Predict NBA Team Winning Record

Mason Chen, Stanford OHS, Palo Alto, CA

Abstract:
This paper is to build an empirical model to predict the NBA team winning percentage based on their team offensive, defensive, and differential statistics by collecting historical data during 2003-2016. The raw data have been standardized through Z transformation to remove mean and large variance bias effect. A multiple linear and step regression model was derived to predict the team winning record. After trimmed the insignificant regression terms, the derived model can predict team winning percent with R-Square > 0.95. The multi-linearity concerns were addressed by looking at the Variance Inflation Factor > 10. The redundant terms were removed to avoid over-fit risk. The regression model has identified 3-point Percentage, Turn Over, and Point per Game most critical to the team offensive efficiency. This observation is consistent with modern basketball. In defense, how to defend the rebound and opponent’s field goal percentage are most critical. Warriors’ 2015-2016 team record has been identified as an extreme outlier since their winning formula and team statistics are significantly different from the remaining 29 teams. The 2nd-order and Interaction Terms were added to enhance the prediction accuracy. The nonlinearity terms have indicated the complexity of the basketball team behaviors. Defense Field Goal% * Defense Point per Game was identified as the most significant interaction term. Which may reflect the Best Defense is the start of a good Offense. The model built based on 2003-2016 data was further validated by the new season 2016-2017. This model can provide NBA coaches and general managers how to draft, recruit, trade, or sign particular players to build a desired Championship team based on the winning % formula. This methodology can be applied to NBA play-off and other major professional sports like baseball, football, hockey, soccer.

Key words: Regression, Sports Analytics, Predictive Model, Statistics

1. Introduction

Sports are big part of our daily life. Every major city has their professional teams and local fans are very supportive to their local professional players as their heroes or role models. National Basketball Associate (NBA) is the largest basketball organization. Each year, each NBA team is fighting for the playoff spot to win the championship. This paper would try to formulate what could be the most deciding factors to formulate how to build a championship team by analyzing historical team statistics. In major professional sports, the coach and team management are looking for ways to win more games to build their championship dynasty (such as Celtics, Lakers, Bulls) in order to attract more fans to support their business. Sports statistical modeling analytics is becoming a critical approach to uncover the winning patterns hidden in sports data collected during each game played. The objective of this paper is to build a statistical model based on the past team offensive, defensive, and differential statistics in order to predict the NBA 2016-2017 Regular Season Team Record. There are several research talks presented in MIT Sloan Sports Analytics Conference. These papers have used intensive Analytics to uncover players’ playing patterns and help coach develop each player in order to create and maximize each player’s values to their specific team. Our paper will provide the predictive methodology to be applied in the end of the regular season (April 2017). In Figure 1, the authors have demonstrated the project scope of this paper: (1) use the 2003-2016 team offensive, defensive, differential statistics (input independent variables Xs) to build a transfer function to predict the 2003-2016 team record (Y); (2) use the same transfer function and 2016-2017 team offensive, defensive, and differential statistics to predict the new 2016-2017 regular season team record.

![Figure 1 Build Predictive Model](image-url)
2. Experimental Section
Author has laid out three subsections: (1) Raw Data Collection, and (2) Apply Z Transformation.

2.1 Raw Data Collection
Team statistics and record were collected\textsuperscript{6,7} from the ESPN Sports NBA Website as shown in Figure 2.

![Figure 2 Team Statistics Raw Data](image)

2.2 Apply Z Standard Score Transformation
Prior to build a predictive model, Z transformation\textsuperscript{8,9} is applied on team offensive, defensive, and differential statistics in Figure 3. Z transformation can eliminate any uneven influence (larger variance) among different team statistics categories in order to build an unbiased model. Otherwise, the predictive model may be dominated by any team statistics with larger variance.

![Figure 3 Z Transformation on Team Statistics](image)

3. Results and Discussion
3.1 Build Multiple Linear Regression Model

The multiple linear regression model was built by using the 2003-2016 Team Statistics and Team Record in Figure 4. ANOVA Table has listed the most significant variables (P-Value) < 0.5 and the responding Regression R-Sq (Adjusted) is 96.75% which indicated the built predictive model is reliable to predict the team record performance based on the few identified team statistics variables.

![Analysis of Variance](image1)

Figure 4 Multiple Linear Regression Model

Main effect regression coefficients and regression equation were listed in Figure 5. However, authors also checked any dependency among the identified input variables (Xs) to assess the multi-collinearity risk. In the VIF\(^10\) (Variance Inflation Factor) column, five VIF index are above 10, which has shown significant concern on the Multi-Collinearity, which may inflate the regression R-Sq (Adjusted) and impact the regression equation.

![Coefficient Table](image2)

Figure 5 Regression Equation and VIF

In addition to VIF, authors also checked the any residual or leverage outlier which may influence the regression model significantly as shown in Figure 6. Observed Data #10 was detected as residual outlier which has standard
residual at 2.66. This Data #10 happens to be the Warriors Team which just broke Bulls’ 72-win record. Warriors has created a new era on emphasizing 3-points, team assistance, and fast offensive flow.

![Fits and Diagnostics for Unusual Observations](image)

<table>
<thead>
<tr>
<th></th>
<th>C1 T</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
<th>C8</th>
<th>C9</th>
<th>C10</th>
<th>C11</th>
<th>C12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>10</td>
<td>0.8900</td>
<td>0.8399</td>
<td>0.0501</td>
<td>2.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6 Detect Regression Outliers

It’s not surprised that the Warriors team statistics and team record is off the predictive regression chart as a residual outlier.

3.2 Improve the Predictive Model

In order to address the model adequacy concerns, authors have first trimmed the less significant terms (P-value above 0.2) as shown in Figure 7. Two terms were dropped from the previous regression model and R-Sq (Adjusted) has actually slightly been improved from 96.75% to 96.82% even R-Sq has been degraded. R-Sq (Adjusted)\(^1\) is a better index to assess the multiple linear regression model. Authors would like to keep the remaining variables with P-values under 0.2 since little impact to trim the regression model further.
Authors further evaluated the model adequacy on the reduced model as shown in Figure 8. The removed two insignificant factors happen to be the higher dependent ones with VIF (Variance Inflation Factor) > 10. Only two factors still with VIF > 10. Though, these two factors are significant with P-values < 0.05. Authors decided to keep these two highly dependent factors in the regression model.

Figure 7 Trim Regression Model

![Analysis of Variance table]

**Figure 8 Evaluate Model Adequacy**

Authors won’t consider adding the quadratic terms to model simple (parsimony). A parsimonious model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible. After built the reduced model, residual analysis was conducted to ensure model is adequate as shown in Figure 9. The predictive regression model has residuals which are normal distribution (Normal Probability Plot, Histogram) in the left-hand side, with equal variance (upper right), and independent (lower right).

![Model Summary table]

![Coefficients table]

![Regression Equation]

431
The following is the Predictive Regression Model Equation. Authors can use this equation and the top 9 factors in the team Offensive, Defensive and Differential Statistics to predict the Team Record Performance for any particular year.

Regression Equation

\[
\text{Win\%} = 0.50695 + 0.01586 \times \text{O-3pt\%}_1 - 0.0427 \times \text{O-TO}_1 + 0.1075 \times \text{O-Pts}_1 - 0.0281 \times \text{D-FG\%}_1 \\
- 0.0189 \times \text{D-RB}_1 + 0.0370 \times \text{D-TO}_1 - 0.0807 \times \text{D-Pts}_1 - 0.0144 \times \text{C-Ast}_1 - 0.0231 \times \text{C-Stl}_1 \\
- 0.01232 \times \text{D-FG\%}_1 \times \text{D-Pts}_1
\]

**4. Conclusions**

Authors have successfully built a predictive model which can predict the NBA Team Winning Record based on the top 9 team statistics records. Authors have addressed the model adequacy such as multi-collinearity (VIF), residual outliers, interaction terms, R-Sq vs. R-Sq (Adjusted), and residual analysis (normality, equal variance, independency). This paper has shown the power of applying the statistical regression model to predict the winning pattern. This paper can be further expanded to consider several situations such as player injury, match-up between two particular teams.

**Acknowledgment**

Thanks to Dr. Ying Huang and Dr. Alex Wu for providing the regression statistics support. Thanks to Mr. Patrick Giuliano for revising our paper manuscript.

**References (To be Added)**