Predicting Wins for College Basketball Teams

Justin Zhang, Abhinn Bhardwaj, James Scott
University of Waterloo
STAT 371
Dr. Matthias Schonlau
December 7, 2021

Abstract

We built a linear regression model that predicts the percentage of wins that a Division I college basketball team is expected to achieve based on annual team-level measures for each team in the league between 2013 and 2020. We selected our model using both backward and forward selection and VIFs to minimize prediction error and multicollinearity. Robust regression was used to minimize the effect of four outliers on coefficient estimates. In the final model, the dependent variable is win percentage and there are 16 independent variables that describe a team's performance. Power7 is an indicator variable that was introduced to determine the effect that being in a highly-funded power 7 or mid-major conference has on a team's win percentage. Win percentage is most negatively sensitive to power7 and most positively sensitive to wins above bubble–the number of wins required to qualify for March Madness. In addition to the high AIC score, the results of three test cases demonstrated the model's strong predictive power. For Gonzaga, the model predicted 98.6% and the actual win percentage was 100%. For Idaho, the predicted value was 21.1% and the actual value was 4.5%. For Utah Valley, these values were 61% and 55%, respectively. Therefore, this model could be useful for betting on seasonal outcomes for Division 1 college basketball teams by tuning the variables according to the better's expectations and the team's past results.

Introduction

Our objective was to build a linear regression model that predicts the percentage of wins that a Division I college basketball team is expected to achieve based on annual team-level measures for each team between 2013 and 2020. Power rating is commonly used by sports enthusiasts to gauge the future success of a team based on the team's win percent, its opponents' win percent, and the win percent of the opponents' of its opponents'. Our prediction model differs from power rating because it uses a team's performance measures to determine the team's expected percentage of wins over one season; it does not depend on the outcomes of a team's opponents. Our model has a clear application to pre-season betting as a gambler could assume that a team's performance will match that of the previous year and make adjustments based on expectations of new players to predict the teams success in the upcoming season. We obtained a dataset that fit our objective from Kaggle, with the original source of the data being barttorvick.com. The dataset contains 22 variables, of which three are categorical and the remaining 19 are measures, and 2455 observations with no missing values. We considered all of the variables in the dataset except for power rating when selecting our model.

Selection Method

We began by analyzing the model diagnostic plots to identify and remove outliers. This ensured our model satisfied the normality assumption. We used both forward and backward selection to choose our final model. We assessed multicollinearity using VIF and removed variables with high VIF scores to settle on a model that balanced both multicollinearity and prediction error, as measured by AIC score.

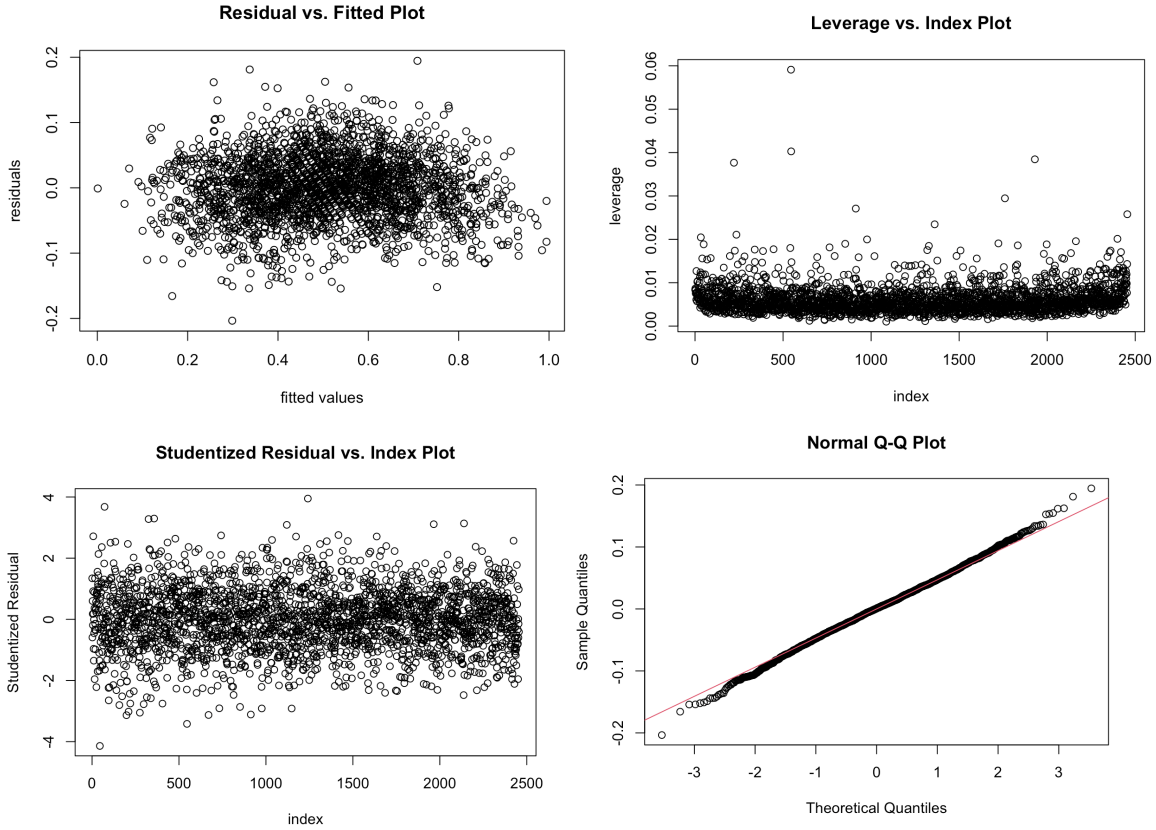
In our final model, we used robust regression to minimize the impact of outliers on the coefficient estimates. The dependent variable was percent of games won in a season (WP) and the independent variables are shown in table 1. Each variable was in our original dataset, except for the last one, Power7, which is an indicator variable for whether a team is in a power conference (ACC, PAC12, BIG10, SEC, BIG12, BigEast, WCC), or in a mid-major (every other conference). This indicator was used since the schools in power conferences have more athletics

funding due to media contracts, thus having more money to spend on facilities and recruiting. This in turn gives them better players and coaches, and hence success. We also considered adding interaction variables between various independent variables and the Power7 indicator variable, but minimal interaction was found (see exhibit 2 for details).

Table 1. List of Independent variables in the full linear regression model

Variable	Description
Win Percentage	Dependent Variable; calculated as Wins/Games
Adjusted offensive efficiency (ADJOE)	Estimate points scored per 100 possessions
Adjusted defensive efficiency (ADJDE)	Estimate of points allowed per 100 possessions
Turnover percentage allowed (TOR)	% of possessions ending in turnover
Turnover percentage committed (TORD)	% of possessions ending with steal
Offensive rebound rate (ORB)	% of possessions offensive rebound obtained
Offensive rebound rate allowed (DRB)	% of possessions offensive rebound given up
Free throw rate (FTR)	% of possessions free throws shot
Free throw rate allowed (FTRD)	% of possessions free throws allowed
Two-point shooting percentage	(2P_O)
Two-point shooting percentage allowed	(2P_D)
Three-point shooting percentage	(3P_O)
Three-point shooting percentage allowed	(3P_D)
Adjusted tempo (ADJ_T)	possessions per 40min against average D1 team
Wins above bubble (WAB)	Wins against teams in march madness
BARTHAG	Probability to beat average D1 team
Effective FG% (EFG_O)	Effective field goal percentage
Effective FG% Allowed (EFG_D)	Effective field goal percentage allowed
Power7	Whether team is in a “power” conference

Model Diagnostics



From the residual vs. fitted plot, we see the constant variance assumption fails to hold. The graph shows that the residual range is far larger (around 0.3) when the fitted value is approximately 50% win percentage, as opposed to smaller values (around 0.1-0.2) when win percentage is a lot higher or lower. From the data, this is because really good teams (win percentage over 85%) perform well in most statistical categories and vice versa for really bad teams (win percentage below 15%) making them easier to predict. On the other hand, average teams excel at some stats while doing poorly at others, and also have variation in win percentage amongst teams with similar metrics. Table 2 illustrates the variance of specific stats for teams with drastically different winning percentages, and helps show why the residuals have smaller magnitude for high and low win percentages. Because of the heteroskedasticity present within our data, we will use robust standard errors for our final model.

Table 2: Range of Various Statistics for Different Team Levels

Statistic	Good Team (>85% WP)	Average Team (35%-65% WP)	Bad Team (<15% WP)
Effective Field Goal % Shot	50-60	43-57	41-49
Effective Field Goal % Allowed	43-50	43-58	50-60
ADJOE	110-130	85-120	75-95

From the leverage vs. index plot, we see four specific points that have leverage above 0.03, which is 1.5-2 times more than the average leverage. These specific points are listed in table 3. Analyzing these points further, we see that the two Grambling State teams have the lowest adjusted offensive rating and 2-point %, and the highest adjusted defensive ratings, all of which are significant outliers in x-space. They also never won a game, the only teams to do so, and so can be considered statistical anomalies. The Wisconsin team had the highest BARTHAG score, and by far the lowest free throw rate allowed of good teams, with other stats very high as well, causing it to also be an outlier. Finally, Savannah State did not have any statistics where they were ranked bottom 10, but they were consistently bad in every single stat, and so is also an outlier in x-space. From these observations, we decided to remove these outliers from the dataset prior to running the regression and model selection.

Table 3: High Leverage Points with Selected Stats

Team	WP	ADJOE	ADJDE	TOR	X2P_O	WAB	Leverage
Grambling State-2015	0%	76.6	121.4	26.8	38.4	-24.3	0.0591
Grambling State-2013	0%	76.7	117.0	26.1	38.4	-22.3	0.0403
Wisconsin	90%	129.1	93.6	12.4	54.8	11.3	0.0384
Savannah State	37.7%	94.6	117.1	19.3	53.7	-15.0	0.0377

From the studentized residuals plot, we notice 3 points that have magnitudes approximately over 4, corresponding to teams from Incarnate Word, and Nicholls State. Though there is no observable reason why these points have high residual, we remove them in order to ensure our sigma estimate is not too large.

From QQplot, the theoretical and observed quantiles are roughly equal, since the points follow the straight line, with both tails skewing slightly due to outliers. However, we can still conclude that the residuals are in fact normally distributed.

Final model diagnostics after the above changes can be found in exhibit 1 of appendix.

VIF Analysis

The first aspect of our model selection is looking at VIF, and multicollinearity between variables. Though there's no benchmarks for what an acceptable VIF score is, anything over 10-20 is considered problematic.

Looking at table 4, we see that the main problems occur between EFG, and X3P for both offense and defense, with VIF between 31-233. It turns out that the effective field goal formula is actually EFG = aX2P + bX3P. Where 'a' is the total shots divided by 2-point shots and 'b' is total shots divided by 3-point shots multiplied by 1.5. This is a linear combination of 2-point and 3-point percentage, and so there is a clear linear relation between these three variables. Hence EFG_O and EFG_D can be predicted by the shooting percentages and so we remove them from our model.

Another incidence of multicollinearity is with the BARTHAG variables, with a VIF of 36.8. BARTHAG measures a team's ability to beat the average D1 team, which is essentially their probability to win any game, and so predicting BARTHAG and win percentage are very

similar. Looking at the explicit formula, it is the pythagorean expectation of ADJOE and ADJDE, which can be approximated linearly, further showing the correlation. Looking at table 5 for the VIF scores after removing these three variables, and every value is roughly below 15, and so our model is now relatively uncorrelated.

Table 4: VIF Scores before Model Adjustment

Stat	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	TORD	ORB	DRB
VIF	29.4	24.7	36.9	148.8	232.4	3.4	3.4	3.1	2.5
Stat	X3P_O	X3P_D	X2P_O	X2P_D	ADJ_T	WAB	Power7		
VIF	31.8	43.2	69.5	124.4	1.2	12.6	2.4		

Table 5: VIF Scores after Model Adjustment

Stat	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	TORD	ORB	DRB
VIF	15.7	13.5	N/A	N/A	N/A	3.2	3.2	2.8	2.3
Stat	X3P_O	X3P_D	X2P_O	X2P_D	ADJ_T	WAB	Power7		
VIF	2.7	2.1	2.6	4.3	1.2	12.6	2.3		

Autocorrelation

Table 6: Durbin Watson Test Statistics

lag	Autocorrelation	D-W Statistic	P-value
1	0.316624	1.366683	0

As can be observed from the Durbin Watson test shown in Table 6, the test statistic is 1.366683 with a corresponding p-value of 0. Given that the p-value is less than 0.05, it can be concluded that the null hypothesis is rejected, and thus the residuals in our model are autocorrelated. The standard accepted range for the Durbin Watson test statistic is between 1.5-2.5, and thus a value of 1.366683 is indicative of a large positive autocorrelation.

Model Interpretation

The final model regresses win percentage on the variables listed in table 7. The coefficient estimates for the model shown are all highly statistically significant with p-values << 0.001. The intercept term, 0.5564, shows that a baseline team will have a win percentage of 55.64%. Furthermore, a one unit increase in each of the independent variables causes an increase in win percentage that is equal to its corresponding estimate.

Six variables predict a greater likelihood of defeat as their values increase. Adjusted offensive efficiency is one of these variables, which is unexpected because it is a measure of points scored. However, this may be because ADJOE is impacted by other offensive metrics and is used to offset potential gains in those categories. The remainder of these variables include

turnover percentage allowed, offensive rebound rate allowed, free throw rate allowed, two-point shooting percentage allowed, and three-point shooting percentage allowed, varying between a 0.2%-1.2% lower win percentage. This is expected, since these allowances imply fewer shot opportunities and success, hence less ability to score.

Eight of the variables predict a greater percentage of wins as they increase. The most sensitive of these variables on a per unit basis are steal rate and wins above bubble with a 1% increase in the former resulting in a 1.7% increase in win percent and an additional win above bubble resulting in a 2.6% increase in win percent. The least sensitive of these variables are offensive rebound rate, free throw rate, and adjusted tempo with a one unit increase resulting in a respective increase in win percent of 0.54%, 0.06%, and 0.024%. The remaining variables in this group are moderately sensitive. These variables are adjusted defensive efficiency, two-point shooting percentage, and three-point shooting percentage, which produce a 1.3%, 1.2%, and 1.1% increase in win percent, respectively, for each unit increase. Contrary to the previous set of variables, these ones imply more scoring opportunities and better conversion rate (shot into score), which means the team will perform better.

Finally, the indicator variable Power7, has a coefficient estimate of 0.0629, means that on average, a Power conference team has a 6.3% higher win percentage than a mid-major team. This makes sense, because when a power conference team plays a mid-major, they normally win the game based on our dataset. Furthermore, there are only a few mid-major teams within the top-50 rankings both currently and historically.

Table 7: Summary Table of Final Model

Stat	Estimate	Std. Error	P value
Intercept	0.5564	0.445	<2.2*e^(-16)
ADJOE	-0.0083	0.0005	<2.2*e^(-16)
ADJDE	0.0098	0.0005	<2.2*e^(-16)
TOR	-0.0128	0.0007	<2.2*e^(-16)
TORD	0.017	0.0007	<2.2*e^(-16)
ORB	0.0054	0.0003	<2.2*e^(-16)
DRB	-0.0107	0.0004	<2.2*e^(-16)
FTR	0.0006	0.0002	0.00085
FTRD	-0.002	0.0002	<2.2*e^(-16)
X2P_O	0.0103	0.0005	<2.2*e^(-16)
X2P_D	-0.0098	0.0005	<2.2*e^(-16)
X3P_O	0.0093	0.0005	<2.2*e^(-16)

X3P_D	-0.0105	0.0005	<2.2*e^(-16)
ADJ_T	0.0024	0.0003	1.473*e^(-15)
WAB	0.0261	0.0005	<2.2*e^(-16)
Power7-Yes	0.0629	0.0031	<2.2*e^(-16)

Applying the Model

To get a sense of how well our model predicts actual outcomes, we will compare the actual outcome for the teams with the highest, lowest, and median win percent in 2021 based on at least 20 games played to the predicted outcome. Gonzaga won every game it played in the most recent season and our model predicted that it won 98.6% of games played. Idaho won just 4.5% of games played and our model predicts that it won 21.1% of games played. Utah Valley had the medin win percentage of 55% and our model predicted that it won 61.0% of its games. Based on this small group of test cases and the low AIC score, we feel confident that our model has strong predictive power and can be useful in a real world setting where there is value in predicting outcomes, such as in sports betting.

Conclusion

As can be observed in our analysis and interpretation sections we have established our claim that the chosen model serves to predict the game win percentage of a division 1 college basketball team based on annually reported performance based statistics. The prediction model accounts for heteroskedasticity in the dataset and high autocorrelation in the residuals by using robust standard errors and by removing variables as necessary, respectively, to build the final model. There are a number of prevailing applications for our model, the most relevant of which include sports betting, specifically surrounding the NCAA Division 1 Men's Basketball elimination tournament commonly known as March Madness. Given that our model does not rely on opponent team statistics, this allows for more precise estimates of a team's performance over the course of a season. Our model is different from the 'moneyline' odds approach employed by most major sports betting platforms, which only provide odds on the outcome of a specific game. Current projections for the sports betting market estimate an approximate 8.33% compound annual growth rate, reaching around USD $179.3 billion market valuation by 2028 (Research, 2021). This projection has been largely attributed to the COVID-19 pandemic (Cohen, 2021) as more people are turning toward a digitally focused lifestyle, and indicates a continued need for models such as ours to better predict long-term team results as compared to the current short-term game predictions offered.

In summary, our model has real world applications and is also adaptable such that it can be modified or used in conjunction with existing tools to better calculate the likelihood of success of basketball teams based on performance data. Apart from being a predictive model, it can also be used to study previously employed team compositions, the associated performances in games, and thus the factors that influence success in basketball games.

Citations

Cohen, J. D. (2021, February 5). *Perspective | sports gambling could be the pandemic's biggest winner*. The Washington Post. Retrieved December 6, 2021, from https://www.washingtonpost.com/outlook/2021/02/05/sports-gambling-could-be-pandemics-biggest-winner/.

Research, Z. M. (2021, September 7). *Global sports betting market to be worth US$ 179.3 billion by 2028 with CAGR of 8.33% during 2021 to 2028 - Zion Market Research*. Global Sports Betting Market to be worth US$ 179.3 billion by 2028 with CAGR of 8.33% during 2021 to 2028 - Zion Market Research. Retrieved December 6, 2021, from https://www.prnewswire.com/news-releases/global-sports-betting-market-to-be-worth-us-179-3-billion-by-2028-with-cagr-of-8-33-during-2021-to-2028--zion-market-research-301370184.html.

Appendix

Exhibit 1: Final Model Diagnostics

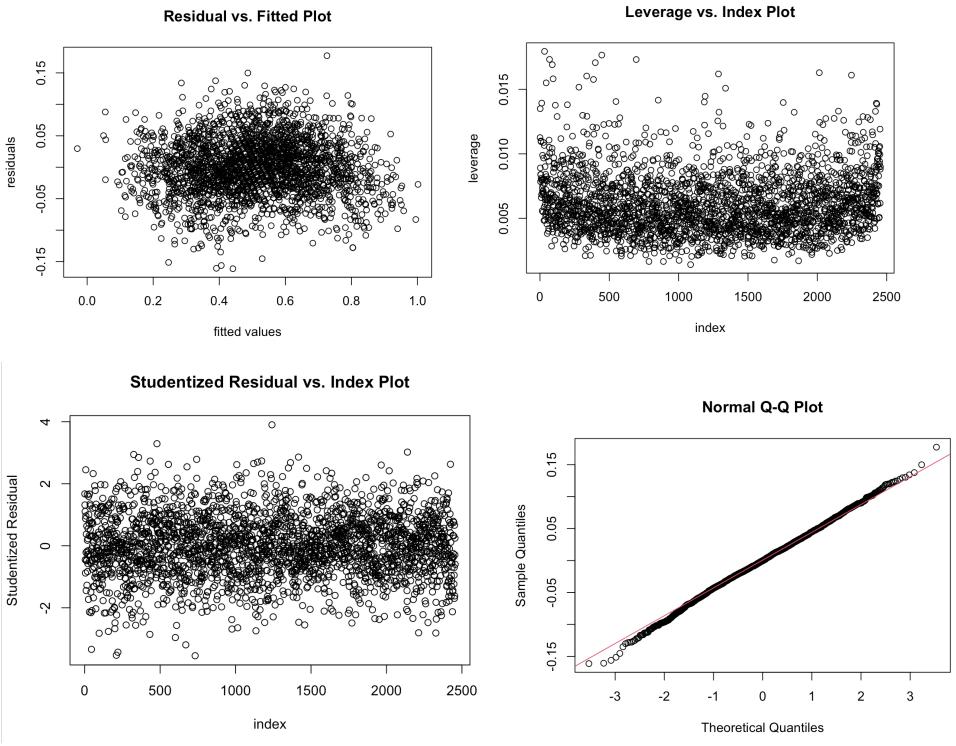


Exhibit 2: Interaction Terms

We examined a possible interaction of form (Power7)*(independent variable) where Power7 is indicator, and other continuous independent variables come from our dataset. No examples of relevant interactions were found, either due to statistical insignificance (high p-value) or low effect on model (negligible coefficient estimate). The graph of WP-ADJOE is below, with Power7 teams in blue, and mid-major teams in red. As we can see, the estimate lines for the two categories are relatively parallel, and hence no interaction.