Regression Analysis of Relationship between Home Team Winning and Other Factors in NBA

QTM 200 - 1

Min Jin

4/26/2021

Introduction

In the field of competitive sports, individual and team statistics not only help coaches and players find meaningful information on their past performances, but also serve as insightful indications about factors that the teams need to focus on to improve their chances of winning in the future. In addition, performance records can also serve as milestones for each team to establish their prestige position in a certain sport, and allow them to break the record to reach higher achievement in the field.

This is especially the case in the National Basketball Association(NBA) where data analysts gather relevant data and prepare coaches and players for upcoming games. Analyses of past-game statistics offers coaches insights about the tactics and efficiency of the players and the team, and help coaches develop and discover new methods that potentially improve the chances of winning by examining those statistics. Moreover, player-level data also showcase the strength and shortcomings of individuals and help them develop better skills to improve their weaknesses.

Thus, basketball's performance data has always been a hot topic among data analysts, and many studies have made significant discoveries regarding factors contributing to win a basketball game at different professional levels (Mikołajec et al, 2013). The two main factors that influence the final outcome of a game have been established by the past research as a team's abilities of playing offense and defense (Ibáñez et al., 2003). Moreover, other studies have also find significant positive correlations between winning a basketball game and shooting efficiency (Akers et al., 1991; Ittenbach and Esters, 1995) as more efficient teams can score more points within the game time. In addition to offensive end, the study by Sampaio and Janeira (2003) also indicates that a team with better defensive plays tend to have higher chances of winning a game since they are good at stopping the other team from scoring.

Therefore, this study takes one step further and tries to understand the relationship between whether or not a home team wins a game and their free throw percentage during that game. Hence the home team outcome is the response variable while free throw percentage is the main independent variable for this study. Additional variables such as number of home team assists and away team's number of rebounds are included in the study to avoid potential problem of omitted variable bias. More on that topic later.

Data Description

The scope of this study focuses on a collection of game statistics in NBA matches from the beginning of the 2004 season to December 2020. The data collection can be found here. From the original dataset, four variables are selected and defined here:

- **HOME_TEAM_WINS**: Did the home team win? (Yes=1, No=0)

- **FT_PCT_home**: Free Throw Percentage of the home team
- **AST_home**: Assists by home team
- **REB_away**: Rebounds by away team

As mentioned in the sections above, one of the most important factors that leads to win a basketball game is found to be the number of successful free throw made during a game(Csataljay et al., 2009). and in previous study researchers also establish free throw as an important and even pivotal factor when determing the outcome of a close game (Kozar et al., 1994). Therefore, free throw percentage is the main predicting variable for the study so that we can examine how the overall free throw percentage affects the outcome of a game. In addition, Gomez et al.(2008) reveals that defensive rebounds is another important factors contributing to win a game since more defensive rebounds means less score by the opponent team. Moreover, Melnik (2001) discovered a significant relationship between winning a game and number of assists. Based on these conclusions, assists by home team and rebounds by away team are included as additional variables for regression analyses to avoid the omitted variable bias.

Research Questions and Expectations

Since winning a basketball game requires a team to score higher points than the opponents at every opportunity and rewarded opportunity (free throw), I suspect that winning is highly associated with the free throw percentage (the relationship is significant), as well as other variables such as number of assists of the home team and number of rebound by the away team. Hence, **final outcome of the game** is the response variable and **free throw percentage** is the main independent variable. **number of home team assists** and **number of away team's rebounds** are additional independent variables that suspected to have correlations with the response variable. Since the response variable is binary, this study uses **binary logistic regression model** for analyses. The research question is whether or not these variables have significant linear relationships.

As teams score more points (having more number of assists) and higher accuracy in free throw (higher free throw percentage), it's reasonable that their chances of winning increases since they are able to make more points during a game (usually last for two hours). In addition, when the away teams have fewer rebounds, it indicates they had fewer chances to score because the balls are not usually in their possessions. Therefore, I anticipate a positive linear relationship between the odds of winning and field goal percentage, and a positive linear relationship between number of home team assists and odds of winning. On the other side, I expect a negative linear relationship between odds of winning and away team's number of rebounds during a game.

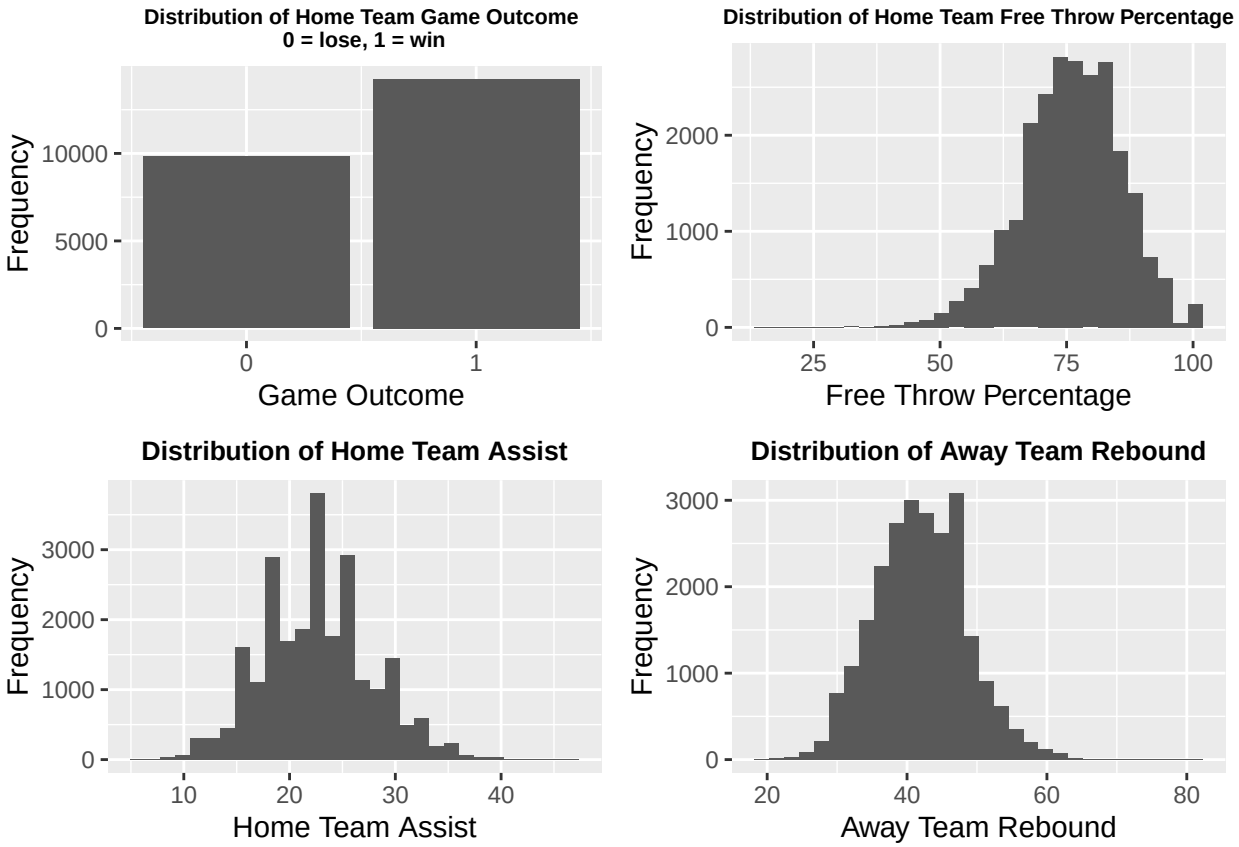
Summary Statistics and Graphs

The table below shows the minimum, median, mean, maximum, and standard deviation of each independent variable.

Table 1: Summary statistics of independent variables, n = 24096

Variables	Min	Median	Mean	Max	Sd
Home team free throw percentage	14.3	76.5	75.867	100	10.045
Home team assists	6.0	22.0	22.604	47	5.157
Away team rebounds	19.0	42.0	41.939	81	6.508

The Histograms below demonstrates the distribution of each independent variable and the response variable.



From the Graph we see that all explanatory variables are about normally distributed. This is because of our large dataset. It's worth noting that distribution of Free throw percentage is little skewed to the left, but that won't influence our models.

Data Assumptions

This study satisfies the assumptions for a binary logistic model as the response variable is binary. Furthermore, each observation in the dataset is independent of each other and are randomly sampled as we assume that each game won't be affected by the previous games. Although there are factors that introduce confounding variables to this model (such as injuries), this is not the focus of the study and will be mentioned in later sections. To check whether the the problem of multicollinearity exists in the independent variables, the following table shows the correlation among all three independent variables.

Table 2: Corrlation table for the independent variables, n = 24096

	FT_PCT_home	AST_home	REB_away
FT_PCT_home	1.0000000	0.0129211	-0.1140075
AST_home	0.0129211	1.0000000	-0.1109585
REB_away	-0.1140075	-0.1109585	1.0000000

From the correlation table, we see that all absolute value of correlations are less than 0.2, which means that there's little or no multicollinearity among the independent variables in this study. Finally, the sample size is large and representative enough. Therefore, the variables satisfy assumptions for logistic model.

Construct Logistic Additive Model

Therefore, we expect to have the following logistic additive models between game outcome and free throw percentage of the home team, assists by home team, and rebound by away team before adding interaction terms:

- **Logistic model 1**: $Outcome = \beta_1 free\_throw\_percentage + \beta_0$
- **Logistic model 2**: $Outcome = \beta_1 free\_throw\_percentage + \beta_2 assist\_home + \beta_0$
- **Logistic model 3**: $Outcome = \beta_1 free\_throw\_percentage + \beta_2 assist\_home + \beta_3 rebound\_away + \beta_0$

construct model 1

By using the logistic model function in r, we get the following output:

```
##
## Call:
## glm(formula = HOME_TEAM_WINS ~ FT_PCT_home, family = binomial(link = "logit"),
##     data = game_var)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5401  -1.3124   0.9466   1.0289   1.5128
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.024973   0.100250  -10.22   <2e-16 ***
## FT_PCT_home  0.018465   0.001315   14.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 32581  on 24095  degrees of freedom
## Residual deviance: 32382  on 24094  degrees of freedom
## AIC: 32386
##
## Number of Fisher Scoring iterations: 4
```

From the summary table, we get the following regression model: $Outcome = 0.018465 free\_throw\_percentage - 1.024973$. From the estimate the model shows that when the free throw percentage of the home team increases by one percentage point, there's expected to be 0.018465 log-odds increase of the the outcome of the game for the home team when holding other variables constant.

Since this is an logistic model, transformation for the coefficients need to be done before explaining in the context. To transform, we use function in r to calculate e^{β_1}, so $e^{0.018465} = 1.018637$. Therefore, the model suggests that one percentage point increases in the home team's free throw percentage increases the odds of winning a game by 1.8637%. From the intercept and $1 - e^{-1.025} = 0.6412035$ we get that when the home team has 0% free throw accuracy, the odds of winning a game decreases by 64.12035% when holding other variables constant.

construct model 2

By using the logistic model function in r, we get the following output:

```
##
## Call:
## glm(formula = HOME_TEAM_WINS ~ FT_PCT_home + AST_home, family = binomial(link = "logit"),
##     data = game_var)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3459  -1.1520   0.6807   0.9926   1.9833
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.122167   0.126879  -32.49   <2e-16 ***
## FT_PCT_home  0.019570   0.001388   14.10   <2e-16 ***
## AST_home     0.135336   0.002993   45.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 32581  on 24095  degrees of freedom
## Residual deviance: 30007  on 24093  degrees of freedom
## AIC: 30013
##
## Number of Fisher Scoring iterations: 4
```

From the summary table, we get the following regression model: $Outcome = 0.019570 free\_throw\_percentage + 0.135336 assist\_home - 4.122167$. This means that when when the free throw percentage of the home team increases by one percentage point, there's expected to be 0.019570 log-odds increase of the the outcome of the game for the home team, and when number of assists of home team increases by one, the log-odds of the outcome of the game increases by 0.135333 when holding other variables constant.

By following the steps in model 1, we get that the odds of winning a game increases by 1.9763%, increases by 14.4921%, or decreases by 98.37906% when home team's free throw accuracy increases by one percentage point, home team's number of assists increases by one, or when neither the home team made any free throw or any assist during a game when holding other variables constant.

construct model 3

By using the logistic model function in r, we get the following output:

```
##
## Call:
## glm(formula = HOME_TEAM_WINS ~ FT_PCT_home + AST_home + REB_away,
##     family = binomial(link = "logit"), data = game_var)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4899  -1.0895   0.6004   0.9570   2.2487
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.445396   0.166490  -2.675  0.00747 **
## FT_PCT_home  0.014945   0.001427  10.473  < 2e-16 ***
```

```
## AST_home     0.132118   0.003065  43.104  < 2e-16 ***
## REB_away    -0.077012   0.002312 -33.311  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 32581  on 24095  degrees of freedom
## Residual deviance: 28814  on 24092  degrees of freedom
## AIC: 28822
##
## Number of Fisher Scoring iterations: 3
```

From the summary table, we get the following regression model: $Outcome = 0.014945 free\_throw\_percentage + 0.132118 assist\_home - 0.077012 away\_team\_rebounds - 0.445396$. This means that the log-odds of the the outcome of the game for the home team increases by 0.014945 or 0.132118 when the home team increases free throw percentage by one percentage point or increase number of assists by one. The log-odds of the outcome of the home game decreases by 0.077012 or 0.445396 when increases the number of away team's rebound by one or having 0% free throw accuracy, no assists for the home team and no rebounds from the away team.

By following the steps in model 1, we get that the odds of winning a game increases by 1.5057%, increases by 14.1243%, decreases by 7.41213%, or decreases by 64.05706% when home team's free throw accuracy increases by one percentage point, home team's number of assists increases by one, away team's number of rebound increases by one, or when the home team didn't make any free throw, any assist, or the away team didn't get any rebound during a game.

Interactive models

The next step construct interactive models for the dataset.

From using R functions, each different interaction term is added to the additive model 3. Therefore, the interactive models are the following: $Outcome = 0.0149289 free\_throw\_percentage + 0.1779922 assist\_home - 0.0529370 away\_team\_rebounds - 0.0010811 AST\_home : REB\_away - 0.445396$

$Outcome = 0.0198655 free\_throw\_percentage + 0.1487982 assist\_home - 0.0770228 away\_team\_rebounds - 0.0002209 AST\_home : REB\_away - 0.8169354$

$Outcome = 0.0240284 free\_throw\_percentage + 0.1320970 assist\_home - 0.0609713 away\_team\_rebounds - 0.0002126 AST\_home : REB\_away - 1.1313023$

Evaluate models fit

Since we have obtained three logistic models, the next step is choose the best-fit model that predicts the data. Since past researches have established that all three variables are all important in determining the outcome of a game (Akers et al., 1991), let's compare the model fit between model3 and the first interactive model (since home assist and away rebound have higher magnitude of log odds) using chi-square test and likelihood ratio test. The output of the anova test is shown below:

```
## Analysis of Deviance Table
##
## Model 1: HOME_TEAM_WINS ~ FT_PCT_home + AST_home + REB_away
## Model 2: HOME_TEAM_WINS ~ FT_PCT_home + AST_home + REB_away + AST_home *
##     REB_away
```

```
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     24092      28814
## 2     24091      28809  1     4.94  0.02624 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Likelihood ratio test
##
## Model 1: HOME_TEAM_WINS ~ FT_PCT_home + AST_home + REB_away
## Model 2: HOME_TEAM_WINS ~ FT_PCT_home + AST_home + REB_away + AST_home *
##     REB_away
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1   4 -14407
## 2   5 -14405  1  4.94    0.02624 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both outputs shows that the interactive model 1 is the best-fit model for the data since it has a p-value that's smaller than the significant level of 0.05, so we reject the null hypothesis and conclude that the interactive model improves our fit over the additive model. Let's also test the fit of the model using The Hosmer-Lemeshow Test using r.

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  game_var$HOME_TEAM_WINS, fitted(model3_co1)
## X-squared = 10.799, df = 8, p-value = 0.2134
```

The output shows a p-value above significant threshold, therefore we can conclude that

$Outcome = 0.0149289 free\_throw\_percentage + 0.1779922 assist\_home - 0.0529370 away\_team\_rebounds - 0.0010811 AST\_home : REB\_away - 0.445396$

is the best fit for the data.

Check Assumptions

Since the best-fit model is determined, assumptions need to be checked to examine any potential factors influencing the accuracy of the model.

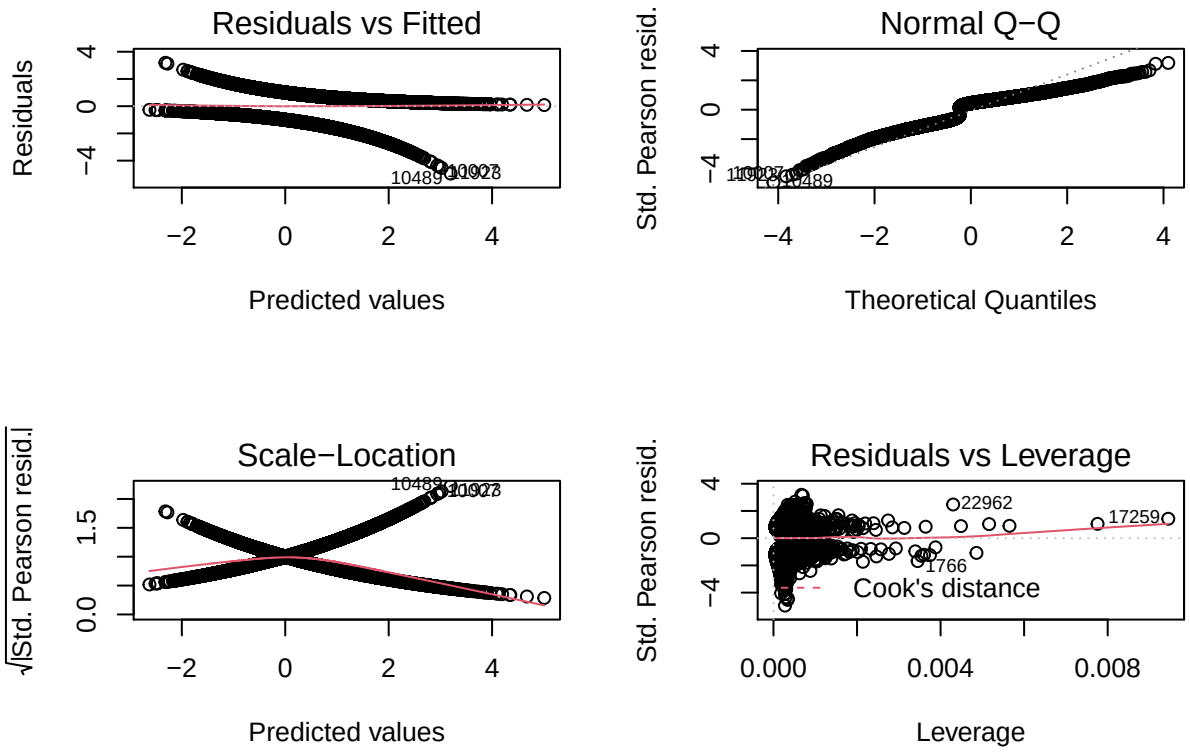
Collinarity using VIF Function

```
##       FT_PCT_home          AST_home          REB_away AST_home:REB_away
##          1.007493         46.590489         23.072193         71.255067
```

The output shows there exists correlation among my varaibles. However, this is explained by the choice of using interactive model.

Plot Residual vs. Fitted, Normal QQ plot, Residual vs. Leverage, and Scale-Location plot

```
# Plot Residual vs. Fitted, Normal QQ plot, Residual vs. Leverage, and Scale-Location plot
par(mfrow=c(2,2))
plot(model3_co1)
```




```
# create bubble plot to check outliers
dev.new(width=5, height=4)
influencePlot(model3_co1)
```

From Residual vs. Fitted graph, a linear relationship between the response variable and the explanatory variables can be observed, and the red horizontal dashed line demonstrates that the model has a constant variance of errors. The pattern can be explained by our large dataset (n = 24096). From the Normal QQ plot we can observe a conform to the dashed line for the residuals. Although the tail of the distribution is bit off the line, it's safe to conclude that the residuals are normally distributed. The Scale-Location plot we can observe that the model satisfy the assumption of homoscedasticity as red line is roughly horizontal. By looking at Residuals vs. Leverage using Cook's distance, we can conclude that there exists outliers (such as observation 22962 and 1766), but these points are not influential to the regression model since all of them are within the minimum of Cook's distance. From `influencePlot` we see there are five outliers in the model.

The only variable that is transformed is the free throw percentage of the home team (`FT_PCT_HOME`) since using it as a number instead of percentage can alter the estimates of log-odds in the model, which leads to extremely large coefficient for free throw percentage and slight increase (1) can leads to 400% increases in the odds of winning a game. This is because as numeric value the range of free throw percentage is between 0 - 1, so one unit increase can be hard to measure in this case. However, when transform each numeric value to percentage, the interpretation of coefficient becomes: odds of winning a basketball game increases as home team's free throw percentage increases by 1 percentage point. This is easier to understand both for readers and in later parts of analyses. Therefore, all assumptions are satisfied for the regression model.

Conclusion and Discussion

Overall, this study examines the relationship between a home team's odds of winning a basketball game and free throw accuracy using an interactive logistic model. Additional variables such as number of assists by the home team and number of rebounds by the away team are also included in the final interactive logistic model after evaluating the fit of each model, whether additive or interactive, with different variables. The model has found significant relationship beteen the response variable and the explanatory variables. Hence, the final regression model is the following:

$Outcome = 0.0149289 free\_throw\_percentage + 0.1779922 assist\_home - 0.0529370 away\_team\_rebounds - 0.0010811 AST\_home : REB\_away - 0.445396$

Each coefficient can be interpret as the following:

- $0.0149289 free\_throw\_percentage$: As home team's free throw percentage increases by one percentage point, the odds of them winning the game increases by $e^{0.0149289} - 1 = 1.015041 - 1 = 1.5041$ when holding other variables constant.
- $0.1779922 assist\_home$: As home team's number of assist increases by one, the odds of them winning the game increases by $e^{0.1779922} - 1 = 1.194816 - 1 = 19.4816$ when holding other variables constant.
- $-0.0529370 away\_team\_rebounds$: As away team's rebound increases by one, the odds of the home team winning decreases by $1 - e^{-0.0529370} = 1 - 0.9484398 = 5.15602$ when holding other variables constant.
- $-0.0010811 AST\_home : REB\_away$ As the away team have multiple rebounds and the home team have multiple assists, the odds of the home team winning the game decreases by $1 - e^{-0.0010811} = 1 - 0.9989195 = 0.10805$ when holding other variables constant.
- constant of - 0.445396 means that when the home team has zero free throw accuracy or assist, and away team has no rebound, the odds of home team winning decreases by $1 - e^{-0.445396} = 1 - 0.6405706 = 35.94294$ when holding other variables constant.

From the model, we can see that assist has the largest effect in increasing the odds of winning the game for a home team when holding other variables constant. This finding is surprising as our initial research question focuses on the relationship between game outcome and home team's free throw percentage. However, this can be justified as past research also concludes asisst to be an critical factor in determining the outcome of a basketball game (Ittenbach and Esters, 1995).

Future study

This model is not yet the perfect model to determine the factors influencing outcome of a basketball game since many other things such as player's physical conditions, food before game, referees' ruling, and more. Therefore, there can also exists some level of confounding variable to our model. For example, the physical condition of a star player on the home team can also have important influence over whether or not the team wins.

In addition, future because of the COVID-19 pandemic, NBA began to hold season 2020 and 2021 inside stadiums with minimum to no live audience. This is the first time in NBA history where no or only few audiences were presented during a live game. Hence, future studies can also build regression models to examine if presence of audience can also affects the performance of the home team.

Reference

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3796832/#b7-jhk-37-145

https://www.tandfonline.com/doi/abs/10.1080/17461390802261470

https://epublications.marquette.edu/account_fac/72/

https://search.proquest.com/openview/6d27f3eac47db103a76e9a41d54d4184/1?pq-origsite=gscholar&cbl=1819738

https://www.tandfonline.com/doi/abs/10.1080/24748668.2003.11868273

https://www.tandfonline.com/doi/abs/10.1080/24748668.2009.11868464

https://pubmed.ncbi.nlm.nih.gov/18756894/

https://pubmed.ncbi.nlm.nih.gov/11361327/

https://journals.sagepub.com/doi/abs/10.2466/pms.1994.78.1.243