*Journal of Applied Statistics & Machine Learning*
Volume 2; Number 1; 2023 : pp. 51-66

# USING THE BRADLEY-TERRY PAIRED COMPARISON MODEL AND LOGISTIC REGRESSION TO SEE IF THERE IS AN ADVANTAGE IN PLAYING AT HOME FOR NBA AND MLB GAMES

**Michael Gonzales and Anwar Hossain***

*Department of Mathematics New Mexico Tech*
**Correspondence: Department of Mathematics, New Mexico Tech, Socorro, NM 87801*

ARTICLEINFO

Received: 02 January 2022
Revised: 22 January 2023
Accepted: 20 February 2023
Online: 10 March 2023

*To cite this paper:*
Michael Gonzales & Anwar Hossain (2023). Using the Bradley-Terry Paired Comparison Model and Logistic Regression to See if there is an Advantage in Playing at Home for NBA and MLB Games. *Journal of Applied Statistics & Machine Learning*. 2(1): pp. 51-66.

ABSTRACT

Home field advantage is quite a common phrase that is used all throughout the world of sports. As the name implies, it leads one to believe that the team (or player in some cases) has an advantage performing on their home field. This analysis looks at whether or not this is the case by using the Bradley Terry Paired Comparison Model to compare the outcomes of NBA games as well as MLB games. The reasoning behind using match ups from the MLB and NBA is because of the limitations of this extension of the Bradley-Terry Paired Comparison Model. This extension does not account for ties. As a result these sports leagues were chosen so no data had to be omitted. The Bradley-Terry Paired Comparison Model found there was in fact an advantage in playing at home for both teams in the MLB and teams in the NBA. A combination of Linear and Logistic Regression models were also used to see how playing at home affected the response variable (points scored) and the log-odds that a team won. While all of these methods supported that home field advantage is a real thing, they were not significant and did not increase the log-odds by a substantial margin. The Bradley-Terry Paired Comparison Model is good at finding if there is an advantage (or disadvantage) in match ups between two teams, but it is held back by some of its limitations.

## 1. INTRODUCTION

In the world of sports, the term home field advantage is used quite often. This term can lead to one believing that the team who plays at home has the ad vantage and is favored to win the match up. O◎en times there are many in stances where an advantage is discussed for one team in the match up. These advantages can range from anything like the home field advantage to size of players/opponents. This analysis is based on the paper of Love [6] where Love uses character "hit box" size to use as an advantage. While Love [6] uses the Bradley-Terry Paired Comparison model to discuss if hit box size is an advan tage, we are going to use it to see if there is an advantage in playing at home for MLB and NBA games. The reason that MLB and NBA is used, is that the Bradley-Terry Paired Comparison model does not account for ties [4] and [6]. Of the more popular professional sports organizations in the United States, MLB (Baseball) and NBA (Basketball) are the only sports that do not allow a tie to be an outcome of the match up. Not allowing for ties allows for the entire use of the data that was used and there are no ignored match ups. Linear and Logistic regression models are also created to see how team location affects the prediction of points and winning the game. This will then be compared to the results of the Bradley-Terry model and see if the results match up.

## 2 DATA OVERVIEW

The MLB data set [9] contains the designation for the home and away teams, the team names, and the scores of the teams at the end of each game for the 2021 season. This data set also includes post season games. The post season games do not need to be discarded since there is a home and away team for these games and were not played on neutral sites. While the MLB data set did not explicitly define which team won and which team lost, the final scores were given which allowed me to determine which team won. To get the data that was necessary for the project, we compared the scores of the teams playing in each match up and placed it as a win for the team with the higher score. This could be done because there is no ties in baseball. If the scores are tied at the end of the normal 9 innings, then the game goes to extra innings until there is a winner. The new data that was created then shows all combinations of home teams and away teams with the number of wins each team had as the away and the home team. If there is a 0 in

**Table 1: MLB Data**

| *Home Team* | *Away Team* | *Home Wins* | *Away Wins* |
|---|---|---|---|
| ANA | ARI | 0 | 0 |
| ANA | ATL | 0 | 0 |
| ANA | BAL | 3 | 0 |
| ANA | BOS | 2 | 1 |
| ANA | CHC | 0 | 0 |
| ANA | CHW | 3 | 1 |

then that means the two teams did not play at all during the season. See table 1 for what this data looks like. This table is just an example and this is the style for all 30 MLB teams. This is the format that was used in the R documentation [4], so to follow the same method as that, the data was sorted in this way.

Table 1: Here we have an image of what the data that was created looks like. In this image, we can see that there were no match ups between ANA (Anaheim) and ARI or ATL (Arizona or Atlanta) but there was three match ups whre Anaheim was the home team and BAL (Baltimore) was the away team. This shows that Anaheim was the winner of all three match ups against Baltimore.

The NBA data set [8] was a little different to work with. This data was also obtained from Kaggle and is for the regular seasons from 2012-2018. Unlike the MLB data, this does not contain postseason data. There were multiple duplicates of data that we did not need to use. This was because the data set that was used also included the referees for the match up and had each game listed twice. After cleaning that up, we compared all of the teams that played each other in the same way that was done for the MLB data. This data set for the NBA however, contains many more statistics such as points, assists, turnovers, rebounds, blocks, steals, etc. These statistics are used for a linear regression model and logistic regression models that we in addition to the Bradley-Terry analysis. Since we am only concerned about the outcome of the game and not the rest of this for the Bradley-Terry model, we only used the outcome of the game to create the data used. The cleaned data was also used in addition to the data formatted in table 2 in the same way as the MLB data in table 1.

**Table 2: NBA Data**

| *Home Team* | *Away Team* | *Home Wins* | *Away Wins* |
|---|---|---|---|
| ATL | BKN | 6 | 4 |
| ATL | BOS | 6 | 7 |
| ATL | CHA | 7 | 3 |
| ATL | CHI | 6 | 3 |
| ATL | CLE | 5 | 4 |
| ATL | DAL | 5 | 1 |

Table 2: Here we have an image of what the data that was created looks like. In this image, we can see that there were ten match ups where ATL (Atlanta) was the home team and played against BKN (Brooklyn). Atlanta won six of these match ups while Brooklyn won four.

## 3. ANALYSIS

In this section we provide the analysis of the data based on linear regression, logistic regression and the Bradley Terry Paired Comparison model.

### 3.1. Difference of Means NBA

We first performed a linear regression model on the NBA data to see what predictors can influence scoring during a game. From anecdotal experience, there is an understanding that the more points scored, the better the likelihood that the team who scores more points wins the match up. The coefficients of this model can be seen in table 3. In this model we can see that team location being home, while not statistically significant with $p$ = 0.2725, does have a positive value. This suggests that there is some significance in playing at home when it comes to scoring points. This can be supported by figure 1. Additionally, performing a difference of means analysis as mentioned in [3] we get the following results. Testing for there being no difference in the mean gives a $p$-value of $p = 1.023 \times 10^{-18}$ and 95% confidence interval for the difference of mean being (–2.20, –1.41). This suggests that the Home team scores more points than the away team, which helps defend that there is an advantage to playing at home for the NBA.

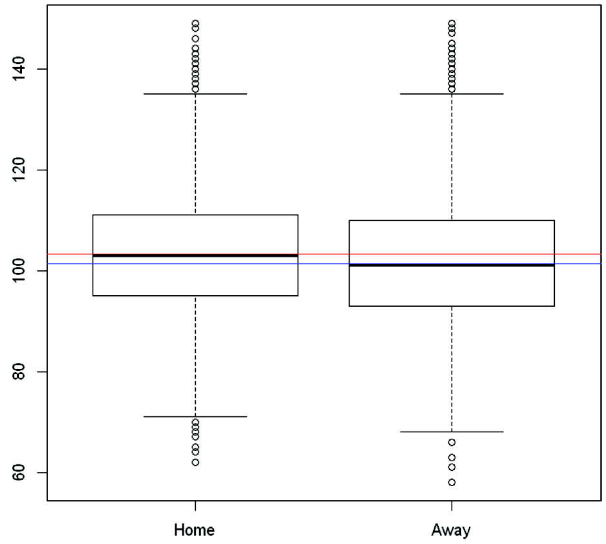

**Figure 1: NBA Points Scored**

Figure 1: Here we have box plots of the scores for home and away teams in the NBA data set. The red line represents the average points for the home team and the blue line represents the average points for the away team.

It should be mentioned here that these results do not necessarily mean that the home team is more likely to win more games over the course of the season. This test was done to find evidence and support that the home team does have an advantage. Now, it can be seen that scoring more points can help you to win the game, but that is not the case. While there is some evidence that the home team scores more points on average, this does not guarantee that the home team is favored to win.

### 3.2 Why Linear Regression?

While there is numerous papers that discuss much more in depth what regression is, we will take a look at what it is and why we used it in this analysis. Regression is a technique that is used in statistics that can be used to study relationships between response variables and predictors [10]. We used a linear regression model here to look at the NBA data and see how the predictors affected the points scored. Linear regression can be represented as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \in_i \tag{1}$$

where $Y_i$ is our response variable and $\beta_j$, for $j = 0, 1, \ldots, p-1$, are parameters. $X_{i,j}$ are known constants and $\in_i$ are independent $N(0, \sigma^2)$ and $i = 1, \ldots, n$ [5]. For this model that was made, the response variable is team points (teamPTS). The predictors are team location (teamLocHome), team days off (teamDay Off), team assists, (teamAST), team turn overs, (teamTO), team steals (team STL), team blocks (teamBLK), team personal fouls (teamPF), team offensive rebounds (teamORB), team defensive rebounds (teamDRB), opponent points, (opptPTS), opponent days off (opptDayOff), opponent assists, (opptAST), op ponent turn overs, (opptTO), opponent steals (opptSTL), opponent blocks (opptBLK), opponent personal fouls (opptPF), opponent offensive rebounds (opptORB), and opponent defensive rebounds (opptDRB). To interpret the val ues of the ß's, these mean that in this model, the predictors impact the number of points that are being predicted in the model. What this means is that a one unit increase in one of our predictors, say predictor $X_i$, then our response variable will be changed by $\beta_i$. For our model, this means that some predictors lead to an increase in points and others will lead to a decrease in points. This is also true if there is a decrease in $X_i$.

The reason that we wanted to look at a linear regression model is that we were curious about how the predictors affected the number of points the team scored. Understanding the game of basketball, it can be inferred that the more points you score, the more likely you are to win. For this reason, we wanted to see how some of the other predictors that are a key part of the game of basketball influence and can affect the scoring of points in a game. To remain fair, the same predictors that were used for the team in the model, these same predictors were used for the opponent. This allows us to see if there is a relationship for scoring points and the other statistics that are kept track of in the game such as those that are listed above.

## 3.3 Linear Regression NBA

Looking at the statistically significant predictors ($Pr(> |t|) < 2.62 \times 10^{-7}$ for this model), it is interesting to see what these numbers indicate. It is easy to see that a teamAST (team assist) contributes to points. An assist in basketball is defined as a pass that leads to a score [2]. Looking at the other coefficients, predictors that define getting possession of the ball have positive coefficients. This includes the predictors like teamSTL (team steal), teamDRB (team defensive rebound), opptTO (opponent turnover ), and opptAST (opponent assist). Predictors that define giving possession of the ball to the other team have negative coefficients. These include teamTO (team turnover), and opptDRB (opponent defensive rebound). The

**Table 3: NBA Linear Regression R Output and Coefficients**

| *Predictors* | *Coefficient Estimates* | *Std. Error* | *t value* | *Pr(> \|t\|)* |
|---|---|---|---|---|
| Intercept | 19.442 | 0.793 | 24.504 | 2*e*–16 |
| teamLocHome | 0.250 | 0.104 | 2.402 | 0.016 |
| teamDayOff | –0.125 | 0.052 | –2.397 | 0.017 |
| teanAST | 0.721 | 0.011 | 63.778 | 2e–16 |
| teamTO | –1.232 | 0.021 | –58.981 | 2e–16 |
| teamSTL | 0.344 | 0.025 | 13.540 | 2e–16 |
| teamBLK | 0.105 | 0.020 | 5.151 | 2.62e–07 |
| teamPF | 0.053 | 0.013 | 4.162 | 3.17e–05 |
| teamORB | 0.002 | 0.013 | 0.012 | 0.904 |
| teamDRB | 1.194 | 0.012 | 96.113 | 2e–16 |
| opptDayOff | 0.059 | 0.052 | 1.133 | 0.257 |
| opptPTS | 0.676 | 0.006 | 111.505 | 2e–16 |
| opptAST | –0.222 | 0.013 | –17.523 | 2e–16 |
| opptTO | 0.915 | 0.022 | 41.612 | 2e–16 |
| opptSTL | –0.066 | 0.026 | –2.600 | 0.009 |
| opptBLK | –0.186 | 0.020 | –9.172 | 2e–16 |
| opptPF | 0.311 | 0.013 | 24.692 | 2e–16 |
| opptORB | 0.014 | 0.013 | 1.031 | 0.303 |
| opptDRB | –1.261 | 0.012 | –105.484 | 2e–16 |

coefficients that show behavior that is unexpected are teamPF (team personal fouls), and opptPF (opponent personal fouls). Another coefficient that is also interesting is the coefficient for teamORB (team offensive rebounds). We can see from the table 3 that this is not a significant predic tor and is very close to 0. Since an offensive rebound gives the team another opportunity to score and keeps the other team from possessing the ball which would give the opponent an opportunity to score, one would think that it would be positive and significant. Opponent personal fouls can be explained because some of these fouls lead to the team that was fouled to shoot free throws which, if made, adds points to the team's score. On that note, team personal fouls can lead to the same thing which could lead to a possession of the team. It is also interesting to note that the coefficient for teamLocHome is positive and somewhat statistically significant and positive. This supports the difference of mean test that the home team scores more points than the opposing team.

Table 3: Here we have the coefficients of the variables for the linear regression model we made to predict the number of points scored in the game. Negative coefficients means that this negatively impacts the points scored of the team and positive coefficients positively increase the points scored.

### 3.4 Why Logistic Regression?

As mentioned in the Linear Regression Section, there are many papers that discuss the more specifics of Logistic Regression models, we will not go in depth but we will talk about it here to provide some background information. Logistic Regression is a parametric method for regression where $Y_i \in \{0, 1\}$ is binary. For a $k$-dimensional covariate $X$, the model is

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 Xi_2 + \ldots + \beta_{p-1} X_{i,p-1} \tag{2}$$

where

$$\text{logit}\,(p) = \log\left(\frac{p}{1-p}\right). \tag{3}$$

The previous two sentences come from Wasserman [10]. Here the logit($p$) is the log-odds function and p is a probability. The predictors that are being used for the logistic regression model are almost the same as the linear regression model. For this model though, instead of the response variable being teamPTS, we used team result, meaning a win or loss, (teamRslt) since this is binary. Those were the only two variables that were switched out. To interpret these coefficients differs than how we interpret the coefficients of linear regression models. Since equation (3) is the log-odds function, this means that it returns the log odds. If we exponentiate both sides of 3, then we end up with

$$\omega = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1}) \tag{4}$$

which gives us the odds [3]. This tells us that a unit change in predictor $X_i$ will lead to an increase in the log odds of our probability by $\beta_i$. From equation (4) we know that the odds is then increased by exp ($\beta_i$). A decrease in the log odds will occur as well as a decrease in the odds.

## 3.5 Logistic Regression NBA

After conducting the difference of mean test and creating the linear regression model, we created a logistic regression model for the NBA data. The model is for predicting wins unlike the linear regression model which was predicting points. The coefficients for this model can be seen in table 4. These coefficients show how each predictor changes the odds ratio [5] for winning the game. From the coefficients, we can see that the team location is not statistically significant in this model ($p$ = 0.20). This model was created using 70% of the data set as a training set. The other 30% was used as a testing set. Fitting this model with the testing set, this model predicted the out come correctly 88.23% of the time with a misclassification error of 11.77%.

Table 4: Here we have the coefficients for the logistic regression model that was created. Negative coefficients mean that this decreases the log odds of winning. Positive coefficients mean that increases the log odds of the team winning.

After looking at this full model, we wanted to see what a model with only the team location looked like. This can be seen in table 5. This model only uses the team location to calculate the odds ratio of winning the game. Since this is the only predictor, it is understandable that in this model, team location is statistically significant. This model shows that team location is important in predicting the outcome. Using the same methods as the model with more predictors, we used 70% of the data to create this model and tested it on the remaining 30%. This model gave an accuracy of 58.29% with a misclassification error of 41.71%.

Table 5: Here we have the coefficients for the logistic regression model that was created using only the team location. Since the coefficient is positive, this model shows that playing at home increases the log odds of playing at home.

**Table 4: NBA Logistic Regression R Output and Coefficients**

| *Predictors* | *Coefficient Estimates* | *Std. Error* | *z value* | *Pr(> \|z\|)* |
|---|---|---|---|---|
| Intercept | 6.054 | 0.559 | 10.826 | $2e-16$ |
| teamLocHome | 0.089 | 0.074 | 1.206 | 0.228 |
| teamDayOff | – 0.052 | 0.039 | – 1.327 | 0.184 |
| teanAST | 0.227 | 0.009 | 24.599 | $2e-16$ |
| teamTO | – 0.445 | 0.017 | – 26.124 | $2e-16$ |
| teamSTL | 0.118 | 0.019 | 6.396 | $1.60e-10$ |
| teamBLK | 0.062 | 0.014 | 4.274 | $1.92e-04$ |
| teamPF | – 0.074 | 0.009 | – 7.948 | $1.90e-15$ |
| teamORB | – 0.013 | 0.009 | – 1.325 | 0.185 |
| teamDRB | 0.420 | 0.012 | 35.232 | $2e-16$ |
| opptDayOff | 0.057 | 0.038 | 1.471 | 0.141 |
| opptPTS | – 0.113 | 0.005 | – 22.455 | $2e-16$ |
| opptAST | – 0.058 | 0.009 | – 6.319 | $7.52e-13$ |
| opptTO | 0.343 | 0.017 | 20.169 | $2e-16$ |
| opptSTL | – 0.014 | 0.019 | – 0.771 | 0.441 |
| opptBLK | – 0.104 | 0.015 | – 7.170 | $7.52e-13$ |
| opptPF | 0.210 | 0.010 | 21.532 | $2e-16$ |
| opptORB | 0.043 | 0.009 | 4.564 | $5.01e-06$ |
| opptDRB | – 0.442 | 0.012 | – 37.135 | $2e-16$ |

**Table 5: NBA Logistic Regression *R* Output and Coefficients Home Field only**

| *Predictors* | *Coefficient Estimates* | *Std. Error* | *z value* | *Pr(> \|z\|)* |
|---|---|---|---|---|
| Intercept | – 0.353 | 0.028 | – 12.51 | $2e-16$ |
| teamLocHome | 0.709 | 0.040 | 17.74 | $2e-16$ |

Comparing these two models to each other, it can be seen that the model with more than the team location predictor is the better model. This is shown with the better misclassification error of only 11.77% while the model using only the team location has a misclassification error of 41.71%. It can also be seen in the AIC values. The AIC value for the "full model" is 5283.5 while the AIC value for the Home Field only model is 14009. AIC stands for the Akaike Information Criterion and is defined as

$$AIC = 2p - 2\ln(\hat{L}) \tag{5}$$

where $p$ is the number of coefficients in the model and $\hat{L}$ is the maximum value of the likelihood function. These values do not tell us how good a model is, but rather allows us to compare how these models are to each other. These values allow us to compare models to each other and tells us the preferred model of the ones that were tested. Since we have a smaller AIC value with the "full model", this is the better of these two models.

### 3.6 What is a Bradley-Terry Paired Comparison Model

Again, there are many papers that discuss in more detail the specifics of the Bradley-Terry Paired Comparison Model. Here we are going to briefly intro duce it for the purposes of explaining it. The standard Bradley-Terry paired comparison model is used to estimate the probability of two teams under com parison. The problem is formulated as

$$P(i > j) = \frac{\pi_i}{\pi_i + \pi_j} \tag{6}$$

From here the likelihood function is given as

$$L(\pi_i) = \prod_{i=1}^{m} \prod_{\substack{j=1 \\ i \neq j}}^{m} \left( \frac{\pi_i}{\pi_i + \pi_j} \right)^{w_{ij}} \tag{7}$$

Here $w_{ij}$ is the number of times that team $i$ has beaten team $j$ and the $\pi_i$ values are positive-valued parameters which can be used to represent 'ability' [4]. As mentioned earlier, one assumption with this model, is that there are no ties. This is the reason that we chose to use NBA and MLB games to look at the advantage of home field. These two sports leagues do not have ties, so there is no special case that we have to look at and no data has to be excluded.

Speaking of home field advantage, the Bradley-Terry model for home field ad vantage is similar to the standard model but it is given as

$$P(i > j) = \begin{cases} \dfrac{\theta\pi_i}{\theta\pi_i + \theta\pi_j} : \text{if } i \text{ has the advantage} \\ \dfrac{\pi_i}{\pi_i + \theta\pi_j} : \text{if } j \text{ has the advantage} \end{cases} \tag{8}$$

Here, $\theta$ is "the amount of multiplicative increase in latent strength a certain team obtains by having an advantage" [6]. In this case, the advantage we are talking about is playing on your home field. This advantage does not always have to be home field though and can be extended out to many other things such as what Love [6] where he used a smaller hitbox as having an advantage.

Now for this model, the Likelihood function is given as

$$L(\pi_i) = \prod_{i=1}^{m} \pi_{j=1}^{m} \left( \frac{\theta\pi_i}{\theta\pi_i + \pi_j} \right)^{w_{ij}^{+}} \left( \frac{\pi_i}{\pi_i + \theta\pi_j} \right)^{w_{ij}^{-}} \tag{9}$$

Here, $w_{ij}^{+}$ are wins by $i$ against $j$ with the advantage and $w_{ij}^{-}$ are wins $i$ made against $j$ without the advantage. Again, there is a drawback here to this method because there is an assumption with this model that there is always an advantage. So, we are assuming that there is always an advantage in this model.

Another good thing about using NBA and MLB games in this case, is that there are very few instances where a game is not played on one of the team's home field. Even playoff games are played at home unlike the Super Bowl in the NFL where it is played at a neutral site determined years before the teams who will be playing there are decided.

We will now briefly talk about the naming of this type of model. As O'Connor and Eskey [7] mention, with paired comparison models, we are checking that $i$ is better than $j$ or in other words we are looking for a preference. As described above, we are using this to see which team is the preferred team to win the match up. This is different than the paired $t$-test, which is used to draw inferences about the population mean [3]. So the name Bradley-Terry Paired Comparison model comes from picking one over the other. In our case we are using it for sports teams. Whereas the paired t-test is used for comparing differences in population means. This paired t-test is actually what we used in section 3.1 and section 3.8.

## 3.7 Bradley-Terry Paired Comparison NBA

Now for the Bradley-Terry paired comparison model, this compares the match ups between all of the teams who play each other which then estimates the strongest team. For this data, teamORL (Orlando) is estimated to be the weakest team (0 since it is reference team) and teamGS (Golden State) is estimated to be the strongest team (1.88). These coefficients are maximum likelihood estimates of $\lambda_2$, $\lambda_3$, ..., $\lambda_{30}$ which are the log-ability. These $\lambda_i$ values are given as $\log(\pi_i)$ [4]. In this case, $\lambda_1$ is set to 0 as the identifying convention and this is for teamORL. Including the at home variable in this analysis takes into account playing at home as having an advantage for the home team. The coefficient for playing at home is is the estimated log odds-multiplier for playing at home [4]. This means that the home team has an estimated odds-multiplier of exp (0.24859) = 1.28 in the home team's favor [4]. This is understandable since of the 7,380 match ups

**Table 6: NBA Bradley-Terry Coefficients**

| *Predictors* | *Coefficient Estimates* | *Std. Error* | *z value* | *Pr(> \|z\|)* |
|---|---|---|---|---|
| teamATL | 0.847 | 0.133 | 6.379 | $1.78e-10$ |
| teamBKN | 0.360 | 0.133 | 2.693 | 0.007 |
| teamBOS | 0.879 | 0.133 | 6.591 | $4.37e-11$ |
| teamCHA | 0.504 | 0.133 | 3.791 | $1.50e-04$ |
| teamCHI | 0.814 | 0.133 | 6.120 | $9.33e-10$ |
| teamCLE | 0.927 | 0.133 | 6.956 | $3.49e-12$ |
| teamDAL | 0.780 | 0.135 | 5.766 | $8.12e-09$ |
| teamDEN | – 0.814 | 0.135 | 6.020 | $1.74e-09$ |
| teanDET | 0.461 | 0.133 | 3.458 | $5.45e-04$ |
| teamGS | 1.882 | 0.144 | 13.050 | $2e-16$ |
| teamHOU | 1.449 | 0.138 | 10.474 | $2e-16$ |
| teamIND | 1.018 | 0.133 | 7.603 | $2.90e-14$ |
| teamLAC | 1.424 | 0.138 | 10.317 | $2e-16$ |
| teamLAL | 0.190 | 0.138 | 1.379 | 0.168 |
| teanMEM | 1.023 | 0.136 | 7.540 | $4.71e-14$ |
| teamMIA | 1.113 | 0.134 | 8.306 | $2e-16$ |
| teamMIL | 0.477 | 0.133 | 3.581 | $3.42e-04$ |
| teamMIN | 0.399 | 0.136 | 2.932 | 0.003 |
| teamNO | 0.617 | 0.135 | 4.554 | $5.27e-06$ |
| teamNY | 0.359 | 0.134 | 2.687 | 0.007 |
| teanOKC | 1.408 | 0.138 | 10.207 | $2e-16$ |
| teamPHI | 0.029 | 0.136 | 0.219 | 0.827 |
| teamPHO | 0.268 | 0.137 | 1.956 | 0.050 |
| teamPOR | 1.052 | 0.140 | 7.759 | $8.57e-15$ |
| teamSA | 1.759 | 0.142 | 12.395 | $2e-16$ |
| teamSAC | 0.261 | 0.137 | 1.903 | 0.057 |
| teanTOR | 1.179 | 0.135 | 8.739 | $2e-16$ |
| teamUTA | 0.837 | 0.135 | 6.195 | $5.85e-10$ |
| teamWAS | 0.802 | 0.133 | 6.036 | $1.58e-09$ |
| at home | 0.249 | 0.025 | 10.065 | $2e-16$ |

in this data set, 4,110 of them were wins for the home team, a home win percentage of 55.7%.

Table 6: Here we have the coefficients for the NBA teams as a result of the Bradley-Terry paired comparison model. The coefficients represent the strength of the team. Larger values suggest that the team is better than the other teams.

The at home variable shows that playing at home does give an advantage to not playing at home.

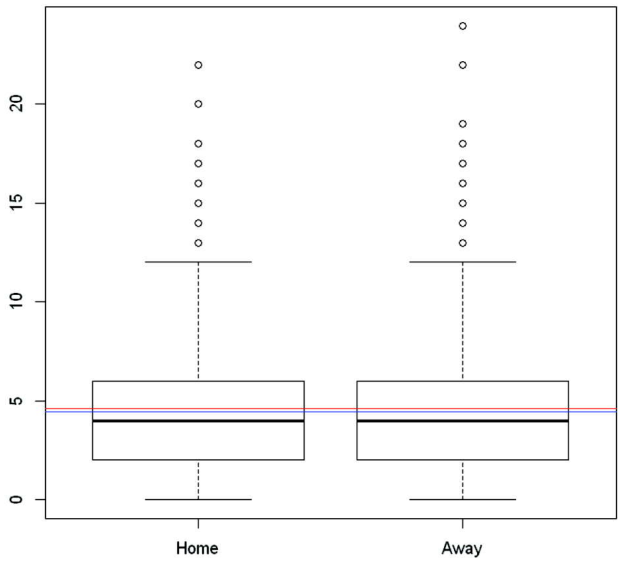

**Figure 2: MLB Points Scored**

## 3.8 Difference of Means MLB

Using the same analysis that we used for the NBA data, we tested for a no difference of mean for the MLB data set also. Using ?? and letting $\mu_H$ and $\mu_A$ represent the average points scored by the home and the away team respectively also. Using this same method, testing for a no difference in the home and away team gives a $p$-value of $p = 0.099$ and 95% confidence interval for the difference of mean being $(-0.329, 0.0266)$. This suggests that the true difference of means lies within this interval. Since 0 is included in this interval, this suggests that there may be no difference in the points that are scored at home vs. being away.

Figure 2: Here we have box plots of the scores for home and away teams in the MLB data set. The red line represents the average points for the home team and the blue line represents the average points for the away team. While these lines do not match with the indicated means of the box plots, the values are the same.

Again, it should be mentioned here that these results do not necessarily mean that the home team is more likely to win by scoring more points. This test was done to find evidence and support that the home team does have an advantage. While there was no evidence that the home team may score more points on average, this should still be clarified. Again, it should be stated that because the average points scored for the home team is larger, this does not mean that the home team will win.

**Table 7: MLB Bradley-Terry Coefficients**

| *Predictors* | *Coefficient Estimates* | *Std. Error* | *z value* | *Pr(> \|z\|)* |
|---|---|---|---|---|
| teamANA | 0.762 | 0.250 | 3.048 | 0.002 |
| teamATL | 0.886 | 0.229 | 3.863 | $1.12e-04$ |
| teamBAL | 0.144 | 0.260 | 0.553 | 0.581 |
| teamBOS | 1.114 | 0.252 | 4.416 | $1.01e-05$ |
| teamCHC | 0.409 | 0.234 | 1.747 | 0.081 |
| teamCHW | 1.031 | 0.253 | 4.066 | $4.78e-05$ |
| teamCIN | 0.676 | 0.233 | 2.894 | 0.004 |
| teamCLE | 0.764 | 0.254 | 3.009 | 0.003 |
| teanCOL | 0.539 | 0.223 | 2.423 | 0.015 |
| teamDET | 0.684 | 0.254 | 2.692 | 0.007 |
| teamFLA | 0.299 | 0.234 | 1.277 | 0.202 |
| teamHOU | 1.168 | 2.462 | 4.744 | $2.1e-06$ |
| teamKCR | 0.628 | 0.254 | 2.471 | 0.013 |
| teamLAD | 1.271 | 0.226 | 5.630 | $1.80e-08$ |
| teanMIL | 1.023 | 0.136 | 7.540 | $7.09e-05$ |
| teamMIN | 0.607 | 0.254 | 2.387 | 0.017 |
| teamNYM | 0.527 | 0.234 | 2.256 | 0.024 |
| teamNYY | 1.085 | 0.255 | 4.255 | $2.09e-05$ |
| teamOAK | 0.961 | 0.250 | 3.848 | $1.19e-04$ |
| teamPHI | 0.643 | 0.233 | 2.758 | 0.006 |
| teanPIT | 0.168 | 0.236 | 0.713 | 0.476 |
| teamSDP | 0.648 | 0.223 | 2.913 | 0.004 |
| teamSEA | 1.035 | 0.249 | 4.147 | $3.37e-05$ |
| teamSFG | 1.292 | 0.228 | 5.658 | $1.53e-08$ |
| teamSTL | 0.825 | 0.233 | 3.533 | $4.11e-04$ |
| teamTBD | 1.251 | 0.256 | 4.889 | $1.01e-06$ |
| teanTEX | 0.344 | 0.252 | 1.366 | 0.172 |
| teamTOR | 1.075 | 0.255 | 4.211 | $2.55e-05$ |
| teamWSN | 0.242 | 0.235 | 1.030 | 0.303 |
| at home | 0.171 | 0.042 | 4.098 | $4.16e-05$ |

## 3.9 Bradley-Terry Paired Comparison MLB

Conducting the Bradley-Terry paired comparison model for the MLB data, we see a similarity to the NBA model. The coefficients can be seen in table 7. Interpreting these results is the same as interpreting them for the NBA model. teamARI (Arizona) is the reference team (coefficient of 0 and not listed in the table) and teamSFG (San Francisco) is the strongest team (coefficient of 1.29). Looking at the coefficients in table 7 for the at.home variable we see a coefficient of 0.17059 which transforming this we can see that $\exp(0.17059) = 1.186$. This is the estimated odds-multiplier in the home team's favor [4]. The estimated odds-multiplier for the MLB model is smaller than the model for the NBA model. We can see this also by looking at the home win percentage. The home win percentage for the MLB model is 54% which is less than the home win percentage than the NBA model.

Table 7: Here we have the coefficients for the MLB teams as a result of the Bradley-Terry paired comparison model. The coefficients represent the strength of the team. Larger values suggest that the team is better than the other teams.

The at home variable shows that playing at home does give an advantage to not playing at home.

### 3.10 Advantages and Disadvantages of the Bradley-Terry Paired Comparison Model

The Bradley-Terry Paired Comparison model is good at detecting advantages or disadvantages in match ups against two opponents. This is useful to see if there is anything that can be determined as "unfair" in many different organizations. While these can be useful and helpful, this model does have some drawbacks. As already mentioned, this particular method can only be used in match ups where ties cannot occur. Additionally, this method can only be used for match ups that are mutually independent. For this analysis, there is an assumption that these match ups are mutually independent of one another when this may in fact not be the case. That is a drawback of this method as there are many factors that can lead to some match ups not being mutually independent.

## 4. SUMMARY AND CONCLUSION

This analysis shows that there is some statistical significance in playing at home for both NBA and MLB match ups. We saw that the team at home has better odds of winning than the team who is away. This analysis shows how significant playing at home is and is probably the reason why teams play an even number of home and away games every season. When looking at other predictors that can impact the course of the game, then we can see that while it still plays a part, the team location is not a statistically significant predictor. When looking at the team location on it's own, we can see that it is significant and that it can be an important role in the outcome of the game. It appears that home field makes a big difference since the win percentages are both greater than 50%. So while looking at additional predictors that directly impact and change the game more than the location of the game, but the location does favor the home team although not statistically significant in those models. The Bradley-Terry paired-comparison model showed that the odds are improved for winning at home. The predictor for team location in our linear regression model for the NBA showed that there is some significance for scoring more points if you are playing at home. This was also supported with the difference of mean test. In the logistic regression model, the team location predictor was not statistically significant but the coefficient was positive which also supports the findings of the Bradley-Terry paired-comparison model.

## 5. ACKNOWLEDGEMENTS

We would like to thank Dr. Gilberto Gonzales-Parra (New Mexico Tech), as well as Dr. Oleg Makhnin (New Mexico Tech) for their contributions, discussions, and editorial notes for this paper.

## *References*

[1] Roger R. Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, mar 1970. https://doi.org/10.1080/01621459.1970.10481082 doi: 10.1080/01621459.1970.10481082.

[2] Matt Errey. Basketball vocabulary. URL: https://www.englishclub.com/vocabulary/sports-basketball.htm.

[3] Daniel Schafer Fred Ramsey. *The Statistical Sleuth: A Course in Methods of Data Analysis*. DUXBURY PR, May 2012. URL: https://www.ebook.de/de/product/16791360/fred_ramsey_daniel_schafer_the_statistical_sleuth_a_course_in_methods_of_data_analysis.html.

[4] David Firth Heather Turner. *Bradley-Terry Models in R: The Bradley Terry2 Package*, February 2020.

[5] Michael Kutner. *Applied linear statistical models*. McGraw-Hill Education (India) Private Limited, New Delhi, 2013.

[6] Jordan R. Love. Predicting match outcomes on saltybet using bradley-terry paired-comparison models. resreport, Montana State University, May 2019.

[7] Thomas R. O'Connor and Michael T. Eskey. Scales and indexes, types of. In *Encyclopedia of Social Measurement*, chapter Comparitive Scales, pages 443–453. Elsevier, 2005. https://doi.org/10.1016/b0-12-369398-5/00434-5 doi:10.1016/b0-12-369398-5/00434-5.

[8] Paul Rossotti. Nba enhanced box score and standings (2012-2018). Kaggle, 2019. URL: https://www.kaggle.com/datasets/pablote/nba-enhanced-stats.

[9] Saurabh Shahane. Major league baseball dataset. Kaggle, May 2021. URL: https://www.kaggle.com/datasets/pablote/nba-enhanced-stats.

[10] Larry Wasserman. *All of Statistics*. Springer New York, Decem ber 2010. URL: h◎ps://www.ebook.de/de/product/13413823/larry_wasserman_all_of_ statistics.html.