College Basketball Dataset Analysis:

Based on Linear Model

Part 1. Introduction

In view of the new trend of data science development, machine learning and data mining techniques often appear in sports projects. As a new interdisciplinary discipline, sports analysis has attracted the attention of scholars in different fields and various groups, and the number of international seminars and related journal papers on sports analysis as the theme has gradually increased. Basketball is a sport that requires comprehensive enumeration of parameters. In order to better evaluate the performance of players and teams and help coaches make better decisions, the use of statistical models can effectively solve this problem based on the comprehensive data set of players, coaches and teams' performance on the basketball court.

To this end, we conduct a modeling analysis of winning rate using a linear regression model based on Andrew Sundberg's College Basketball Dataset, looking for the key factors that influence a team to win a game.

Part 2. Explanatory Data Analysis

2.1 Variables Introduction

The data include Division I college basketball seasons from 2013 to 2019, and all variables are shown in Table 1.

Table 1 Variables Introduction

Variables	Explanation
TEAM	The Division I college basketball school
CONF	The Athletic Conference in which the school participates in
G	Number of games played
W	Number of games won
PWIN	W/G, odds of winning
ADJOE	Adjusted Offensive Efficiency (An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average Division I defense)
ADJDE	Adjusted Defensive Efficiency (An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have

	against the average Division I offense)
BARTHAG	Power Rating (Chance of beating an average Division I team)
EFG_O	Effective Field Goal Percentage Shot
EFG_D	Effective Field Goal Percentage Allowed
TOR	Turnover Percentage Allowed (Turnover Rate)
TORD	Turnover Percentage Committed (Steal Rate)
ORB	Offensive Rebound Rate
DRB	Offensive Rebound Rate Allowed
FTR	Free Throw Rate (How often the given team shoots Free Throws)
FTRD	Free Throw Rate Allowed
2P_O	Two-Point Shooting Percentage
2P_D	Two-Point Shooting Percentage Allowed
3P_O	Three-Point Shooting Percentage
3P_D	Three-Point Shooting Percentage Allowed
ADJ_T	Adjusted Tempo (An estimate of the tempo (possessions per 40 minutes) a team would have against the team that wants to play at an average Division I tempo)
WAB	Wins Above Bubble (The bubble refers to the cut off between making the NCAA March Madness Tournament and not making it)
POSTSEASON	Round where the given team was eliminated or where their season ended
SEED	Seed in the NCAA March Madness Tournament

2.2 Descriptive Statistics

After removing the two variables with too many missing values (POSTSEASON and SEED), the descriptive statistics are shown in Table 2.

Table 2 Descriptive Statistics

vars	n	mean	sd	min	max	se
*TEAM**	2455	178.81	102.52	1.00	355.00	2.069
*CONF**	2455	17.94	10.54	1.00	35.00	0.213
G	2455	31.49	2.66	15.00	40.00	0.054
W	2455	16.28	6.61	0.00	38.00	0.133
PWIN	2455	0.51	0.18	0.0	0.97	0.000
ADJOE	2455	103.30	7.38	76.60	129.10	0.149
ADJDE	2455	103.30	6.61	84.00	124.00	0.133
BARTHAG	2455	0.49	0.26	0.01	0.98	0.005
EFG_O	2455	49.81	3.14	39.20	59.80	0.063

EFG_D	2455	50.00	2.94	39.60	59.50	0.059
TOR	2455	18.76	2.09	11.90	27.10	0.042
TORD	2455	18.69	2.20	10.20	28.50	0.044
ORB	2455	29.88	4.13	15.00	43.60	0.083
DRB	2455	30.08	3.15	18.40	40.40	0.064
FTR	2455	35.99	5.25	21.60	58.60	0.106
FTRD	2455	36.27	6.25	21.80	60.70	0.126
X2P_O	2455	48.80	3.38	37.70	62.60	0.068
X2P_D	2455	48.98	3.34	37.70	61.20	0.067
X3P_O	2455	34.41	2.79	24.90	44.10	0.056
X3P_D	2455	34.60	2.42	27.10	43.10	0.049
ADJ_T	2455	67.81	3.28	57.20	83.40	0.066
WAB	2455	-7.80	6.97	-25.20	13.10	0.141
YEAR	2455	2016.01	2.00	2013.00	2019.00	0.040

A sample matrixplot using the VIM::matrixplot() function is shown in Figure 1.

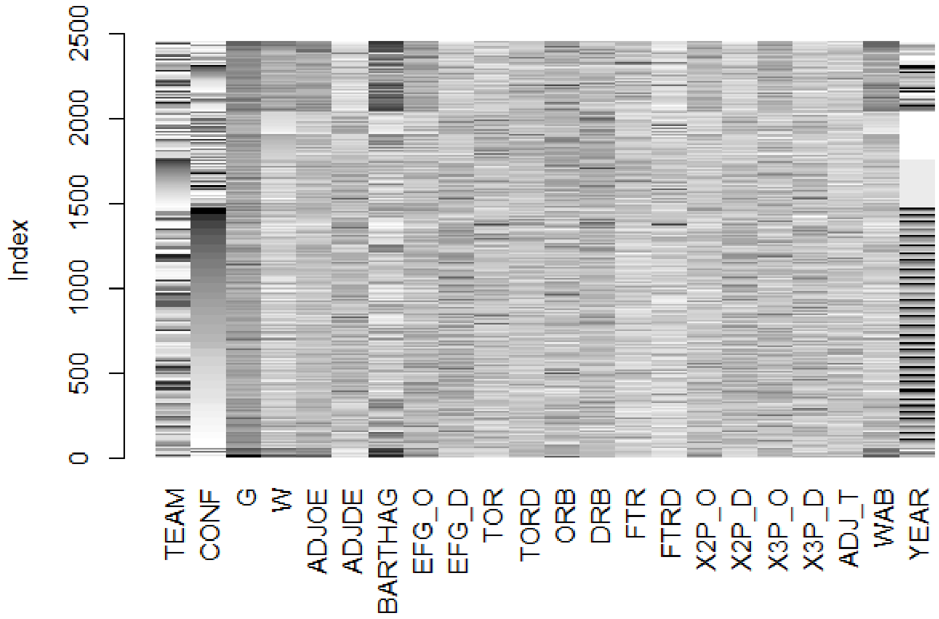

Figure 1 Sample matrix diagram

2.3 Data Exploration

The thermal map of correlation between variables is shown in Figure 2.

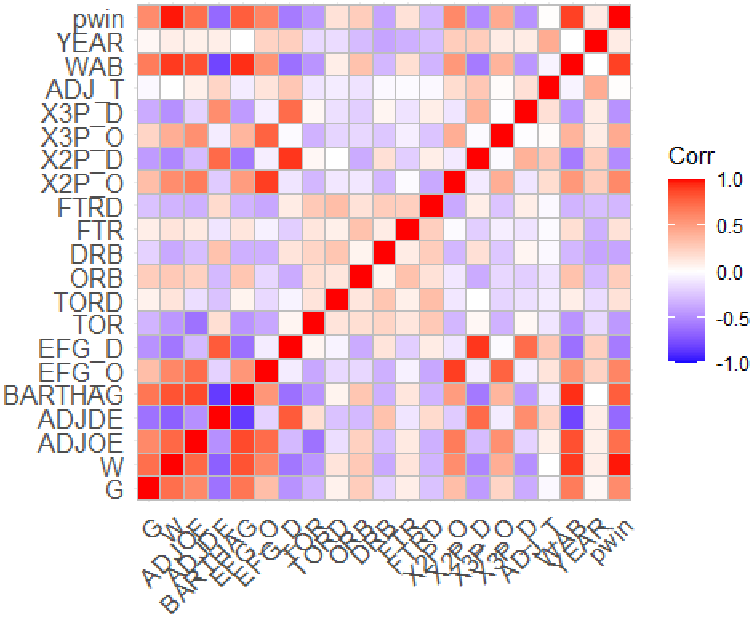

Figure 2 Correlation coefficient matrix heat map

In later modeling, we view the *pwin* variable (winning rate) as either the explained variable or the predictive variable. Therefore, a bar chart of correlation coefficients (in descending absolute value order) between each variable and the winning rate is shown below. It can be found that WAB, BARTHAG, ADJOE and other variables have a high correlation with winning rate.

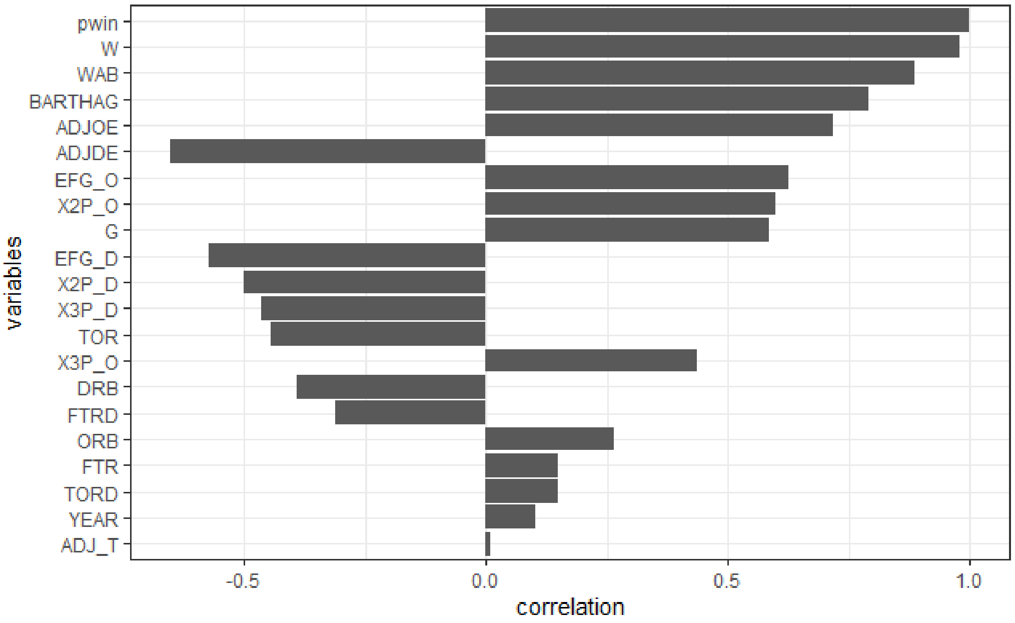

Figure 3 Correlation between each variable and winning rate

Part 3.Linear Model

3.1 Model Building

In order to get a reasonable explanation for the winning rate, we use the stepwise regression method to obtain the best model. All variables are first used for the regression of winning rate, and the benchmark regression results are shown in Column 1 of Table 3. In the second step, we delete the variables with insignificant coefficients in the benchmark regression and run the regression again, and the results are shown in the second column. Finally, we remove variables that may cause bidirectional causality problems (such as *w, g, wab*, etc.), and the regression results are shown in the third column.

To sum up, the third column of Table 3 is the final model. It can be found that the regression coefficients of all independent variables are significant at the statistical level of 1%. Among them, barthag has the greatest influence on winning rate with a coefficient of 0.208. This is followed by the efg_o variable with a coefficient of 0.030.

Table 3 Regression Results

	(1)	(2)	(3)
VARIABLES	**pwin**	**pwin**	**pwin**
g	-0.018***	-0.018***	
	(-80.688)	(-81.001)	
w	0.032***	0.032***	
	(142.146)	(143.379)	
adjoe	-0.003***	-0.003***	-0.006***
	(-11.516)	(-11.516)	(-5.654)
adjde	0.002***	0.002***	0.004***
	(7.228)	(7.303)	(4.543)
barthag	0.111***	0.111***	0.208***
	(16.140)	(16.158)	(6.267)
efg_o	0.003***	0.002***	0.030***
	(2.915)	(6.244)	(27.656)
efg_d	0.001	-0.001*	-0.025***
	(0.753)	(-1.876)	(-22.260)
tor	-0.002***	-0.002***	-0.025***
	(-6.256)	(-6.147)	(-20.602)
tord	0.001**	0.001**	0.023***
	(2.390)	(2.574)	(22.888)
orb	0.001***	0.001***	0.010***
	(4.850)	(4.685)	(17.162)
drb	-0.000*	-0.000*	-0.012***
	(-1.912)	(-1.876)	(-18.650)
ftr	0.000***	0.000***	0.002***
	(5.551)	(5.422)	(7.185)
ftrd	-0.000	-0.000	
	(-0.175)	(-0.218)	
x2p_o	-0.001		
	(-1.462)		
x2p_d	-0.001		

	(-1.027)		
x3p_o	-0.001		
	(-1.396)		
x3p_d	-0.001		
	(-1.273)		
adj_t	0.000	0.000	0.004***
	(1.217)	(1.054)	(8.163)
wab	-0.001***	-0.001***	
	(-5.532)	(-5.653)	
year	0.001***	0.001***	
	(3.571)	(4.007)	
Constant	-0.993**	-1.130***	0.040
	(-2.376)	(-2.789)	(0.544)
Observations	2,455	2,455	2,455
R-squared	0.994	0.994	0.849

*Note: *** p<0.01, ** p<0.05, * p<0.1. T-statistics in parentheses.*

3.2 Model Validation

We mainly test the model residuals. Figure 4 shows that the model residuals are basically independent of the estimated values, which satisfies the basic assumption of the linear regression model on the residuals.

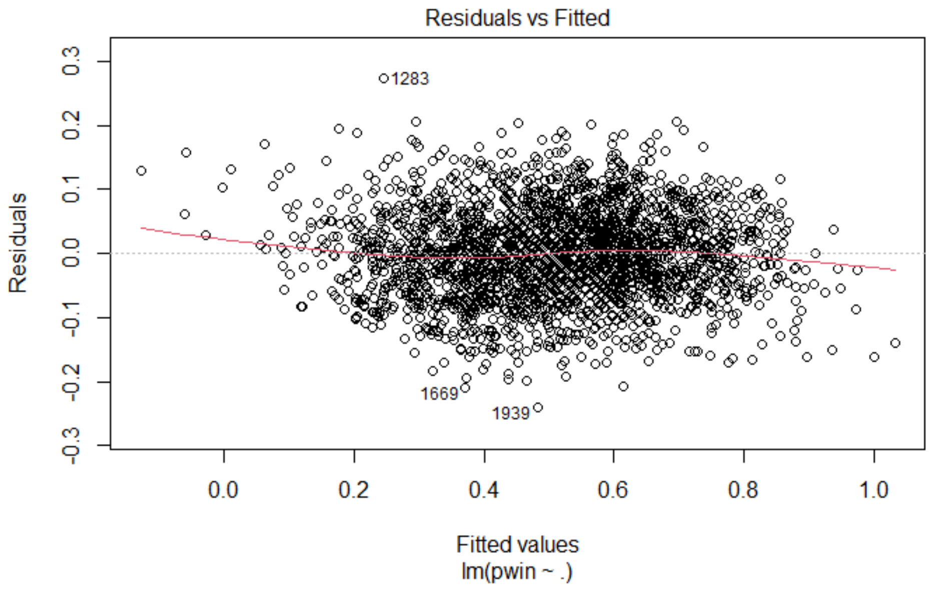

Figure 4 Residuals vs fitted

The distribution of the residuals is tested below. In Figure 5, the scatter is basically distributed on the y=x line, indicating that the residual basically follows the normal distribution, indicating that the model extracts enough information.

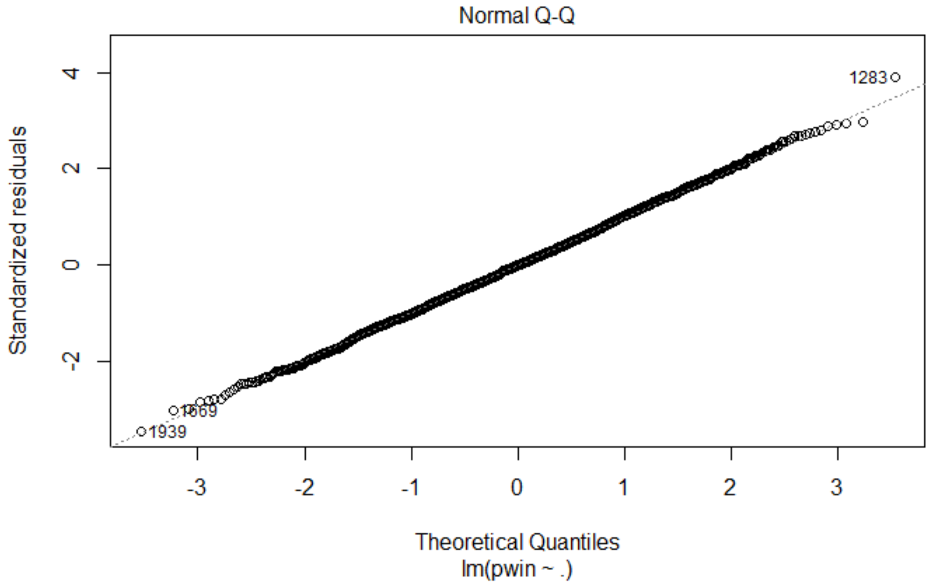

Figure 5 Residuals QQ-plot

Part 4. Results

Based on Andrew Sundberg's College Basketball Dataset, this paper uses multiple linear regression model to model and analyze the winning rate. The results show that Power Rating and Effective Field Goal Percentage Shot have significant positive effects on winning rate. Turnover Percentage and Offensive Rebound Rate Allowed have a significant negative effect on winning rate. Therefore, a team in the training, should try to consider the above four factors to improve the winning rate.

Reference

Sundberg. “College Basketball Dataset”. Kaggle(2021).

https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset