STAT 428 Group 6 Written Report

Simulating NBA Match Results and Predicting NBA Playoff Teams

Zepeng Xiao, zepengx2
Shuogong, shuog2
Xiaoping Hua, xh3
Dongfan Li, dongfan2
Sixin Ma, sixinma2

April 28, 2019

Abstract

The main purpose of our project is simulating match results and predicting NBA playoff teams for season 2017-18 based on this season's real data, and comparing our results with this season's real data to calculate our method's accuracy.

Introduction

The National Basketball Association (NBA) is one of the most worldwide popular sporting events that is held every year in the world.

- It has two conferences, East and West, with 15 teams in each division.
- 30 teams fight in the league fight for a champion in a year-long season from October to June every year.
- Each season is split into regular season and playoffs.
- Each team has 82 matches to play in their regular season
- Teams in the same conference would play more often than those.
- Only the top 8 teams in each division are able to enter the playoffs of the season.
- Ranks are determined ascending by a special statistic called Games Behind:
- When team a is the leading team of the conference.

$$Team B's games behind = \frac{(Team A's Wins - Team A's Losses) - (Team B's Wins - Team B's Losses)}{2}$$

Thus which 16 teams would enter the playoffs every season is one of the biggest mystery in NBA every season.

Our goal, as introduced above, is simulating match results and predicting NBA playoff teams. In other words, we are going to predict which eight teams in each conference would get the lowest games behind by the end of the regular season.

The focus is on the past 17-18 season. We will try to predict based on prior part of this season's real data, and comparing our results with this season's real data in the rest season to calculate our method's accuracy. If we found our prediction is reliable in this season, we have prepared data in the past 10 years for test and verification.

Methods

Generalized Linear Regression

Permutation Test

The linear model we used to predict the game result includes predictors that might be correlated to each other. For example, `homeTeamRate` and `homeTeamlast10` seem to be positively correlated in common sense. Therefore, to improve our regression fit, we want to thorougly examine the correlations among the variables and proceed to use ridge regression if the variables are confirmed to be correlated. The variables for testing are,

- `homeTeamRate`
- `awayTeamRate`
- `homeTeamlast10`
- `awayTeamlast10`
- `histRate`

A permutation test essentially checks if X and Y have the same distribution by doing the following procedure,

1. Observe a test statistic for the null hypothesis, H_0
2. For each replicated $b = 1, 2, ..., B$:
   - Generate a random permutation π
   - Generate a new test statistic from the random permutation
3. Get a Monte Carlo estimate of the p-value by calculating the probability of obtaining a new test statistic that is more extreme than the observed test statistic in the B replicates
4. Reject H_0 at a significance level α if the p-value is less than α.

In our case, we can inherit the idea of permutation test and adapt it to paired data to check the correlation in between. If there is truly no association between X and Y, the distribution of $(X_i, Y_{\pi(i)})$ will be the same as that of (X_i, Y_i), where $\pi(i)$ is the i-th element of a permutation π of $1, 2, ..., 13$. We implement this idea by randomly permuting Y and pairing it with a fixed X and get a p-value for testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$, where ρ is either the Pearson correlation coefficient or the Spearman correlation coefficient:

- The *Pearson Method* evaluates the **linear** relationship between two continuous variables, where a change in one variable is associated with a **proportional** change in the other variable.
- The *Spearman Method* is based on the ranked values for each variable rather than the raw data. It evaluates the **monotonic** relationship between two continuous or ordinal variables, where the variables tend to change together, but not necessarily at a constant rate. It is more general than the Pearson method.

To determine which correlation coefficient to use for which pair of comparison, we plot variables against each another and added some noise using `jitter()` to visualize the trend of relationship and decide which method to use.

According to the plots, *Home Rate vs. Home Last 10*, *Away Rate vs. Away Last 10*, *Home Last 10 vs. Historical Rate*, and *Away Last 10 vs. Historical Rate* appear to be linear. We use the Pearson method for these pairs. For the rest pairs, trends are not that obvious. Therefore, we use the more genearl Spearman method.

Ridge Regression

Results

Linear Regrssion

Permutation Test

By performing permutation tests, we obtained the following table, which shows all the pairs we tested, the method we used, the p-value for the permutation test, and the decisions based on the p-values.

Pairs	Method	p-value	Decision
Home Rate vs. Home Last 10	Pearson	9.999e-05	Correlated
Away Rate vs. Away Last 10	Pearson	9.999e-05	Correlated
Home Rate vs. Historical Rate	Spearman	9.999e-05	Correlated
Away Rate vs. Historical Rate	Spearman	9.999e-05	Correlated
Home Last 10 vs. Historical Rate	Pearson	9.999e-05	Correlated
Away Last 10 vs. Historical Rate	Pearson	9.999e-05	Correlated
Home Rate vs. Away Rate	Spearman	0.4131587	Not Correlated
Home Last 10 vs. Away Last 10	Spearman	9.999e-05	Correlated
Home Rate vs. Away Last 10	Spearman	0.02689731	Correlated
Away Rate vs. Home Last 10	Spearman	0.02329767	Correlated

We found that all pairs of variables have correlation except *Home Rate vs. Away Rate*. This observation suggests that there exists problem of collinearity among the predictors. It leads us to further examine the VIF (Variance Inflation Factor) of the predictors and fit a Ridge Regression model to remedy the problem of collinearity.

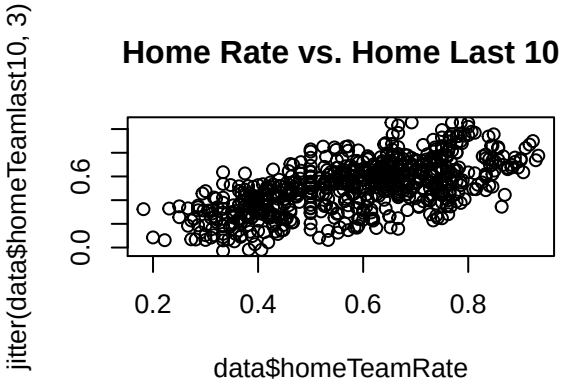
Ridge Regression

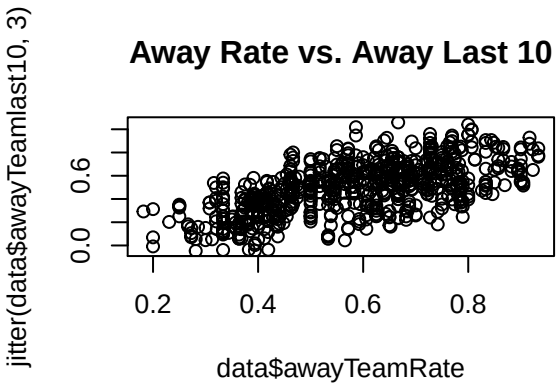
Discussion

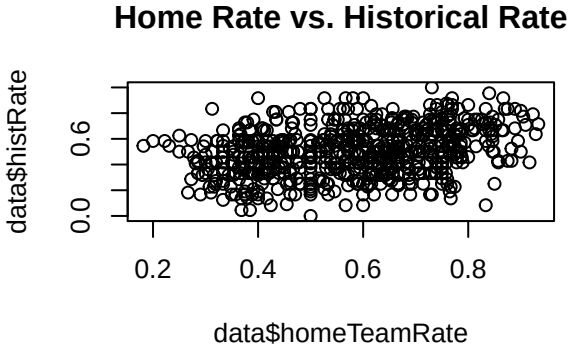
The test for correlation is not very accurate because the "Last 10" data (i.e. `homeTeamlast10` and `awayTeamlast10`) are highly categorical (discrete) because based on their calculation criteria: the number of times the team wins, which is an integer, divided by 10. When the "Last 10" data is paired with the other continuous variables, such as `homeTeamRate` and `histRate`, the correlation between a continuous variable and a somewhat discrete variable cannot be simply determined by correlation coefficients.

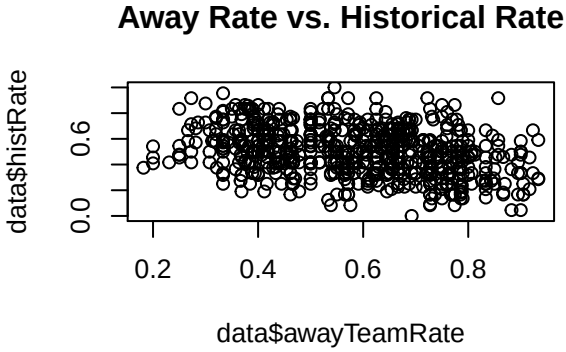
Appendix

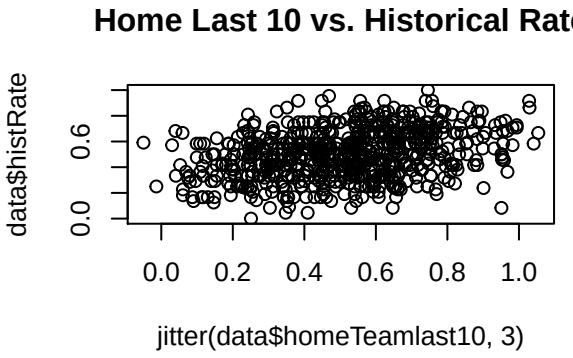
Scatterplots for Correlation

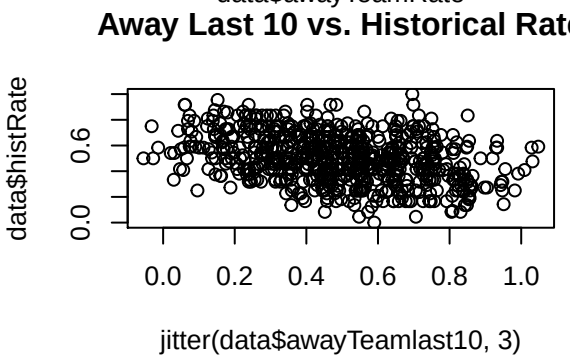


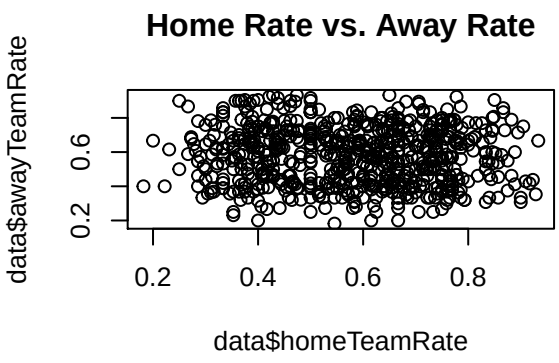


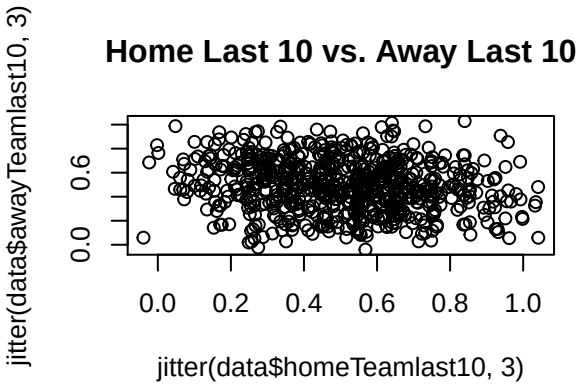


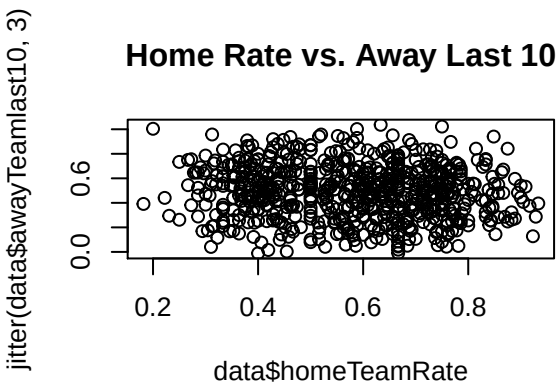


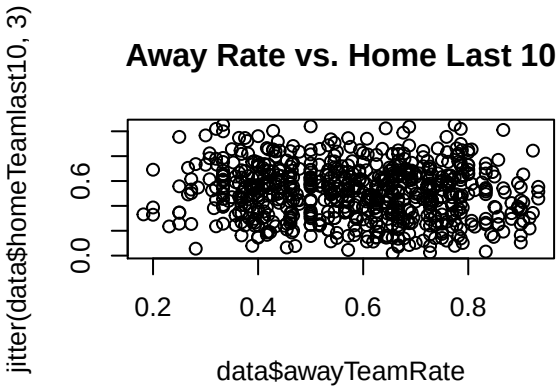












Permutation Test with either correlation coefficient

```
perm_test = function(X, Y, B = 10000, method = "pearson") {
  nu = seq_along(X)
  reps = numeric(B)

  if (method == "pearson") { # Pearson Method - default
    rho0 = abs(cor(X, Y))
    for ( i in 1:B ) {
      perm = sample(nu, size = length(X), replace = FALSE)
      X1 = X[perm]
      reps[i] = abs(cor(X1, Y))
    }
    pval = mean(c(rho0, reps) >= rho0)
  } else if (method == "spearman") { # Spearman Method
    rho0 = cor(X, Y, method = "spearman")
    t0 = abs(rho0*sqrt((length(X) - 2)/(1 - rho0^2)))
    for ( i in 1:B ) {
      perm = sample(nu, size = length(X), replace = FALSE)
      X1 = X[perm]
      rho = cor(X1, Y, method = "spearman")
      reps[i] = abs(rho*sqrt((length(X) - 2)/(1 - rho^2)))
    }
    pval = mean(c(t0, reps) >= t0)
  }

  return(pval)
}

# running the tests
perm_test(data$homeTeamRate, data$homeTeamlast10, 10000, "pearson")
perm_test(data$awayTeamRate, data$awayTeamlast10, 10000, "pearson")
perm_test(data$homeTeamRate, data$histRate, 10000, "spearman")
perm_test(data$awayTeamRate, data$histRate, 10000, "spearman")
perm_test(data$homeTeamlast10, data$histRate, 10000, "spearman")
perm_test(data$awayTeamlast10, data$histRate, 10000, "spearman")
perm_test(data$homeTeamRate, data$awayTeamRate, 10000, "spearman")
perm_test(data$homeTeamlast10, data$awayTeamlast10, 10000, "spearman")
perm_test(data$homeTeamRate, data$awayTeamlast10, 10000, "spearman")
perm_test(data$awayTeamRate, data$homeTeamlast10, 10000, "spearman")
```