STAT 346 Final Project: Defense Wins in the NBA

Alvaro Aleman, Tony Liu, Reid Pryzant

Due 7 December, 2014 5pm.

Introduction and Context

There has been a recent movement towards the use of advanced statistical analysis to track and project player and overall team performance in many professional sports. In particular, the paper "An Economic Evaluation of the *Moneyball* Hypothesis" by Hakes and Sauer [1] provides an analysis of Major League Baseball (MLB) player performance and identified which individual player statistics were more crucial to team success. Subsequently Hakes and Sauer determined which player attributes were undervalued within the MLB market based on salary.

We wish to conduct a similar analysis with data from the National Basketball Associaton (NBA). Our goal is to identify which attributes contribute more to overall team success and attempt to predict team success using winning percentage. In this report, we construct a multiple linear regression model on 30 years of NBA team data to predict this winning percentage. After selecting the best subset of predictors, we diagnosed our model and data for the presence of multicolinearity, influential observations, goodness of fit, and autocorrelation. Our final model suggests that defensive tactics like creating turnovers and reducing opponent's field goal percentage are the most important contributors to team performance.

Data Collection

We obtained all data from Basketball-Reference.com. Initially, we began with individual player data [2], but we soon realized that we would have difficulty associating individual player statistics with overall team success given the large number of players belonging to each team. To get around this problem, we collected team statistics (presented in basketball-reference on a per season basis) with the intention of regressing on winning percentage [3]. For each team, we combined three datasets from basketball-reference: "Team Stats", "Opponent Stats", and "Miscellaneous Stats". We first collected a raw team data set from a single year of observations. The size of the data set was problematic, as we had 66 potential predictors of winning percentage with only 30 observations. We needed to collect more data to increase our sample size, but we did not want to introduce potential correlations between the same NBA teams from consecutive seasons. So, we decided to pull data from seasons a decade apart. This would minimize any potential correlations between teams that have the same players on their roster for more than one season. Our final data set contains 109 data points.

Data Trimming

Before we began to analyze the dataset for outliers and correlated predictors, we trimmed the dataset of numerous predictors. First, we removed pre-calculated statistics from out dataset. Basketball-reference.com provides some statistics such as Pythagorean Wins and Simple Rating System, which are not observed data and are calculations meant to predict the exact response

we were trying to regress on (Wins or Winning Percentage). Additionally, we trimmed out data that was uniform across the entire dataset and all observations such as number of games played. This is because these observations do not add any additional predictive power as they are the same across all teams.

There are multiple observations within the dataset that are likely to be highly correlated because they are interrelated, such as Free Throws Made (FT), Free Throws Attempted, Free Throw Rate, etc. However, we chose not to preemptively remove these variables. We did not know which variables would be the best at predicting our response, so we decided to remove them from the model later (see our Multicolinearity section).

Model Selection

Initial Selection

We began by running a stepwise procedure with both AIC and BIC, regressing on Winning Percentage (Win.Pct.). The AIC stepwise procedure produced a model with an AIC of 661.362, an R^2 value of 0.936, and 31 predictors, while the BIC stepwise procedure produced a model with a BIC of 735.927, an R^2 of 0.927, and 25 predictors (26 estimated $\beta - coefficients$). The discrepancy between the number of predictors makes sense as the penalty term for BIC makes that model selection criteria favor smaller models with fewer predictors. Since the R^2 values are almost identical, we chose to use the model produced by BIC selection criteria as it is more parsimonious.

Multicollinearity

With such a large number of predictors (and with many of them related), we knew there could be potential multicollinearity problems with our model fit. In the heat map of the predictors within our BIC model (Figure 2), we saw numerous dark-colored cells indicating a large amount of correlation between many predictors. The lower left corner in particular indicates that those variables had large correlation coefficients.

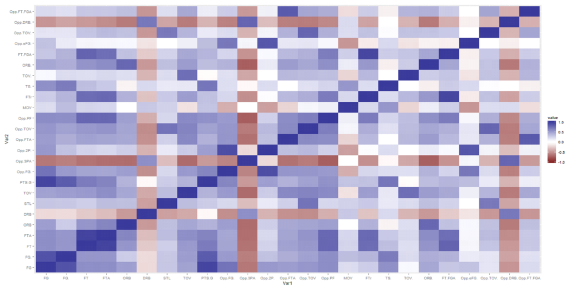

Figure 1: Heat Map of Stepwise BIC Model

152.09	**27.82**	**76.67**	**58.72**
10.39	8.36	4.74	9.02
100.46	**169.22**	**37.35**	**74.89**
10.82	**12.83**	6.53	**114.96**
38.14	**63.72**	**18.39**	**26.03**
39.75	**60.36**	**19.49**	**17.70**
12.07			

Figure 2: Varience Inflation Factors of each predictor

This was an indication that there was potential multicollinearity within our model. After calculating the Variance Inflation Factor for every predictor within our model, we found many predictors with VIFs above the standard threshold of 10 (figure 2, bolded). Since we had many predictors within our model (many of them measuring similar observations), we decided to run an iterative procedure where:

1. The predictor with the largest VIF value was removed from the model
2. VIF for each predictor was calculated for new model

3. Process was continued until all VIF values were below 10

In the end, we removed seven predictors from our model: FG, MOV, FT, Opp.FG%, PTS.G, FTr, and Opp.TOV. We believe this was the correct decision because many of the predictors remaining in the model still accounted for the removed predictors (Opp.TOV% remained in the model, which is directly related to OPP.TOV, FG% remained in the model, which is direclty related to FG, etc.). We decided on this procedure instead of using Principle Component Analysis because we wanted to mantain the interpretive power of our model.

Our final model (Figure 3) contains 18 predictor variables, and has an R^2 value of 0.8361.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-162.9799	41.6159	-3.92	0.0002
FG.	371.7905	79.0055	4.71	0.0000
FTA	-0.0047	0.0056	-0.84	0.4030
ORB	0.0002	0.0093	0.02	0.9863
DRB	0.0372	0.0085	4.37	0.0000
STL	0.0109	0.0129	0.85	0.3978
TOV	-0.0634	0.0121	-5.25	0.0000
Opp.3PA	0.0034	0.0031	1.08	0.2821
Opp.2P.	-91.7831	82.7913	-1.11	0.2706
Opp.FTA	-0.0094	0.0071	-1.33	0.1869
Opp.PF	0.0161	0.0073	2.22	0.0292
TS.	57.3054	77.7321	0.74	0.4629
TOV.	1.3681	1.1220	1.22	0.2259
ORB.	1.2498	0.3737	3.34	0.0012
FT.FGA	32.5868	43.3050	0.75	0.4537
Opp.eFG.	-167.7166	85.7232	-1.96	0.0535
Opp.TOV.	4.2475	0.9169	4.63	0.0000
Opp.DRB.	0.0053	0.3905	0.01	0.9892
Opp.FT.FGA	-34.9382	59.2646	-0.59	0.5570

Figure 3: Model after Removing Correlated Predictors

The corresponding heat map for this model (figure 4) is an indication that the correlation issues have been resolved, with no dark-colored cells other than the diagonal. Additionally, none one of the associated varience inflation factors were greater than 10 (figure 5).

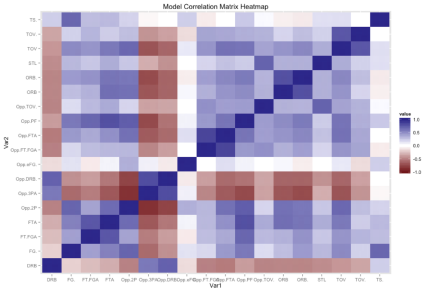

Figure 4: Heat Map without Correlated Predictors

8.91	5.89	4.96	4.18	3.33
6.97	8.28	8.46	9.23	4.79
5.76	4.15	5.30	3.46	7.59
3.48	4.71	7.19		

Figure 5: Variance Inflation Factor of each predictor

Model Diagnostics

Initial Examination

The normal probability plot (Figure 6) is roughly linear. However, the sample quantiles deviate from their theoretical values at both the lower and upper ends of the plot. This indicates that some normality assumptions may have been violated along with a higher than average variance.

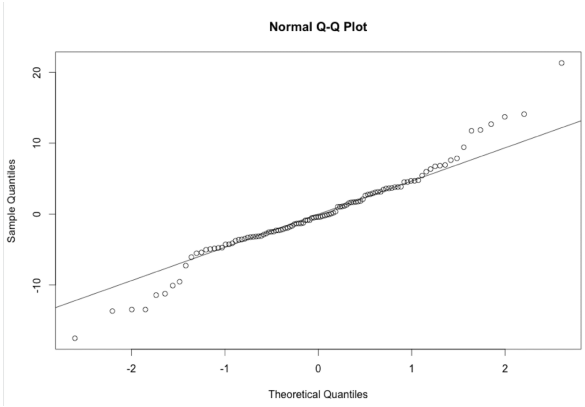

Figure 6: Normal Probability Plot

The residual plot (Figure 7) reveals some problems with our model fit. Though the residuals are fairly uniformly distributed in the horizontal band around zero, there are quite a few very large and very small residuals, indicating that there are potential outliers in our data set. These outliers give the residuals a slight funnel shape. We decided to conduct tests to see if there are any particularly influential observations.

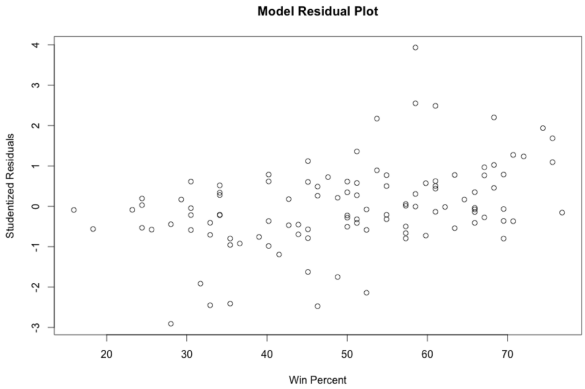

Figure 7: Residual Plot

Outlier Pruning

We checked for influential observations using DFFITS as a measure. Our dataset is small with 109 observations, so our threshold for influential observations was if the magnitude of the DFFITS value is greater than one.

After calculating DFFITS, we found 12 observations that were deemed influential. Since all 12 observations came from the same season, which was an era where the style of play was much different than other eras, we decided that it would be a good idea to remove these outlying observations from our data set. We sill have 97 observations within our dataset, which is sufficiently large.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-123.0632	28.7548	-4.28	0.0001
FG.	215.7360	66.6219	3.24	0.0018
FTA	-0.0063	0.0044	-1.41	0.1629
ORB	0.0032	0.0112	0.29	0.7745
DRB	0.0174	0.0070	2.47	0.0157
STL	0.0149	0.0086	1.73	0.0881
TOV	-0.0448	0.0089	-5.01	0.0000
Opp.3PA	0.0021	0.0021	1.01	0.3177
Opp.2P.	-75.7590	57.8018	-1.31	0.1938
Opp.FTA	-0.0112	0.0052	-2.16	0.0340
Opp.PF	0.0173	0.0058	3.00	0.0037
TS.	272.7688	62.3920	4.37	0.0000
TOV.	-0.1107	0.8373	-0.13	0.8952
ORB.	1.3475	0.4130	3.26	0.0016
FT.FGA	27.3128	37.5877	0.73	0.4696
Opp.eFG.	-288.0600	57.8258	-4.98	0.0000
Opp.TOV.	3.4704	0.6565	5.29	0.0000
Opp.DRB.	0.3630	0.2588	1.40	0.1647
Opp.FT.FGA	-42.4587	43.9276	-0.97	0.3368

Figure 8: Model with Outlying Observations Removed

Predictor Reselection

Note that many of the predictors in our model are insignificant when taken individually (Figure 8) after we have removed the outlying observations. The t-tests for ORB, Opp.3PA, Opp.2p., TOV., FT.FGA, Opp.DRB., and Opp.FT.FGA all yielded p-values above $\alpha = 0.05$. To test whether these predictors belong in the model when taken together, we performed the following partial F-test:

$$H_0 : \beta_{ORB} = \beta_{Opp.3PA} = \beta_{Opp.2p.} = \beta_{TOV.} = \beta_{FT.FGA} = \beta_{Opp.DRB.} = \beta_{Opp.FT.FGA} = 0$$
$$H_A : At\ least\ one\ \beta\ coefficient\ \neq\ 0$$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	87	1350.41				
2	78	1188.78	9	161.63	**1.18**	0.3206

Under the null hypothesis, $F*$ should follow $F(7, 87)$. The p-value for our observed statistic (1.18) is 0.32, implying that we do not have sufficient evidence to reject H_0 and that these predictors should be removed.

Once we refit the final model with the new data, we get a much better normal probability plot that is almost completely linear, as well as much better residual plot with no patterning or outlying residuals (Figures 9 and 10). Additionally, our model fit has improved in the absence of outliers and insignificant predictors: our new R^2 is 0.936.

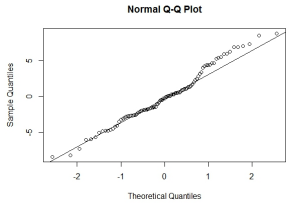

Figure 9: Normal Probability Plot

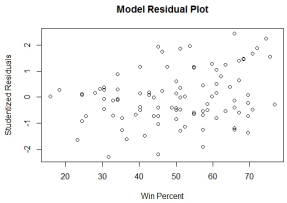

Figure 10: Residual Plot

Transformation

The diagnostic plots indicate that now the model is appropriate for the data, but we also ran a Box-Cox transformation just to ensure that no further improvements can be made to the model (Figure 11).

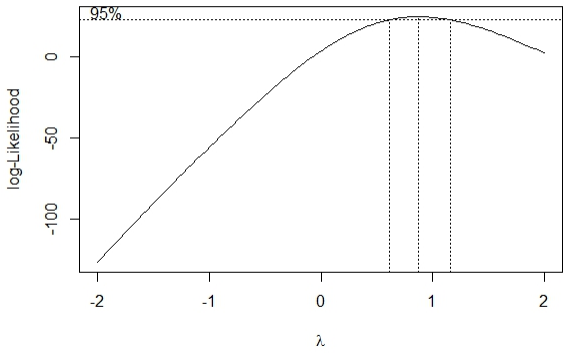

Figure 11: Box-Cox Plot

As shown by the Box-Cox plot, no transformations are needed to improve the model fit.

Autocorrelation

Because our team data comes from successive NBA seasons, there is the possibility that correlations exist through time in our predictors. We attempted to combat this problem by spacing out each generation of teams by 10 years. When win percentage is plotted against teams (sorted by season year), there are no discernable patterns or trends in the data (figure 12).

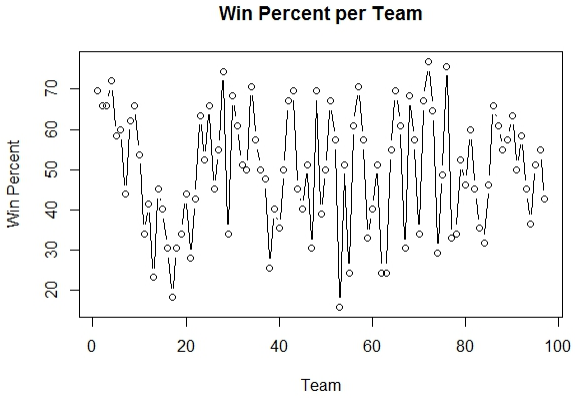

Figure 12: Plot of Win.Pct Over Teams, sorted by Time

Additionally, the autocorrelation function (ACF) plot indicates that there is little autocorrelation between our predictors (figure 13). The only lags with significant autocorrelation are 1, 12, and 13. However, the magnitide of the ACF for these lags is barely beyond the 95% confidence threshold for autocorrelation, implying that the relationships are tenuous at best.

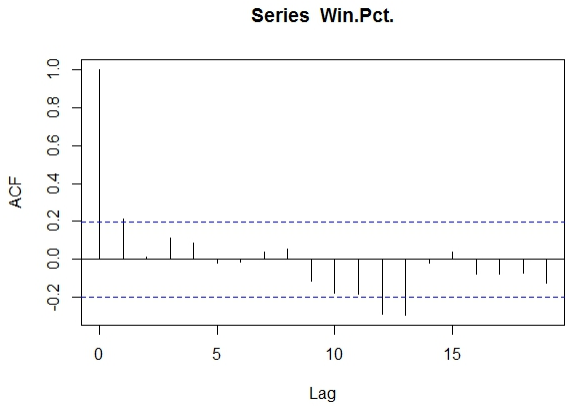

Figure 13: ACF Plot

To test whether the true autocorrelation is greator than 0, we performed the following Durbin-Watson test with a significance threshold of $\alpha = 0.01$:

$$H_0 : autocorrelation\ (\rho) = 0$$
$$H_A : autocorrelation\ (\rho) > 0$$

Our observed test statistic was $D = 1.73$, which translates to a p-value of 0.059. Because this is larger than our α, we fail to reject H_0 and conclude that the true autocorrelation in our data is not different from 0.

Results and Conclusion

The Final Model

Our final model is:

$$\begin{aligned}\widehat{Win.Pct.} = &-111.03 + 185.2x_{FG.} - 0.02x_{DRB} - 0.05x_{TOV} - 0.02x_{Opp.FTA} \\ &+ 0.01x_{Opp.PF} + 307.99x_{TS.} + 1.35x_{ORB.} - 343.34x_{Opp.eFG.} + 4.39x_{Opp.TOV.}\end{aligned}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-111.0257	20.3248	-5.46	0.0000
FG.	185.2398	39.4401	4.70	0.0000
DRB	0.0236	0.0046	5.10	0.0000
TOV	-0.0486	0.0042	-11.47	0.0000
Opp.FTA	-0.0161	0.0022	-7.34	0.0000
Opp.PF	0.0083	0.0034	2.42	0.0176
TS.	307.9899	48.1088	6.40	0.0000
ORB.	1.3519	0.1631	8.29	0.0000
Opp.eFG.	-343.2354	25.1585	-13.64	0.0000
Opp.TOV.	4.3875	0.4083	10.75	0.0000

Figure 14: Our Final Model with 9 Predictors

Model Verification

We performed the following global F-test to validate our entire model:

$$H_0 : All\ \beta\ coefficients = 0$$
$$H_A : At\ least\ one\ \beta\ coefficient\ \neq\ 0$$

Source	Df	SS	MS	F
Regression	9	19807.74	2200.86	
Error	87	1350.4	15.52	
Total	96	21158.14		**141.8**

Under the null hypothesis, F^* should follow $F(10, 87)$. The p-value for $F^* = 141.8$ is 2.2×10^{-16}, giving us sufficient evidence to reject the null hypothesis and conclude that at least one β-coefficient is nonzero. Thus, our model fit is appropriate.

Discussion

The coefficients in our final model had both expected results and a few surprises. Due to the scaling of the percentage variables, the shooting statistics were calculated between the range of 0 and 1, while winning percentage and rebound percentages ranged from 0 to 100, so we have to be careful when interpreting the size of the effect for each predictor.

With this in mind, the predictor variable with the highest coefficient was actually opponent turnover percentage. A one percent increase in $x_{Opp.turnover\%}$ was associated with a 4.4 percentage point increase to predicted winning percentage, on average. This makes sense as it would give a team more possession of the ball, having more chances to score while the opponent has less. Opponent's effective field goal percentage had the second largest coefficient. A one percent increase in opponent effective field goal percentage, which, due to the scaling, corresponds with a .01 increase in the predictor variable, is predicted to decrease winning percentage by approximately 3.43 percent.

The most signifcant predictors affecting team performance concerned defensive statistics, such as reducing an opponent's effective field goal percentage and increasing their turnover rate. Thus a team's performance on defense is more indicative of their success. Using this model, in order to maximize winning percentage a team must focus on two aspects of basketball:

1. Creating turnovers in order to maximize possession and thus shooting opportunities.
2. Reducing the efficiency of opponent's shots

These results seem to confirm the common basketbal adage: "Defense wins championships".

References

[1] J. K. Hakes and R. D. Sauer. An Economic Evaluation of the *Moneyball* Analysis. *Journal of Economic Perspectives*, Volume 20, Number 3 (173-185), 2006.

[2] 2013-14 NBA Stats: Per Game. *Basketball-Reference.com - Basketball Statistics and History.* Sports Reference LLC. `http://www.basketball-reference.com/leagues/NBA_2014_per_game.html` 21 November, 2014.

[3] NBA & ABA League Index. *Basketball-Reference.com - Basketball Statistics and History.* Sports Reference LLC. `http://www.basketball-reference.com/leagues/` 28 November, 2014.