*Open Journal of Statistics*, 2013, 3, 225-230
http://dx.doi.org/10.4236/ojs.2013.34026 Published Online September 2013 (http://www.scirp.org/journal/ojs)

# Determining Factors Influencing the Outcome of College Basketball Games

**Rhonda Magel*, Samuel Unruh**
Department of Statistics, North Dakota State University, Fargo, USA
Email: *Rhonda.magel@ndsu.edu

Received June 6, 2013; revised July 6, 2013; accepted July 13, 2013



## 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 to which these statistics 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 models were then used with data from the 2011-2012 Season to verify the accuracy of the models. The point spread model was also used with 2013 March Madness game statistics.

**Keywords:** NCAA Men's Basketball; March Madness

## 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 USA. During the 2011-2012 basketball season, a total of 27,691,051 people attended 5,335 total Division I men's basketball games [1]. 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 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 [2].

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?

One variable that does appear to affect the outcome of a college basketball game is whether or not a team is playing at home. Home-court advantage in college basketball has been studied by Harville and Smith [3]. Harville and Smith collected data on 1678 Division I college basketball games played during the 1991-1992 regular season. They estimated home court advantage given to teams playing at home as compared to a neutral site to be 4.68 ± 0.28 points. They also found no positive or negative relationship between a strong home court advantage and the performance of the team. Namely, good teams could have a strong or weak home court advantage and poor teams could have a strong or weak home court advantage.

Schwertman, Schenk, and Holbrook [4] have worked on developing probability models to estimate the probability of any given team winning their regional tournament advancing to the "Final Four". To develop their probability models, they considered 600 games from the NCAA tournaments held in 1985 to 1994 NCAA. One variable entered into their models was the team's overall seed in the tournament. This would not be known when trying to predict a winner of a basketball game during the regular season.

Oliver [5] analyzed both NCAA and NBA basketball games and determined four factors of basketball games that teams should work on to increase their probability of winning the game. The four factors that he identified were: increase shooting percentage, increase the number

*Corresponding author.

of offensive rebounds, decrease the number of turnovers, and increase the number of free throw attempts along with increasing the free throw percentage. Oliver [5] noted that rebounding appeared to have a lesser role than the other factors in the NBA, but not necessarily at the college level.

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, along with the weights of these key factors. 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.

## 2. Methods Used

Two regression methods were used to develop models to determine key factors explaining outcomes in Division I men's college basketball games. Least squares regression was used in developing a model to explain the point spread in the final scores of a basketball game. Logistic regression was used in developing a model to estimate the probability of winning a game knowing the values of the key factors. A random sample consisting of 150 games chosen from both the 2009-2010 Season and the 2010-2011 Season was taken to use in developing the two models. For each of these seasons, 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 in the sample to 280.

For each of the 280 games in the sample, the team that was randomly selected to be in the sample is referred to as the "team of interest", and all of the statistics for that game are recorded are in the order the value for the "team of interest" minus the value for the "opposing team". Point spread was the dependent variable collected for the least squares model. If the value of point spread was 10, this meant that the "team of interest" had won the game by 10 points. If the value of point spread was −2, this meant that the "team of interest had lost the game by 2 points. The dependent variable collected for the logistic regression model was in terms of whether the "team of interest'" won the game (recorded as a "1"), or whether the "team of interest" lost the game (recorded as a "0"). Game statistics were retrieved from the NCAA website [6].

The following is a list of all the independent variables considered for entry into either of the two models:

- Home indicator variable for the "team of interest";
- Difference in number of free throws attempted;
- Difference in number of offensive rebounds;
- Difference in number of defensive rebounds;
- Difference in number of assists;
- Difference in number of blocks;
- Difference in number of players fouled out;
- Difference in number of fouls committed by starters;
- Difference in number of turnovers committed;
- Difference in number of steals;
- Difference in number of fouls; and
- Difference in number of field goals attempted.

The differences listed above are always in the order "team of interest" minus "opposing team".

### 2.1. Development of Point Spread Model

The dependent variable in this model was the point spread between the "team of interest" and the "opposing team". Stepwise regression was used in SAS to help develop the model with the $\alpha$ value set at 0.10 for entry and exit. The twelve variables previously mentioned were considered for entry into the model. Stepwise regression selected the six variables as given in **Table 1** for the model [7,8].

The model was developed with the intercept set to zero. This is because if the point spread of the game between Team A and Team B was 10, then the point spread between Team B and Team A, should be −10. While the stepwise regression procedure did select field goals attempts (differences) and offensive rebounds (differences) to add to the model, the addition of these two variables, did not contribute very much to the overall R-square (0.0109 and 0.0017, respectively). For this reason, these two variables were removed from consideration in the model and the model was refit using the remaining four variables. The parameter estimates for this regression model are given in **Table 2**.

The final least squares regression model involving point spread, Y as the response variable is then given by

**Table 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 | <0.001 |
| 2 | Free Throw Attempts | | 0.1022 | 0.6940 | 92.90 | <0.001 |
| 3 | Defensive Rebounds | | 0.0504 | 0.7445 | 54.64 | <0.001 |
| 4 | Turnovers | | 0.1579 | 0.9024 | 446.3 | <0.001 |
| 5 | Field Goal Attempts | | 0.0109 | 0.9132 | 34.52 | <0.001 |
| 6 | Offensive Rebounds | | 0.0017 | 0.9150 | 5.58 | 0.019 |

 

**Table 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 (DA)** | 1.48572 | 0.06552 | 514.23 | <0.0001 |
| **Assists (A)** | 0.58736 | 0.06961 | 71.21 | <0.0001 |
| **Turnovers (TO)** | −1.60131 | 0.07580 | 446.29 | <0.0001 |

$$\mathrm{Y} = 0.062(\mathrm{FTA}) + 1.49(\mathrm{DR}) + 0.587(\mathrm{A}) - 1.601(\mathrm{TO}) \quad (1)$$

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 receives more than their opponent worth an increase of 1.49 points. The assumptions of the model were verified by examining residual plots.

## 2.2. Development of Logistic Model

The dependent variable was coded "0" if the "team of interest" lost and "1" if the "team of interest" won the game. A stepwise logistic regression analysis was conducted with the same eleven difference variables under consideration as before in the point spread model in addition to the indicator variable for home and away. Both the $\alpha$ entry level and the exit level were set to 0.10. As in the point spread model, the $\beta_0$ term in the model was assumed to be zero since if the model estimates the probability of Team A winning the game to be 0.60 when Team A and Team B are playing, the probability of Team B winning the game should be estimated at 0.40.

The same four variables that were selected in the end for the point spread model were also selected in the logistic regression model. A summary of the stepwise selection procedure for the logistic regression model is given in **Table 3**. To verify that the logistic model was a good fit for the data, a Hosmer-Lemeshow goodness-of-fit test was conducted. The P-value for this test was found to be 0.9149. This indicates that the hypothesis of the model being a good fit can't be rejected.

The parameter estimates associated with each of the variables in the logistic regression model are given in **Table 4**. Given the parameter estimates for the logistic regression model, the final model for estimating the probability of victory, is as follows:

$$\pi(\mathrm{FTA},\mathrm{DR},\mathrm{A},\mathrm{TO}) = \exp\left(0.123(\mathrm{FTA}) + 0.488(\mathrm{DR}) + 0.363(\mathrm{A}) - 0.474(\mathrm{TO})\right)\Big/\mathrm{D}* \quad (2)$$
$$\mathrm{D}* = 1 + \exp\left(0.123(\mathrm{FTA}) + 0.488(\mathrm{DR}) + 0.363(\mathrm{A}) - 0.474(\mathrm{TO})\right)$$

# 3. 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 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 logistic regression model from Equation (2), UC Riverside had a projected probability of victory of:

$$\pi = \exp\left(0.123(-1) + 0.488(5) + 0.363(7) - 0.434(4)\right)\Big/\mathrm{D}* = 0.958$$
$$\mathrm{D}* = 1 + \exp\left(0.123(-1) + 0.488(5) + 0.363(7) - 0.434(4)\right)$$

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

Using the least squares regression model from Equation (1), UC Riverside had a predicted point spread of:

$$\mathrm{y} = 0.062(-1) + 1.49(5) + 0.587(7) - 1.601(4) = 5.09$$

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 5 points with a score of 73 to 68.

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. If the point spread for the least squares model was estimated to be greater than 0, then a win was predicted for the "team of interest". The accuracy of each model is noted in **Table 6**.

As is shown in **Table 6**, both the logistic regression

 

**Table 3. Summary of stepwise selection for logistic regression model.**

| Step | Effect Entered | DF | Score Chi-Square | P Value |
|---|---|---|---|---|
| **1** | **Assists** | 1 | 101.7818 | <0.0001 |
| **2** | **Free Throw Attempts** | 1 | 78.7975 | <0.0001 |
| **3** | **Defensive Rebounds** | 1 | 21.1658 | <0.0001 |
| **4** | **Turnovers** | 1 | 29.2154 | <0.0001 |

**Table 4. Parameter estimates for logistic regression model.**

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | P Value |
|---|---|---|---|---|---|
| Free Throw Attempts | 1 | 0.1233 | 0.0358 | 11.8676 | 0.0006 |
| Defensive Rebounds | 1 | 0.4875 | 0.0909 | 28.7681 | <0.0001 |
| Assists | 1 | 0.3629 | 0.0840 | 18.6609 | <0.0001 |
| Turnovers | 1 | −0.4737 | 0.1002 | 22.3472 | <0.0001 |

**Table 5. Example data entry.**

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

**Table 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 |

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 94% (125/132 * 100), indicating that the variables identified are indeed significant to determining wins and losses in a Division I college basketball game.

## Use of Point Spread Model during March Madness

The point spread model was used on the basketball games in March Madness 2013. The model did a great job overall in estimating the point spread of a basketball game once the following four statistics are known where the differences are between the two teams playing as stated in the point spread model: differences in assists; differences in the number of free throw attempts; differences in the number of defensive rebounds; and differences in the number of turnovers. The results from the Final 4 games and Championship game are given using the point spread model as well as two interesting results along the way when there was a surprise winner as far as seeding was concerned. All game statistics were referenced from the ESPN website [9].

Iowa State played Notre Dame in Round 2 with Iowa State being seeded as 10 and Notre Dame being seeded as 7. Iowa State won the game. The final score and the four game statistics used in both models are given in **Table 7**.

The point spread prediction equation gives a point spread of 17.48 points.

$$Y = 0.062(-1) + 1.49(-4) + 0.59(10) + 1.60(11) = 17.48$$

The actual point spread between Iowa State and Notre Dame was 18 points.

Harvard played New Mexico in Round 2 with Harvard seeded as 14 and New Mexico seeded as 3. Harvard won the game. The final score and the four game statistics used in both models are given in **Table 8**.

The point spread prediction equation gives a points spread of 5.77 points.

$$Y = 0.062(-4) + 1.49(5) + 0.59(3) - 1.60(2) = 5.77$$

The actual point spread between Harvard and New Mexico was 6 points.

The four teams making it to the Final Four in 2013 were Syracuse, Michigan, Wichita State, and Louisville. Syracuse played Michigan in one of the games. Wichita State played Louisville in the other. The game statistics for the Syracuse and Michigan game are given in **Table 9**. The game statistics for the Wichita State and Louisville game are given in **Table 10**.

The point spread prediction equation for the game between Syracuse and Michigan gives a point spread of −4.41 points and is calculated below

$$Y = 0.062(-9) + 1.49(-1) + 0.59(-4) + 1.60(0) = -4.41.$$

The actual point spread was −5 points.

The point spread prediction equation for the game between Wichita State and Louisville gives a point spread of −1.74 points.

$$Y = 0.062(-5) + 1.49(0) + 3(0.59) - 1.60(2) = -1.74.$$

The actual point spread was −4 points.

The championship game was played between Louisville and Michigan. The final score and the four game statistics used in the models are given in **Table 11**.

The point spread prediction equation gives a point

 

**Table 7. Round 2 game (example).**

| Teams | Iowa State (10) | Notre Dame (7) | Difference (Iowa S-Notre D) |
|---|---|---|---|
| Final Score | 76 | 58 | 18 |
| Free Throws | 12 | 13 | −1 |
| Def. Rebounds | 22 | 26 | −4 |
| Assists | 19 | 9 | 10 |
| Turnovers | 6 | 17 | −11 |

**Table 8. Round 2 game (example).**

| Teams | Harvard (14) | New Mexico (3) | Difference (Harvard-New Mex) |
|---|---|---|---|
| Final Score | 68 | 62 | 6 |
| Free Throws | 20 | 24 | −4 |
| Defensive Rebounds | 23 | 18 | 5 |
| Assists | 11 | 8 | 3 |
| Turnovers | 13 | 11 | 2 |

**Table 9. Final 4 game 1.**

| Teams | Syracuse (4) | Michigan (4) | Difference (Syracuse-Michigan) |
|---|---|---|---|
| Final Score | 56 | 61 | −5 |
| Free Throws | 11 | 20 | −9 |
| Defensive Rebounds | 23 | 24 | −1 |
| Assists | 13 | 17 | −4 |
| Turnovers | 10 | 10 | 0 |

**Table 10. Final 4 game 2.**

| **Teams** | **Wichita State (9)** | **Louisville (1)** | **Difference (Wichita State-Louisville)** |
|---|---|---|---|
| **Final Score** | 68 | 72 | −4 |
| **Free Throws** | 24 | 29 | −5 |
| **Defensive Rebounds** | 22 | 22 | 0 |
| **Assists** | 13 | 10 | 3 |
| **Turnovers** | 11 | 9 | 2 |

spread of 5.24 points.

$$Y = 0.062(-2) + 1.49(-2) + 0.59(6) - 1.60(-3) = 5.24.$$

The actual point spread was 6 points.

**Table 11. Championship game**

| Teams | Louisville (1) | Michigan (4) | Difference (Louisville-Michigan) |
|---|---|---|---|
| Final Score | 82 | 76 | 6 |
| Free Throws | 23 | 25 | −2 |
| Defensive Rebounds | 17 | 19 | −2 |
| Assists | 18 | 12 | 6 |
| Turnovers | 9 | 12 | −3 |

## 4. Using 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 to the game being played were collected for both the "team of interest" and the "opposing team" for each of the significant variables already identified.

**Table 12** represents data for a randomly selected game between Air Force and San Diego State played on

January 21, 2012 with Air Force being selected as the "team of interest". For each of the teams, the 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 and Equation (1), the predicted point spread was:

$$y = 0.061(1.5) + 1.486(-2.5) + 0.587(-2) - 1.601(0) = -4.80$$

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 and Equation (2), the projected probability of victory for Air Force was given by:

Again, since $\pi < 0.5$, the game would be predicted as a loss for Air Force. In this instance, both models correctly predicted the game, as the outcome was a 13 point loss for Air Force.

This process was repeated for the 100 games selected randomly from the 2011-2012 Season, 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 13**.

As can be seen from **Table 13**, both the logistic and point spread models struggled to predict future games

$$\pi = \exp\left(0.123(1.5) + 0.488(-2.5) + 0.363(-2) - 0.474(0)\right) / D^* = 0.147$$

$$D^* = 1 + \exp\left(0.123(1.5) + 0.488(-2.5) + 0.363(-2) - 0.474(0)\right)$$

 

**Table 12. 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 State | 15 | 24 | 14 | 6 |
| San Diego State | 20 | 29 | 16 | 16 |
| San Diego State | 18 | 28 | 17 | 14 |
| San Diego State | 32 | 23 | 13 | 10 |

| Team | FTA | DR | A | TO |
|---|---|---|---|---|
| Air Force | 20.5 | 23.5 | 13 | 12 |
| San Diego State | 19 | 26 | 15 | 12 |
| Difference | 1.5 | −2.5 | −2 | 0 |

**Table 13. 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 | 59 | 4 | 63 |
| | Loss | 3 | 66 | 69 |
| | Total | 50 | 50 | 100 |

based on prior game median data. The logistic regression model correctly predicted 68/100 = 68% of games, while the point spread model correctly predicted 64/100 = 64% of games. A random sample of 75 games was also taken from the 2012-2013 Season. The same procedures were used to try to predict the outcome of a basketball game using both the logistic and point spread models estimating the variable differences in the model by considering the past four games played by both teams and taking the differences of the medians of the four factors used in both models. Overall, approximately the same results were obtained as when a sample of the 2011-2012 games was taken. The point spread model correctly predicted 62.67% of the games correctly, while the logistic model predicted 66.67% of the games correctly.

## 5. Conclusion

Four factors were identified which influence the outcome of a college basketball game. These factors were differences in assists, difference in free throw attempts, differences in defensive rebounds and differences in turnovers. Turnovers had the largest effect. This was followed by defensive rebounds and then assists. If the actual differences are known between the teams, two models were developed which do a great job in predicting the outcome of a basketball game as far as point spread of the game and also estimating the probability that a particular team wins the game. These actual differences will not be known ahead of time. An attempt was made to estimate the actual differences ahead of time by using the differences between the medians of these values for the four previous games that each of the teams had played. When these differences were put into either of the two models, the models had approximately a 62% to 68% chance of correctly predicting the winner of the basketball game.

## REFERENCES

[1] A. S. Malik, O. Boyko, N. Atkar and W. F. Young, "A Comparative Study of MR Imaging Profile of Titanium Pedicle Screws," *Acta Radiologica*, Vol. 42, No. 3, 2001, pp. 291-293. doi:10.1080/028418501127346846

[2] T. Hu and J. P. Desai, "Soft-Tissue Material Properties under Large Deformation: Strain Rate Effect," *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, San Francisco, 1-5 September 2004, pp. 2758-2761.

[3] R. Ortega, A. Loria and R. Kelly, "A Semiglobally Stable Output Feedback PI2D Regulator for Robot Manipulators," *IEEE Transactions on Automatic Control*, Vol. 40, No. 8, 1995, pp. 1432-1436. doi:10.1109/9.402235

[4] E. Wit and J. McClure, "Statistics for Microarrays: Design, Analysis, and Inference," 5th Edition, John Wiley & Sons Ltd., Chichester, 2004. doi:10.1002/0470011084

[5] A. S. Prasad, "Clinical and Biochemical Spectrum of Zinc Deficiency in Human Subjects," In: A. S. Prasad, Ed., *Clinical, Biochemical and Nutritional Aspects of Trace Elements*, Alan R. Liss, Inc., New York, 1982, pp. 5-15.

[6] B. M. S. Giambastiani, "Evoluzione Idrologica ed Idrogeologica Della Pineta di San Vitale (Ravenna)," PhD. Thesis, Bologna University, Bologna, 2007.

[7] J. K. Wu, "Two Problems of Computer Mechanics Program System," *Proceedings of Finite Element Analysis and CAD*, Peking University Press, Beijing, 1994, pp. 9-15.

[8] L. Honeycutt, "Communication and Design Course," 1998. http://dcr.rpi.edu/commdesign/class1.html

[9] O. Wright and W. Wright, "Flying-Machine," US Patent No. 821393, 1906.

 