NORTHERN ILLINOIS UNIVERSITY

Building the Winning Percentage Model to Predict Regular Season Results of NBA Teams Based on Regression and Time Series Analysis of Common Basketball Statistics

A Thesis Submitted to the

University Honors Program

In Partial Fulfillment of the

Requirements of the Baccalaureate Degree

With Upper Division Honors

Department Of

Statistics

By

Sinong Ou

DeKalb, Illinois

May 13, 2017

University Honors Program

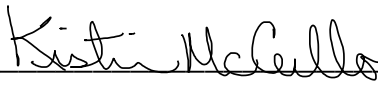
Capstone Approval Page

Capstone Title (print or type)

Building the Winning Percentage Model to Predict Regular Season Results of NBA Teams Based on Regression and Time Series Analysis of Common Basketball Statistics

Student Name (print or type) _____Sinong Ou_____

Faculty Supervisor (print or type) ___Kristin McCullough_____

Faculty Approval Signature 

Department of (print or type) _____Statistics_____

Date of Approval (print or type) _____May 10, 2017_____

Building the Winning Percentage Model to Predict Regular Season Results of NBA Teams Based on Regression and Time Series Analysis of Common Basketball Statistics

Sinong Ou

Prof. Kristin McCullough

May 3rd, 2017

Building the Winning Percentage Model to Predict Regular Season Results of NBA Teams Based on Regression and Time Series Analysis of Common Basketball Statistics

ABSTRACT

With the trend to apply statistics to predict sport games, the purpose of this paper is to find a model that can help to predict the percentage of games won for NBA teams during a season based on data for team and individual player performance. Multiple linear regression is used to build a predictive model, while time series analysis is used to assist with model selection. Great care is taken here, because statistical software will build a model regardless of collinearity, which means the model contains highly correlated variables, and despite whether regression assumptions are met. A general model that can predict all teams' performance is found. The model basically fits every team, and even the worst predictions look decent. However, each team has its own philosophy, so each has different significant factors. Thus models built for individual teams perform better.

TABLE OF CONTENTS

INTRODUCTION AND BACKGROUND

In 2001, Billy Beane (manager of the Oakland Athletics, an MLB team) used statistics in order to win many games at the third lowest operating expense. This was beyond anyone's expectations and it broke the stereotype of, "pay more, win more." More and more teams started to believe that data was truly helpful for a team's operation. Data scientists were hired by teams all over the world to find indices that contribute to game-winning and effectively optimize every dollar they spent. Hidden figures, hidden rules, hidden tactics, and hidden data became the key to success. Provided statistics can be used to determine outcomes in baseball, this paper will look at statistics for basketball. For example, in today's NBA, every team has at least two data analysts, whose daily statistics directly affects the manager's decisions. The NBA has no model as complete and accurate as Beane built for baseball. This paper will give a complete overview of how multiple linear regression models can predict the percentage of games won during a season for teams by finding relationships between varieties of indices. Teams have been through different reconfigurations, such as when a manager resigns or a superstar player leaves. Different restructuring conditions create variables in how teams react. Three teams, Chicago Bulls, San Antonio Spurs, and Golden State Warriors, are chosen to demonstrate the precision of the predictive models and offer an interpretation of the data.

It is known that basketball was created in the USA, where the most contributions to the development of basketball has been made. The NBA (created seventy years ago) has introduced new elements to the game. As examples, the NBA has introduced the "24-second rule," "48-minute rule," "three-second defense area rule," and "three pointers." The trends in tactics have been changing, as well. For instance, Steven Curry led the Golden State Warriors, as a shooting team, to win the championship for the first time in 2015. Nobody believed that a shooting team

was able to win the championship before that. In short, the factors that affect the game have been changing. “Professional intuition” has gradually become insufficient for figuring out the right way to win the game. Data, instead, takes over that responsibility. It is possible, using multiple linear regression, for the data to drive results.

Under the stereotype of a good center guaranteeing a championship, basketball was considered as a sport designed for tall players. However, Michael Jordan (who is not a center) broke this stereotype by reigning in the league for 14 years. As a score guard, Jordan dunked, blocked, and stole every ball, victory, and championship from those proud centers. Prior to Michael Jordan, nobody showed such powerful scoring ability or constant stamina. Jordan is now considered a god of basketball because of his skills.

In the 1979-1980 season, 3-point shots were introduced to the NBA. When first introduced, coaches felt 3-pointers lacked tactical forethought and were lazy. The stereotype is “ the closer players are to the basket, the more likely they will be to put the ball into the basket.” With the success of shooting teams in early 21st century and the Golden States Warriors winning the championship, critics were silenced. They know that they are not aware of the new trends that the data has shown. Since the Golden State Warriors’ amazing championship win, data mining and analyzing have become the NBA’s general managers’ most important daily job. Another example is in the 2016-2017 season Daryl Morey, a great statistician and graduate of MIT and the general manager of the Houston Rockets, led his team to win the 3rd most games while spending comparably little money. Data can help a manager and a team to make wiser decisions.

MATERIALS AND METHODS

Today, big data is used in selecting the most significant indices when building a model. In the NBA win-percentages display how a team performs. The purpose for this research is to find variables, or indices, which contribute to the prediction of win percentage. Multiple linear regression and time series analysis are the statistical methods that will be used. Variables/indices considered are described in Table 1. These indices are chosen, because they are well established and data is collected on them for all teams in the NBA.

Multiple linear regression examines how multiple independent variables $X_1, X_2, \dots, X_n$ are related to a single dependent variable Y. It is a statistical tool used to create a model to predict an outcome. The model will have the form:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \beta_3 \mathrm{X}_3 + \cdots + \beta_\mathrm{n} X_n + E,$$

where E is a term representing any error in the model. To predict our dependent variable, win percentage, regression will identify the strength of the effect of an independent variable or group of independent variables, taken from Table 1, on win percentage. In this research, SAS will be used to run the regression and to build the model, i.e. find appropriate independent variables for the model and to estimate their regression parameters $\beta_i, i = 0,1, \dots, n$.

In SAS, three ways of model selection (forward selection, backward elimination, and stepwise regression) will specify the variables/indices that significantly help to predict win percentage. Model selection is the task of selecting a statistical model from a set of candidate models, given data. In the forward selection approach, one adds variables to the model one at a time. At each step, each variable that is not already in the model is tested for inclusion and the variable with the most significant contribution is added to the model, so long as it's p-value is

Variables/Indices	Abbreviations
Team Name	Team
Win Percentage	win_
Field Goals	FG
Field Goals Attempted	FGA
Field Goals Percentage	FG_
Two Points Made	threeP
Two Points Attempted	threePA
Two Points Percentage	threeP_
Three Points Made	twoP
Three Points Attempted	twoPA
Three Points Percentage	twoP_
Free Throw Made	FT
Free Throw Attempted	FTA
Free Throw Percentage	FT_
Offensive Rebound	ORB
Defensive Rebound	DRB
Total Rebound	TRB
Assistant	AST
Steal	STL
Block	BLK
Turnover	TOV
Personal Foul	PF
Points	PTS
Age	Age
Margin of Victory	MOV
Strength of Schedule	SOS
Simple Rating System	SRS
Offensice Rating	Ortg
Defensive Rating	DRtg
Pace	Pace
Free Throw Attempted Rate	FTr
Three-point Attempted Rate	threePAr
True-shooting Percentage	TS_
Effective Field Goals Percentage	eFG_
Turnover Rate	TOV_
Rebound Rate	RB_
Free Throw/Field Goals Attempted	FT_FGA
Season	Season

Table 1: List of variables/indices included in the data set.

below some pre-set level, for example 0.05. A p-value is a number that measures the significance. Low p-values are desired. Unlike forward selection, backward elimination begins

with the full least squares model containing all predictors and then iteratively removes the least useful predictor, one-at-a-time. Stepwise selection is a method that allows moves in either direction, dropping or adding variables at the various steps.

Additionally, it is important to verify that standard regression assumptions are met. If they are not, then the model will not produce accurate estimates for win percentage. For example, this verification is partially done through residual analysis. Residual is another word for error when using regression. The residuals must be normally distributed with a mean of zero and a constant variance. When SAS runs the model selections, the resulting output will show graphs that can help to test those regression assumptions. Another concern when applying multiple linear regression is collinearity, which means the model contains highly correlated independent variables. Correlation coefficients are computed to check for this. Values close to 1 or -1 indicate that there is a strong linear association between variables. Note that we want high correlation between win percentage and the independent variables but not between pairs of the independent variables. If collinearity is present, the model built by SAS might be inaccurate due to unstable estimates and inflated standard errors for the regression parameters.

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time and plotted via line charts. Excel can be used to create time series plots for all variables listed in Table 1. By studying the plot for win percentage for an individual team, big changes and/or trends can be identified. Variables that may have had an effect are those whose plots match that of the win percentage. Here, time series plots will help to improve and modify the model.

COMPUTATION, ANALYSIS, AND OBSERVATIONS

I. Multiple Linear Regression Analysis to Build a General Model

The correlation between pairs of variables from Table 1 was computed using SAS. Part of the SAS output is shown below in Figure 1. The top row shows the correlation between win percentage and the independent variables/indices. The data represents the level of correspondence between the indices. The Indices that had a correlation greater than 0.7 were selected to build the model. Those indices were: PTS, FG, FG_, FGA, MOV, ORB, DRB, TRB, AST, BLK, STL, twoP, twoP_, eFG_, SRS, and TS_. Several interaction and higher order terms were considered for the model as well; see Table 2. The correlation of each with win percentage was computed, terms that had a correlation greater than 0.7 were selected as additional indices to use to build the model. Those indices were: k1, k2, k3, k4, k5, k6, k7, k8, and k9.

	win_	FG	FGA	FG_	threeP	threePA	threeP_	twoP	twoPA	twoP_	FT	FTA	FT_	ORB	DRB	TRB	AST	STL	BLK	TOV
win_	1.00000	0.75805 <.0001	-0.14339 <.0001	0.82143 <.0001	0.32982 <.0001	0.18444 <.0001	0.39323 <.0001	0.45586 <.0001	-0.22916 <.0001	0.81658 <.0001	0.44171 <.0001	0.40416 <.0001	0.21785 <.0001	-0.06492 0.0360	0.68460 <.0001	0.50241 <.0001	0.66247 <.0001	0.35112 <.0001	0.47585 <.0001	-0.29587 <.0001
FG	0.75805 <.0001	1.00000	0.27883 <.0001	0.77883 <.0001	0.07872 0.0109	-0.05998 0.0527	0.27520 <.0001	0.81052 <.0001	0.22618 <.0001	0.72948 <.0001	-0.12615 <.0001	-0.13788 <.0001	0.02583 0.4045	0.09799 0.0015	0.54548 <.0001	0.48374 <.0001	0.74598 <.0001	0.40558 <.0001	0.37831 <.0001	-0.38797 <.0001
FGA	-0.14339 <.0001	0.27883 <.0001	1.00000	-0.38378 <.0001	-0.08459 0.0062	-0.03227 0.2975	-0.13328 <.0001	0.28735 <.0001	0.67307 <.0001	-0.40843 <.0001	-0.59126 <.0001	-0.58318 <.0001	-0.14236 <.0001	0.63055 <.0001	-0.41808 <.0001	0.02367 0.4450	0.02692 0.3849	0.46750 <.0001	-0.32210 <.0001	-0.58364 <.0001
FG_	0.82143 <.0001	0.77883 <.0001	-0.38378 <.0001	1.00000	0.12772 <.0001	-0.03934 0.2040	0.34927 <.0001	0.59338 <.0001	-0.21933 <.0001	0.96734 <.0001	0.26381 <.0001	0.24678 <.0001	0.11822 0.0001	-0.31630 <.0001	0.79758 <.0001	0.45042 <.0001	0.69998 <.0001	0.08380 0.0067	0.57232 <.0001	0.00891 0.7738
threeP	0.32982 <.0001	0.07872 0.0109	-0.08459 0.0062	0.12772 <.0001	1.00000	0.95521 <.0001	0.45187 <.0001	-0.51929 <.0001	-0.76151 <.0001	0.25623 <.0001	-0.01060 0.7323	-0.02139 0.4900	0.04120 0.1834	-0.20593 <.0001	0.00840 0.7864	-0.10956 0.0004	0.20467 <.0001	0.04177 0.1775	0.18420 <.0001	-0.10435 0.0007
threePA	0.18444 <.0001	-0.05998 0.0527	-0.03227 0.2975	-0.03934 0.2040	0.95521 <.0001	1.00000	0.27141 <.0001	-0.61285 <.0001	-0.76075 <.0001	0.13654 <.0001	-0.04745 0.1255	-0.04371 0.1581	-0.01414 0.6481	-0.13751 <.0001	-0.11048 0.0003	-0.16477 <.0001	0.08207 0.0080	0.01247 0.6874	0.10757 0.0005	-0.07789 0.0118
threeP_	0.39323 <.0001	0.27520 <.0001	-0.13328 <.0001	0.34927 <.0001	0.45187 <.0001	0.27141 <.0001	1.00000	-0.02859 0.3561	-0.28678 <.0001	0.29925 <.0001	0.08795 0.0045	0.04527 0.1438	0.15745 <.0001	-0.22318 <.0001	0.21245 <.0001	0.04156 0.1796	0.30127 <.0001	0.06532 0.0348	0.20534 <.0001	-0.10488 0.0007
twoP	0.45586 <.0001	0.81052 <.0001	0.28735 <.0001	0.59338 <.0001	-0.51929 <.0001	-0.61285 <.0001	-0.02859 0.3561	1.00000	0.64059 <.0001	0.47531 <.0001	-0.10238 0.0009	-0.10628 0.0006	-0.00167 0.9570	0.20318 <.0001	0.46302 <.0001	0.47830 <.0001	0.51995 <.0001	0.32201 <.0001	0.21767 <.0001	-0.27017 <.0001
twoPA	-0.22916 <.0001	0.22618 <.0001	0.67307 <.0001	-0.21933 <.0001	-0.76151 <.0001	-0.76075 <.0001	-0.28678 <.0001	0.64059 <.0001	1.00000	-0.36557 <.0001	-0.34899 <.0001	-0.34674 <.0001	-0.08124 0.0086	0.51087 <.0001	-0.18911 <.0001	0.13759 <.0001	-0.04272 0.1678	0.29450 <.0001	-0.28799 <.0001	-0.32115 <.0001
twoP_	0.81658 <.0001	0.72948 <.0001	-0.40843 <.0001	0.96734 <.0001	0.25623 <.0001	0.13654 <.0001	0.29925 <.0001	0.47531 <.0001	-0.36557 <.0001	1.00000	0.26607 <.0001	0.26039 <.0001	0.08464 0.0062	-0.32982 <.0001	0.77161 <.0001	0.42245 <.0001	0.68524 <.0001	0.06134 0.0475	0.58404 <.0001	0.02826 0.3616
FT	0.44171 <.0001	-0.12615 <.0001	-0.59126 <.0001	0.26381 <.0001	-0.01060 0.7323	-0.04745 0.1255	0.08795 0.0045	-0.10238 0.0009	-0.34899 <.0001	0.26607 <.0001	1.00000	0.95163 <.0001	0.34592 <.0001	-0.11999 0.0001	0.43215 <.0001	0.27372 <.0001	-0.00574 0.8532	0.04255 0.1695	0.16889 <.0001	0.06349 0.0403

Figure 1: Matrix of correlation coefficients produces using SAS.

Index	Shortcut	Index	Shortcut	Index	Shortcut
FG*FG_	i1	FT*FTA	i21	twoP*twoP	l3
FG*twoP	i2	FT*PF	i22	twoP_*twoP_	l4
FG*twoP_	i3	FT*FT_	i23	AST*AST	l5
FG*AST	i4	FT*FGA	i24	PST*PST	l6
FG*PTS	i5	DRB*TRB	i25	MOV*MOV	l7
FG*MOV	i6	DRB*PTS	i26	SRS*SRS	l8
FG*SRS	i7	DRB*MOV	i27	eFG_*eFG_	l9
FG*eFG_	i8	DRB*SRS	i28	FT*FT	l10
FG_*twoP_	i9	DRB*eFG_	i29	ORtg*ORtg	l11
FG_*DRB	i10	PTS*MOV	i30	DRtg*DRtg	l12
FG_*AST	i11	PTS*SRS	i31	threeP*ORB*FT	k1
FG_*PTS	i12	PTS*Ortg	i32	threeP*DRB*FT	k2
FG_*MOV	i13	PTS*eFG_	i33	FGA*AST*threePA	k3
FG_*SRS	i14	MOV*SRS	i34	STL*BLK*PTS	k4
FG_*eFG_	i15	ORtg*MOV	i35	TOV*PF	k5
AST*twoP_	i16	eFG_*MOV	i36	threePA*ORB*FTA	k6
PTS*twoP_	i17	ORtg*TS_	i37	threeP_*ORB*FT_	k7
MOV*twoP_	i18	ason*threeP/	i38	threeP_*ORB*FTA	k8
TS_*twoP_	i19	FG*FG	l1	threePA*ORB*FT_	k9
eFG_*twoP_	i20	FG_*FG_	l2	TRB*STL*BLK	k10

Table 2: New indices created from the variables in Table 1.

Next, the three methods of model selection, forward, backward and stepwise, were used to find the most appropriate indices among the selected indices to build the model. An example of the SAS output obtained when performing model selection is shown in Figure 2. When performing forward selection, SAS determined that the variables that should be in the predictive model are: MOV, FGA, BLK, PTS, eFG_, STL, TS_, k3, k4, k6, k8. Look to the fourth and the last column, which contain the coefficients of determination and p-values, respectively. The level of significance was set at 0.50, meaning variables with p-values no more than 0.5 are the best choices. Thus, all indices in the table below are considered the best variables. The coefficient of determination measures how much of the variation in win percentage is explained by the chosen indices. All values shown in the SAS output are close to one, which is desirable.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	MOV	1	0.9423	0.9423	32.0456	17024.1	<.0001
2	FGA	2	0.0012	0.9435	12.0017	21.85	<.0001
3	BLK	3	0.0004	0.9439	7.4110	6.57	0.0105
4	PTS	4	0.0003	0.9442	3.0592	6.36	0.0118
5	eFG_	5	0.0001	0.9443	3.1866	1.88	0.1709
6	k8	6	0.0001	0.9444	2.7814	2.42	0.1205
7	k6	7	0.0001	0.9446	2.6365	2.16	0.1423
8	k4	8	0.0001	0.9447	2.6478	2.00	0.1575
9	STL	9	0.0001	0.9448	2.8496	1.81	0.1787
10	AST	10	0.0001	0.9449	2.8668	2.00	0.1578
11	TS_	11	0.0000	0.9449	4.1201	0.75	0.3859
12	k3	12	0.0000	0.9449	5.5100	0.61	0.4332

Figure 2: SAS output showing the order in which variables were selected to enter the model using the forward selection model selection technique.

The following graphs in Figure 3 are part of the forward selection SAS ouptut and show whether the normality assumption for the residuals is met when using the selected model. As an example, the graph in the top left corner shows the residuals plotted against the predicted values for win percentage. It shows a random scatter, which is a good. It indicates that the residuals have a constant variance. The plot below it is a normal probability plot. Most of the points are on or around the line, which also indicates that the residuals are normally distributed. The graph in the bottom left corner shows the distribution of the residuals. Their distribution has the same shape as a normal distribution (symmetric bell curve), which enhances the proof of normality.

Figures 4 and 5 show parts of the SAS output that results when performing backward elimination. SAS determined that the variables that should be in the predictive model are: PTS,

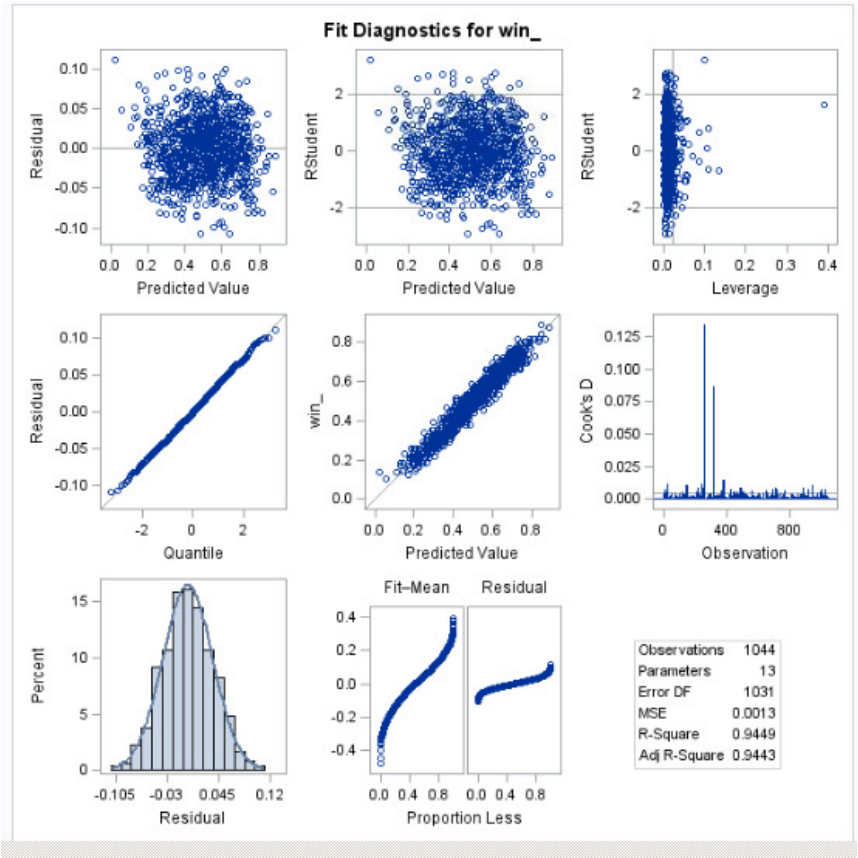


Figure 3: Residual analysis plots from SAS output for the model chosen through forward selection.

FG_, MOV, BLK, twoP, and eFG. Residual analysis shows that regression assumption are met here. When using stepwise regression, SAS determined that the variables that should be in the predictive model are: MOV, FGA, BLK, and PTS. The significance level was set much lower than 0.50 when using these two methods. Figure 6 shows the SAS output where SAS fits the model using the variables selected through stepwise regression. The estimates for the regression parameters are given. Note that they were computed using data for all teams. The general model built is:

$$Win\_ = 0.49986 + 0.01172 \times PTS + 0.00152 \times FGA + 0.01977 \times MOV + 0.00287 \times BLK$$

Figure 7 contains the residual analysis for this model. Regression assumption are met here, as well.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	k1	24	0.0000	0.9452	24.0218	0.02	0.8828
2	k2	23	0.0000	0.9452	22.0485	0.03	0.8700
3	SRS	22	0.0000	0.9452	20.0782	0.03	0.8633
4	k9	21	0.0000	0.9452	18.1304	0.05	0.8190
5	k7	20	0.0000	0.9452	16.2332	0.10	0.7481
6	ORB	19	0.0000	0.9452	14.5029	0.27	0.6028
7	FG	18	0.0000	0.9452	12.8673	0.37	0.5452
8	FGA	17	0.0000	0.9452	10.8835	0.02	0.8986
9	k5	16	0.0000	0.9452	9.2568	0.38	0.5399
10	twoP_	15	0.0000	0.9451	7.9364	0.68	0.4081
11	k3	14	0.0000	0.9451	6.6577	0.73	0.3941
12	TS_	13	0.0000	0.9451	5.2419	0.59	0.4430
13	TRB	12	0.0001	0.9450	4.3081	1.08	0.3000
14	DRB	11	0.0000	0.9450	2.4710	0.16	0.6853
15	AST	10	0.0001	0.9449	2.5137	2.06	0.1513
16	STL	9	0.0001	0.9448	2.4305	1.93	0.1648
17	k4	8	0.0001	0.9447	2.1501	1.73	0.1884
18	k6	7	0.0001	0.9446	2.5201	2.39	0.1228
19	k8	6	0.0001	0.9444	2.9991	2.49	0.1147

Figure 4: SAS output, for backward elimination, showing a list of variables removed, in order, from a model containing all selected indices.

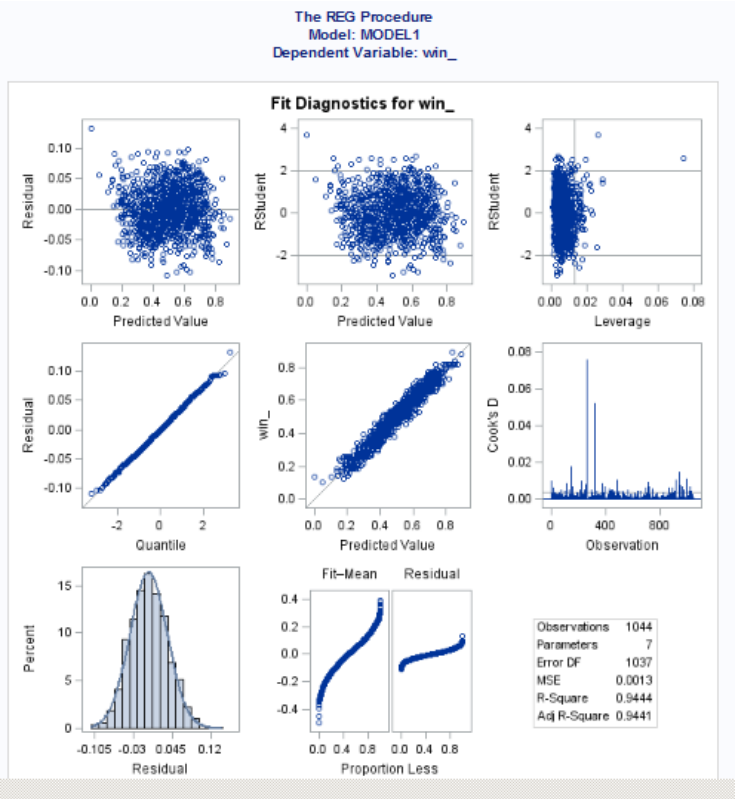


Figure 5: Residual analysis plots from SAS output for the model chosen through backward elimination.

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.49986	0.00114	260.23137	192912	<.0001
PTS	0.01172	0.00465	0.00858	6.36	0.0118
FGA	-0.00152	0.00040213	0.01940	14.38	0.0002
MOV	0.01977	0.00489	0.02206	16.35	<.0001
BLK	0.00287	0.00110	0.00922	6.84	0.0091

Figure 6: SAS output containing information about the fitted multiple linear regression model chosen through stepwise regression.

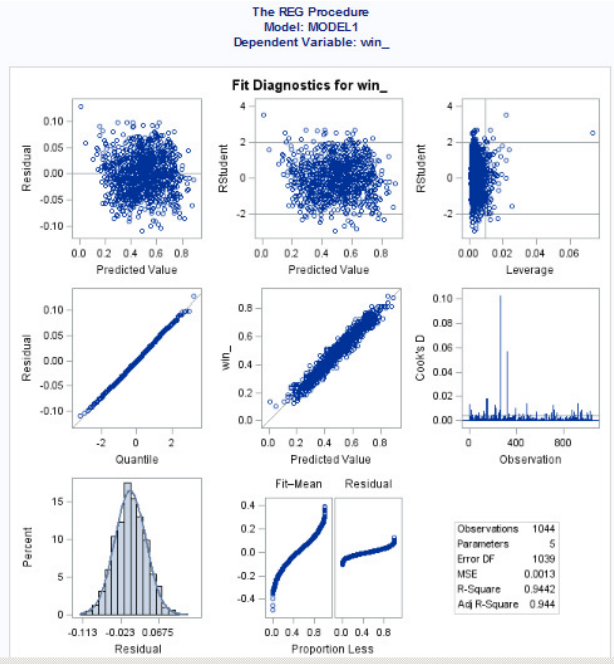


Figure 7: Residual analysis plots from SAS output for the model chosen through stepwise regression.

Then the stepwise regression model is applied to the data just for the Chicago Bulls to see if the model fits the team's winning percentage over the past 38 seasons. Figure 8 shows a comparison of the time series plot for win percentage and the predicted win percentage using the regression model. The model has produced results very similar to the Chicago Bulls' actual history. As the model is tested on other teams, the data can vary significantly as shown in Figure 9. When considering the reasons behind the imprecise predictions for the Golden State Warriors, the highly correlated indices could cause collinearity. Figure 10 reveals that the selected model contains two highly correlated indices: PTS and MOV. Certainly, they cannot be kept in the model at the same time when their interaction index is not significantly helpful. Thus, this model cannot be successfully applied to all teams.

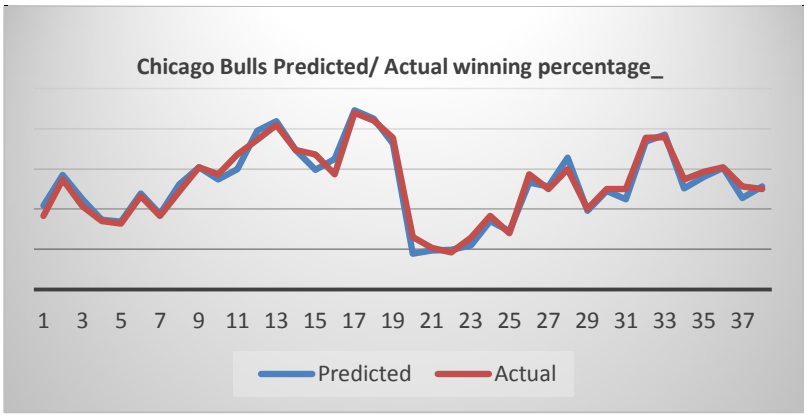


Figure 8: Time series plot for win percentage of the Chicago Bulls with regression model overlaid, showing the predicted win percentages.

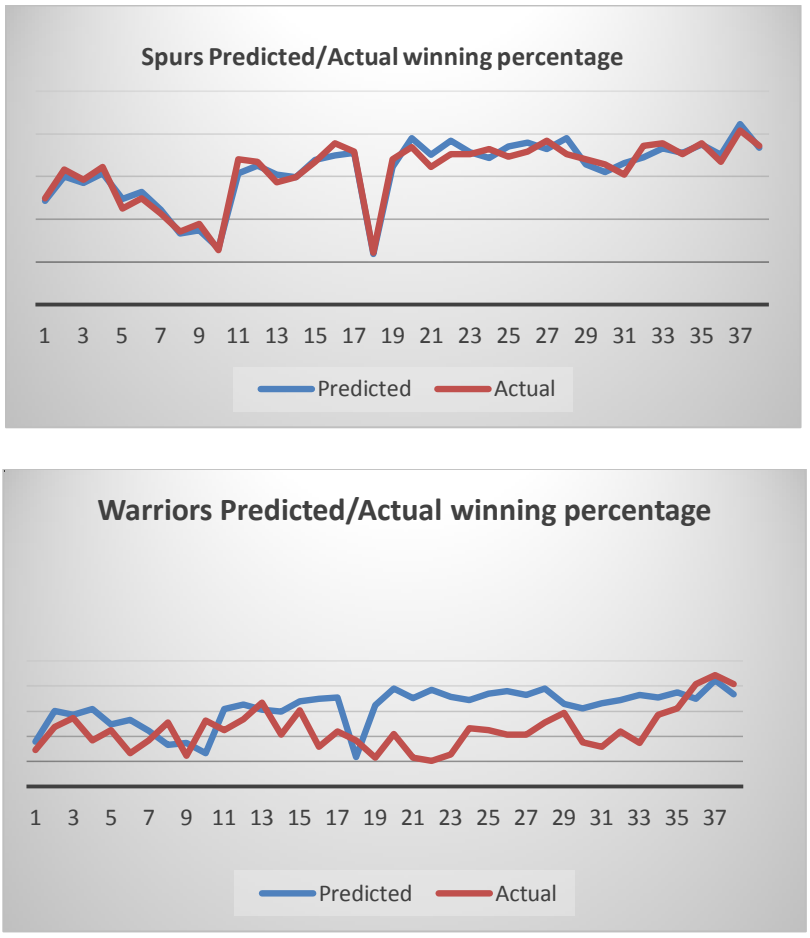


Figure 9: Time series plot for win percentage of the San Antonio Spurs and the Golden State Warriors with regression model overlaid, showing the predicted win percentages.

Pearson Correlation Coefficients, N = 1044 Prob > |r| under H0: Rho=0				
	PTS	FGA	MOV	BLK
PTS	1.00000	-0.10973 0.0004	0.99873 <.0001	0.46214 <.0001
FGA	-0.10973 0.0004	1.00000	-0.11247 0.0003	-0.32210 <.0001
MOV	0.99873 <.0001	-0.11247 0.0003	1.00000	0.46421 <.0001
BLK	0.46214 <.0001	-0.32210 <.0001	0.46421 <.0001	1.00000

Figure 10: Matrix of correlation coefficients.

The time series analysis should be applied to find the model for each team because every team manager builds a team with their own philosophy. Some managers like to build the team based on superstars. Others build their teams based on a reputable coach. Some like to attract players with great ability regardless of their personality. Another likes their players to be self-disciplined and great play-executors. Some like to run and shoot; but others like letting the center score. The next step is analyzing each team's data to find a model fitting each team.

II. Model modification

For model modification for individual teams, the time series plots of all indices need to be compared with that of the winning percentage. Choose indices that display similarities to the win percentage plot. Some graphs are shown in Figure 11 for the Chicago Bulls. Here is the principal for choosing similarity of plot graphs. In the time series graph of the win percentages, from the 19^{th} to the 20^{th} seasons, there is a big drop. This results from Michael Jordan, Scottie

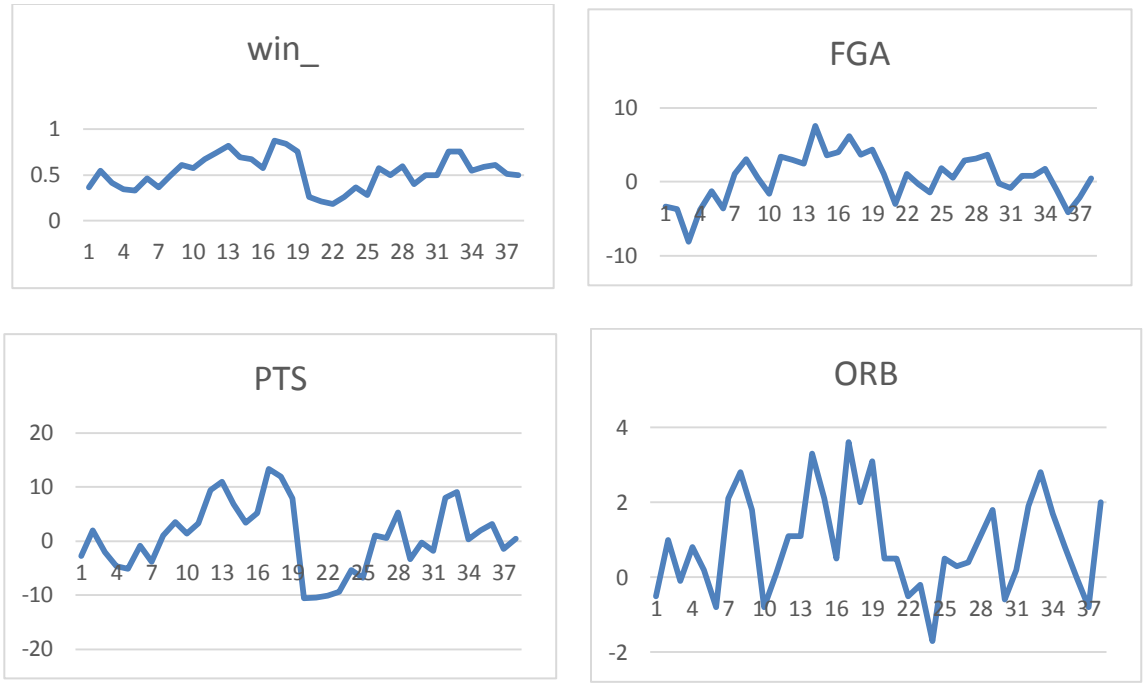


Figure 11: Time series plot for the Chicago Bulls of win percentage and the variables FGA, PTS, and ORB.

Pippen and Dennis Rodman's leaving. It is known that Jordan and Pippen are best scorers, so their leaving causes the drop in points. Rodman's leaving causes the drop in offensive rebound. The time series plots for win percentage and points clearly show the effect of big changes to the team. Those indices where the time series graph matches the interpretation would be significant to win percentage.

The chosen indices from the time series are then run through SAS to find their estimated regression parameters. The resulting fitted regression model is shown in Figure 12. The error in the fitted model might result from collinearity. Reduce the collineratiy of the model by eliminating highly correlated indices as before. The new model is:

$$win\_ = -0.002 \times FGA + 0.0008 \times AST + 0.00321 \times BLK + 0.03256 \times MOV - 0.1757 \times eFG\_$$

The improvement can be seen in Figure 13. This modified model fits the data well.

Note that this model can be applied successfully to other teams, too. It is applied to the San Antonio Spurs and the Golden State Warriors in Figure 14. Note that this model can be applied successfully to other teams. It is applied to the San Antonio Spurs and the Golden State Warriors in Figure 14. In fact the model fits well for most teams. The worst predictions were for the Los Angeles Clippers. See Figure 15. Even here the predictions are still very good for most seasons.

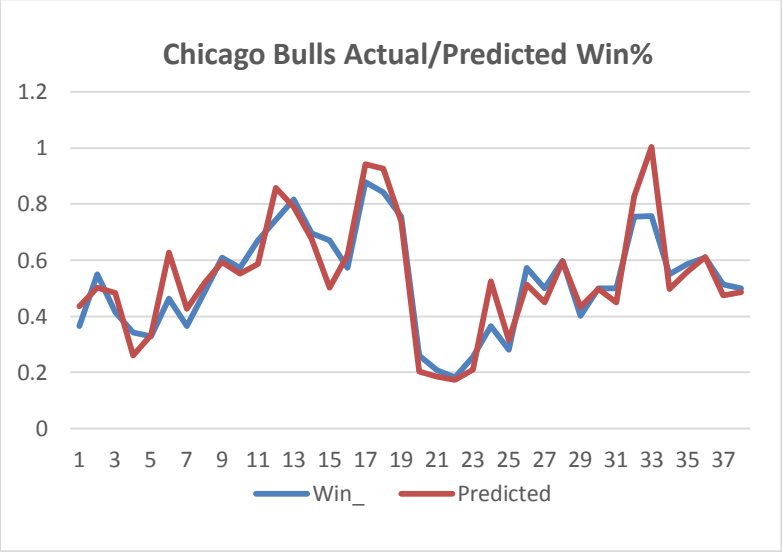


Figure 12: Time series plot for win percentage of the Chicago Bulls with regression model built specifically for the Bulls overlaid.

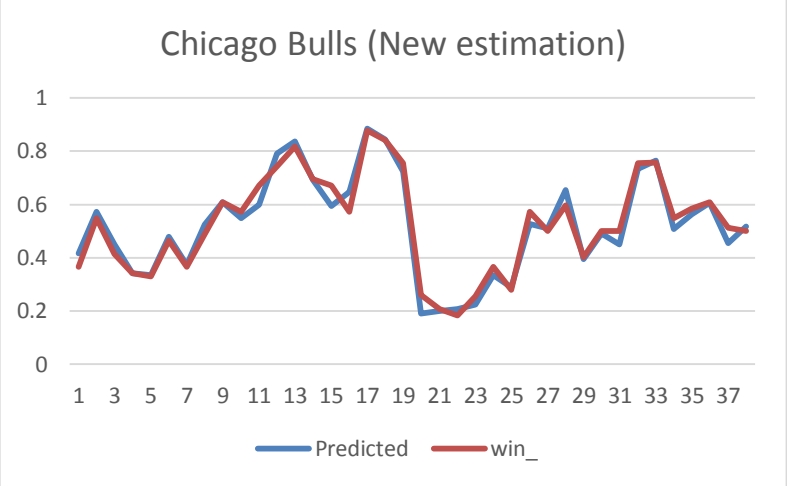
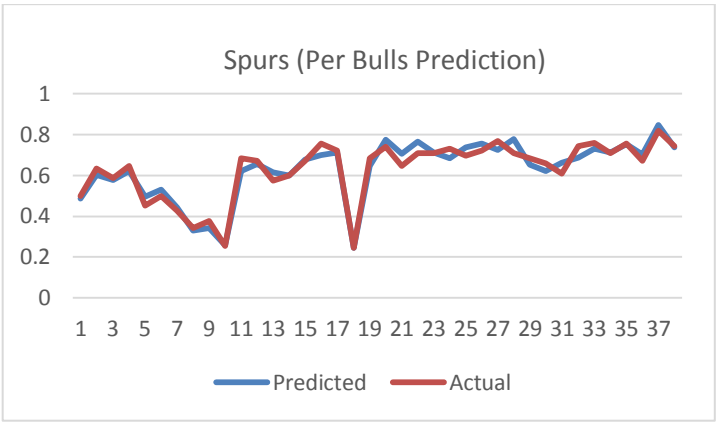


Figure 13: Time series plot for win percentage of the Chicago Bulls with the modified regression model (reduced collinearity) overlaid.



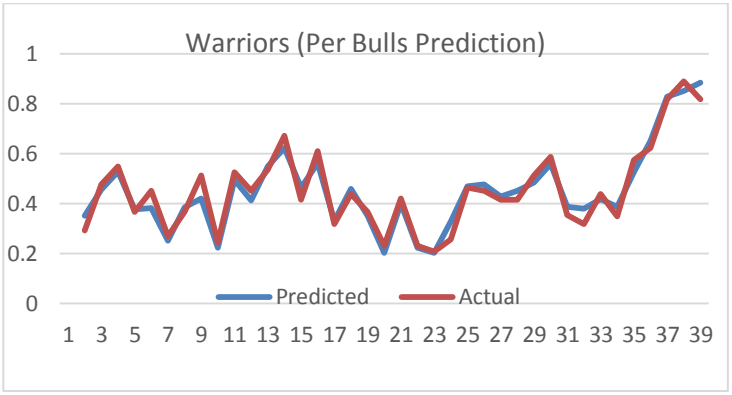


Figure 14: Time series plot for win percentage of the San Antonio Spurs and Golden Sate Warriors with the (Chicago Bulls') regression model overlaid.

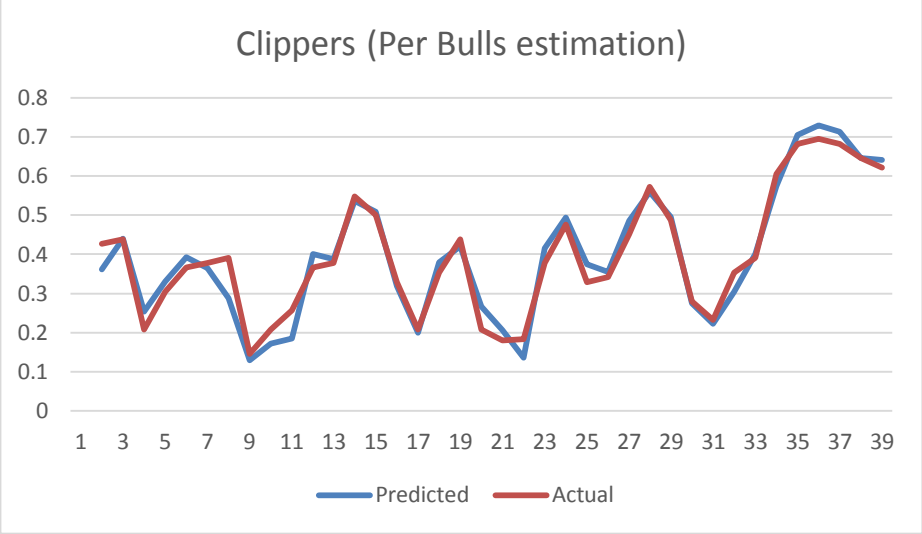


Figure 15: Time series plot for win percentage of the Los Angeles Clippers with the (Chicago Bulls') regression model overlaid.

Because different teams have different philosophies, building models based on individual teams is best. For example, some teams rely on a super center player, so its win percentage is affected by rebounds and blocks, while some teams rely on a run and shoot strategy, so its win percentage depends on three-point attempted or three-point percentage. Thus, rather than use the Chicago Bull's model for the Los Angeles Clippers, a more accurate model can be made using data for the Clippers. After using time series plots to choose indices and then fitting the model in SAS, we obtain the following model for the Clippers:

$$\text{win\_} = 0.25 + 1.87 \times \text{FG\_} - 1.063 \times \text{twoP\_} - 0.018 \times \text{DRB} + 0.011 \times \text{TRB} + 0.012 \times \text{BLK} - 0.1 \times \text{PF} + 0.024 \times \text{PTS} + 0.0088 \times \text{Age}$$

Figure 16 shows the superiority of this model. Note that collinearity is present in this model.

The standard error of the estimates of the regression parameters was checked. The indices FG_, twoP_, DRB, and TRB had variance inflation factors greater than 10, which is too high.

Looking at the variance inflation factor is another way to check for collinearity, similar to checking correlation coefficients. After removing collinearity the model for the Clippers becomes:

$$\text{win\_} = 0.26367 + .01001 \times \text{BLK} - -.00531 \times \text{PF} + 0.02581 \times \text{PTS} + 0.0082 \times \text{Age}$$

However, in this case, the new model does not fit the data as well. See Figure 17. So the original Clippers model is deemed best.

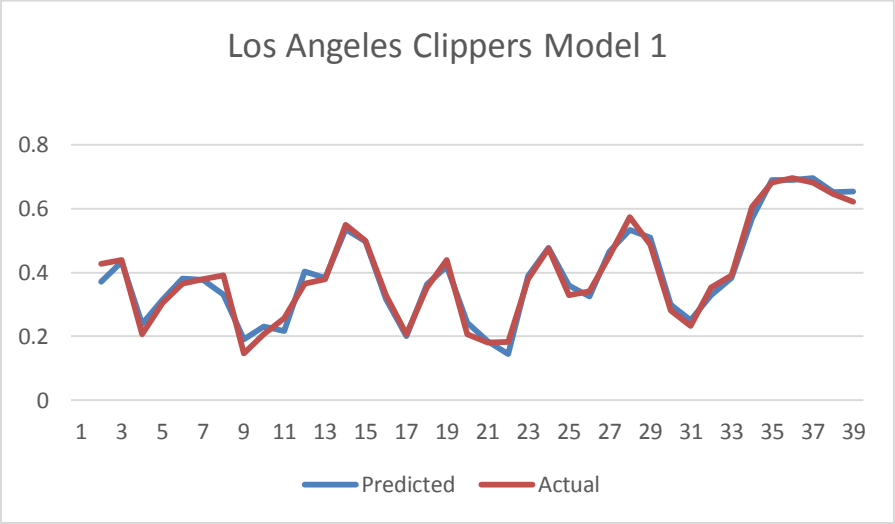


Figure 16: Time series plot for win percentage of the Los Angeles Clippers with regression model built specifically for the Clippers overlaid.

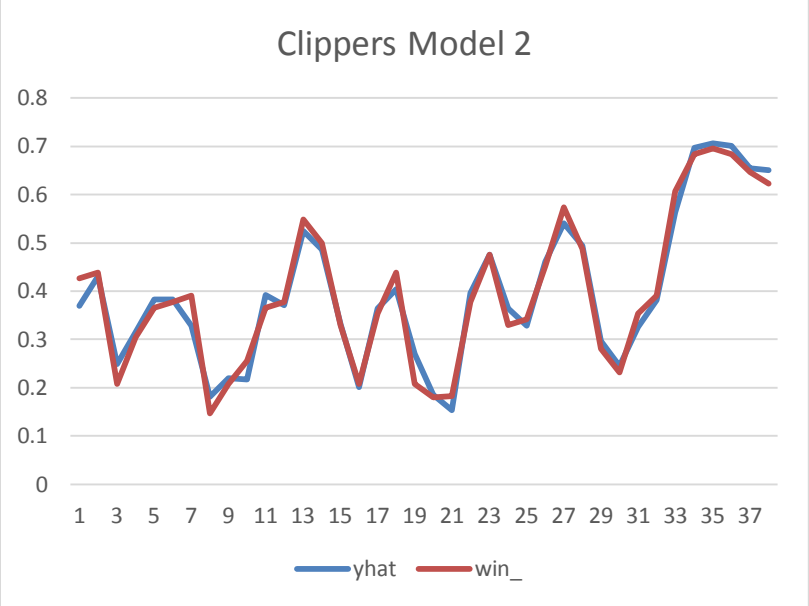


Figure 17: Time series plot for win percentage of the Los Angeles Clippers with the modified regression model (reduced collinearity) overlaid.

CONCLUSION

A general model to predict the percentage of wins a team will have during a season can be built through multiple linear regression. Using model selection techniques, indices that significantly help to predict win percentage can be determined. Different models result from the different model selection techniques. Using the stepwise regression model selection, the indices that significantly help to predict win percentage were determined to be points, field goals attempted, margin of victory, and block. The general fitted regression model is:

$$Win\_ = 0.49986 + 0.01172 \times PTS + 0.00152 \times FGA + 0.01977 \times MOV + 0.00287 \times BLK$$

The model works well, as shown in the example Chicago Bulls and San Antonio Spurs examples.

However, looking at an individual team, there could be different significant indices unique to that team. For example, some managers like to build the team based on an exception player. That player's strengths and weaknesses will influence which indices need to be included in a model to predict win percentage. Time series analysis was used to help determine which indices were significant for individual teams. By looking at the time series plot of win percentage, trends can be found. Then by comparing the time series plots for the indices with the plot for win percentage, indices that contribute to a change in trend are identified.

Models for the Chicago Bulls and the Los Angeles Clippers were built using time series and multiple linear regression. The indices included in the models differ from each other and from the general model. Thus, the best way of predicting the win percentage of each team is analyzing each team, individually, i.e. the model for a team should be built upon its own data.

WORK CITATION

Kleinbaum, David G., Lawrence L. Kupper, Azhar Nizam and Eli S. Rosenberg. Applied Regression Analysis and Other Multivariable methods. United States of America: Cengage Learning, 2014. Print.

Hicks, Charles R., and Kenneth V. Turner, Jr. Fundamental Concepts in the Design of Experiments. Oxford: Oxford University Press, 1997. Print.

Cryer, Jonathan D., and Kung-Sik Chan. Time Series Analysis With Applications in R. United States of America: Springer, 2008. Print.

APPENDIX

Codes used in SAS

```
/*Bring data into SAS*/
data NBA ;
input Team $ win_ FG FGA FG_ threeP threePA threeP_ twoP twoPA twoP_ FT FTA FT_
      ORB DRB TRB AST STL BLK TOV PF PTS Age MOV SOS SRS ORtg DRtg
Pace FTr threePAr TS_ eFG_ TOV_ RB_ FT_FGA Season;

i1=FG*FG_;
i2=FG*twoP;
i3=FG*twoP_;
i4=FG*AST;
i5=FG*PTS;
i6=FG*MOV;
i7=FG*SRS;
i8=FG*eFG_;
i9=FG_*twoP_;
i10=FG_*DRB;
i11=FG_*AST;
i12=FG_*PTS;
i13=FG_*MOV;
i14=FG_*SRS;
i15=FG_*eFG_;
i16=twoP_*AST;
i17=twoP_*PTS;
i18=twoP_*MOV;
i19=twoP_*TS_;
i20=twoP_*eFG_;
i21=FT*FTA;
i22=FT*PF;
i23=FT*FT_;
i24=FT*FGA;
i25=DRB*TRB;
i26=DRB*PTS;
i27=DRB*MOV;
i28=DRB*SRS;
i29=DRB*eFG_;
i30=PTS*MOV;
i31=PTS*SRS;
i32=PTS*ORtg;
i33=PTS*eFG_;
i34=MOV*SRS;
i35=ORtg*MOV;
i36=eFG_*MOV;
i37=ORtg*TS_;
i38=season*threePAr;
l1=FG*FG;
l2=FG_*FG_;
l3=twoP*twoP;
l4=twoP_*twoP_;
l5=AST*AST;
l6=PTS*PTS;
l7=MOV*MOV;
l8=SRS*SRS;
l9=eFG_*eFG_;
l10=FT*FT;
l11=ORtg*ORtg;
l12=DRtg*DRtg;
k1=threeP*ORB*FT;
k2=threeP*DRB*FT;
k3=FGA*AST*threePA;
k4=STL*BLK*PTS;
k5=TOV*PF;
k6=threePA*ORB*FTA;
k7=threeP_*ORB*FT_;
k8=threeP_*ORB*FTA;
k9=threePA*ORB*FT_;
k10=TRB*STL*BLK;
k11=threeP_*threePAr;
k12=threeP*threePAr;
cards;
(One team example)
WAS  0.597560976  1.7  2  0.009  -0.7  -2.4  0.008  2.4  4.4  0.002  -0.9
     -2.1  0.033  -0.4  0.2  -0.2  1.1  1.1  -0.5  -1.2  1.6  1.9  26
     1.8  -0.45  1.36  111.2  109.3  97.3  0.254  0.285  0.564  0.004  -1  -51.
     -0.014  38
;
run;
proc print data = NBA;
run;
```

```
/*Confidence interval*/
proc sgplot data= NBA;
  vbar win_ / response=win_ stat=mean limitstat=stddev limits=both;
run;

/*Compute correlation*/
proc corr data = NBA;
var win_ FG--k10;
run;

/* see interaction*/
proc reg data = NBA;
Model win_=PTS FG FG_ FGA MOV ORB DRB TRB AST BLK STL twoP twoP_ eFG_ SRS TS_ k1 k2 k3
k4 k5 k6 k7 k8 k9/selection = forward;
run;

/*model selection*/
proc reg data = NBA;
Model win_=PTS FG FG_ FGA MOV ORB DRB TRB AST BLK STL twoP twoP_ eFG_ SRS TS_ k1 k2 k3
k4 k5 k6 k7 k8 k9/selection= backward;
run;

proc reg data = NBA;
Model win_=PTS FG FG_ FGA MOV ORB DRB TRB AST BLK STL twoP twoP_ eFG_ SRS TS_ k1 k2 k3
k4 k5 k6 k7 k8 k9/selection = stepwise;
run;

proc reg data = NBA;
Model win_= PTS FG FG_ FGA MOV ORB DRB TRB AST BLK STL twoP twoP_ eFG_ SRS TS_ k1 k2 k3
k4 k5 k6 k7 k8 k9;
run;

proc glm data = NBA;
Model win_= PTS FG FG_ FGA MOV AST BLK STL eFG_ TS_ k8;
run;

proc corr data = NBA;
var PTS FGA MOV BLK;
run;

proc reg data = NBA;
Model win_= PTS FG_ threeP threeP_ twoP twoP_ FT DRB TOV PF;
run;

proc reg data = NBA;
Model win_= PTS FG_ threeP threeP_ twoP twoP_ FT DRB TOV PF/selection=forward;
run;

proc reg data = NBA;
Model win_= PTS FG_ threeP threeP_ twoP twoP_ FT DRB TOV PF/selection=backward;
run;
```

```
proc reg data = NBA;
Model win_= PTS FG_ threeP threeP_ twoP twoP_ FT DRB TOV PF/selection=stepwise;
run;

proc reg data=NBA;
model win_ = FGA FG_;
run;

proc reg data=NBA;
model win_ = twoP twoP_;
run;

proc reg data=NBA;
model win_ = threeP threeP_;
run;

proc reg data=NBA;
model win_ = PF FT;
run;

/*GSW*/
proc glm data=NBA;
model win_ = PTS  AST twoP_ BLK FG_;
run;

proc corr data=NBA;
Var PTS threeP_ threeP MOV twoP_ AST BLK FG_  SRS eFG_ threePAr k11 k12 DRB;
run;
```