parts of football that point to winning is not the final objective though. We wanted to quantify those effects. Now, a coach, player, or fan does not have to simply say that more first downs for their team is beneficial, they can point to an empirical model and say that their team can expect almost three extra points for every two more first downs they acquire. Or they can say that their odds of winning the game will decrease by a factor of 0.27 when they have one more turnover than their opponent. These are quantitative measurements that teams and organizations can use to make decisions on strategy, player personnel, or staff. As a team owner, hiring a coach that has a reputation for calling plays that increase passing efficiency may be a better choice than one who is more prone to calling running plays. The former maximizes the chance of winning, while the latter is focused in an area that is not as crucial.

Most of the results we have presented come down to being efficient in football. At the season level, the more drives you score on, the more likely you are to make the playoffs. That means you are not spending a lot of your offensive possessions moving the ball a little bit, only to punt it away and give the ball back to your opponent. Turnovers are another measure of efficiency, the most inefficient play you can have is to immediately forfeit possession of the ball to the opponent and end any chance of scoring points.

One other variable that was consistently included in our models was total plays. The interesting aspect of this statistic is that for all of the models where it was included, it had a negative effect as it increased. For the point spread model, every two extra plays over your opponent resulted in a loss of about one point in the score margin. One possible explanation for this is offensive efficiency. Those teams that are winning and scoring an abundance of points are doing so with less total plays. An inefficient team will use more plays to travel the same distance to score points than one which is efficient.

It is also important to consider the amount that each of these significant variables contributes. The question of offense versus defense, or the old axiom that defense wins championships, is absolutely something that should be confirmed or discredited. The truth is, there are no universal rules that say either offense or defense is better than the other. It needs to be examined on a statistic by statistic basis, and once the value of each is quantified, it can be exploited to benefit of the teams that are willing to do so.

When it comes to turnovers, it is better for your team to create them, than try to prevent committing them. Either way, both will affect the outcome of the game, but you can expect a larger return for creating a turnover then preventing one. This can lead to different strategies, such as being more aggressive on defense by trying to intercept more passes, or trying to cause more fumbles.

On the other hand, when you consider passing productivity, it is more important to have an offense that produces more yards per pass than it is to have a defense that prevents a larger yard per pass value. So this, along with the fact that offensive turnovers do not hurt as much, would suggest an approach that is more aggressive in passing offense. An example would be calling more plays that will result in greater passing yardage, even if the risk of an interception is slightly increased.

With all of this information available to each team and the public, it is important that NFL teams use it effectively. There is still some resistance throughout the league to use analytics for improved decision making off the field, and enhanced performance on it. Hopefully, the continued analysis and application of statistics in sports will convince those who are in the position of making decisions that soon they will be at a disadvantage if they fail to make the most of these underlying tendencies in the game of football.

REFERENCES CITED

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

Agresti, A. [2002]. *Categorical Data Analysis* (2nd ed.) New York: Wiley.

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

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

Burke, M. [2013]. "How the National Football League can reach $25 billion in annual revenues". Forbes: SportsMoney, www.forbes.com. August 17.

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.

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.

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.

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.

Hansen, B.E. [2007]. "Notes and Comments: Least Squares Model Averaging". *Econometrica*. Vol. 75, No. 4, pp. 1175-1189.

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.

Hosmer, D.W. & Lemeshow, S. [2000]. *Applied Logistic Regression* (2nd ed.). New York: John Wiley & Sons, Inc.

Liao, D. & Valliant, R. [2012]. "Variance Inflation Factors in the Analysis of Complex Survey Data". *Survey Methodology*. Vol. 38, No. 1, pp. 53-62.

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.

Luker, R. [2011]. "Fan base complexity sets up intense competition for attention". Sports Business Journal. www.sportsbusinessdaily.com. June 27.

NFL Scores. [2009-2013]. Retrieved July 22, 2014, from ESPN.com: http://scores.espn.go.com/nfl/scoreboard.

Rencher, A.C. [2002]. *Methods of Multivariate Analysis* (2nd ed.) New York: Wiley.

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.

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.

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

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

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.

APPENDIX A. LIST OF COLLECTED GAME VARIABLES

- Score
- Win Percentage in previous games
- Total First Downs ^{T1}
- Passing First Downs ^{P1}
- Rushing First Downs ^{P1}
- First Downs From Penalties ^{P1}
- Total Yards ^{T2}
- Total Plays ^{P2}
- Yards Per Play ^{P2}
- Total Passing Yards ^{T3}
- Passing Attempts ^{P3}
- Yards Per Pass Attempts ^{P3}
- Total Rushing Yards ^{T4}
- Rushing Attempts ^{P4}
- Yards Per Rush ^{P4}
- Number of Penalties
- Total Penalty Yards
- Turnovers ^{T5}
- Fumbles ^{P5}
- Interceptions ^{P5}
- 3rd Down Percentage ^{T6}
- 3rd Down Conversions ^{P6}
- 3rd Down Attempts ^{P6}
- 4th Down Percentage ^{T7}
- 4th Down Conversions ^{P7}
- 4th Down Attempts ^{P7}
- Defensive Touchdowns
- Time of Possession
- Times Sacked
- Yards lost to Sacks
- Red Zone Percentage ^{T8}
- Red Zone Scores ^{P8}
- Red Zone Attempts ^{P8}
- Average Kick Return
- Average Punt Return

^{T} Indicates a variable that measures the team total in a game

^{P} Indicates the partial variables combined to calculate the matching total variable

The corresponding totals and parts are never jointly considered as initial variables for a model, but in separate models containing all other total or partial measurements.

APPENDIX B. LIST OF COLLECTED SEASON VARIABLES

- Total Points Scored
- Total Yards ^{T1}
- Total Plays ^{P1}
- Yards Per Play ^{P1}
- Turnovers ^{T2}
- Fumbles Lost ^{P2}
- Interceptions ^{P2}
- Total First Downs ^{T3}
- Passing First Downs ^{P3}
- Rushing First Downs ^{P3}
- First Downs From Penalties ^{P3}
- Passing Yards ^{T4}
- Passing Attempts ^{P4}
- Yards Per Pass Attempt ^{P4}
- Completions
- Passing Touchdowns
- Rushing Yards ^{T5}
- Rushing Attempts ^{P5}
- Yards Per Rush ^{P5}
- Rushing Touchdowns
- Scoring Percentage
- Turnover Percentage
- Passing Yards Per Passing Touchdown
- Rushing Yards Per Rushing Touchdown
- Penalty Yards ^{T6}
- Number of Penalties ^{P6}
- Average Yards Per Penalty ^{P6}

^{T} Indicates a variable that measures the team total for a season

^{P} Indicates the partial variables combined to calculate the matching total variable

The corresponding totals and parts are never jointly considered as initial variables for a model, but in separate models containing all other total or partial measurements.

APPENDIX C. SAS CODE

```
*Collect raw data from ESPN boxscores;
%macro scrape(season, week, site);

filename football url "http://scores.espn.go.com/nfl/boxscore?gameId=&site";

* first pass, keep everything;
data source2;
      format nfl $2000.;
      infile football lrecl=3276700 delimiter=">";
      input nfl $ @@;
run;

data win;
      set source2;
      if index(nfl,'</p')then temp1 = scan(nfl,1,'<');
      if temp1^='';
      Win=scan(temp1,1,'-');
      Win=compress(Win,'(');
      Loss=scan(temp1,2,'-');
      Loss=scan(Loss,1,',');
      Tie=scan(temp1,3,'-');
      Tie=scan(Tie,1,',');
      drop temp1;
run;

data formatted1;
      set source2;
      if index(nfl,'</td') then temp = scan(nfl,1,'<');
      if index(nfl,'</tr') then temp = scan(nfl,1,'<');
      if index(nfl, '</a') then temp = scan(nfl,1,'<');
      if index(nfl, '</th') then temp = scan(nfl,1,'<');
      if temp = "/td" then delete;
      if temp = "/th" then delete;
      if temp = "/tr" then delete;
      if temp = "/a" then delete;
      if temp ^= '';
      if temp = "Total Drives" then miss=1000;
            miss+1;
      if 1000<miss<1004 then delete;
      if temp = "1st Downs" then ctr=1000;
      if temp = "Possession" then ctr=2000;
            ctr+1;
      drop nfl;
      if 0<ctr<18 then delete;
      if 41<ctr<1000 then delete;
run;

data formatted;
      set formatted1;
      find=find(temp, 'Returns');
      if find>0 then count=3000;
      bad=find(temp, 'Kicking');
      if bad>0 then count=3500;
      count+1;
      if count>=3500 then delete;
      if count<3000 then delete;
run;

data formatted3;
      set formatted;
      if temp="Team" then counter=4000;
            counter+1;
      if 4000<counter<4007 then temp2=temp;
      if temp2 ^='';
            obs+1;
      drop find bad counter;
run;

data reset;
  merge formatted1 formatted1(firstobs=1 keep=temp rename=(temp=Date));
  merge formatted1 formatted1(firstobs=2 keep=temp rename=(temp=Away));
  merge formatted1 formatted1(firstobs=3 keep=temp rename=(temp=Home));
  merge win win(firstobs=3 keep=Win rename=(Win=AwayWins));
  merge win win(firstobs=4 keep=Win rename=(Win=HomeWins));
  merge win win(firstobs=3 keep=Loss rename=(Loss=AwayLosses));
  merge win win(firstobs=4 keep=Loss rename=(Loss=HomeLosses));
  merge win win(firstobs=3 keep=Tie rename=(Tie=AwayTie));
  merge win win(firstobs=4 keep=Tie rename=(Tie=HomeTie));
  merge formatted1 formatted1(firstobs=8 keep=temp rename=(temp=OT));
  merge formatted1 formatted1(firstobs=15 keep=temp rename=(temp=AwayScr));
  merge formatted1 formatted1(firstobs=17 keep=temp rename=(temp=AwayScrOT));
  merge formatted1 formatted1(firstobs=24 keep=temp rename=(temp=HomeScrOT));
  merge formatted1 formatted1(firstobs=21 keep=temp rename=(temp=HomeScr));
  merge formatted1 formatted1(firstobs=26 keep=temp rename=(temp=AwayFD));
  merge formatted1 formatted1(firstobs=27 keep=temp rename=(temp=HomeFD));
  merge formatted1 formatted1(firstobs=28 keep=temp rename=(temp=AwFDPass));
  merge formatted1 formatted1(firstobs=29 keep=temp rename=(temp=HoFDPass));
  merge formatted1 formatted1(firstobs=30 keep=temp rename=(temp=AwFDRun));
  merge formatted1 formatted1(firstobs=31 keep=temp rename=(temp=HoFDRun));
  merge formatted1 formatted1(firstobs=32 keep=temp rename=(temp=AwFDPen));
  merge formatted1 formatted1(firstobs=33 keep=temp rename=(temp=HoFDPen));
  merge formatted1 formatted1(firstobs=34 keep=temp rename=(temp=Aw3Dwn));
  merge formatted1 formatted1(firstobs=35 keep=temp rename=(temp=Ho3Dwn));
  merge formatted1 formatted1(firstobs=36 keep=temp rename=(temp=Aw4Dwn));
  merge formatted1 formatted1(firstobs=37 keep=temp rename=(temp=Ho4Dwn));
  merge formatted1 formatted1(firstobs=39 keep=temp
                             rename=(temp=AwTotalPlay));
  merge formatted1 formatted1(firstobs=40 keep=temp
                             rename=(temp=HoTotalPlay));
  merge formatted1 formatted1(firstobs=42 keep=temp
                             rename=(temp=AwTotalYard));
  merge formatted1 formatted1(firstobs=43 keep=temp
                             rename=(temp=HoTotalYard));
  merge formatted1 formatted1(firstobs=45 keep=temp rename=(temp=AwYPPlay));
  merge formatted1 formatted1(firstobs=46 keep=temp rename=(temp=HoYPPlay));
  merge formatted1 formatted1(firstobs=48 keep=temp rename=(temp=AwPassing));
  merge formatted1 formatted1(firstobs=49 keep=temp rename=(temp=HoPassing));
  merge formatted1 formatted1(firstobs=52 keep=temp rename=(temp=AwYPPass));
  merge formatted1 formatted1(firstobs=53 keep=temp rename=(temp=HoYPPass));
  merge formatted1 formatted1(firstobs=56 keep=temp rename=(temp=AwSack));
  merge formatted1 formatted1(firstobs=57 keep=temp rename=(temp=HoSack));
  merge formatted1 formatted1(firstobs=59 keep=temp rename=(temp=AwRush));
  merge formatted1 formatted1(firstobs=60 keep=temp rename=(temp=HoRush));
  merge formatted1 formatted1(firstobs=61 keep=temp
                              rename=(temp=AwRushAtmpt));
  merge formatted1 formatted1(firstobs=62 keep=temp
                              rename=(temp=HoRushAtmpt));
  merge formatted1 formatted1(firstobs=63 keep=temp rename=(temp=AwYPRush));
  merge formatted1 formatted1(firstobs=64 keep=temp rename=(temp=HoYPRush));
  merge formatted1 formatted1(firstobs=66 keep=temp rename=(temp=AwRedZone));
  merge formatted1 formatted1(firstobs=67 keep=temp rename=(temp=HoRedZone));
  merge formatted1 formatted1(firstobs=69 keep=temp rename=(temp=AwayPen));
  merge formatted1 formatted1(firstobs=70 keep=temp rename=(temp=HomePen));
  merge formatted1 formatted1(firstobs=72 keep=temp rename=(temp=AwayTurn));
  merge formatted1 formatted1(firstobs=73 keep=temp rename=(temp=HomeTurn));
  merge formatted1 formatted1(firstobs=74 keep=temp rename=(temp=AwFumble));
  merge formatted1 formatted1(firstobs=75 keep=temp rename=(temp=HoFumble));
  merge formatted1 formatted1(firstobs=76 keep=temp rename=(temp=AwInt));
  merge formatted1 formatted1(firstobs=77 keep=temp rename=(temp=HoInt));
  merge formatted1 formatted1(firstobs=79 keep=temp rename=(temp=AwDefTD));
  merge formatted1 formatted1(firstobs=80 keep=temp rename=(temp=HoDefTD));
  merge formatted1 formatted1(firstobs=82 keep=temp rename=(temp=AwPoss));
  merge formatted1 formatted1(firstobs=83 keep=temp rename=(temp=HoPoss));
  merge formatted3 formatted3(firstobs=4 keep=temp2 rename=(temp2=AwAveKR));
  merge formatted3 formatted3(firstobs=10 keep=temp2 rename=(temp2=HoAveKR));
  merge formatted3 formatted3(firstobs=16 keep=temp2 rename=(temp2=AwAvePR));
  merge formatted3 formatted3(firstobs=22 keep=temp2 rename=(temp2=HoAvePR));
  merge formatted3 formatted3(firstobs=1 keep=obs);
run;

data first;
      set reset;
      if obs>1 then delete;
      drop temp ctr obs temp2 count Win Loss Tie nfl;
run;

data test1;
      set first;
      if (OT = "OT") then AwayScr=AwayScrOT;
      if (OT = "OT") then HomeScr=HomeScrOT;
      AwayWins=input(AwayWins, 2.);
      AwayLosses=input(AwayLosses, 2.);
      HomeWins=input(HomeWins, 2.);
      HomeLosses=input(HomeLosses, 2.);
      drop AwayScrOT HomeScrOT;
run;

data test2;
      set test1;
      AwPen=scan(AwayPen,1, '-');
      AwPenYard=scan(AwayPen,2,'-');
      HoPen=scan(HomePen,1, '-');
      HoPenYard=scan(HomePen,2,'-');
      AwSacked=scan(AwSack,1,'-');
      AwSackYds=scan(AwSack,2,'-');
      HoSacked=scan(HoSack,1,'-');
      HoSackYds=scan(HoSack,2,'-');
      AwRZScr=scan(AwRedZone,1,'-');
      AwRZAtt=scan(AwRedZone,2,'-');
      HoRZScr=scan(HoRedZone,1,'-');
      HoRZAtt=scan(HoRedZone,2,'-');
      AwRZPer=round(AwRZScr/AwRZAtt,0.0001);
      HoRZPer=round(HORZScr/HoRZAtt,0.0001);
      Aw3DCon=scan(Aw3Dwn,1,'-');
      Aw3DAtt=scan(Aw3Dwn,2,'-');
      Aw4DCon=scan(Aw4Dwn,1,'-');
      Aw4DAtt=scan(Aw4Dwn,2,'-');
      Ho3DCon=scan(Ho3Dwn,1,'-');
      Ho3DAtt=scan(Ho3Dwn,2,'-');
      Ho4DCon=scan(Ho4Dwn,1,'-');
      Ho4DAtt=scan(Ho4Dwn,2,'-');
      Aw3DPer=round(Aw3DCon/Aw3DAtt,0.0001);
      Aw4DPer=round(Aw4DCon/Aw4DAtt,0.0001);
      Ho3DPer=round(Ho3DCon/Ho3DAtt,0.0001);
      Ho4DPer=round(Ho4DCon/Ho4DAtt,0.0001);
      Date=substr(Date,12,12);
      Date=compress(Date, ',');
      AwayScr=input(AwayScr, 2.);
      HomeScr=input(HomeScr, 2.);
      if AwayTie^=1 then AwayTie=0;
      if HomeTie^=1 then HomeTie=0;
      if AwayScr>HomeScr then AwayWins=AwayWins-1;
      if AwayScr>HomeScr then HomeLosses=HomeLosses-1;
      if AwayScr<HomeScr then AwayLosses=AwayLosses-1;
      if AwayScr<HomeScr then HomeWins=HomeWins-1;
      if AwayScr=HomeScr then AwayTie=AwayTie-1;
      if AwayScr=HomeScr then HomeTie=HomeTie-1;

run;

* write output to text file, append;
data _null_;
      set test2;
      file "C:\Users\joseph.roith\Desktop\Code\Data\&season\&week..csv"
            dlm=',' mod;
      put Date Away Home AwayWins AwayLosses AwayTie HomeWins HomeLosses
          HomeTie AwayScr HomeScr AwayFD HomeFD AwFDPass HoFDPass AwFDRun
          HoFDRun AwFDPen HoFDPen AwTotalPlay HoTotalPlay AwTotalYard
          HoTotalYard AwYPPlay HoYPPlay AwPassing HoPassing AwYPPass HoYPPass
          AwRush HoRush AwRushAtmpt HoRushAtmpt AwYPRush HoYPRush AwPen
          AwPenYard HoPen HoPenYard AwayTurn HomeTurn AwFumble HoFumble AwInt
          HoInt Aw3DCon Aw3DAtt Aw3DPer Aw4DCon Aw4DAtt Aw4DPer Ho3DCon
          Ho3DAtt Ho3DPer Ho4DCon Ho4DAtt Ho4DPer AwDefTD HoDefTD AwPoss
          HoPoss AwSacked AwSackYds HoSacked HoSackYds AwRZScr AwRZAtt
          AwRZPer HoRZScr HoRZAtt HoRZPer AwAveKR HoAveKR AwAvePR HoAvePR;
run;

%mend;

*OLS Regression SAS Code;

proc import
datafile='C:\Users\joseph.roith\Dropbox\Code\Data\Marginals\train_margins_cle
an.xlsx'
      out=nflmargins replace dbms=xlsx;
```

```
run;

proc import datafile=
  'C:\Users\Joe Roith\Dropbox\Code\Data\Marginals\train_margins_clean.xlsx'
  out=nflmargins replace dbms=xlsx;
run;

 ods graphics on;

proc univariate;
      histogram;
run;

/*Stepwise with stat totals */;
proc reg data=nflmargins;
  model ScoreM = WinPerM--TotalYdM PassingM--PenaltyM TurnoverM
                 _3DPerM--SackedM
                  /selection=stepwise
                  slentry=0.10
                  slstay=0.15
                  rsquare;
  output p=p r=r;
  plot residual.*predicted. / cmallows cookd;
run;

/*Stepwise with per play stats*/;
proc reg data=nflmargins;
  model ScoreM = WinPerM FirstDownM--YPPassM YPRushM PenYardsM--TurnoverM
                 SackYardsM--AvePRM
                  /selection=stepwise
                  slentry=0.10
                  slstay=0.15
                  rsquare;
  output p=p r=r;
  plot residual.*predicted. / cmallows cookd;
run;

/*Final Point Spread Model*/;
proc reg data=nflmargins plot=all;
  model ScoreM = FirstDownM TotalPlayM YPPassM TurnoverM _3DPerM
                 SackYardsM;
  output p=p r=r;
run;

*Logistic Regression SAS Code;
/*logistic regression game*/;

proc logistic data=nflmargins outest=betas covout plots=all;
  model win(event='1')= WinPerM FirstDownM--FDPenM TotalYdM--AvePRM
                        / selection=s
                        slentry=0.25
                        slstay=0.20
                        details lackfit scale=none rsquare;
  output out=pred p=phat lower=lcl upper=ucl
  predprob=(individual crossvalidate);
run;
```

```
/*Final Game Logistic Model*/;
proc logistic data=nflmargins covout plots=all;
  model win(event='1')= FirstDownM TotalPlayM YPPassM PenYardsM TurnoverM
                        _3DPerM SackYardsM
                        / details rsquare lackfit;
  output out=pred p=phat;
run;

/*NFL seasons logistic regression*/;
proc import datafile=
  'C:\Users\Joe Roith\Dropbox\Code\Data\Marginals\nfl_seasons.xlsx'
      out=nflseason replace dbms=xlsx;
run;

data newseason;
  set nflseason;
  PassYds_TD=PassYds/PassTD;
  RushYds_TD=RushYds/RushTD;
run;

proc logistic data=newseason;
  model Playoffs(event='1')= Yds--Y_P Y_A TO_--AveragePenalty PassYds_TD
                             RushYds_TD / selection=s
                             slentry=0.25
                             slstay=0.20
                             details lackfit scale=none rsquare;
  output out=pred p=phat lower=lcl upper=ucl
  predprob=(individual crossvalidate);
run;

proc logistic data=newseason covout plots=all;
  model Playoffs(event='1')= TO_ Ply PenaltyYards PassYds_TD RushYds_TD /
                             noint details rsquare lackfit;
  output out=pred p=phat;
run;

*Discriminant Analysis SAS Code;
/* Individual Game Analysis*/;

proc import datafile=
  'C:\Users\Joe Roith\Dropbox\Code\Data\Marginals\train_margins_clean.xlsx'
      out=nflmargins replace dbms=xlsx;
  run;

proc stepdisc data=nflmargins slentry=0.25 slstay=0.20;
      class Homewin;
      var WinPerM FirstDownM--TotalYdM PassingM--TurnoverM SackedM;
run;

proc stepdisc data=nflmargins slentry=0.25 slstay=0.20;
      class Homewin;
      var FirstDownM TotalPlayM--YPRushM PenYardsM SackYardsM--AveKRM;
run;
```

```
/*Final Game Discrim Model*/;
proc discrim data=nflmargins crossvalidate pool=yes method=normal;
      class Homewin;
      var YPPassM TurnoverM _3DPerM PenaltyM SackedM TotalPlayM FirstDownM;
run;

/*Season Analysis*/;
proc import datafile=
  'C:\Users\Joe Roith\Dropbox\Code\Data\Marginals\nfl_seasons_std.csv'
      out=nflseasons replace dbms=csv;
  run;

data newseason;
  set nflseasons;
  PassYds_TD=pptd;
  RushYds_TD=rptd;
run;

proc stepdisc data=newseason slentry=0.25 slstay=0.20;
      class Playoffs;
      var Yds--Y_P TO_--AveragePenalty PassYds_TD RushYds_TD;
run;

/*Final Season Discrim Model*/;

proc discrim data=newseason crossvalidate pool=yes;
      class Playoffs;
      var TO_ Yds RushYds_TD PassYds_TD PenaltyYards;
      priors '0'=0.625 '1'=0.375;
run;

/*No margins Analyze Offense vs Defense*/;

proc import datafile=
  'C:\Users\Joe Roith\Dropbox\Code\Data\nomargins_ind_train.xlsx'
      out=nflgames_ind replace dbms=xlsx;
run;

proc stepdisc data=nflgames_ind slentry=0.25 slstay=0.20;
      class Win;
      var FDGain--SackYdsGain;
run;

proc import datafile=
  'C:\Users\joseph.roith\Dropbox\Code\Data\nomargins_ind_std.csv'
      out=newgames_ind replace dbms=csv;
run;

/*Final Model Off vs Def*/;

proc discrim data=newgames_ind crossvalidate pool=yes;
      class Win;
      var TurnoverCom--FDGive;
run;
```

APPENDIX D. R CODE

```
### Mixed home and away data set ###

mix<-read.csv("C:\\Users\\joseph.roith\\Dropbox\\Code\\
                  Data\\nomargins_ind_train.csv",header=T)
attach(mix)
names(mix)
v<-mix[,c(21,22,13,14,23,24,26,28,5,6)]

ave.v<-apply(v,2,mean)
sd.v<-apply(v,2,sd)
v.std<-matrix(nrow=nrow(v),ncol=ncol(v))

for(i in 1:ncol(v)){
  v.std[,i]<-(v[,i]-ave.v[i])/sd.v[i]}

v.std<-as.data.frame(v.std)
names(v.std)<-names(v)
v.std<-cbind(Win,v.std)

write.csv(v.std,"C:\\Users\\joseph.roith\\Dropbox\\Code\\Data\\nomargins_ind_
std.csv")

###   Proportional Odds Model   ###

data<-read.csv("C:\\Users\\joseph.roith\\Dropbox\\Code\\Data\\Marginals\\
                  train_margins_clean.csv", header=T)
attach(data)

library(MASS)
library(faraway)
library(nnet)

fact<-cut(ScoreM,breaks=c(-60,-10,0,10,60),
            labels=c("StrongAway","WeakAway","WeakHome","StrongHome"))

##Proportional Odds ##
pomod<-step(polr(fact~WinPerM+FirstDownM+FDPassM+FDRunM+FDPenM+TotalPlayM
                  +TotalYdM+YPPlayM+PassingM+YPPassM+RushM+YPRushM+
                  PenaltyM+PenYardsM+TurnoverM+FumbleM+IntM+X3DPerM+
                  SackedM+SackYardsM+AveKRM+AvePRM,data))

pmod1<-step(polr(fact~FirstDownM+SackYardsM+TurnoverM+X3DPerM+RushM+
                  PassingM+PenYardsM+TotalPlayM,data))
summary(pmod1)
pmod1$deviance;pmod1$edf
pchisq(deviance(pmod1)-deviance(mult),mult$edf-pmod1$edf,lower=F)

##Examples of In-Game Marginals ##
awayadv<-data.frame(FirstDownM=-5,SackYardsM=-10,TurnoverM=-1,
                    X3DPerM=-10,RushM=-50,PassingM=-75,PenYardsM=15,
                    TotalPlayM=-4)
homeadv<-data.frame(FirstDownM=10,SackYardsM=15,TurnoverM=2,
                    X3DPerM=20,RushM=-50,PassingM=125,
                    PenYardsM=-30,TotalPlayM=-5)
noadv<-data.frame(FirstDownM=0,SackYardsM=0,TurnoverM=0,
                  X3DPerM=0,RushM=0,PassingM=0,PenYardsM=0,TotalPlayM=0)
home<-predict(pmod1,homeadv,type="probs");home
away<-predict(pmod1,awayadv,type="probs");away
neutral<-predict(pmod1,noadv,type="probs");neutral

par(mfrow=c(3,1))
t<-seq(-5,5,0.05)
plot(t,dlogis(t),type="l",xlab="",ylab="Density",main="No Advantage")
abline(v=c(-3.6261,-0.2148,3.4949))
plot(t,dlogis(t),type="l",xlab="",ylab="Density",main="Away Team Small
                  Advantage")
abline(v=c(-3.6261,-0.2148,3.4949)-as.numeric(awayadv)%*%pmod1$coef)
plot(t,dlogis(t),type="l",xlab="",ylab="Density",main="Home Team Moderate
                  Advantage")
abline(v=c(-3.6261,-0.2148,3.4949)-as.numeric(homeadv)%*%pmod1$coef)

##Fitted values and New data Predictions ##
newdata<-read.csv("C:\\Users\\joseph.roith\\Dropbox\\Code\\Data\\2013-
                  14\\Season13.csv",header=T)
predict(pmod1,newdata[1:4,],type="probs")

pred<-predict(pmod1,newdata,type="probs")

fact.new<-cut(newdata$ScoreM,breaks=c(-60,-10,0,10,60),labels=c(-2,-1,1,2))

##Selects the Level with the highest probability of occuring##
index<-vector()
for(i in 1:nrow(pred)){
      index[i]<-which(pred[i,]==max(pred[i,]))}

##New Category Level Classification##
levels<-function(x){
  for(i in 1:length(x)){
      if (x[i]==1) x[i]<-(-2)
      if (x[i]==2) x[i]<-(-1)
      if (x[i]==3) x[i]<- 1
      if (x[i]==4) x[i]<- 2}
      return(x)}

real<-as.numeric(fact.new)          ##Actual Level observed
diff<-real-index
diff
sum(diff==0)/length(diff)     ##Proportion of correct categories predicted
sum(abs(diff)<=1)/length(diff)
prod<-levels(real)*levels(index)
prod
sum(prod>0)/length(prod)      ##Proportion of correct game outcomes
```

```
##Split graph of probabilities ##
colorgraph<-function(x){
t<-seq(-5,5,0.05)
a<-c(qlogis(x[1]),qlogis(x[1]+x[2]),qlogis(x[1]+x[2]+x[3]),qlogis(1-x[4]))
q1<-dlogis(t,0,1)
q2<-0.000001*dlogis(t,0,1)
shade1<-seq(-5,a[1],0.01)
shade2<-seq(a[1],a[2],0.01)
shade3<-seq(a[2],a[3],0.01)
shade4<-seq(a[3],6,0.01)
r<-as.numeric(rank(x))

plot(t,q1,type="l",xlab="",ylab="Density",main="Proportional Odds Model")
points(t,q2,type="l",col="black")
polygon(c(shade1,rev(shade1)),c(dlogis(shade1,0,1),0.000001*dlogis(rev(shade1
           ),0,1)),col="gray")
polygon(c(shade2,rev(shade2)),c(dlogis(shade2,0,1),0.000001*dlogis(rev(shade2
           ),0,1)),col="gray40")
polygon(c(shade3,rev(shade3)),c(dlogis(shade3,0,1),0.000001*dlogis(rev(shade3
           ),0,1)),col="gray30")
polygon(c(shade4,rev(shade4)),c(dlogis(shade4,0,1),0.000001*dlogis(rev(shade4
           ),0,1)),col="gray20")}

noadv<-data.frame(FirstDownM=0,SackYardsM=0,TurnoverM=0,
                  X3DPerM=0,RushM=0,PassingM=0,PenYardsM=0,TotalPlayM=0)
noadvpred<-predict(pmod1,noadv,type="probs");
colorgraph(noadvpred)

awayadv<-data.frame(FirstDownM=-5,SackYardsM=-10,TurnoverM=-1,X3DPerM=-
10,RushM=-50,PassingM=-75,PenYardsM=15,TotalPlayM=-4)
awayadvpred<-predict(pmod1,awayadv,type="probs")
colorgraph(awayadvpred)

###   Simulations          ###

## Chose first 50 games

i<-sample(65:257,50,replace=F)

library(TTR)
library(faraway)

##get for and against for all teams##
teams<-read.csv("C:\\Users\\Joe Roith\\Dropbox\\Code\\Data\\2013-
                  14\\allteams.csv",header=T)
attach(teams)

season<-function(data,team){
  out<-data.frame(Team=numeric(),Opp=numeric(),Week=numeric(),
      Game=numeric(),Score=numeric(),FirstDown=numeric(),TotalPlay=numeric(),
      YPPass=numeric(),YPRush=numeric(),PenaltyYards=numeric(),
      Turnover=numeric(),X3DPer=numeric(),SackYards=numeric(),
      Rush=numeric(),Pass=numeric(),OScore=numeric(),OFirstDown=numeric(),
      OTotalPlay=numeric(),OYPPass=numeric(),OYPRush=numeric(),
      OPenaltyYards=numeric(),OTurnover=numeric(),OX3DPer=numeric(),
      OSackYards=numeric(),ORush=numeric(),OPass=numeric(),
      stringsAsFactors=T)
  a<-data[Away==team,]
  h<-data[Home==team,]
      for(i in 1:8){
      out[i,]<-
a[i,c(2,3,4,5,16,18,26,34,40,43,46,54,69,36,32,17,19,27,35,41,45,47,60,71,37,
33)]
}
      for(i in 1:8){
      out[i+8,]<-
h[i,c(3,2,4,5,17,19,27,35,41,45,47,60,71,37,33,16,18,26,34,40,43,46,54,69,36,
32)]
}
  total<-rbind(a,h)
  return(out)

}

Niners<-season(teams,"49ers");Niners<-Niners[order(Niners$Week),]
Bears<-season(teams,"Bears");Bears<-Bears[order(Bears$Week),]
Bengals<-season(teams,"Bengals");Bengals<-Bengals[order(Bengals$Week),]
Bills<-season(teams,"Bills");Bills<-Bills[order(Bills$Week),]
Broncos<-season(teams,"Broncos");Broncos<-Broncos[order(Broncos$Week),]
Browns<-season(teams,"Browns");Browns<-Browns[order(Browns$Week),]
Buccaneers<-season(teams,"Buccaneers");Buccaneers<-
Buccaneers[order(Buccaneers$Week),]
Cardinals<-season(teams,"Cardinals");Cardinals<-
Cardinals[order(Cardinals$Week),]
Chargers<-season(teams,"Chargers");Chargers<-Chargers[order(Chargers$Week),]
Chiefs<-season(teams,"Chiefs");Chiefs<-Chiefs[order(Chiefs$Week),]
Colts<-season(teams,"Colts");Colts<-Colts[order(Colts$Week),]
Cowboys<-season(teams,"Cowboys");Cowboys<-Cowboys[order(Cowboys$Week),]
Dolphins<-season(teams,"Dolphins");Dolphins<-Dolphins[order(Dolphins$Week),]
Eagles<-season(teams,"Eagles");Eagles<-Eagles[order(Eagles$Week),]
Falcons<-season(teams,"Falcons");Falcons<-Falcons[order(Falcons$Week),]
Giants<-season(teams,"Giants");Giants<-Giants[order(Giants$Week),]
Jaguars<-season(teams,"Jaguars");Jaguars<-Jaguars[order(Jaguars$Week),]
Jets<-season(teams,"Jets");Jets<-Jets[order(Jets$Week),]
Lions<-season(teams,"Lions");Lions<-Lions[order(Lions$Week),]
Packers<-season(teams,"Packers");Packers<-Packers[order(Packers$Week),]
Panthers<-season(teams,"Panthers");Panthers<-Panthers[order(Panthers$Week),]
Patriots<-season(teams,"Patriots");Patriots<-Patriots[order(Patriots$Week),]
Raiders<-season(teams,"Raiders");Raiders<-Raiders[order(Raiders$Week),]
Rams<-season(teams,"Rams");Rams<-Rams[order(Rams$Week),]
Ravens<-season(teams,"Ravens");Ravens<-Ravens[order(Ravens$Week),]
Redskins<-season(teams,"Redskins");Redskins<-Redskins[order(Redskins$Week),]
Saints<-season(teams,"Saints");Saints<-Saints[order(Saints$Week),]
Seahawks<-season(teams,"Seahawks");Seahawks<-Seahawks[order(Seahawks$Week),]
Steelers<-season(teams,"Steelers");Steelers<-Steelers[order(Steelers$Week),]
```

```
Texans<-season(teams,"Texans");Texans<-Texans[order(Texans$Week),]
Titans<-season(teams,"Titans");Titans<-Titans[order(Titans$Week),]
Vikings<-season(teams,"Vikings");Vikings<-Vikings[order(Vikings$Week),]


##Parameters function ##

simulate<-function(x,week,n){
mean<-as.data.frame(matrix(NA,nrow=16,ncol=ncol(x)))
sd<-as.data.frame(matrix(NA,nrow=16,ncol=ncol(x)))
sim<-as.data.frame(matrix(NA,nrow=n,ncol=ncol(x)))
      for(i in 1:ncol(x)){
      mean[,i]<-runMean(x[,i],n=1,cumulative=T)
      sd[,i]<-runSD(x[,i],n=1,cumulative=T)
}
names(mean)<-names(x);names(sd)<-names(x)
p<-list(mean=mean,sd=sd)
      for(j in 1:ncol(x)){
      sim[,j]<-rnorm(n,mean[week-1,j],sd[week-1,j])
}
names(sim)<-names(x)
return(sim)
}

##Create mix of Offense and Defense Margins##
offndef<-function(a,b){
  d<-matrix(ncol=10,nrow=nrow(a))
  for(i in 1:10){
      d[,i]<-((a[,i+5]+b[,i+16])/2)-((b[,i+5]+a[,i+16])/2)
}
d<-as.data.frame(d)
names(d)<-names(a[6:15])
return(d)
}

##Simulate model outcomes ##
game_sim<-function(home,away,week,n){
  h<-simulate(home,week,n)
  a<-simulate(away,week,n)
  m<-h-a
  d<-offndef(h,a)
  z<-m[,c(6,13,11,12,14,15,10,7)]
  names(z)<-
c("FirstDownM","SackYardsM","TurnoverM","X3DPerM","RushM","PassingM","PenYard
sM","TotalPlayM")
  v<-d[,c(1,8,6,7,9,10,5,2)]
  names(v)<-
c("FirstDownM","SackYardsM","TurnoverM","X3DPerM","RushM","PassingM","PenYard
sM","TotalPlayM")
  ptsprd<-1.00306+1.37997*m$FirstDown-0.53459*m$TotalPlay+1.00567*m$YPPass-
            3.88568*m$Turnover+17.715*m$X3DPer-0.12464*m$SackYards
  odps<-1.00306+1.37997*d$FirstDown-0.53459*d$TotalPlay+1.00567*d$YPPass-
            3.88568*d$Turnover+17.715*d$X3DPer-0.12464*d$SackYards
  logis<-.2128+.2865*m$FirstDown-.1074*m$TotalPlay+.4646*m$YPPass-
            .0191*m$PenaltyYards-1.3074*m$Turnover+7.04*m$X3DPer-
            .0459*m$SackYards
  odlog<-.2128+.2865*d$FirstDown-.1074*d$TotalPlay+.4646*d$YPPass-
            .0191*d$PenaltyYards-1.3074*d$Turnover+7.04*d$X3DPer-
            .0459*d$SackYards
  discrim0_off<-(-0.66717)-.19287*m$YPPass+.46682*m$Turnover-
            .03388*m$X3DPer+.05034*m$Penalty+.14643*m$Sacked+.01927*
            m$TotalPlay-.04828*m$FirstDown
  discrim1_off<-(-.70289)+.17067*m$YPPass-.48546*m$Turnover+.02957*m$X3DPer-
            .07674*m$Penalty-.12365*m$Sacked-
            .04553*m$TotalPlay+.10926*m$FirstDown
  discrim0_od<-(-0.66717)-.19287*d$YPPass+.46682*d$Turnover-
            .03388*d$X3DPer+.05034*d$Penalty+.14643*d$Sacked+.01927*
            d$TotalPlay-.04828*d$FirstDown
  discrim1_od<-(-.70289)+.17067*d$YPPass-.48546*d$Turnover+.02957*d$X3DPer-
            .07674*d$Penalty-.12365*d$Sacked-
            .04553*d$TotalPlay+.10926*d$FirstDown
  prop_off<-predict(pmod1,z,type="probs")
  prop_od<-predict(pmod1,v,type="probs")
  mean_po_off<-apply(prop_off,2,mean)
  mean_po_od<-apply(prop_od,2,mean)
      out<-c(mean(ptsprd),(sum(ptsprd>0)/n),mean(odps),(sum(odps>0)/n),
             mean(ilogit(logis)),mean(ilogit(odlog)),mean(discrim0_off),
             mean(discrim1_off),mean(discrim0_od),mean(discrim1_od))
      names(out)<-c("Est PS Off","Home Prob PS Off","Est PS Both","Home Prob
                     PS Both","Logist Off","Logist Both","D0 Off","D1
                     Off","D0 OD","D1 OD")
      out<-cbind(t(out),mean_po_off[1],mean_po_off[2],mean_po_off[3],
                  mean_po_off[4],mean_po_od[1],mean_po_od[2],mean_po_od[3],
                  mean_po_od[4])
      return(out)
}

## Sim set 1##
g1<-game_sim(Niners,Rams,13,10000)
g2<-game_sim(Bills,Chiefs,9,10000)
            ...
g49<-game_sim(Vikings,Eagles,15,10000)
g50<-game_sim(Vikings,Packers,8,10000)

total<-rbind(g1,g2,g3,g4,g5,g6,g7,g8,g9,g10,g11,g12,g13,g14,g15,
             g16,g17,g18,g19,g20,g21,g22,g23,g24,g25,g26,g27,g28,
             g29,g30,g31,g32,g33,g34,g35,g36,g37,g38,g39,g40,g41,
             g42,g43,g44,g45,g46,g47,g48,g49,g50)
total

write.csv(total,"C:\\Users\\joseph.roith\\Dropbox\\Code\\Data\\2013-
14\\team\\simulations.csv")
```

```
## Sim set 2##

g1<-game_sim(Bengals,Patriots,5,10000)
g2<-game_sim(Falcons,Jets,5,10000)
            ...
g50<-game_sim(Saints,Buccaneers,17,10000)

total<-rbind(g1,g2,g3,g4,g5,g6,g7,g8,g9,g10,g11,g12,g13,g14,g15,
             g16,g17,g18,g19,g20,g21,g22,g23,g24,g25,g26,g27,g28,
             g29,g30,g31,g32,g33,g34,g35,g36,g37,g38,g39,g40,g41,
             g42,g43,g44,g45,g46,g47,g48,g49,g50)
total

write.csv(total,"C:\\Users\\joseph.roith\\Dropbox\\Code\\Data\\2013-
14\\team\\simulations2.csv")

##Example in Section 4.5###

cumtot<-function(x,week,n){
mean<-as.data.frame(matrix(NA,nrow=16,ncol=ncol(x)))
sd<-as.data.frame(matrix(NA,nrow=16,ncol=ncol(x)))
sim<-as.data.frame(matrix(NA,nrow=n,ncol=ncol(x)))
      for(i in 1:ncol(x)){
      mean[,i]<-runMean(x[,i],n=1,cumulative=T)
      sd[,i]<-runSD(x[,i],n=1,cumulative=T)
}
names(mean)<-names(x);names(sd)<-names(x)
p<-list(mean=mean,sd=sd)
      for(j in 1:ncol(x)){
      sim[,j]<-rnorm(n,mean[week-1,j],sd[week-1,j])
}
names(sim)<-names(x)
return(p)
}

##Simulate point spread using offense only Example ##
game_ps_off<-function(home,away,week,n){
  h<-simulate(home,week,n)
  a<-simulate(away,week,n)
  m<-h-a
  d<-offndef(h,a)
  ptsprd<-1.00306+1.37997*d$FirstDown-0.53459*d$TotalPlay+1.00567*d$YPPass-
3.88568*d$Turnover+17.715*d$X3DPer-0.12464*d$SackYards
      h<-hist(ptsprd,breaks=75,plot=F)
      divide<-cut(h$breaks, c(-Inf,-.00001,Inf))
      plot(h, col=c("gray","gray40")[divide],main="Point Spread (Off and Def
Marginals)",xlab="Points")
      out<-c((sum(ptsprd>0)/n),mean(ptsprd),(sum(ptsprd<=-10)/n))
      names(out)<-c("Home Probability","Estimated Point Spread","Actual")
      return(out)}

game_ps_off(Bills,Chiefs,9,10000)
```