March Madness: An Examination of the NCAA Men's' Basketball Championship

December 8, 2015

Ellie Li

Chris Matthews

Brad Gazdag

I. Introduction

Every year, 68 of the finest collegiate basketball teams come together for a single elimination tournament. The compelling event, nicknamed "March Madness," is a tournament sponsored by the NCAA and garners an incredible viewership both in the U.S. and around the world. Due to the popularity of the event and relative ease of creating a "bracket" or list of predicted outcomes of each consecutive round, betting on the outcomes of games is a huge event. There are many ways to predict outcomes, however a number of individuals simply use pre-assigned seeding information to assign value to each team. Other, "more professional" betters employ various statistical analyses to predict success. Given the highly improbable nature of successfully predicting each game outcome, there has never, to the general public's knowledge, been a perfect bracket. More interestingly, the uncertain nature of this tournament, including the almost expected occurrence of upsets (or when a lower seed beats a higher seed or the seed closest to one), makes the ability to distinguish the quality and predictive value of certain characteristics or variables would be valuable.

History of the Tournament

The tournament began as an eight-team contest, but the NCAA and selection committee have changed the format and structure of this event on numerous occasions to eventually derive the modern day platform. Starting in 1985 and termed the "Modern Era," the new platform set 64 teams in this single elimination playoff. The only other change occurred in 2001, which now sees 4 additional "play in" games that feature 8 teams playing for the right to participate in the first round of 64 in the tournament. A selection committee meets before the tournament to place teams into four regions, which vary in name each year normally depending on the location of

later round games. In 2015, the regions were Midwest, West, East, and South Teams are placed in each bracket based on the minimization of travel, however the formula for choosing which teams make the tournament is two fold. First, there are 31 teams that earn automatic bids received for winning respective conferences during that year. For example, teams that win the SEC or ACC conference tournament receive an automatic entry into the tournament regardless of record or expected success. This only increases viewership and makes each conference tournament game more compelling. The remaining 37 teams are granted entry based on various factors unknown to the general public but more than likely related to overall quality of the particular team's season. "High seeds" are arbitrarily defined as seeds 1-3 whereas seeds 4-16 are considered "low seeds." Any seed in a region should be of similar quality to a corresponding seed in another region and games are played in a specific format. The idea suggests the higher seeds will advance to each successive round given their designation as a superior team, however that is hardly ever a given. Interestingly enough, if an underdog (a team with a higher numerical seed indicating lower quality) wins against a lower seed team, that team will assume the higher seed perks going forward.

II. Literature Review

Due to the previously stated popularity of this event and the massive albeit illicit betting market surrounding this tournament, there have been a multitude of studies with the attempt to create a successful, predictive model for the March Madness tournament. Wright compiled a large data set with the goal of finding variables to create models with higher predictive values than a simple model relying only on seeding. Wright created two models, probit regression and OLS (ordinary least squares) regression, with a dummy dependent variable for win (1=margin of

victory greater than zero; 0=otherwise; note games cannot end in ties) and margin of victory for the two models respectively. The independent variables included statistical factors like offensive efficiency, Sagarin rank, win percent, and other high level variables. According to the probit model, win percentage and defensive efficiency are the strongest predictors of success while other seemingly important variables like a coach's experience level or points per game are negligible. However, Wright identifies endogeneity between some of the variables, which might throw off the coefficients. Similarly, the OLS model shows high coefficients and statistical significance of win percentage, Sagarin model, and defensive efficiency. It also showed opponent's seed as a positive factor (in that the higher the seed, the weaker the opponent).

In the Jacobson paper, the author places high value on seedings although taking careful time to explain the inefficiency of seeding in games between high seeds. Jacobson et al describes seeding as a product of win percentage, which is nearly identical between top teams. Thus, he took careful consideration to differentiate the number of occasions where teams of that nature (high seeded) would play head to head. While the nature of seeding as a predictive tool is powerful in the first few rounds, and the past winners of the tournament have predominantly been top seeds, the exclusive use of seeding is undesirable relative to other more extensive models.

III. Data Sample

The data that we have used for our research are historical statistics compiled for teams from the NCAA tournament from the years 1986 - 2015. The data collected are made up of 74 variables for each matchup, listed in Appendix I. About half of the variables for each matchup refer to the given team, and half refer to the opponent (denoted as "opp_variable name").

We have data on the teams and outcomes from 1901 matchups over the 30 year period that we are examining. It is important to note that the dataset represents each matchup of teams rather than each individual team isolated from their opponent. The dependent variable for analysis is margin of victory for the higher seed in any given matchup, with a negative number therefore representing a win by the lower seed. We also create a dummy variable called "win" that takes a value of 1 if the margin of victory is greater than zero.

We obtained the data 1986-2010 from the Chris Wright, the author of Statistical Predictors of March Madness: An Examination of the NCAA Men's' Basketball Championship.[1] Then, we added data from 2011 to 2015 using data from http://www.sports-reference.com/ and https://www.teamrankings.com to make our analysis more relevant to the current trend. Due to the lack of availability data online, we haven't been able to compile time played by each class since 2011. Despite the continuous effort of replicating the data for the OLS and Probit variables, we couldn't perfectly replicate the coefficients and t-statistic from Wright's publication. In Table 1 and 2, we include the OLS and Probit regressions with the 1986-2011 data with all of their coefficients and p-values matching up with their publication.

In our replication process, we extracted data from 2011-2015 which allows us to use the previous year data to estimate outcomes for the upcoming NCAAB 68 team tournament in 2016. For most of the variables, we were able to extract the data from the sources that they cited, mainly sports reference. However, for the coaching variables and the round variable there, was some difficulty. Even though for every variable we had to extract each data point by hand for each team, the coaching data was difficult to obtain. For the coaching data, we had to use the

Sports Reference dataset to code through the terminology to sum the coaching tenure at the college, total coaching seasons and if the coach has been to a final four or won a championship. So for our replication with the additional five years, we received some different results.

When comparing our updated model with the five most recent years (2011-2015) added, most of our results were pretty similar to their coefficients in the OLS model except for the favored team seed variable. When regressing win margin on all of the variables in Table 1 we found the t-stat and p-value of favorite team seeding to be highly insignificant. For the most part, all of our other variables match the sign of their coefficients with our new data. In our results we received a higher p-value since these underdogs like Harvard and Stanford in recent years have been going to sweet sixteen and elite eight more than before.

For our comparison of summary of statistics, for the most part our variables means, std dev, min and max match up with their summary statistics. We ran our summary with our new updated data with 2011-2015 data, again the seeding variable has some discrepancies with our data. In Table 7, our maximum is 16 since the new plan in 2001 allowed 8 teams to "play in" or take part in a preliminary round to get into the first round of 64. In these specific games, teams ranked 12, 16, 11 and 15 are favorites, so the mean and min and max is skewed upwards for our model.

Evaluation of Wright's Model

One of the first things we wanted to do in terms of improving their model was dropping some variables that were particularly insignificant in order to develop a model with more significant variables. We understand that we don't want to throw out a variable because it is insignificant, but by Table 3, it clearly shows of opp_win variable being highly insignificant.

However, when testing for the variable seed, we found the variable to be significant. We obviously ran the F-test to test for the joint significance amongst the independent variables, but the p-value was 0. We ran multiple F-tests altering different groups of variables gauging our results. We noticed that opp_win has an extremely high F statistic and statistically insignificant. We dropped these variables and ran a restricted model regression, and found that the adjusted R-squared increased when these variables were omitted. The F-stat for the variables are shown in Tables 4 and 5.

Further, as discussed in class, the test for heteroskedasticity, robust, is simple enough and worth implementing. Running regressions with robust can correct for the possibility that the error term varies depending on the actual observation. The robust command does have an impact in reducing the standard errors of the variables, but not in a noticeable amount. This increase suggests these variables have little variance and thus need no correction for heteroskedasticity, which intuitively makes sense when considering the variables and how many observations there are in the regression. The final regression with the test for heteroskedasticity is shown in Table 4.

We also developed some rationale for why we did not want to include the variables that we subsequently dropped. Regarding opp_scoredpergame and points_allowed, we did not believe this would properly explain underdog teams for all conferences because teams in inferior conferences are able to play increasingly inferior teams and thus “stuff the stat sheet” in the regular season. Since our dependent variable is based on win margin for the favorite teams, which tend to be from the power five conferences, we theorized that the efficient metric would be the best way to determine the effect of the opponent level of play.

Initially for our improvement, we found that the opposing points scored and allowed per game were inadequate for measuring opposing teams. After performing their model, we wanted to test for multicollinearity for the independent coefficients since we know that this is a kitchen sink model. In Table 2 one can see that opposing points scored and allowed per game have high inflation factors. We understand that it is generally agreed upon that a value greater than 10 for the VIFs suggests a problem with multicollinearity. Since opp pts scored and allowed per game are near 10 then we need to drop these variables and add new variables to explain for the opposing team's quality of play.

Ken Pomeroy is a famous statistician who has his own college basketball college rating models where he uses the variable of luck to explain seeding. Luck defines a team that simply seems to win close games. Thus, in the same vein as Pomeroy, we developed a simple indicator variable to highlight teams with luck. This was created by using a dummy variable luck and oppLuck that equals to one if a team (or opponent given our model) wins by less than one possession. The idea of luck is highly significant in opp_team which is intuitive in that people love to see a "Cinderella story." It seems that every year there is one team that wins supposedly unwinnable games; the most recent of which is Florida Gulf Coast. Advancing to the Sweet Sixteen in 2013, after having beaten both Georgetown and SDSU, despite a measly seeding of 15. They epitomize the cinderella story and more than likely used a bit of luck along the way.

Additionally, we noticed high levels of collinearity between variables like ppg and offensive efficiency. Given that efficiency is a variable created to describe how well a team scores, passes, defends, etc based on offensive or defensive, we decided to create a variable called "Elite" that acts as a dummy variable highlight teams that are both above average in terms

of offensive efficiency (this means their efficiency was greater than 106) and had a defensive efficiency or efficiency allowed lower than the average (efficiency allowed is better as it approaches zero). We created this variable using a simple dummy variable to highlight teams that were numerically above average in both categories. Regressing Elite and opposite with luck, oppoluck, and sos against win yields positive results that are all significant.

Replication and Improvement of Probit Model

The probit model will use the binary variable "win" as its output variable which is generated if winmargin is greater than zero. Please see the side by side comparison between the original model and our model. Comparing to the original model, even though we got different coefficients due to the addition of data information. The scale and magnitude are in line with the original model. The effect of win percentage seems to have a large, positive, statistically significant effect on the outcome, which makes sense as a team that won more during the regular season is typically a better team, all else held equal. Opponent win percentage also shows significant impact, in opposite direction. In addition, Sagarin rankings have a statistically significant effect on the outcomes of games. A team's Sagarin rank has a negative effect on the outcome of "win", which indicates better ranked teams will have lower value for their Sagarin ranking. Taking variables that are statistically significant, the results are shown in Table 7.

As we said above, in the Jacobson paper, the author places high value on seeding although taking careful time to explain the inefficiency of seeding in games between high seeds. He goes so far as to remove individual games that feature two of the same seed. Jacobson describes seeding as a product of win percentage, which is nearly identical between top teams. Even though he did not present his model in the paper, we tried to replicate a model to see how

results change as teams advance to further rounds in Table 11. As shown, seed has a negative impact which make senses intuitively, higher seed team has lower possibility to win. The seed coefficient is negative which makes sense given that the greater one's seed the more likely that team is to lose. The same goes for opponent seed. However, when we added rounds to the regression, we noticed that seeding was only negative up until round four. This coincides with the idea held by many march madness statisticians that once participants reach a certain stage of the tournament, seeding no longer holds predictive value. Even though the p-values indicate that the rounds variables are insignificant, we still wanted to show the contrast in the two probit models.

2. OLS Model Improvement

We mainly improved the model by removing variables experiencing both high collinearity and/or statistical insignificance. We then sought to create additional variables that represented trends found in the data. The most interesting variable, Elite, is a dummy variable describing teams that are both above average in offensive and defensive efficiency. These variables go a long way to predict trends similar to PPG, number of steals, and turnovers forced with adding unnecessary variables to the model. We regressed this variable with win, elite, oppoelite, and sos and found it to be both significant and having predictive value. Teams that are "elite" are considered so due to above average ability on both offensive and defensive assignments. While this may negate ridiculous offenses that win games through high octane shooting competitions, the idea of a complete team more often than not becomes victorious near the end of March. We created this dummy variable to equal 1 if the team had efficiency higher than 100 and had lower than 93 for defensive efficiency.

As for luck, the idea originated from going into “kill mode” in the fourth quarter by great players from around the league. One specific example mentioned in the article is Magic Johnson who seemed to will his team to wins. Luck is an indicator variable that highlights teams who win games with less than a one possession margin of victory. While this may be less telling than elite or win percent, it is still significant at a high level and provides insight into the inherent luck we have in sports. The luck coefficient was based off of Pomeroy’s model which was previously mentioned so we calculated the dummy variable by taking the difference between ppg and ppg allowed and if it was below three than the team was considered to be lucky.

The last variable we decided to add was opp_TotalGamesPlayed. We need to find a variable that represented the experience of the underdog teams while indirectly explaining if the teams went far in their conference tournaments without skewing the other coefficients drastically. We had variables like automatic bid and RPI, but those are very collinear with other variables already in the model. Once using the total games played for opposing teams, the adjusted R-squared increased dramatically creating a better R-squared and Adj. R-Squared in our model.

As we can see in Table 14, our improved model is better than Wright’s second OLS model. The adjusted R-squared while there is no increase of multicollinearity since the new variables do not have a VIF near 10. We originally included the opp_effeciency variable, but then we noticed that the variable was statistically insignificant so then by creating the dummy variable. We also tested for joint significance and they are still statistically significant. Lastly, in Table 14 we created a scatter plot comparing the predicted win margin with the actual win

margin and the distribution between the two variables are similar to Wright's graph in his dissertation.

Conclusion

Overall, through our process of manipulating variables, functional forms, and different forms of regression (OLS, Robust, Adjusted for Autocorrelation) we have determined that the thesis of the paper is correct: People cannot accurately predict the winners of the brackets for the NCAA tournament. While we were able to develop variables that were statistically significant with betas that agree with Wright's theory, this is because we worked to achieve these results. As we ran many regressions with other variables like rebounds, steals, and ppg squared, we produced regressions with widely different results. If one sets out to develop a model that shows no variables can explain win margin or the possibility of favorite team winning, they will likely succeed. If instead one wants to prove that a specific variable can explain winning trends they will eventually find a combination of relevant variables that make this possible. For the challenge of having a perfect bracket, which involves many different plays, calls and player matchups, no number of measurable statistics can fully explain why people can't create the perfect bracket. It is also important to remember that we were recreating a multiple regression and probit model, spanning different time periods. For example, the regression we were recreating had seed as negative and statistically insignificant, and our results indicate semi-strong statistical significance. Depending on which variables were added or omitted, however, the p-value for seed fluctuated measurably. This shows that even the most significant variables fluctuate since the seeding determining the match ups can be hindered due to adding too many variables, adding poor variables, or suffer due to omitted variable bias. As psychology and other

social sciences progress in terms of data collection, which has occurred over time, perhaps some of this difficulty can be alleviated, but currently, successful bracketology seems to be outside of the scope of econometric analysis.

Table 1

```
. reg winmargin seed winpercent percentofwinsaway winsinlast10 SagarinSRS ppg ppgallowed  opp_seed opp_win opp_winsinlast10 opp_
> percentofwinsaway opp_SagarinSRS opp_PointsScoredPerGame opp_PointsAllowedPerGame
```

Source	SS	df	MS		
Model	89277.424	14	6376.95885	Number of obs	= 1642
Residual	197718.725	1627	121.523494	F(14, 1627)	= 52.48
				Prob > F	= 0.0000
				R-squared	= 0.3111
				Adj R-squared	= 0.3051
Total	286996.149	1641	174.891011	Root MSE	= 11.024

winmargin	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
seed	.0808577	.1964698	0.41	0.681	-.3045027	.4662182
winpercent	27.60882	6.885143	4.01	0.000	14.10414	41.1135
percentofwinsaway	-5.247796	3.026672	-1.73	0.083	-11.18438	.6887892
winsinlast10	-.1773286	.2607841	-0.68	0.497	-.6888365	.3341793
SagarinSRS	-.1350727	.0290317	-4.65	0.000	-.192016	-.0781293
ppg	.239992	.107922	2.22	0.026	.0283112	.4516728
ppgallowed	-.1986524	.1192253	-1.67	0.096	-.4325036	.0351988
opp_seed	.4627659	.1188931	3.89	0.000	.2295663	.6959655
opp_win	1.725508	6.087557	0.28	0.777	-10.21477	13.66578
opp_winsinlast10	-.2134614	.2367856	-0.90	0.367	-.6778981	.2509753
opp_percentofwinsaway	8.215328	3.192998	2.57	0.010	1.952508	14.47815
opp_SagarinSRS	.0669828	.0093869	7.14	0.000	.048571	.0853945
opp_PointsScoredPerGame	-.3711312	.1293075	-2.87	0.004	-.624758	-.1175044
opp_PointsAllowedPerGame	.3940096	.1371753	2.87	0.004	.1249508	.6630684
_cons	-24.62957	7.534106	-3.27	0.001	-39.40714	-9.852005

Table 2

```
. vif
```

Variable	VIF	1/VIF
o~redPerGame	10.33	0.096824
o~wedPerGame	10.07	0.099315
ppg	6.86	0.145766
ppgallowed	5.56	0.179767
winpercent	4.98	0.200855
opp_win	3.85	0.260014
opp_Sagari~S	3.43	0.291299
seed	3.01	0.331809
opp_seed	2.99	0.334894
SagarinSRS	2.29	0.435740
winsinlast10	2.02	0.495519
opp_winsi~10	1.94	0.514168
opp_percen~y	1.40	0.714444
percentofw~y	1.30	0.766423
Mean VIF	4.29	

Table3

```
. correlate winmargin seed winpercent percentofwinsaway winsinlast10 SagarinSRS ppg ppgallowed  opp_seed opp_win opp_winsinlast10 opp_percentofwinsaway opp_Sag
> arinSRS opp_PointsScoredPerGame opp_PointsAllowedPerGame
(obs=1642)
```

	winmar~n	seed	winper~t	percen~y	winsi~10	Sagari~S	ppg	ppgall~d	opp_seed	opp_win	opp_w~10	opp_pe~y	opp_Sa~S	o~Scor~e	o~Allo~e
winmargin	1.0000														
seed	-0.2505	1.0000													
winpercent	0.2619	-0.7397	1.0000												
percentofw~y	-0.0368	-0.0188	0.1632	1.0000											
winsinlast10	0.1570	-0.4956	0.6998	0.1236	1.0000										
SagarinSRS	-0.2677	0.6553	-0.5182	0.1167	-0.2797	1.0000									
ppg	0.1782	-0.4096	0.4069	-0.1268	0.2288	-0.4083	1.0000								
ppgallowed	-0.0042	0.0270	-0.1355	-0.1666	-0.1136	0.0458	0.7389	1.0000							
opp_seed	0.3287	0.1813	-0.2093	0.0186	-0.1440	0.1477	-0.1356	0.0010	1.0000						
opp_win	-0.2061	-0.0858	0.0530	-0.0695	0.0149	-0.0717	-0.0010	-0.0491	-0.3810	1.0000					
opp_winsi~10	0.0850	-0.0957	0.0476	-0.0121	0.0076	-0.0792	0.0004	-0.0288	0.1874	0.4943	1.0000				
opp_percen~y	0.1401	0.0415	-0.0693	0.2395	-0.0756	0.0354	-0.0929	-0.0501	0.3224	0.1322	0.3323	1.0000			
opp_Sagari~S	0.4580	-0.1006	0.0414	-0.0918	0.0298	-0.0505	0.0379	0.0214	0.7420	-0.3627	0.2510	0.2560	1.0000		
o~redPerGame	-0.0954	-0.0741	0.0456	-0.0603	0.0457	-0.0386	0.2236	0.2524	-0.2029	0.2565	0.0552	-0.0923	-0.1831	1.0000	
o~wedPerGame	0.0811	-0.0490	0.0302	-0.0437	0.0448	-0.0135	0.2459	0.2942	0.0556	-0.2036	-0.1114	-0.0777	0.1265	0.8305	1.0000

```
.
```

Table 4

```
. test seed

 ( 1)  seed = 0

       F(  1,  1885) =    4.04
            Prob > F =    0.0446
```

```
 test opp_win

 ( 1)  opp_win = 0

       F(  1,  1627) =    0.08
            Prob > F =    0.7769
```

Table 5

```
. test seed winpercent percentofwinsaway winsinlast10 SagarinSRS ppg ppgallowed op
> p_seed opp_win opp_winsinlast10 opp_percentofwinsaway opp_SagarinSRS opp_PointsS
> coredPerGame opp_PointsAllowedPerGame

 ( 1)  seed = 0
 ( 2)  winpercent = 0
 ( 3)  percentofwinsaway = 0
 ( 4)  winsinlast10 = 0
 ( 5)  SagarinSRS = 0
 ( 6)  ppg = 0
 ( 7)  ppgallowed = 0
 ( 8)  opp_seed = 0
 ( 9)  opp_win = 0
 (10)  opp_winsinlast10 = 0
 (11)  opp_percentofwinsaway = 0
 (12)  opp_SagarinSRS = 0
 (13)  opp_PointsScoredPerGame = 0
 (14)  opp_PointsAllowedPerGame = 0

       F( 14,  1885) =   56.67
            Prob > F =    0.0000
```

Table 6

Figure 8 – Results from OLS regression for Model Two

OLS Results

variable	coefficient	std error	t	P > |t|
seed	0.156348	0.215345	0.73	0.468
win percentage	30.44454	6.941815	4.39	0.000
win in last ten	-0.24395	0.267238	-0.91	0.361
percent of wins away	-14.0896	4.079614	-3.45	0.001
Sagarin rank	-0.12991	0.031335	-4.15	0.000
ppg	0.196113	0.109627	1.79	0.074
ppg allowed	-0.14389	0.121059	-1.19	0.235
opp_seed	0.277245	0.126024	2.20	0.028
opp_win percentage	2.061132	6.134739	0.34	0.737
opp_win in last ten	-0.22487	0.242445	-0.93	0.354
opp_percent of wins away	4.641649	3.578983	1.30	0.195
opp_Sagarin rank	0.086361	0.010208	8.46	0.000
opp_ppg	-0.28903	0.131985	-2.19	0.029
opp_ppg allowed	0.315573	0.139416	2.26	0.024
consant	-23.5521	7.596111	3.10	0.002

Adj R^2 = 0.3186

Table 7

```
. summ
```

Variable	Obs	Mean	Std. Dev.	Min	Max
year	1642	1998.53	7.518451	1986	2011
team	0				
seed	1642	3.515225	2.404561	1	16
opponent	0				
winmargin	1642	7.514616	13.22464	-37	103
previousto~t	1642	.7527406	.43296	-1	1
tournament~k	1642	.6912302	.4621267	0	1
winpercent	1642	.7872202	.0881905	.516129	.9714286
winsinlast10	1642	7.551766	1.482397	3	10
percentofw~y	1642	.3140496	.1027013	.04	.9166667
coachtenure	1642	9.399513	7.592209	1	39
totalseaso~g	1642	15.83861	8.663815	1	39
coachfinal~s	1642	1.130329	1.910571	0	11
coachchamp~s	1642	.3282582	.6850314	0	4

coachchamp~s	1642	.3282582	.6850314	0	4
coachNBAdr~s	1575	12.28698	13.54334	0	70
automaticbid	1575	.375873	.4845013	0	1
RPI	693	14.67677	13.58667	1	169
SagarinSRS	1642	14.97247	14.20007	-5.8	207
SOS	1642	33.35194	31.53445	-4.93	252
TotalGames	1642	34.23934	2.756958	26	41
ppg	1642	78.49569	6.604467	36.55172	102.9
ppgallowed	1642	67.59848	5.383356	49	88.5
Efficiency	949	109.9979	4.468144	92.68919	125.464
Efficiency~d	949	94.12873	4.059057	79.92938	106.776
RbsPerGame	949	40.73259	4.616381	26.58621	57.8
RbsPerGame~d	949	35.70864	3.104723	25.5	46.2
StealsPerG~e	949	7.807432	1.567878	4.129032	13.5
StealsPerG~d	882	6.498488	.9926416	3.5	10.46667
BlocksPerG~e	949	4.796924	2.225309	1.777778	17
FG	949	47.05638	2.575697	40.2	57
pt	949	36.32792	2.453554	25.8	42.9
TS	949	55.62497	2.176034	49.5	61.8
freethrowr~e	949	37.60273	6.210176	15.908	56.3
assistsper~e	949	15.53878	1.829174	11.2	20.7
freshmenpl~e	882	12.97337	5.635338	0	30.50746
sophomorep~e	882	18.788	6.896557	0	37.21875
juniorplay~e	882	20.98255	7.911384	0	36
seniorplay~e	882	21.56933	8.793533	0	37.4
AM	1642	1998.53	7.518451	1986	2011
AN	0				
opp_seed	1642	9.803898	3.955177	1	16
opp_previo~t	1575	.4546032	.498093	0	1
opp_coachf~s	1642	.3349574	.9096187	0	9
opp_coachc~s	1642	.0943971	.338805	0	3
opp_coachN~s	1575	5.983492	9.869272	0	69
opp_automa~d	1575	.4761905	.4995914	0	1
opp_RPI	693	51.44589	42.45438	1	262
opp_Sagari~S	1642	57.08272	53.71337	-13.37	305
opp_SOS	1642	95.20195	83.19973	-10.31	317
opp_TotalG~s	1642	32.80024	2.485605	27	41
o~redPerGame	1642	75.17454	6.763312	50.3871	122.4
o~wedPerGame	1642	68.12209	6.294938	48.3	108.1
opp_Effici~y	949	106.7995	4.914253	85.2	128.9377
opp_Effici~d	949	96.10497	4.079141	80.04976	115.9233
opp_RbsPer~e	949	39.16571	4.673051	22.82759	58.6
opp_RbsPer~d	949	35.79779	3.206229	25.2	49
opp_Steals~e	949	7.690317	1.599702	3.870968	13.9
opp_Steals~d	882	6.551482	1.009668	3.5	9.878788
opp_Blocks~e	949	4.198951	2.218645	.75	16.2
opp_FG	949	46.01644	2.527387	38.7	55.9
opp_3pt	949	35.91581	2.7061	25.1	43
opp_TS	949	54.76143	2.368328	45.8	62
opp_freeth~e	949	37.31707	6.397797	15.908	65.4
opp_assist~e	949	14.6608	1.701216	9.9	20.3
opp_freshm~e	882	12.88502	5.721206	0	30.50746
opp_sophom~e	882	18.59906	6.739932	0	36.13333
opp_junior~e	882	20.8284	7.966002	0	37.3
opp_senior~e	882	22.08924	9.226352	0	39.40625
Round	1642	1.900122	1.194039	0	6
Luck	1642	.272838	.445554	0	1
OppLuck	1642	.4013398	.4903189	0	1
Eff	1642	.2399513	.4271836	0	1
OppEff	1642	.1126675	.3162822	0	1

Table 7

```
. probit win seed winpercent percentofwinsaway winsinlast10 SagarinSRS ppg ppgallo
> wed opp_seed opp_win opp_winsinlast10 opp_percentofwinsaway opp_SagarinSRS opp_P
> ointsScoredPerGame opp_PointsAllowedPerGame

Iteration 0:   log likelihood = -1138.6698
Iteration 1:   log likelihood = -958.10815
Iteration 2:   log likelihood = -943.52206
Iteration 3:   log likelihood = -943.34067
Iteration 4:   log likelihood = -943.34061

Probit regression                                 Number of obs     =       1900
                                                  LR chi2(14)       =     390.66
                                                  Prob > chi2       =     0.0000
Log likelihood = -943.34061                       Pseudo R2         =     0.1715
```

win	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
seed	-.0136701	.0184311	-0.74	0.458	-.0497944	.0224541
winpercent	5.52904	.8153697	6.78	0.000	3.930945	7.127135
percentofwins~y	-1.249165	.2974033	-4.20	0.000	-1.832065	-.6662655
winsinlast10	-.0448964	.0306935	-1.46	0.144	-.1050544	.0152617
SagarinSRS	-.0146764	.0029226	-5.02	0.000	-.0204047	-.0089482
ppg	-.0216766	.0128439	-1.69	0.091	-.0468502	.0034969
ppgallowed	.0257594	.0140959	1.83	0.068	-.001868	.0533869
opp_seed	.0446754	.0121876	3.67	0.000	.0207881	.0685628
opp_win	-3.628489	.799835	-4.54	0.000	-5.196137	-2.060841
opp_winsinla~10	.0167487	.0279096	0.60	0.548	-.0379532	.0714506
opp_percentof~y	1.906115	.3339976	5.71	0.000	1.251492	2.560738
opp_SagarinSRS	.0083224	.0014332	5.81	0.000	.0055132	.0111315
opp_PointsSco~e	.0044777	.0151195	0.30	0.767	-.025156	.0341115
opp_PointsAll~e	-.0113346	.0161617	-0.70	0.483	-.0430109	.0203418
_cons	-1.214716	.9166627	-1.33	0.185	-3.011342	.5819094

Table 8

analysis are presented below, in Figure 6.

Figure 6 – Results from the Probit Regression for Model Two

Probit Results				
variable	**coefficient**	**std error**	**z**	**P > |z|**
seed	0.0667	0.02832	2.35	0.019
win percentage	7.39	0.949	7.79	0.000
win in last ten	-0.0459	0.03633	-1.26	0.206
percent of wins away	-3.83	0.570	-6.72	0.000
Sagarin rank	-0.0218	0.00372	-5.85	0.000
ppg	-0.0521	0.01445	-3.61	0.000
ppg allowed	0.0619	0.01621	3.82	0.000
opp_seed	-0.0198	0.01826	-1.09	0.277
opp_win percentage	-3.93	0.909	-4.32	0.000
opp_win in last ten	-0.0188	0.03389	-0.56	0.579
opp_percent of wins away	1.81	0.520	3.48	0.001
opp_Sagarin rank	0.0171	0.00223	7.66	0.000
opp_ppg	0.0390	0.01824	2.14	0.033
opp_ppg allowed	-0.0482	0.01943	-2.48	0.013
constant	-1.54	1.045	-1.47	0.141

Pseudo $R^2 = 0.221$[11]

Table 9

Figure A – Summary Statistics of variables used in Model One

variable	observations	mean	std dev	min	max
seed	1575	3.48	2.346	1	12
win percent	1575	0.788	0.0882	0.516	0.971
wins in last ten	1575	7.56	1.487	3	10
Sagarin ranking	1575	14.9	14.45	1	207
ppg	1575	78.7	6.62	36.5	102.9
ppg allowed	1575	67.8	5.41	49	88.5
off efficiency	882	110.1	4.25	94	121.6
def efficiency	882	94.1	3.94	84.7	106.2
TS percentage	882	55.7	2.18	49.5	61.8
assists per game	882	15.6	1.82	11.2	20.7
coach final fours	1575	1.09	1.87	0	10
opp_seed	1575	9.79	1.869	1	16
opp_win percent	1575	0.709	0.0874	0.367	0.95
opp_wins in last ten	1575	7.25	1.601	3	10
opp_Sagarin ranking	1575	59.1	53.91	1	305
opp_ppg	1575	75.3	6.79	50.4	122.4
opp_ppg allowed	1575	68.4	6.36	48.3	108.1
opp_off efficiency	882	106.9	4.53	85.2	119.4
opp_def efficiency	882	96.1	3.84	83.1	110.2
opp_TS percentage	882	54.8	2.37	45.8	62
opp_assists per game	882	14.7	1.69	10.1	20.3
opp_coach final fours	1575	0.331	0.9055	0	9

Table 10

```
. probit win seed opp_seed

Iteration 0:   log likelihood = -1138.6698
Iteration 1:   log likelihood = -1038.5648
Iteration 2:   log likelihood = -1037.9397
Iteration 3:   log likelihood = -1037.9397

Probit regression                                 Number of obs     =      1,900
                                                  LR chi2(2)        =     201.46
                                                  Prob > chi2       =     0.0000
Log likelihood = -1037.9397                       Pseudo R2         =     0.0885

------------------------------------------------------------------------------
         win |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        seed |  -.1229828   .0122532   -10.04   0.000    -.1469986   -.0989669
    opp_seed |   .0960948    .008518    11.28   0.000     .0793999    .1127898
       _cons |    .145045   .0852176     1.70   0.089    -.0219785    .3120685
------------------------------------------------------------------------------
```

Table 11

```
. probit win seed opp_seed i.round

Iteration 0:   log likelihood = -1138.6698
Iteration 1:   log likelihood = -1035.7316
Iteration 2:   log likelihood = -1034.9673
Iteration 3:   log likelihood = -1034.9672

Probit regression                                 Number of obs     =      1,900
                                                  LR chi2(8)        =     207.41
                                                  Prob > chi2       =     0.0000
Log likelihood = -1034.9672                       Pseudo R2         =     0.0911

------------------------------------------------------------------------------
         win |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        seed |  -.1169646    .015123    -7.73   0.000    -.1466052    -.087324
    opp_seed |   .1098509   .0129988     8.45   0.000     .0843737    .1353282
             |
       round |
           1 |  -.2546229   .3538915    -0.72   0.472    -.9482375    .4389917
           2 |  -.1967107   .3782171    -0.52   0.603    -.9380026    .5445812
           3 |  -.0400503   .3977276    -0.10   0.920     -.819582    .7394815
           4 |  -.1761627   .4184471    -0.42   0.674     -.996304    .6439785
           5 |   .0317736   .4412394     0.07   0.943    -.8330398     .896587
           6 |   .1901048   .4776344     0.40   0.691    -.7460414    1.126251
             |
       _cons |   .1803695   .4422624     0.41   0.683    -.6864489    1.047188
------------------------------------------------------------------------------
```

Improvement Table 12

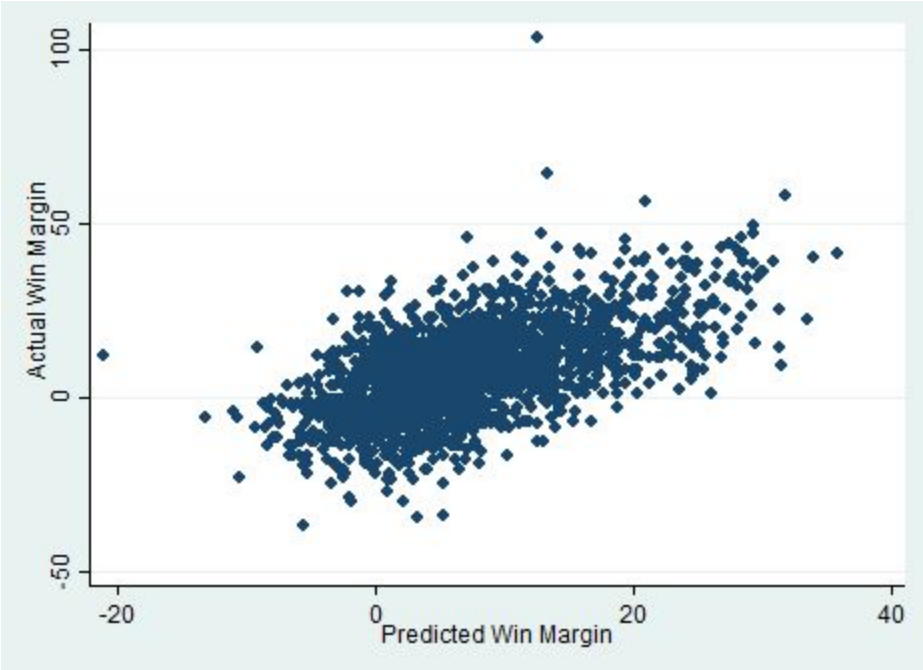
```
. reg winmargin seed winpercent percentofwinsaway SagarinSRS ppg ppgallowed opp_seed opp_perc
> entofwinsaway opp_SagarinSRS OppLuck OppEff opp_TotalGames

      Source |       SS       df       MS              Number of obs =    1901
-------------+------------------------------           F( 12,  1888) =   77.68
       Model |  109004.797     12  9083.73308           Prob > F      =  0.0000
    Residual |  220766.493   1888  116.931405           R-squared     =  0.3305
-------------+------------------------------           Adj R-squared =  0.3263
       Total |  329771.29    1900  173.563837           Root MSE      =  10.813

------------------------------------------------------------------------------
   winmargin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        seed |  -.3392461   .1472576    -2.30   0.021    -.6280509   -.0504413
  winpercent |   15.89531   5.314239     2.99   0.003     5.472909    26.31771
percentofwi~y |   .8745195   2.199369     0.40   0.691    -3.438929    5.187968
  SagarinSRS |  -.1139429   .0240611    -4.74   0.000     -.161132   -.0667537
         ppg |   .4245928   .0972854     4.36   0.000     .2337946    .615391
   ppgallowed |  -.5173101    .106543    -4.86   0.000    -.7262644   -.3083557
    opp_seed |   .4604878   .0892899     5.16   0.000     .2853706    .6356051
opp_percento~y |   8.738879   2.283348     3.83   0.000     4.260727    13.21703
opp_SagarinSRS |   .0405671   .0071173     5.70   0.000     .0266085    .0545257
     OppLuck |   2.854478   .5780964     4.94   0.000     1.720703    3.988253
      OppEff |  -.9170458   .7936316    -1.16   0.248    -2.473533    .6394413
opp_TotalGames |  -1.190204   .1189725   -10.00   0.000    -1.423536   -.9568732
       _cons |    27.7501   7.338111     3.78   0.000     13.35844    42.14176
------------------------------------------------------------------------------
```

Table 13

```
. vif
```

Variable	VIF	1/VIF
ppg	6.70	0.149227
ppgallowed	5.60	0.178613
winpercent	3.71	0.269860
seed	2.43	0.412109
opp_Sagari~S	2.26	0.441911
opp_seed	2.13	0.468888
percentofw~y	2.00	0.500478
SagarinSRS	1.68	0.594281
opp_TotalG~s	1.55	0.645473
opp_percen~y	1.52	0.657384
OppLuck	1.30	0.769795
OppEff	1.15	0.866847
Mean VIF	2.67	

Table 14



Bibliography

"CBB at Sports Reference." College Basketball at Sports-Reference.com. N.p., n.d. Web. 08 Dec. 2015.

Chris, Wright. Statistical Predictors of March Madness: An Examination of the NCAA Men's' Basketball Championship (n.d.): n. pag. 30 Apr. 2012. Web. 1 Dec. 2015.

Jacobson, S. H., Nikolaev, A. G., King, D.M., Lee, A. J., 2011, "Seed distributions for the NCAA Men's Basketball Tournament", OMEGA, 39(6), 719-724.

"NCAA College Basketball Betting Picks & NCAAB Picks 2015." NCAA College Basketball Betting Picks (NCAAB Picks). N.p., n.d. Web. 08 Dec. 2015.

Rudy, Kevin. "Analyzing." "Luck" in College Basketball: Part II. N.p., 14 Mar. 2014. Web. 08 Dec. 2015.