DEVELOPMENT OF A PREDICTION MODEL FOR THE NCAA DIVISION-I FOOTBALL CHAMPIONSHIP SUBDIVISION

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

By

Joseph Charles Long

In Partial Fulfillment
for the Degree of
MASTERS OF SCIENCE

Major Department:
Statistics

April 2013

Fargo, North Dakota

North Dakota State University
Graduate School

Title

DEVELOPMENT OF A PREDICTION MODEL FOR THE NCAA DIVISION-I FOOTBALL CHAMPIONSHIP SUBDIVISION

By

Joseph Charles Long

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

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Rhonda Magel
Chair

Dr. Seung Won Hyun

Dr. Gang Shen

Dr. Fariz Huseynov

Approved:

04/05/2013	Dr. Rhonda Magel
Date	Department Chair

ABSTRACT

This thesis investigates which in-game team statistics are most significant in determining the outcome in a NCAA Division-I Football Championship Subdivision (FCS) game. The data was analyzed using logistic and OLS regression techniques to create models that explained the outcome of the past games. The models were then used to predict games where the actual in-game statistics were unknown. A random sample of games from the 2012 NCAA Division-I FCS regular season was used to test the accuracy of the models when used to predict future games. Various techniques were used to estimate the in-game statistics in the models for each individual team in order to predict future games. The most accurate technique consisted of using three game medians with respect to total yards gained by the teams in consideration. This technique correctly predicted 78.85% of the games in the sample data set when used with the logistic regression model.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Rhonda Magel, who aided me in the process of developing this thesis. Her advice and knowledge of sport statistics was a useful tool. I would also like to thank my other committee members; Dr. Seung Won Hyun, Dr. Gang Shen and Dr. Fariz Huseynov for providing review and feedback on this thesis.

Finally I would like to thank my friends and family for always being there for me and believing in me. They have been a key factor to my achievements in life. I would also like to give a special thanks to my friends in the statistics department for sharing their thoughts and ideas on sports modeling and also helping me stay motivated during the process of writing this thesis.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1. INTRODUCTION

Watching and participating in sports has been a major past time in the United States for decades. Sports enthusiasts which include coaches, players and fans have always analyzed and debated on what are the most significant factors that attribute to a team’s victory over their opponent. Sports enthusiasts will make claims such as “the best defense is a strong offense” or some might debate vice versa that “the best offense is a strong defense”. This debate boils down to does having a strong offense win a game or is it more important to have a strong defense to win a game or is it a balance between the two. People have made strong debates and speculations on both sides for many years. It has not been until recently where there has been advanced research by statisticians and mathematicians in the field of sports analytics. Their research attempts to tackle the debate of what are the most influential variables that affect a team’s chances of winning a game.

When the topic of sports analytics is brought up, Bill James is usually a name that pops into most of the conversations. James is a famous baseball writer and statistician who was the first to analyze the game of baseball differently than any other with his uses of Sabermetrics. Since 1977, James has written dozens of books on baseball history and statistics. ESPN True Hoops writer Kevin Arnovitz [2012] claims “James' legacy in sports will be as the godfather of advanced stats”. MITnews writer Peter Dizikes [2013] said in a recent article “The popularity of sports analytics owes a lot to Lewis’ 2003 book “Moneyball,” which illuminated how the Oakland Athletics used the 1980s-era insights of pioneering baseball analyst Bill James to compete with wealthier teams”. This book was also turned into a movie in 2011 with the same title “Moneyball”.

Sports analytics conventions have also been popping up around the United States during this last boost in popularity of sports analytics. One mentionable conference which has been receiving popularity across the sports industry is the MIT Sports Analytics Conference which was first held in 2007. According to Dizikes [2013] this year's conference had 30 panels, a research-paper competition, and representatives of more than 90 professional teams in attendance, and revolved around the theme of "big data".

This "big data" term is a key component of the evolution of research in the field of sports analytics. When Bill James first started working with baseball in 1980's he was limited to box scores that were printed in the newspaper. Now the amount of data that is collected on sports is endless and easily accessible through the internet. The increased processing speed of computers have also aided as a major factor since more complex analysis and simulations can be quickly done. The topics and papers discussed at this conference were aimed at answering the question that was introduce at the beginning of this thesis, which factors are the most significant in determining the outcome of a sporting event.

While these questions are being answered by advanced statistical research, more professional and amateur sports are seeing the significance of this type of analysis. This is influencing them to start employing statisticians to work for their team in order to gain an advantage over other teams.   Daryl Morey general manager of the NBA's Houston Rockets was quoted saying "I think what teams are going to be running in 10 years will be totally different," Morey was talking about the way offensive styles in basketball are evolving based on statistical feedback Dizikes [2013].

This thesis will focus on investigating which in-game team statistics are most significant in determining the victor in a NCAA Division-I Football Championship Subdivision (FCS) game. The NCAA Division-I Football Championship Subdivision is a division of college football that operates in the National Collegiate Athletic Association (NCAA). The NCAA has three main divisions. They are denoted as Division-I, II, and III. Division-I is split up into two subdivisions. The first subdivision is known as the Football Bowl Subdivision (FBS) and the second is known as the Football Championship Subdivision (FCS). The FCS is unique in that teams in this subdivision will play both FBS and Division-II teams throughout the season. Therefore, when analyzing what factors are the most significant in determining the outcome of a NCAA Division-I FCS game, games selected from our sample from games played by FCS teams might also include FBS and Division-II teams.

There has been research by Magel and Childress [2012], which shows that turnover margin is significant in determining the winner in a professional football. Other statistics that have been brought up in debates of their significance and will be explored in this thesis are; third down conversions, first down conversions, number of penalties, total penalty yards, kick returns, kick return yards, pass completion, number of possessions, number of plays, rushing yards, passing yards, time of possession, and number of defensive sacks.

The investigation will be done by using both logistic and ordinary least squares regression analysis applied to past data sampled from three seasons of NCAA Division-I Football Championship Subdivision (FCS) games. Once the most significant models are obtained we will then try to predict future games where in-game team statistics are unknown. We will use the 2012 NCAA Division-I Football Championship Subdivision regular season which is the most recent season to test the accuracy of the use of these models to predict future games.

Chapter 2 discusses previous research on significant factors in football games and also research on predicting future outcomes of sporting events. Chapter 3 will discusses the methodology that is used to collect the data and conduct the analysis of significant factors. Chapter 4 gives the results of the methods used in Chapter 3. Chapter 5 will include an example of how to use the model to make predictions for futures games and also the accuracy of the different models when used to predict future games. Chapter 6 with finish with a conclusion of the thesis and a discussion of future work.

CHAPTER 2. SURVEY OF LITERATURE

There has been a lot of research in sports analytics that range from developing prediction models for game outcomes, to developing ranking systems that remove the bias from coach's polls. In a literature review, three sources were found to be of interest that had research that analyzed the factors that influence football games.

Magel and Childress [2012] conducted an analysis on the National Football League examining the outcome effects of turnover margin. Their study used logistic regression techniques to analyze 1,783 regular season games from the 2001 to 2007 seasons. They fit the logistic model estimating the probability of winning the game using four independent variables; season, week, home-field advantage and turnover margin. Their analysis found that that turnover margin and home field advantage were both significant in determining the outcome of games. Using the 2008 NFL regular season games to test the accuracy of their model, they found their model had an accuracy of 72.16% in predicting the outcome of games with only the use turnover margin and home-field advantage. This confirms the significance of both turnover margin and home-field advantage in the NFL.

In a study done by Harville [1977] of using linear-model methodology applied to point spread to rank high school and college football teams, Harville focused on home-field advantage and mean performance levels of teams. He tested his methodology on the 1975 college football regular season. The games he considered for his analysis consisted of Division-I teams, and non-Division-I teams that had at least one or more Division-I opponents in their schedule. Over all his analysis covered 1,024 games. He found home-field advantage contributed to a 3.4 point advantage over the away team.

A study on the Canadian Football League done by Willoughby [2002] used logistic regression models to classify the important factors of a team's overall success. Willoughby focused his analysis on three different teams. He chose three different teams in order in analyze the difference between a "very good" team an "average" team and a "poor" team. The statistics Willoughby used from games were; the difference in rushing yardage, the difference in passing yardage, the difference in the number of interceptions, the difference in the number of fumble recoveries and the difference in the number of quarterback sacks. He found that these game statistics were in fact significant in determining the outcome of a game, though the significance of the variables varied based on the strengths of the teams. Willoughby tested the models on the data he used to developed them and found with the in-game statistics known they predicted 85.9% for the "very good" team, 90.2% for the "average" team and 78.8% for the "poor" team.

CHAPTER 3. METHODOLOGY

The overall goal of this study was to determine which in-game statistics are significant in determining the winner of a NCAA Division-I FCS game. This was done by fitting two types of models using in-game statistics. Logistic regression was used to estimate the probability of winning a game and ordinary least squares regression was used to estimate the point spread of a game when the in-game statistics are known. A secondary goal of the study was then to take these models and attempt to predict future games by estimating the unknown in-game statistics.

Data was collected from three regular seasons of NCAA Division-I FCS games which included the 2009, 2010 and 2011 seasons. Since collecting data for every single game would be time consuming a stratified random sample of 228 games was collected. Five out of the thirteen conferences in the NCAA Division-I FCS were randomly sampled. Within each of the five conferences that were chosen, a random sample of half of the teams in each conference was then selected. For each year and each team, four regular season games were randomly sampled from game 1 through game 11 and the values of the variables we were interested in were noted. Due to the fact that the sampling of games was based on individual teams, there is a problem of sampling the same game more than once. If this happened the following game was then used for the sample. The games that were sampled are shown in Table 3.1 with their respective conference, team and year.

The data was collected from three websites. In-game statistics were collected from two websites due to the fact that the two sites reported different kinds of in-game statistics that the other site did not collect. The two sites that were used were ESPN College Football Score Board [ESPN.com] and the official NCAA Football Statistics Database [NCAA.org]. The third source was from a computer ranking website [CompughterRatings.com] which was used to collect computer generated rankings of the teams.

Table 3.1. Games Sampled for Model Fitting

Big South Conference			
Team	**2009 Games**	**2010 Games**	**2011 Games**
Charleston Southern	3, 5, 7 , 9	2, 7, 8, 10	2, 4, 7, 10
Coastal Carolina	3, 4, 7, 11	1, 4, 7, 11	2, 4, 6, 9
Gardner-Webb	3, 9, 10, 11	4, 5, 9, 11	1, 2, 5, 6
Ohio Valley Conference			
Team	**2009 Games**	**2010 Games**	**2011 Games**
Eastern Illinois	2, 3, 9, 11	2, 5, 7, 11	4, 7, 9, 10
Eastern Kentucky	1, 3, 6, 9	3, 7, 9, 10	3, 6, 7, 11
Jacksonville State	1, 2, 4, 8	3, 4, 7, 11	2, 7, 8, 10
Tennessee-Martin	3, 6, 8, 10	1, 5, 6, 10	4, 8, 9, 10
Patriot League			
Team	**2009 Games**	**2010 Games**	**2011 Games**
Bucknell	4, 6, 8, 9	1, 2, 6, 8	4, 6, 7,11
Fordham	1, 3, 9, 10	1, 3, 6, 10	3, 4, 8, 10
GeorgeTown	3, 4, 5, 11	2, 4, 7, 11	2, 4, 7, 8
Southern Conference			
Team	**2009 Games**	**2010 Games**	**2011 Games**
Chattanooga	1, 2, 9, 11	1, 2, 8, 9	1, 4, 5, 9
Elon	2, 3 , 9 , 11	1, 3, 8, 10	1, 6, 8, 9
Georgia Southern	1, 6, 8, 10	1, 3, 4, 6	1, 4, 6, 11
Samford	1, 2, 4, 6	2, 3, 7, 11	2, 6, 10, 11
Southwestern Athletic Conference			
Team	**2009 Games**	**2010 Games**	**2011 Games**
Alabama State	3, 7, 8, 10	3, 4, 5, 10	2, 6, 7, 9
Alcorn State	1, 7, 9, 11	5, 7, 9, 11	3, 5, 9, 11
Grambling State	2, 4, 6, 11	4, 5, 6, 11	1, 4, 8, 9
Jackson State	1, 3, 8, 11	1, 3, 4, 6	2, 3, 7, 11
Southern University	3, 8, 9, 11	2, 5, 6, 11	4, 5, 7, 11

ESPN College Football Score Board [ESPN.com] was used to collect the following in-game statistics for each team; points scored, first downs, third down efficiency, total yards, rushing yards, passing yards, pass completions, pass attempts, rushing attempts, penalties, penalty yards, turnovers, possession time, punt returns, and punt return yards. An additional indicator variable was used to denote whether a team was home or away. This variable was giving the response of "1" for home and "0" for away. The second source of in-games statistics was the official NCAA Football Statistics Database [NCAA.org]. This was used to collect the amount of sacks a team's defense had for a given game which was not on the ESPN College Football Score Board.

Field goals and kick returns were omitted from the analysis since they are directly related to the score of the game. This was due to the fact that a field goal gives a team three points and a kick return happens after the other team scores. Therefore, having these statistics in the analysis would take away from our main goal of determining which in-game statistics are most influential at determining the outcome of a game. For instance if one team has fewer kick returns and more field goals than another team, this would mean that they have scored more points in the game.

The third source of data was from CompughterRatings.com [CompughterRatings.com]. This site was used to collect the rank of a team going into the game they were playing. This website was developed by Steve Pugh. The website uses an advanced mathematical model to rate sports teams in various competitive sports. The model uses a combination of the least squares method and the maximum likelihood technique to generate the rankings of the teams. Pugh's website has the ability to rank teams in each individual division and also has the ability to rank teams across all division. Since our stratified random sample includes games that have NCAA Division-I FCS, NCAA Division-I FBS and NCAA Division-II teams, the rankings that combined all division were used. This data will be used to compare the accuracy of the models at predicting future games. It will also be used in combination with in-game statistics and the weekly rankings in developing a model to predict future games.

With the sample data set of 228 games, four models were fitted to the data to help explain the outcome of an individual game. The explanatory variables used in all of the models were the difference in the statistics that we collected for each team. The variables that were created from the initial data set are:

- The difference between time of possession (diff_top);
- The difference between offensive yards (diff_yards);
- The difference between rushing yards (diff_rushyards);
- The difference between passing yards (diff_passyards);
- The difference between interceptions (diff_int);
- The difference between fumbles (diff_fumble);
- The difference between turnovers, turnover margin (TOM);
- The difference between penalties (diff_pen);
- The difference between penalty yards (diff_pen_yards);
- The difference between percentage of pass completions (diff_pass_comp);
- The difference between defensive sacks (diff_sacks);
- The difference between first downs (diff_first);
- The difference between third down conversion percentage (diff_3rd_pct);
- The difference between punt returns (diffpunt_return);
- The difference between punt return yards (diff_puntyards);
- The difference between the ratio of punt return yards over punt returns (diffave_puntyards);
- The difference between the ratio of total offensive yards over by total offensive attempts (diffave_yardsperplay).

Two of the models estimated the probability of winning the game using logistic regression. These two models used a binary response variable of "1" for a team winning a game and "0" for a team losing the game. The difference between the two models is one was fitted using only the in-game statistics and the other one was fitted using in-game statistics and the computer generated rankings. We will denote the model with only in-game statistics as Model 1 and the model with in-game statistics and the computer rankings as Model 2. This was done in order to see if the addition of a ranking system would improve the results of the model. In order to help select the significant in-game statistics to use in the models, a stepwise model selection procedure was used in SAS 9.3, using α value for entry of .25 and exit of .2.

The other two models estimated the point spread of the game using ordinary least squares regression. The first model used only in-game statistics as the explanatory variables and the second model used in-game statistics and the addition of the computer generated ranking system. We will denote the model using only in-game statistics as Model 3 and the model with in-game statistics and computer rankings as Model 4. Model selection was also done by the stepwise model selection procedure in SAS 9.3, using α value for entry of .25 and exit of .2.

Once we determined the models that best fit the sample data, we then attempted to use the models to predict future games without knowing the actual in-game statistics. Since we are trying to predict future games with these models, we need a way to estimate the in-game statistics. In order to do this, we will look at the past three games a team has played and explore methods that give us the most accurate predictions of the outcomes.

CHAPTER 4. RESULTS

We will first discuss the results of Model 1 which uses logistic regression to fit the probability of a team winning a game and only considers in-game statistics as the independent variables. With the aid of stepwise selection with an α level of .25 for entry and .2 for exit, we narrowed down the variables that were significant at a .05 level of significance. The final variables that were selected for this model are; turnover margin, difference in pass completion, difference in defensive sacks, difference in third down percentage, difference in punt returns and difference in average yards per play. Since this model only contains difference of in-game statistics, the model was fit using no intercept. The theory behind this is that if the two teams have virtually equal in-game statistics, then the game would end in a tie and therefore the intercept would be zero. Using Hosmer and Lemeshow Goodness-of-Fit test where the null hypothesis is that the model is a good fit for the data and the alternative hypothesis is the model is not a good fit, we get a p-value of 0.9663 which indicates the model is a good fit for the data, parameter estimates for this regression model are listed in Table 4.1

Table 4.1. Parameter Estimates for Model 1

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
TOM	1	0.9673	0.1825	28.1057	<.0001
diff_pass_comp	1	6.3759	1.8778	11.5295	0.0007
diff_sacks	1	0.6768	0.1565	18.7002	<.0001
diff_3rd_pct	1	5.0762	1.5456	10.7868	0.001
diffpunt_return	1	0.5857	0.1689	12.0194	0.0005
diffave_yardsperplay	1	0.5001	0.1744	8.221	0.0041

To interpret the estimates of the parameters in the model we will use odd ratios which are reported in Table 4.2. Turnover margin has a parameter estimate of 0.9673 this yields an odds ratio of 2.631 for an additional turnover with a 95% confidence limit of (1.84, 3.762). This means that each additional turnover a team recovers would increase the odds of winning for that team by a factor of 2.631. The odd ratio shows that turnovers have a major impact on a game; therefore it is important for a team to focus on stopping turnovers on offense and causing turnovers on defense.

Table 4.2. Odds Ratios for Model 1

	Odds Ratio Estimates		
Effect	**Point Estimate**	**95% Wald**	
		Confidence Limits	
TOM	2.631	1.84	3.762
diff_pass_comp	587.532	14.814	>999.999
diff_sacks	1.968	1.448	2.674
diff_3rd_pct	160.158	7.744	>999.999
diffpunt_return	1.796	1.29	2.501
diffave_yardsperplay	1.649	1.171	2.321

An interesting aspect of this model is that it did not include the home indicator variable. This is of interest because of past research showing that home field advantage is significant in determining the outcome for football games. To investigate this, the home variable was included in the model and it was shown to have a negative coefficient which goes against the theory of home field advantage. The cause of this is due to multicollinearity which infers home field advantage is already explained by the in-game statistics that are selected. These results are reported in Table 4.3.

Table 4.3. Addition of Home into Model 1

Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Home	1	-0.3329	0.389	0.7324	0.3921
TOM	1	0.9748	0.1828	28.4259	<.0001
diff_pass_comp	1	6.696	1.9175	12.1941	0.0005
diff_sacks	1	0.6936	0.1595	18.9165	<.0001
diff_3rd_pct	1	5.1684	1.5829	10.6613	0.0011
diffpunt_return	1	0.5966	0.1718	12.0511	0.0005
diffave_yardsperplay	1	0.503	0.175	8.2632	0.004

Let us now consider Model 2 which uses logistic regression to estimate the probability of a team winning a game considering the use of in-game statistics and the computer generated rankings. With the aid of stepwise selection with an α level of .25 for entry and .2 for exit, we narrowed down the variables that were significant at a .05 level of significance. As expected with the addition of the computer generated rankings increases the predictability of the model and also selects the same variables as Model 1. Using Hosmer and Lemeshow Goodness-of-Fit test we get a p-value of 0.4927 which indicates this model is a good fit for the data. The parameter estimates for this regression model are listed in Table 4.4 and the odds ratios are reported in Table 4.5.

Table 4.4. Parameter Estimates for Model 2

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Difference_in_rank	1	0.00824	0.00235	12.2622	0.0005
TOM	1	1.1234	0.2223	25.54	<.0001
diff_pass_comp	1	7.5924	2.1403	12.5843	0.0004
diff_sacks	1	0.7363	0.1774	17.222	<.0001
diff_3rd_pct	1	4.8726	1.6103	9.1565	0.0025
diffpunt_return	1	0.6185	0.1955	10.0089	0.0016
diffave_yardsperplay	1	0.4522	0.2085	4.7029	0.0301

Table 4.5. Odds Ratios for Model 2

	Odds Ratio Estimates		
Effect	**Point Estimate**	**95% Wald**	
		Confidence Limits	
Difference_in_rank	1.008	1.004	1.013
TOM	3.075	1.989	4.755
diff_pass_comp	>999.999	29.892	>999.999
diff_sacks	2.088	1.475	2.957
diff_3rd_pct	130.667	5.565	>999.999
diffpunt_return	1.856	1.265	2.723
diffave_yardsperplay	1.572	1.044	2.365

Let us now look at Model 3 which uses ordinary least squares regression to estimate the point spread of the games using only the in-game statistics as the independent variables. With the aid of stepwise selection with a α level of .25 for entry and .2 for exit, we narrowed down the variables that were significant at a .05 level of significance. The final variables that were selected for this model are; turnover margin, difference in penalties, difference in first downs, difference in 3rd down percentage, difference in punt returns, difference in punt yards and difference in average yards per play. Since this model only contains difference of in-game statistics, the model was fit using no intercept. With this technique we developed a model that yields an adjusted R^2 of 0.8306, which mean that 83.06% of the total variation of point spread is explained by the model suggesting that it is a good fit for modeling the point spread of a game. The parameter estimates for this regression model are listed in Table 4.6.

Table 4.6. Parameter Estimates for Model 3

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
TOM	**1**	3.87225	0.28928	13.39	<.0001
diff_pen	**1**	-0.51624	0.14291	-3.61	0.0004
diff_first	**1**	0.38999	0.08768	4.45	<.0001
diff_3rd_pct	**1**	20.75532	3.39939	6.11	<.0001
diffpunt_return	**1**	1.00627	0.4034	2.49	0.0133
Diff_puntyards	**1**	0.06932	0.02319	2.99	0.0031
diffave_yardsperplay	**1**	4.49276	0.34299	13.1	<.0001

This model also does not select the variable home. To test to see if home field advantage is already explained by these selected variables we will put it in the model. As you can see in Table 4.7 home is giving a negative coefficient of -.10467 and a p-value of 0.9004 suggesting that there is multicollinearity due to the fact that the variables in the model already explain home-field advantage.

Table 4.7. Addition of Home into Model 3

Parameter	DF	Estimate	Standar Error	t Value	Pr > |t|
Home	**1**	-0.10467	0.83497	-0.13	0.9004
TOM	**1**	3.87194	0.28994	13.35	<.0001
diff_pen	**1**	-0.51754	0.1436	-3.6	0.0004
diff_first	**1**	0.39112	0.08834	4.43	<.0001
diff_3rd_pct	**1**	20.72965	3.41314	6.07	<.0001
diffpunt_return	**1**	1.00882	0.40481	2.49	0.0134
Diff_puntyards	**1**	0.0694	0.02325	2.99	0.0032
diffave_yardsperplay	**1**	4.49244	0.34376	13.07	<.0001

Now considering Model 4 which uses ordinary least squares regression to estimate the point spread of the games using the in-game statistics along with the computer generated rankings. With the aid of stepwise selection with an α level of .25 for entry and .2 for exit, we narrowed down the variables that were significant at a .05 level of significance. With this technique we developed a model that yields an adjusted R^2 of 0.8448. The parameter estimates for this regression model are listed in Table 4.8.

Table 4.8. Parameter Estimates for Model 4 Without Difference in Punt Return

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Difference_in_rank	**1**	0.02099	0.00404	5.2	<.0001
TOM	**1**	3.58041	0.26797	13.36	<.0001
diff_pen	**1**	-0.50676	0.13658	-3.71	3E-04
diff_first	**1**	0.35191	0.084	4.19	<.0001
diff_3rd_pct	**1**	19.88392	3.25922	6.1	<.0001
Diff_puntyards	**1**	0.09692	0.01716	5.65	<.0001
diffave_yardsperplay	**1**	4.22699	0.33249	12.71	<.0001

In the comparison of Model 3 and Model 4 you can see when the difference in rank is added to the model it kicks out the difference in punt returns. To make Model 3 and Model 4 comparable we will put difference in punt returns into model 4 and see if the results are justifiable. After adding difference in punt returns back into the model, it brings the adjusted R^2 up to 0.8462 and give a p-value of 0.0868 for difference in punt returns, therefore we can justify adding this variable back into the model even though its p-value is greater than our desired level or 0.05. The parameter estimates for this regression model are listed in Table 4.9.

Table 4.9. Parameter Estimates for Model 4

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Difference_in_rank	**1**	0.01975	0.00408	4.83	<.0001
TOM	**1**	3.71258	0.27763	13.37	<.0001
diff_pen	**1**	-0.49136	0.13627	-3.61	4E-04
diff_first	**1**	0.33225	0.0844	3.94	1E-04
diff_3rd_pct	**1**	19.68779	3.24687	6.06	<.0001
Diff_puntyards	**1**	0.07278	0.02211	3.29	0.001
diffpunt_return	**1**	0.67187	0.39058	1.72	0.087
diffave_yardsperplay	**1**	4.13958	0.33491	12.36	<.0001

CHAPTER 5. MODEL PREDICTON & ACCURACY

In attempt to test the accuracy of predicting future games with the four models we have developed in this study; we will look at the past three games a team has played and explore a method that gives us the most accurate prediction of the actual game outcome. We will only be able to predict one game in advance with this model, because of this method.

To obtain data to test the accuracy of the models used for future predictions, we took a stratified random sample of 52 games from the 2012 NCAA Division-I FCS regular season. For this stratified random sample we considered all thirteen conferences in the NCAA Division-I FCS. We then randomly sampled two teams in each conference. Once we had two teams chosen for each conference we then randomly sampled two games for each team in which to predict results. Since we are using the last three games as a way to estimate the in-game statistics, the random sample only contains game 4 through game 11. We also made sure that we did not have duplicate games in our test sample. Sampled games are reported in Table 5.1.

Table 5.1. Sampled Games from 2012 NCAA Division I FCS

Conference	**1st Team & Games**	**2nd Team & Games**
Big Sky Conference	Idaho State - 5, 10	Northern Colorado- 6, 11
Big South Conference	Gardner-Webb- 4, 6	VMI- 6, 8
Colonial Athletic Association	Delaware- 9, 11	James Madison- 9, 10
Ivy League	Columbia Lions- 6, 9	Cornell- 7, 10
Mid-Eastern Athletic Conference	Norfolk State- 8, 10	North Carolina Central- 8, 11
Missouri Valley Football Conference	Illinois State- 4, 5	Missouri State- 5, 10
Northeast Conference	Bryant University- 4, 8	Wagner- 9, 10
Ohio Valley Conference	Austin Peay- 7, 11	Murray State- 4, 7
Patriot League	Georgetown- 6, 7	Holy Cross- 6, 11
Pioneer Football League	Drake- 8, 10	Morehead State- 4, 6
Southern Conference	Citadel- 4, 9	Georgia Southern- 5, 7
Southland Conference	Lamar- 5, 11	McNeese State- 8, 11
Southwestern Athletic Conference	Alabama A&M- 5, 9	Grambling State- 6, 7

When using the models obtained from logistic regression techniques the model will estimate a probability between 0 and 1. Therefore, we will classify a game as a win if the model produced a probability greater than 0.5. If the model produces a probability less than 0.5 then we will classify the outcome of the game as a loss. When considering the models obtained for ordinary least squares regression, we will consider the game a win if the model estimates a point spread greater than zero and we will consider it a loss if the model estimates a point spread less than zero.

Several methods were used in attempt to estimate the in-game statistics these include the following: using an average of the last three in-game statistics; using a two game average of in-game statistics; using the last game as a predictor of in-game statistics; using the median of the last three in-game statistics; using the sum of the last three in-game statistics; using the maximum of the last three in-game statistics and using the minimum of the last three in-game statistics. We then tried estimating the in-game statistics for each team by considering the past three games each team under consideration played and then collecting the in-game statistics for each team where that team had the median amount of total yards. This last method ended up being our best attempt to estimate the future in-game statistics.

To better explain this method we will illustrate an example using Model 3, which is the point spread model that was fitted using ordinary least squares regression. We will predict the outcome of a game played on November, 3rd 2012 were Missouri State University played North Dakota State University at Missouri State University (Game 10 for both teams). First, we will look at the total yards of the past three games for each team in consideration.

Table 5.2. Median Total Yards Example

Game	MSU Total Yards	NDSU Total Yards
9	350	386
8	305	385
7	283	294

You can see by looking at Table 5.2, that Missouri State's game with the median total yards gained for the past three games occurred in game number 8. Therefore, we will use game number 8 to collect the in-game statistics that will be used in Model 3 for MSU. By using the same method you can see that game number 8 for North Dakota state is also the game that has the median amount of total yards gained; thus we will use game number 8 to collect the in-game statistics for North Dakota State. These in-game statistics are shown in Table 5.3. The difference of the in-game statistics which are the variables that are used in Model 3 are shown in Table 5.4.

Table 5.3. In-game Statistics for MSU vs. NDSU

Team	Game	First	3rd Down Pct	pen	Turnovers	punt return	punt yards	Average Yards Per Play
MSU	8	17	53.33%	6	3	4	49	9.53
NDSU	8	21	57.14%	7	0	4	132	9.87

Table 5.4. Difference of In-game Statistics for MSU vs. NDSU

Diff_first	diff_3rd_pct	diff_pen	TOM	diffpunt_return	diff_puntyards	diffave_yardsperplay
-4	-3.81%	-1	-3	0	-83	-1.24

Now using Model 3 we obtain the following equation for predicting the point spread of this game:

Point Spread = TOM*3.87225 - diff_pen*0.51624 + diff_first*0.38999 +

diff_3^{rd}_pct*20.75532 + diffpunt_return*1.00627 + diff_puntyards*0.06932 +

diffave_yardsperplay*4.49276

Plugging in our estimated in-game statistics from the median total yards method, we obtain the following:

Point Spread = (-3)*3.87225 – (-1)*0.51624 + (-4)*0.38999 + (-0.0381)*20.75532 + (0)*1.00627 + (-83)*0.06932 + (-1.24)*4.49276 = -24.77

Therefore, the predicted point spread for the game is -24.77 since the estimated point spread is less than zero the prediction for this game would be that MSU losses the game. In looking at the actual score of this game we found MSU loses to NDSU 17 to 21, giving an actual point spread of -4. In this case the model correctly predicted the winner of the game but overestimated the margin of victory.

Now that we have gone through an example of how to use the models to predict future games, we can discuss the accuracy of the models for predicting future games. The highest prediction accuracy was from Model 2 yielding and accuracy of 78.89% which was the logistic regression model that used both in-game statistics and difference in rank. The next accurate model was Model 1 which was the logistic regression model that only considered in-game statistics yielding an accuracy of 75%. Model 4, the ordinary least squares regression model that included both in-game statistics and difference in rank, yielded an accuracy of 73.08%. The least accurate model was Model 3, which was the ordinary least squares method which only considered in-game statistics which yielded an accuracy of 69.23%. The accuracy results are shown in Table 5.5.

Table 5.5. Accuracy Testing Results

Models	**Correct**	**Incorrect**	**Prediction Accuracy**
Model 1	39	13	75.00%
Model 2	41	11	78.85%
Model 3	36	16	69.23%
Model 4	38	14	73.08%
Computer Ranking	37	15	71.15%
Home	35	17	67.31%

In the comparison of these results to the computer ranking which correctly predicted 71.15% of the games it is confirmed that using the three game medians to predict future games with the four models we have developed a suitable way to predict future games. The 95% confidence intervals for the accuracy results are reported in Table

Table 5.6. Confidence Intervals for Accuracy Results

	95% Confidence Intrevals	
	Lower bound	**Upper Bound**
Model 1	63.23%	86.77%
Model 2	67.75%	89.95%
Model 3	56.69%	81.78%
Model 4	61.02%	85.13%
Computer Ranking	58.84%	83.47%
Home	54.56%	80.06%

CHAPTER 6. CONCLUSIONS

After fitting the model when only considering in-game statistics, it was found that with the logistic model that predicted win probability, there are six statistics that are highly significant in determining the outcome of a game in the NCAA Division-I Football Championship Subdivision. These variables are; turnover margin, pass completion percentage, number of sacks a team's defense has, percent of third down completions, number of punt returns, and average offensive yards per play. Therefore, if a team focuses on these key in-game statistics they will increase their probability of winning a football game.

In the analysis of the ordinary least squares regression that predicts the point spread, there are seven statistics that are highly significant in determining the outcome of a game in the NCAA Division-I Football Championship Subdivision. These variables are turnover margin, number of penalties, number of first downs, number of punt returns, number of punt return yards, percent of third down completions, and average offensive yards per play. These variables are different than the variables that predict the probability of winning a game. The key difference could be attributed to the fact that the logistic model does not account for the actual score of a game. It just predicts whether a team lost or won. The point spread model takes into account the magnitude of the victory.

One model may be more attractive than another one, depending on who is using the model. If a coach was looking at the models, they would probably be more interested in looking at the point spread model since they would be interested in obtaining the largest victory over a team. Therefore, a coach would focus on increasing their team's turnover margin in order to increase the number of possessions they have and also at the same time take away scoring chances from the other team. Increase the amount of first downs, which is intuitive since the more first down you have, the further you move the ball down the field. Increase the number of punt returns the team receives in a game. This can be done by strengthening the team's defense, since a team will only receive a punt return if they stop the other team's defense. A coach would also want to focus on his special teams, because if the team can gain more yards off of the punt returns, they have less distance to travel down the field to score. It would also be important for a coach to instill the importance of converting third downs to his team. Lastly, make smart plays so the team can increase its average yards per play.

The other side of the sport scene is the sports gamblers. Gamblers would be most interested in the logistics regression model if their primary interest is to bet on whether a team wins or loses. Predictions are made a lot easier if you are only considered with whether you win or lose, because you do not have to deal with the variance of point spread which makes modeling the outcome of games harder. A good example of this is the results of the four different models we obtained in this study. The point spread models using ordinary least square regression had the lowest accuracy in predicting future games. The logistic win probability models had the highest accuracy with the best model yielding 78.89% accuracy.

Areas of future research would be further investigating the future game prediction models. Home field advantage could be estimated when constructing our models based on the in-game statistics. In the models we used, home field advantage was already incorporated in them. When predicting future games based off the three games medians, we will know whether or not the team is playing at home and can incorporate this into our prediction rather than just relying on past in-game statistics based on both home and away games. Therefore, one could assume you could increase the accuracy of the predictions if one controlled for home field advantage when trying to predict future games off of past in-game statistics.

It would also be interesting to develop a prediction model that only relied on rankings of team in-game statistics. The NCAA reports weekly rankings of team statistics such as; rushing offense, passing offense, pass defense, rushing defense, turnover margin, sacks, ect. Using these team rankings would take away the problem of multicollinearity when trying to estimate the effect of home field advantage when using in-game statistics in the model.

REFERENCES

Arnovitz, Kevin. [2012]. “On Bill James, the writer”. ESPN, http://espn.go.com/blog/truehoop/post/_/id/37682/on-bill-james-the-writer

Compughterratings. [2009-2012]. Retrieved February 16, 2013, from Compughterratings.com: http://www.compughterratings.com/CFB/ratings

Dizikes, D. [2013]. “Sports analytics: a real game-changer”. MITnews, http://web.mit.edu/newsoffice/2013/sloan-sports-analytics-conference-2013-0304.html

ESPN Scoreboard. [2009-2012]. Retrieved February 16, 2013, from ESPN.com: http://scores.espn.go.com/ncf/scoreboard?confId=40&seasonYear=2011&seasonType=2&weekNumber=2

Harville, D. [1977]. “The use of linear-model methodology to rate high school or college Teams”. *Journal of the American Statistical Association*, Vol 72, Issue 358.

Lewis, M. [2003] *Moneyball*. New York, NY: W.W. Norton & Company Inc.

Magel, R. C., Childress, G. [2012]. “Examining the Outcome Effects of the Turnover Margin in Professional Football”. *International Journal of Sports Science and Engineering*, Vol 6, Issue 3.

NCAA Stats. [2009-2012]. Retrieved February 16, 2013 from NCAA.org: http://web1.ncaa.org/stats/StatsSrv/rankings?doWhat=archive&sportCode=MFB

Willoughby, K. A. [2002] “Winning Games in Canadian Football: A Logistic Regression Analysis”. *The College Mathematics Journal*, Vol 33, Issue 3.

APPENDIX. SAS CODE

```
proc import out = WORK.FCS datafile= "G:\Grad_Classes\Thesis\2013\21March2013 FCS
DATA SET.xlsx"
          dbms=xlsx replace;
    sheet="sheet1";
    getnames=yes;
run;
*********/Logistic models/********************;
**/Model-1 Stepwise with all variables/**;
proc logistic data=FCS ;
          model win(event='1')=  week year home diff_top diff_yards diffrushyard diffpassyard
diff_int diff_fumble TOM diff_pen diff_pen_yards diff_pass_comp diff_sacks diff_first
diff_3rd_pct diffpunt_return diff_puntyards diffave_puntreturns diffave_yardsperplay
/noint selection=stepwise SLENTRY=.25 SLSTAY=.2;
Run;
**/Model-1 stepwise with reduced variables/**;
proc logistic data=FCS ;
          model win(event='1')= week year home diff_top ball_first  homemiles diff_sacks
diff_yards tom diff_pen diff_pen_yards diff_pass_comp diff_sack_yard diff_first diff_3rd_pct
diffpunt_return Diff_puntyards diffave_yardsperplay
/noint selection=stepwise SLENTRY=.25 SLSTAY=.05;
Run;
***/Model-1/***;
proc logistic data=FCS;
          model win(event='1')= TOM diff_pass_comp diff_sacks diff_3rd_pct diffpunt_return
diffave_yardsperplay/noint lackfit;
run;
***/Mode-1 with home added/***;
proc logistic data=FCS;
          model win(event='1')= home TOM diff_pass_comp diff_sacks diff_3rd_pct
diffpunt_return diffave_yardsperplay/noint lackfit;
run;
**************/ Logistic Prection with rank/*********************;
**/Model-2 stepwise /**;
proc logistic data=FCS ;
          model win(event='1') = Difference_in_rank week year home diff_top diff_yards
diffrushyard diffpassyard diff_int diff_fumble TOM diff_pen diff_pen_yards diff_pass_comp
diff_sacks diff_first diff_3rd_pct diffpunt_return diff_puntyards diffave_puntreturns
diffave_yardsperplay
/noint selection=stepwise SLENTRY=.25 SLSTAY=.2;
run;
**/Model-2 stepwise with reduced variables/**;
proc logistic data=FCS ;
          model win(event='1') =  Difference_in_rank tom  diff_sacks diff_pen_yards
diff_pass_comp diff_first diff_3rd_pct diffpunt_return Diff_puntyards   diffave_yardsperplay
```

```
/noint selection=stepwise SLENTRY=.25 SLSTAY=.05;
run;
***********/Model-2/******;
proc logistic data=FCS ;
        model win(event='1')= Difference_in_rank tom diff_pass_comp diff_sacks diff_3rd_pct
diffpunt_return diffave_yardsperplay/noint lackfit;
run;
*****/Point Spread model/**************;
**/Model-3 stepwise with all variables/**;
proc stepwise data=FCS;
        model spread = week year home diff_top diff_yards diffrushyard diffpassyard diff_int
diff_fumble TOM diff_pen diff_pen_yards diff_pass_comp diff_sacks diff_first diff_3rd_pct
diffpunt_return diff_puntyards diffave_puntreturns diffave_yardsperplay
/noint SLENTRY=.25 SLSTAY=.2;
run;
**/Model-3 stepwise with reduced variables/**;
proc stepwise data=FCS;
        model spread =   week home ball_first diff_top diff_sacks tom diff_pen diff_pen_yards
diff_first diff_3rd_pct diffpunt_return Diff_puntyards diffave_yardsperplay
/noint SLENTRY=.35 SLSTAY=.05;
run;
***/Model 3/***;
proc reg data=FCS;
        model spread = TOM diff_pen diff_first diff_3rd_pct diffpunt_return diff_puntyards
diffave_yardsperplay/noint;
run;
***/Model 3 with Home added/***;
proc reg data=FCS;
        model spread = TOM diff_pen diff_first diff_3rd_pct diffpunt_return diff_puntyards
diffave_yardsperplay/noint;
run;
***************/ Point spread Prection with rank/**********************;
**/Model-4 stepwise with all variables/**;
proc stepwise data=FCS;
        model spread = week year diff_top Difference_in_rank home ball_first  diff_yards tom
diff_pen diff_pen_yards diff_pass_comp diff_sack_yard diff_first diff_3rd_pct diffpunt_return
Diff_puntyards homemiles homeabove_375 homeabove_361 homemiles_256 homeabove_398
diffave_yardsperplay
/noint SLENTRY=.35 SLSTAY=.2;
run;
**/Model-4 stepwise with reduced variables/**;
proc stepwise data=FCS;
        model spread = diff_top diff_sacks Difference_in_rank home ball_first tom diff_pen
diff_pen_yards  diff_first diff_3rd_pct Diff_puntyards  diffpunt_return diffave_yardsperplay
/noint SLENTRY=.35 SLSTAY=.05;
run;
```

```
****/Model 4/*********;
proc reg data=fcs;
	model spread = Difference_in_rank TOM diff_pen Diff_first diff_3rd_pct diff_puntyards
diffpunt_return diffave_yardsperplay/noint;
run;
```