ANALYSIS OF SIGNIFICANT FACTORS IN DIVISION I MEN'S COLLEGE BASKETBALL AND DEVELOPMENT OF A PREDICTIVE MODEL

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

By

Samuel Paul Unruh

In Partial Fulfillment
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

April 2013

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Analysis of Significant Factors in Division I Men's College Basketball and Development of a Predictive Model

By

Samuel Paul Unruh

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. Ronald Degges

Dr. Megan Orr

Dr. Edward Deckard

Approved:

4/3/2013	Dr. Rhonda Magel
Date	Department Chair

ABSTRACT

While a number of statistics are collected during an NCAA Division I men's college basketball game, it is potentially of interest to universities, coaches, players, and fans which of these variables are most significant in determining wins and losses. To this end, statistics were collected from two seasons of games and analyzed using logistic and least squares regression methods. The differences between the two competing teams in four common statistics were found to be significant to determining victory: assists, free throw attempts, defensive rebounds, and turnovers. The logistic and least squares models were then used with data from the 2011-2012 season to verify the accuracy of the models. To determine the accuracy of the models in predicting future game outcomes, four prior game median statistics were collected for teams competing in a sample of games from 2011-2012, with the differences taken and used in the models.

ACKNOWLEDGMENTS

First, I must express my appreciation to the faculty members of the NDSU Statistics department. Before any work on this thesis was even attempted, every one of you provided wonderful instruction, support, and guidance to help me grasp the concepts put into place here. Thank you for the opportunity to continue my studies here.

I also owe a great deal of thanks to my parents and sister. You have always provided a great deal of support for any endeavor I’ve undertaken, and my time at graduate school has been no different.

Finally, I cannot express enough gratitude to my lovely wife, Kayla. Your confidence in me and constant support allowed me to get through the most difficult times of these past two years. You believed that I could do this before I did and none of this would have been possible without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

With 347 teams playing across 49 states (all but Alaska) in the 2012-2013 season, NCAA Division I men's college basketball is one of the most popular and widespread sports in the country. During the 2011-2012 basketball season, a total of 27,691,051 people attended 5,335 total Division I men's basketball games (NCAA, 2012). To add to its popularity, the NCAA tournament in March and April of every season attracts incredible national attention. The 2011 NCAA tournament drew the highest television rankings for an opening week of the tournament in 20 years, in addition to the 2.4 million unique visitors to the NCAA's website, where tournament games can be streamed live to computers or smartphones (NCAA, 2011).

With such a large amount of popularity and attention being paid to the sport, a number of statistics are kept at every single game for use by universities, coaches, players, and casual fans. However, with such an abundance of information, questions naturally arise – which of these statistics is the most important? What does my team need to do well to improve its chances of winning a contest? What are my team's chances of winning an upcoming game?

The primary objective of this work will be to determine key factors that explain victory or defeat in a Division I men's college basketball game. This work can benefit coaches, teams, and even casual fans, as they can then focus on these principal areas of the game as they tend to lead to victories.

A secondary objective of this work will be to identify if a model can be developed using the significant factors that are identified to predict the outcomes of future games using previous game data from the teams involved in the contest. A model such as this could be of use to coaches and teams who are approaching an upcoming game. If a team is approaching a future game knowing that they are weak in a key area when compared to their future opponent, they can make necessary changes to game plans to either improve that weakness or focus on other skills to offset that weakness.

The primary focus of this work was placed on differences in statistics between the two teams involved in a basketball game. The reason behind this is that, to win a basketball game, a team does not necessarily need to do well, simply better than their opponent.

CHAPTER 2. REVIEW OF LITERATURE

With college basketball having such national popularity, and a number of statistics kept regularly, it naturally attracts a great deal of statistical attention and analysis. In reviewing previous works regarding the topic of significant factors in college basketball, three were found that related to this work.

In 1994, David Harville and Michael H. Smith conducted an analysis to determine whether or not home-court advantage was a significant factor in college basketball, and if so, determine the advantage in points it gave a team over playing at a neutral site. Game data was restricted to regular season games, and consisted of 1,678 games played during the 1991-1992 Division I college basketball season.

It was found that home-court advantage varied from team to team, but estimated the advantage given to teams playing at home as compared to a neutral site game to be 4.68 ± 0.28 points. They also discovered that while home-court advantage varied among teams, there was no positive or negative relationship between having a strong home-court advantage and overall performance level of a team – good teams could have lower home-court advantages when compared to all teams, and poor teams could have strong home-court advantages.

Neil Schwertman, Kathryn Schenk, and Brett Holbrook conducted research published in 1996 regarding the development of probability models for NCAA regional basketball tournaments. This work attempted to estimate the probability of any given team winning their regional tournament, thus advancing to the 'Final Four'. To do this, probability models were developed using NCAA regional tournament games from 1985-1994, a total of 600 games. However, the independent variable under consideration in these probability models was a team's seed in the NCAA tournament, information that would not be known during a randomly played regular season contest.

In 2004, Dean Oliver performed analysis related to what he referred to as the "four factors". These factors were elements of the game that he felt teams needed to successfully execute to increase their probabilities of winning a basketball game. This analysis was based on data from both NBA and NCAA basketball games. The factors were identified in his work to be shooting percentage, offensive rebounds, turnovers, and high numbers of free throw attempts combined with high free throw percentage. Also mentioned was the fact that none of these statistics will tell how well a player creates a good shot, which Oliver felt is a critical factor in any basketball game.

On rebounding in particular, Oliver noted that rebounding does not appear to be as valuable as shooting, getting to the free throw line, and good ball control in the NBA. However, he mentioned that in high school and college basketball, rebounding may play a larger role in influencing a team's probability of victory.

CHAPTER 3. METHODS

3.1. Data Collection

Regression methods were used to determine significant factors affecting the outcomes of NCAA Division I men's basketball games. Two models were developed; one using a logistic regression approach with responses recorded as a '1' for a win and '0' for a loss, and the second utilizing least squares regression with point spread as a response.

The data needed for the identification of significant factors was collected from a random sample of box scores provided by the NCAA (NCAA 2013). This random sample consisted of 150 games chosen from both the 2009-2010 and 2010-2011 seasons. For each season, 30 teams were selected at random, and from those teams, five games of data were selected. For the 2009-2010 season, games 7, 13, 15, 23, and 26 were selected. For the 2010-2011 season, games 5, 11, 15, 19, and 21 were selected. Any game that was played against a non-Division I opponent was discarded from consideration, along with any neutral site games, bringing the total number of games observed to 280.

For each game, it was observed whether or not the team that was selected randomly (hereafter referred to as the 'team of interest') was playing at home or on the road, with home being recorded as '1' and road recorded as a '0'. Also recorded was the point spread with respect to the team of interest, with positive values indicating they had won the game and negative indicating that they had lost. For example, if the team of interest had lost to their opponent by 10 points, point spread would be recorded as '-10'. To go along with the point spread variable, an indicator variable was kept separately signifying whether or not the team of interest won, classified as a '1', or lost, classified as a '0'.

To go along with win/loss, home/away, and point spread data, the following variables in Table 3.1 were collected for both the team of interest and their opponent. Following the collection of the variables for each of the games, the differences of all variables were taken. For example, if the team of interest committed five turnovers and their opponent committed seven, the difference would be recorded as '-2'. Only the differences would be under consideration by the models developed, since the primary interest is in comparing two teams.

Table 3.1. Variables Under Consideration

Number of Free Throws Attempted	Number of Players Fouled Out
Number of Offensive Rebounds	Number of Fouls Committed by Starters
Number of Defensive Rebounds	Number of Turnovers Committed
Number of Assists	Number of Steals
Number of Blocks	Number of Fouls
Home/Away Indicator Variable	Number of Field Goals Attempted

3.2. Identification of Significant Factors and Development of Initial Models

3.2.1. Development of Point Spread Model

To determine the significant factors that help predict win and loss, the method of least squares regression was utilized. The response variable for this model was point spread with respect to the team of interest, where positive values indicate a win for the team of interest and negative values indicate a loss. To select the significant independent variables, all differences were placed under consideration. The decision was reached to not include an intercept in the model, due to the nature of the data. Since only differences were under consideration, if the team of interest and their opponent were a perfect duplicate of each other in every regard, all differences would equal zero, and one would expect a tie game (a point spread of zero); this necessitates an intercept of zero as well.

The generalized least squares model to fit for this case will be $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is a vector of point spreads with respect to the team of interest, $\boldsymbol{x}$ will be a matrix consisting of independent significant factors to determine point spread, $\boldsymbol{\beta}$ is the vector of coefficients corresponding to the independent factors, and $\boldsymbol{\varepsilon}$ consisting of random error.

To determine the least squares model to be used, stepwise selection was utilized with an α value of .10 for both entry and exit. The differences of the variables listed in Table 3.1 for the team of interest and their opponent were considered for entry in the model. This method begins with none of the independent variables being considered. It then begins adding variables to the model one at a time if they are significant at the .10 level with the most significant variable being added first. Each time a variable is entered into the model, the selection method then rechecks to see if all variables are still significant at the .10 level. If at least one variable is no longer significant, the variable least significant is then removed from the model. This process continues until all variables are significant at the .10 level, and no further variables can be added or removed. (SAS Institute Inc., 2010)

3.2.2. Development of Logistic Regression Model

Logistic regression fits well with the nature of this data, as games can be easily coded 0 for a loss and 1 for a win. To go along with the point spread model utilizing least squares regression, a logistic regression model was also fit to the data, using win/loss as the response variable. The logistic regression model will take on the form $\pi(\boldsymbol{x_i}) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$ where $\boldsymbol{x}_i'\boldsymbol{\beta} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ (Abraham & Ledolter, 2006).

Again, similarly to the development of the point spread model, no intercept will be used in the development of the logistic regression model. If a team were to play a perfect copy of themselves in a game, all differences would equal zero, meaning $\boldsymbol{x_i} = \mathbf{0}$, and $\pi(\boldsymbol{x}_i) = 0.50$, as it should.

To determine the significant variables in developing a logistic regression model, stepwise selection was used, with $\alpha = .10$ to enter and exit consideration. Significance of individual variables was determined using a Wald test, where:

$$t = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$$

Small p-values for this test indicate that β_j is significantly different from zero, and therefore x_j should be left in the model.

3.3. Verification of Significant Factors

Following the development of both the least squares regression point spread model and the logistic regression model, new data was collected to verify that the models were performing well in predicting wins given new data not associated with the development of the models. Both models were verified with a random sample of data from the 2011-2012 season, which was not used in development of either the point spread or logistic model.

Additional data was collected from the NCAA website in the following way. Similar to the previous data collection, 30 teams were selected at random, and 5 games for each team randomly selected. Games numbered 2, 8, 12, 13, and 17 were selected for each team. All games played at neutral sites or against non-Division I opponents were discarded, for a total of 132 games. The point spread with respect to the team of interest was recorded, along with a binary indicator for win/loss. Each of the previously identified significant variables was then recorded for both the team of interest and their opponent, and the difference taken and recorded.

To verify that the identified significant variables were indeed important in determining wins and losses for the point spread model, the variables for each game were placed in the least squares model. The estimated response y_i was then observed. If $y_i > 0$, indicating the point spread was predicted to be positive, the game was coded as a win for the team of interest. If $y_i < 0$, indicating the point spread was predicted to be negative with respect to the team of interest, it was coded as a loss. The predicted win or loss of the game was then compared to the actual win or loss of the game, and the results tabulated.

To verify the variables were important in determining wins and losses using the logistic regression model, a similar process was followed. For each of the games collected, the identified significant variables were placed in the logistic regression equation, and the predicted probability of victory $\pi(\boldsymbol{x_i})$ determined. If $\pi(\boldsymbol{x_i}) > 0.50$, the game was coded as a victory for the team of interest, since they were predicted to have the better probability of victory than their opponent. Likewise, if $\pi(\boldsymbol{x_i}) < 0.50$, the game was coded as a defeat for the team of interest. The predicted win or loss of each game was then compared to the actual result, and the accuracy noted.

3.4. Accuracy of Initial Models in Predicting Future Games

A secondary goal of this work was to determine if either of the models developed could be used in predicting future games based on previous games for both the team of interest and their opponent. To determine this accuracy, data was collected from the NCAA for the 2011-2012 season, with a completely random sample of 100 Division I games being selected. For each game selected in the sample, data was collected on the identified significant variables for four previous games for each the team of interest and their opponent. The median of the four previous games' significant factors was then calculated for each team. Four game medians were used as opposed to four game averages as the averages were more sensitive to a skewed value, i.e., a team performing exceptionally well or poorly for one of the four previous games.

After the medians for each statistic had been found for the team of interest and their opponent, the differences of these medians was calculated, and the differences placed into both the point spread and logistic regression models to find the predicted value for point spread, y_i, and probability of victory, $\pi(\boldsymbol{x_i})$. The predicted point spread and probability of victory were then determined, with a win predicted if $y_i > 0$ for the point spread model and a win predicted if $\pi(\boldsymbol{x_i}) > 0.50$ for the logistic regression model. The predicted values for victory were then compared with the actual record of win/loss for each game, and the results compared for each model.

3.5. Development of Predictive Model Using Prior Games and Accuracy

In an attempt to improve the accuracy of using the identified significant factors in predicting future game outcomes, a new predictive model was developed. Two models were developed similarly to previous steps, a least squares regression model with point spread as the response variable and a logistic regression model with win(1)/loss(0) as the response variable. The independent variables used to develop both of these models consisted of the four game medians of previous games' statistics, rather than single game values used to develop the initial models.

After development of the new predictive models, the accuracy of the least squares and logistic regression models was checked against a random sample of 75 games from the 2012-2013 season. For each of the 75 randomly selected games, statistics for the significant factors were collected for the four previous games of both the team of interest and their opponent. Medians of these statistics were then found, and the difference taken and placed into the new predictive models developed to find a predicted point spread, y_i, and predicted probability of victory for the team of interest, $\pi(\boldsymbol{x_i})$. If $y_i > 0$, a predicted win for the point spread model was coded. If $\pi(\boldsymbol{x_i}) > 0.50$, a predicted win for the logistic regression model was coded. These were then compared against the actual win/loss values for each of the 75 games, and the accuracy of both models noted.

CHAPTER 4. RESULTS

4.1. Identification of Significant Factors and Development of Initial Models

4.1.1. Development of Point Spread Model

To develop the point spread model, the method of least squares regression was used with point spread as the dependent variable, and the full list of independent variables under consideration. Stepwise selection methods were used to identify variables that were significant with $\alpha = .10$ to enter or exit the model, and the results are summarized in Table 4.1.

Table 4.1. Summary of Stepwise Selection for Point Spread Model

Step	Variable Entered	Variable Removed	Partial R-Square	Model R-Square	F Value	P Value
1	Assists		0.5918	0.5918	404.50	<.0001
2	Free Throw Attempts		0.1022	0.6940	92.90	<.0001
3	Defensive Rebounds		0.0504	0.7445	54.64	<.0001
4	Turnovers		0.1579	0.9024	446.29	<.0001
5	Field Goal Attempts		0.0109	0.9132	34.52	<.0001
6	Offensive Rebounds		0.0017	0.9150	5.58	0.0188

While the stepwise selection method did select six variables as being significant below the $\alpha = .10$ level, the variables for the differences in field goal attempts and offensive rebounds do not contribute a great deal to the overall model r-square (.0109 and .0017, respectively). For this reason, they were removed from consideration and the model refit using assists, free throw attempts, defensive rebounds, and turnovers. The parameter estimates for this regression model are listed in Table 4.2.

Table 4.2. Point Spread Model Parameter Estimates

Variable	Parameter Estimate	Standard Error	F Value	P Value
Free Throw Attempts (FTA)	0.06175	0.03256	3.60	0.0589
Defensive Rebounds (DR)	1.48572	0.06552	514.23	<.0001
Assists (A)	0.58736	0.06961	71.21	<.0001
Turnovers (TO)	-1.60131	0.07580	446.29	<.0001

The final least squares regression model involving point spread as the response variable is then given by $y = 0.06175(FTA) + 1.48572(DR) + 0.58736(A) - 1.60131(TO)$. The coefficients indicate that the most influential factor in determining point spread is the difference in turnovers. For each turnover a team commits more than their opponent, the model indicates a loss of 1.6 points. Similarly, the difference in defensive rebounds is very influential, with each defensive rebound a team acquires more than their opponent worth an increase of 1.49 points.

To verify that the regression model satisfies the assumptions of residuals following a normal distribution with a mean of zero and a constant variance across all residuals, the following diagnostic plots were assembled. As the residuals satisfy all assumptions, it is assumed the model is valid. The diagnostic plots for the residuals are shown in Figure 4.1.

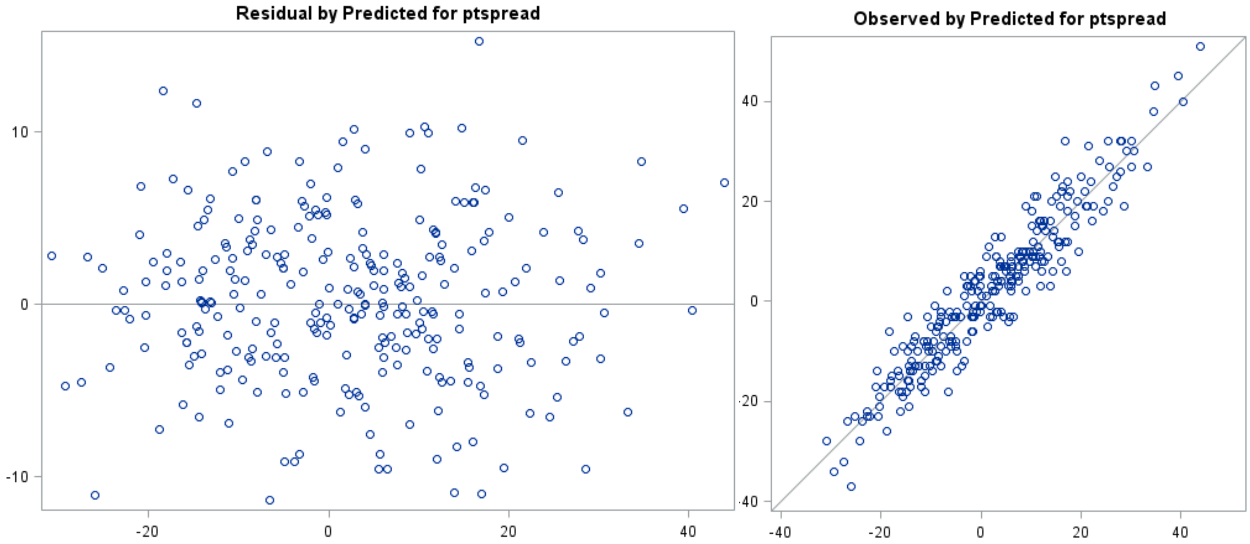

Figure 4.1. Residual Diagnostic Plots for Point Spread Model

4.1.2. Development of Logistic Regression Model

The development of the logistic regression model uses win, coded as a '1', or loss, coded as a '0', as the response variable, with all independent variables initially under consideration. Stepwise selection was used to isolate only significant variables, with an α level of .10 for entry and exit.

Table 4.3. Summary of Stepwise Selection for Logistic Regression Model

Step	Effect Entered	DF	Score Chi-Square	P Value
1	**Assists**	1	101.7818	<.0001
2	**Free Throw Attempts**	1	78.7975	<.0001
3	**Defensive Rebounds**	1	21.1658	<.0001
4	**Turnovers**	1	29.2154	<.0001

The same four variables that were selected as significant in the point spread least squares regression model are also significant in the logistic regression model. To verify a good fit for the logistic regression model, a Hosmer-Lemeshow goodness-of-fit test was conducted, revealing a p-value of $.9149$. Therefore, it cannot be rejected that this logistic regression model provides a good fit for explaining wins and losses.

Table 4.4. Parameter Estimates for Logistic Regression Model

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	P Value
Free Throw Attempts ($\boldsymbol{FTA}$)	1	0.1233	0.0358	11.8676	0.0006
Defensive Rebounds ($\boldsymbol{DR}$)	1	0.4875	0.0909	28.7681	<.0001
Assists ($\boldsymbol{A}$)	1	0.3629	0.0840	18.6609	<.0001
Turnovers ($\boldsymbol{TO}$)	1	-0.4737	0.1002	22.3472	<.0001

Given the parameter estimates for the logistic regression model, the final model for estimated probability of victory is as follows:

$$\pi(FTA, DR, A, TO) = \frac{e^{0.1233(FTA)+0.4875(DR)+0.3629(A)-0.4737(TO)}}{1+e^{0.1233(FTA)+0.4875(DR)+0.3629(A)-0.4737(TO)}}$$

4.2. Verification of Significant Factors

To verify that indeed the variables identified in both the point spread and logistic regression models are significant, the models were used with data from the 2011-2012 season that was not used in the creation of either model. The differences between the two teams were calculated and used in the model to compare predicted victories with actual victories.

Table 4.5. Example Data Entry

Team A	Team B	Point Spread	Win?	FTA	DR	A	TO
UC Riverside	UTSA	5	1	-1	5	7	4

Table 4.5 represents a data entry from a game played between UC Riverside and UTSA on December 28, 2011. All columns are calculated with respect to UC Riverside (the team of

interest), meaning UC Riverside won by 5 points, had 1 fewer free throw attempt, 5 more defensive rebounds, 7 more assists, and committed 4 more turnovers than UTSA.

Using the least squares regression model already developed, UC Riverside had a predicted point spread of:

$$y = 0.06175(-1) + 1.48572(5) + 0.58736(7) - 1.60131(4) = 5.07$$

Since the predicted point spread is greater than zero, this game was coded as a (correctly) predicted win for UC Riverside, who won the game by a score of 73-68.

Using the logistic regression model, UC Riverside had a projected probability of victory of:

$$\pi(FTA, DR, A, TO) = \frac{e^{0.1233(-1)+0.4875(5)+0.3629(7)-0.4737(4)}}{1 + e^{0.1233(-1)+0.4875(5)+0.3629(7)-0.4737(4)}} = 0.951$$

Since this projected probability of victory is greater than 0.50, this game was also coded as a predicted win for UC Riverside.

This process was then repeated for a sample of 132 games, with the number of predicted victories and defeats from each model being compared to the actual victories and defeats from the sample of games. The accuracy of each model is noted in Table 4.6.

Table 4.6. Accuracy of Original Models

Logistic		Predicted		
		Win	Loss	Total
Actual	Win	60	3	63
	Loss	4	65	69
	Total	64	68	132

Point Spread		Predicted		
		Win	Loss	Total
Actual	Win	59	4	63
	Loss	3	66	69
	Total	62	70	132

As is shown in Table 4.6, both the logistic regression and point spread models are highly accurate at predicting the winner of games based on the identified significant factors. Both models had an accuracy of $\frac{125}{132} = 94.7\%$, indicating that the variables identified are indeed significant to determining wins and losses in a Division I college basketball game.

4.3. Accuracy of Initial Models in Predicting Future Games

Next, to determine if the logistic or point spread models were useful in predicting games in advance of being played, a sample of 100 games from the 2011-2012 season was used. Game statistics from four games prior were collected for both the team of interest and their opponent for each of the significant variables already identified.

Table 4.7. 4 Game Median Example

4 game Statistics				
Team	FTA	DR	A	TO
Air Force	18	22	12	16
Air Force	19	25	11	9
Air Force	22	20	14	7
Air Force	22	28	21	15
San Diego St.	15	24	14	6
San Diego St.	20	29	16	16
San Diego St.	18	28	17	14
San Diego St.	32	23	13	10

Medians				
Team	FTA	DR	A	TO
Air Force	20.5	23.5	13	12
San Diego St.	19	26	15	12
Difference	**1.5**	**-2.5**	**-2**	**0**

Table 4.7 represents data for a randomly selected game between Air Force and San Diego St. played on January 21, 2012. For each of the teams, the significant statistics were collected for the previous four games they had played. Then for each team, the medians were found, and the differences taken. Using the differences of the medians, the predicted point spread was:

$$y = 0.06175(1.5) + 1.48572(-2.5) + 0.58736(-2) - 1.60131(0) = -4.796$$

Since the projected point spread was less than zero, the game would be predicted (in this case, correctly) as a loss for Air Force.

Using the differences of the medians, the projected probability of victory for Air Force was given by:

$$\pi(FTA, DR, A, TO) = \frac{e^{0.1233(1.5)+0.4875(-2.5)+0.3629(-2)-0.4737(0)}}{1 + e^{0.1233(1.5)+0.4875(-2.5)+0.3629(-2)-0.4737(0)}} = 0.147$$

Again, since $\pi(\boldsymbol{x}) < 0.50$, the game would be predicted as a loss by Air Force. In this instance, both models correctly predicted the game, as the outcome was a 13 point loss by Air Force.

This process was repeated for the 100 games selected randomly from the 2011-2012 seasons, and the accuracy of predicting future games recorded for both the logistic regression model and point spread least squares regression model. The accuracy of both models is noted in Table 4.8.

Table 4.8. Accuracy in Predicting Future Games by Original Models

Logistic		Predicted		
		Win	Loss	Total
Actual	Win	33	15	48
	Loss	17	35	52
	Total	50	50	100

Point Spread		Predicted		
		Win	Loss	Total
Actual	Win	29	19	48
	Loss	17	35	52
	Total	46	54	100

As can be seen from Table 4.8, both the logistic and point spread models struggled to predict future games based on prior game median data. The logistic regression model correctly predicted $\frac{68}{100} = 68\%$ of games, while the point spread model correctly predicted $\frac{64}{100} = 64\%$ of games. While this may seem like a moderately acceptable percentage, simply picking the home team to win in every game resulted in a $\frac{66}{100} = 66\%$ accuracy rate for predicting games from this sample.

4.4. Development of Predictive Models and Accuracy in Predicting Future Games

While the original logistic and point spread models did not do an outstanding job at predicting future games, it was not wholly unexpected. They were not created expressly for that purpose. To further explore the concept, the four previous game medians that were calculated from the previous section were used as independent variables in developing new models; a new predictive point spread least squares regression model and new predictive logistic regression model would be developed using this data.

To go along with these four independent variables, a fifth was added – an indicator variable for home, coded as a 1, or away, coded as a 0. The rationale behind this is that, while looking at the box score for a game that has already been played, as done in prior steps, which team was the home team and which was away is explained fairly well by the four statistics chosen as significant. However, going into a game that has yet to be played, there is no possibility that the four game median statistics will predict which will be home and which will be away.

4.4.1. Development of Predictive Least Squares Model

The predictive least squares model would be generated using point spread as the dependent response variable, with five independent variables: home/away, and the differences of the median statistics calculated for free throw attempts, defensive rebounds, assists, and turnovers. Stepwise selection method was again employed here, with a slightly more generous value of $\alpha = .15$ to enter or exit the model.

While both of the variables selected by the stepwise selection procedure are significant at the $\alpha = .15$ level, it is worth noting this model produced a very low value of r-square, 0.1239. This indicates that very little variation in point spread is explained by the model.

Table 4.9. Parameter Estimates for Predictive Point Spread Model

Variable	Parameter Estimate	Standard Error	F Value	P Value
Home	4.87870	1.75742	7.71	0.0066
Turnovers	-0.91125	0.39675	5.28	0.0238

Therefore, the predictive point spread model is given by:

$$y = 4.8787(Home) - 0.91125(TO)$$

4.4.2. Development of Predictive Logistic Regression Model

The predictive logistic regression model was formulated using win/loss as the response variable, with the differences of the medians of the four significant statistics as the independent variables, along with home/away. Stepwise selection method was used to determine significant variables, with $\alpha = .15$ to enter or exit the model.

Table 4.10. Parameter Estimates for Predictive Logistic Regression Model

Parameter	DF	Estimate	Wald Chi-Square	P Value
Assists	1	0.1156	3.4059	0.0650
Turnovers	1	-0.1239	3.8905	0.0486

From Table 4.10, it can be seen that the selected predictive logistic regression model is given by the following equation:

$$\pi(A, TO) = \frac{e^{0.1156(A) - 0.1239(TO)}}{1 + e^{0.1156(A) - 0.1239(TO)}}$$

Both variables selected by the stepwise selection procedure are significant below the $\alpha = .15$ level. A Hosmer-Lemeshow test for goodness-of-fit was also conducted, yielding a p-value of 0.1273. At the standard α value of $.05$, goodness-of-fit cannot be rejected; however, it does indicate the model may not be a very good fit.

4.4.3. Accuracy of Predictive Models

To determine the accuracy of the predictive models, a random sample of 75 games was selected from the 2012-2013 season. The procedure for data collection was similar to that used in Section 4.3 of this work. For each of the 75 games, the accuracy of the predictive models was assessed, along with the original point spread and logistic model accuracy over the same sample for sake of comparison.

Table 4.11. Prediction Accuracy across All Models 2012-2013

Predictive Point Spread		Predicted		
		Win	Loss	Total
Actual	Win	31	13	44
	Loss	18	13	31
	Total	49	26	75
Overall Accuracy			58.67%	

Predictive Logistic Regression		Predicted		
		Win	Loss	Total
Actual	Win	28	16	44
	Loss	8	23	31
	Total	36	39	75
Overall Accuracy			68%	

Original Point Spread		Predicted		
		Win	Loss	Total
Actual	Win	26	18	44
	Loss	10	21	31
	Total	36	39	75
Overall Accuracy			62.67%	

Original Logistic Regression		Predicted		
		Win	Loss	Total
Actual	Win	28	16	44
	Loss	9	22	31
	Total	37	38	75
Overall Accuracy			66.67%	

Table 4.11 indicates that, over this sample, the predictive models that were developed using four game medians, along with home/away information, did not perform significantly better than the original models developed using single game information.

CHAPTER 5. CONCLUSIONS

The primary objective of this work was to identify the key factor or factors that most heavily influenced a team's propensity to win or lose a Division I men's college basketball game. Using least squares regression with point spread as a response variable, it was shown that outperforming an opponent in each of four factors influence a team's likelihood of winning a basketball game: free throw attempts, defensive rebounds, assists, and turnovers. Likewise, using logistic regression with win/loss as a response variable, these four variables were determined to be significant in affecting a team's probability of winning a Division I college basketball game. Using data from the 2011-2012 season not used in development of either model, it was shown that these four variables are indeed highly influential in affecting victory for a given team.

A secondary goal of this work was to determine if these four key variables could be used to predict a game prior to its occurrence using previous game data for each team involved in the contest. It was determined that while the original models developed, both point spread and logistic, were moderately adequate at predicting future game outcomes (64% and 68%, respectively), they remained no better than simply predicting the home team to win every contest over the same sample (66% accurate).

In an attempt to improve the accuracy, new predictive models were developed using prior four game medians and a home/away indicator variable as independent variables. However, both the new predictive logistic regression model and predictive point spread model failed to improve on the accuracy of the original models.

This seems to indicate that while the four variables identified are very significant in explaining the outcome of a game, it is difficult to estimate the future values of these variables ahead of a game occurring. Future research could involve the development of a predictive model with the inclusion of additional team information, such as strength of schedule, RPI, etc. Inclusion of this data may increase the accuracy of the predictive models.

WORKS CITED

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

Harville, D., & Smith, M. (1994). *The Home-Court Advantage: How Large Is It, and Does It Vary From Team to Team?*. The American Statistician, 48(1), 22-28.

NCAA. (2011, March 22). *Men's basketball tournament ratings continue to rise*. Retrieved from http://www.ncaa.org/wps/wcm/connect/public/NCAA/Resources/Latest+News/2011/March/Mens+basketball+tournament+ratings+continue+to+rise (Last Accessed 26 March 2013)

NCAA. (2012). *2012 NCAA Men's Basketball Attendance*. Retrieved from http://fs.ncaa.org/Docs/stats/m_basketball_RB/Reports/attend/2012.pdf (Last Accessed 26 March 2013)

NCAA. (2013). *Game by game stats*. Retrieved from http://stats.ncaa.org/team/inst_team_list (Last Accessed 26 March 2013)

Oliver, D. (2004). *Basketball on paper*. (1st ed.). Dulles, VA: Brassey's, Inc.

SAS Institute Inc. (2010). Model-selection methods. *SAS/STAT(R) 9.22 User's Guide.* Cary, NC.

Schwertman, N., Schenk, K., & Holbrook, B. (1996). More Probability Models for the NCAA Regional Basketball Tournaments. *The American Statistician*, 50(1), 34-38.

APPENDIX. SAS CODE

```
/* Imports original data */
proc import datafile='C:\Users\samuel.unruh\Documents\Thesis\Final Thesis
Files\Final Original Data.csv'
      out=games;
run;

/* Generates least squares regression point spread model */
Proc Reg Data=Games plots=all;
      Model ptspread = fta offreb defreb assists turnovers steals blocks
                       fouls starterfouls fouledout fga home
            /selection=stepwise slentry=.10 slstay=.10 noint;
            output out=reg r=res cookd=cooks;
run;

/* Generates logistic regression model */
proc logistic data=Games plots=all;
       model win(event='1')= fta offreb defreb assists turnovers steals
       blocks fouls starterfouls fouledout fga
       / selection=stepwise ctable rsquare noint lackfit slstay=.10
       slentry=.10;
            output out=test p=prob;
run;

/* Imports previous four game median data */
proc import datafile='C:\Users\samuel.unruh\Documents\Thesis\Final Thesis
Files\Final Predictive Data.csv'
      out=predictive;
run;

/* Generates new predictive logistic regression model */
proc logistic data=predictive plots=all;
       model win(event='1')= home ftamed defrebmed assistmed turnovermed
            /selection=stepwise noint ctable slstay=.15 slentry=.15 lackfit;
            output out=test p=prob ;
run;

/* Generates new predictive point spread model */
Proc reg Data=predictive plots=all;
      Model spread = home ftamed defrebmed assistmed turnovermed
            /selection=stepwise slentry=.15 slstay=.15 details adjrsq noint;
run;
```