Stat 4201 Project: Is Carmelo Anthony helping the Knicks?

Mengqi Zong < *mz2326@columbia.edu* >

May 1, 2012

1 Introduction

How to measure the productivity of an individual participating in a team sport? This has always been an interesting and important question.

In this paper, we will explore methods of linking the player's statistics in the National Basketball Association (NBA) to team wins. Specifically, we will measure the productivity of New York Knicks player Carmelo Anthony to his team based on 18 consecutive games of Knicks from February 6th 2012 to March 12th 2012.

2 Background

Carmelo Klyan Anthony is an American professional basketball player who currently plays for the New York Knicks in NBA. Since entering the NBA, Anthony has emerged as one of the most well-known and skilled players in the league. Now, he is being considered as the superstar of New York Knicks.

Despite his excellent basketball skills, Anthony has been long criticized for not being able to bring wins to his team. The level of the doubt becomes severe than ever after March 12th 2012, that Knicks suffered a 6 straight lost.

On February 6th, New York Knicks player Jeremy Lin was promoted to the starting lineup at that day's game of New York Knicks versus New Jersey Nets. Since that game, due to Lin's excellent performance, Knicks produced a winning streak.

However, it is worth pointing out that Carmelo Anthony got injured in the February 6th's game and only played 5 minutes. Later, he missed the next 7 games. During Anthony's absence, Knicks won 7 of the 8 games. However, after Anthony's return, Knicks only won 2 of 10 games from Feb 20th to March 12th. All 18 game records are shown in Table-1.

Date	Opponent	Scores	W/L	Date	Opponent	Scores	W/L
Feb. 6th	UT	99-88	W	Feb. 20th	NJ	92-100	L
Feb. 8th	WAS	107-93	W	Feb. 22nd	ATL	99-82	W
Feb. 10th	LAL	92-85	W	Feb. 23rd	MIA	88-102	L
Feb. 11th	MIN	100-98	W	Feb. 29th	CLE	120-103	W
Feb. 14th	TOR	90-87	W	Mar. 4th	BOS	111-115	L
Feb. 15th	SAC	100-85	W	Mar. 6th	DAL	85-89	L
Feb. 17th	NO	85-89	L	Mar. 7th	SA	105-118	L
Feb. 19th	DAL	104-97	W	Mar. 9th	MIL	114-119	L
				Mar. 11th	PHI	94-106	L
				Mar. 12th	CHI	99-104	L

Table 1: Brief game records

After Knicks's losing streak, the media blame Anthony for Knick's bad performance. We will try to figure out if it is Anthony's fault.

3 Model 1: Categorical Data Analysis

3.1 Data Processing

The data can be displayed as a 2×2 table of counts, which lists the numbers of games falling in each cross-classification of a row factor (Anthony is absent or Anthony is available) and a column factor (Knicks won or Knicks lost). The categorical data is shown in Table-2.

3.2 Fisher's Exact Test

We can apply Fisher's Exact Test to the data to test whether the two probabilities of winning when Anthony is absent or not are the same. Here is the output from R:

	Knicks Won	Knicks Lost
Anthony is absent	7	1
Anthony is available	2	8

Table 2: Categorized game records

```
Fisher's Exact Test for Count Data

data:  win.data
p-value = 0.01522
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
    1.533351 1396.017102
sample estimates:
odds ratio
  21.39941
```

As we can see, the p-value is 0.01522. This indicates that the two probabilities of winning when Anthony is absent or not are not likely to be the same. More specifically, when Anthony is available, the Knicks tends to lose more.

3.3 A closer look at this model

There are certain things we have to consider about this model:

1. The sample size maybe too small to make an accurate prediction.

   There are only 18 sample units in total. And the 95 percent confidence interval of the odds ratio is $[1.53, 1396.02]$. The confidence interval is way too wide, which means that the result is not accurate enough.

2. The two probabilities of winning are different does not necessarily imply that this is all Carmelo Anthony's fault.

   Even if the Fisher's exact test is reliable, we still can't draw the conclusion that Knicks's bad performance is all Anthony's fault. Because there are many factors that affect the winning and losing of a basketball team we haven't considered in this model:

- The strength of the opponents.

  Apparently, the probability of winning when playing against weak opponents is higher than that of when playing against strong opponents. In this model, the underlying assumption is that the strength of the opponents in the two time periods are roughly the same.

- Home court advantage.

  It has been widely believed that the team playing at their home court has an advantage. In this model, the underlying assumption is there's no home court advantage out there.

- Teammates' performance.

  Basketball is a team sports, the performance of Anthony's teammates will surely affect the probability of winning and losing of the Knicks.

To sum up, this model is an easy and straight forward model. However, it fails to consider many factors that affects the results of a game.

4 Observations

A much detailed version of the game records is shown in Table-3. Note that whether the opponent is a strong or not is based on my empirical experience. From the records, we draw the follow observations.

Observation 1. *This is only 2 of 8 strong opponents when Anthony is absent, but there are 5 of 10 strong opponents when Anthony is back.*

This indicates that the strength of opponents for the two time period is not the same.

Observation 2. *When Anthony is absent, Knicks played against its 2 strong opponents in New York with a home court advantage. When Anthony is avaible, they played against its 5 strong oppoents out of New York without a home court advantage.*

To sum up, observations show that we should take the strength of the opponents and home court advantage into consideration to achieve more accuarte result.

Date	Opponent	Scores	Home	Strong OP	W/L
Feb. 6th	UT	99-88	Home	False	W
Feb. 8th	WAS	107-93	Away	False	W
Feb. 10th	LAL	92-85	Home	True	W
Feb. 11th	MIN	100-98	Away	False	W
Feb. 14th	TOR	90-87	Away	False	W
Feb. 15th	SAC	100-85	Home	False	W
Feb. 17th	NO	85-89	Home	False	L
Feb. 19th	DAL	104-97	Home	True	W
Feb. 20th	NJ	92-100	Home	False	L
Feb. 22nd	ATL	99-82	Home	False	W
Feb. 23rd	MIA	88-102	Away	True	L
Feb. 29th	CLE	120-103	Home	False	W
Mar. 4th	BOS	111-115	Away	True	L
Mar. 6th	DAL	85-89	Away	True	L
Mar. 7th	SA	105-118	Away	True	L
Mar. 9th	MIL	114-119	Away	False	L
Mar. 11th	PHI	94-106	Home	False	L
Mar. 12th	CHI	99-104	Away	True	L

Table 3: A much detailed version of the game records

5 Model 2: Logistic Regression

5.1 Data Processing

The main problem of the data processing is how to measure the strength of an oppoent. We will use the opponent's winning percentage till the day of that game to indicate the strength of this opponent.

Example 3. *Till Feb. 6th, Utah Jazz played 31 games and won 15 of them. The winning percentage of Utah at Feb. 6th is*

$$p = 15 \div 31 = 0.48$$

One concern about this method is that each team's schedule is different, so it is possible that the winning percentage defined can not accurately indicate the strength of a team. However, we will show that this is not a big problem. Till Feb. 6th, all

teams have played at least 30 games and there are 30 teams in the league. This means that it is quite possible all teams have played against each other at least once. So the schedule doesn't affect much here.

The data used in this model is shown in Table-4

Number	W/L	PCT	Home	Anthony
1	W	0.48	Home	Absent
2	W	0.22	Away	Absent
3	W	0.46	Home	Absent
4	W	0.44	Away	Absent
5	W	0.30	Away	Absent
6	W	0.28	Home	Absent
7	L	0.26	Home	Absent
8	W	0.60	Home	Absent
9	L	0.32	Home	Available
10	W	0.55	Home	Available
11	L	0.82	Away	Available
12	W	0.38	Home	Available
13	L	0.57	Away	Available
14	L	0.61	Away	Available
15	L	0.73	Away	Available
16	L	0.44	Away	Available
17	L	0.58	Home	Available
18	L	0.75	Away	Available

Table 4: data for logit regression

5.2 The Logit Regression

Let π denote the Knicks' probability of winning. We build the following logit regression model:

$$logit(\pi) = \beta_0 + \beta_1 \cdot Home + \beta_2 \cdot PCT + \beta_3 \cdot CA$$

In this model, Home is an indicator variable that denote if Knicks is the home team. PCT is the opponents winning percentage. CA is the indicator variable that

denote if Carmelo Anthony is available in this game.

Here is the output from R:

```
Call:
glm(formula = Win ~ Home + PCT + CA)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95772  -0.12527  -0.02737   0.18760   0.68669

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7988     0.3489   2.289   0.0381 *
Home          0.1857     0.2023   0.918   0.3742
PCT          -0.1045     0.6909  -0.151   0.8819
CA           -0.6129     0.2333  -2.627   0.0199 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for gaussian family taken to be 0.1651435)

    Null deviance: 4.500  on 17  degrees of freedom
Residual deviance: 2.312  on 14  degrees of freedom
AIC: 24.141

Number of Fisher Scoring iterations: 2
```

We can see that the p-value of CA is 0.0199, which indicates that CA is an important variable in this model. Also, the drop-in-deviance test gives the similar result:

```
Model:
Win ~ Home + PCT + CA
       Df Deviance    AIC scaled dev. Pr(>Chi)
<none>      2.3120 24.141
Home    1   2.4511 23.193      1.0518 0.305099
PCT     1   2.3158 22.171      0.0294 0.863845
CA      1   3.4518 29.355      7.2140 0.007234 **
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

So we get

$$logit(\pi) = 0.7988 + 0.1857 \cdot Home - 0.1045 \cdot PCT - 0.6129 \cdot CA$$

Since the coefficient of CA is -0.6129, this indicates that Carmelo Anthony has a negative effect on Knicks. This means that when Carmelo Anthony is available, Knicks tends to lose more.

5.3 A closer look at this model

Here are some comments about this model:

1. The sample size may not be enough.

   Recall that in class, we learned that the rule of thumb for the sample size of a regression is 1 variable needs 10 sample units. Since there are 3 variables in the logit model, the ideal sample size should be at least 30. However, there are only 18 sample units. This may lead to an inaccurate variable estimation.

2. The data are not independent.

   The assumption of regression requires all data are independent from each other. However, the independent assumption doesn't hold here because in each season two teams will play against each other multiple times. For example, Knicks played against Dallas Mavericks at Feb. 19th. Then the two team played against each other again at Mar. 6th.

   We will show that this doesn't affect our model much since the only team that appeared multiple time is Dallas Mavericks for twice.

3. Model team wins as binary responses may not be sufficient.

   Since each game's situation differs from each other, the binary responses may not be a good model to check if Anthony is helping the Knicks. Suppose Anthony is helping the Knicks, but the opponent is the second time period is too strong for Knicks to get a win. In this case, we will incorrectly draw the conclusion that Anthony has a negative effect on the team.

4. Teammates' performance is not concerned in the model.

   We will show that this is not a big deal. Because it is rational to assume that all player's performance is consistent.

To sum up, basically, the logit regression is a good model to show if Carmelo Anthony is helping the Knicks.

6 Conclusion

Based on the results given by the two models, we draw the conclusion that Carmelo Anthony is not helping the Knicks during Feb. 20th to Mar. 12th.

Appendices

The R code is listed below:

```
#model 1
win.data = matrix(c(7,1,2,8), 2, 2)
fisher = fisher.test(win.data)

#model 2
data.m2 = read.table("knicks.csv",header=T, sep=",")
data.m2 = data.m2[1:18,]
attach(data.m2)
CA = 1 * (Player == "Anthony")
fit.m2 = glm(Win~Home + PCT + CA)
drop1(fit.m2, test="Chisq")
```