3 Pythagorean expectation and the NBA

Jiannan Xu[*]

17 July 2020

Pythagorean Expectation and the NBA

The NBA is the most popular basketball league in the world, and consists of 30 teams playing an 82 game regular season followed by playoffs to determine the champion. In terms of scale, this data looks much more like MLB data than the IPL data we just looked at.

Basketball resembles cricket in one way - the scores are much higher than in baseball. However, the points difference between winning and losing teams tend to be relatively small.

Let's see what we find this time. We follow the same procedure.

```
# Load the packages

options(warn = -1)
library("readxl",quietly = TRUE)
library("tidyverse",quietly = TRUE)
library("dplyr",quietly = TRUE)
library("ggplot2",quietly = TRUE)
```

```
# Now we import the data, which comes in the form of
# a list of games played in the 2018 season.
# We print out the list of variables names in the dataframe
# Load the data and see what it looks like

NBA = read.csv('NBA_Games.csv')
head(NBA)
```

```
##            CITY             TEAM_NAME    TEAM_ID  GAME_ID NICKNAME
## 1 Oklahoma City Oklahoma City Thunder 1610612760 11300001  Thunder
## 2       Chicago         Chicago Bulls 1610612741 11300002    Bulls
## 3       Indiana        Indiana Pacers 1610612754 11300002   Pacers
## 4   New Orleans  New Orleans Pelicans 1610612740 11300003 Pelicans
## 5       Houston       Houston Rockets 1610612745 11300003  Rockets
## 6  Golden State Golden State Warriors 1610612744 11300004 Warriors
##        STATE YEAR_FOUNDED SEASON_ID TEAM_ABBREVIATION GAME_DATE
## 1   Oklahoma         1967     12013               OKC 10/5/2013
## 2   Illinois         1966     12013               CHI 10/5/2013
## 3    Indiana         1976     12013               IND 10/5/2013
```

[*] ansonxjn@umich.edu; Master student at Department of Statistics, University of Michigan.

```
## 4  Louisiana         2002     12013               NOP 10/5/2013
## 5      Texas         1967     12013               HOU 10/5/2013
## 6 California         1946     12013               GSW 10/5/2013
##       MATCHUP WL MIN PTS PTSAGN FGM FGA FG_PCT FG3M FG3A FG3_PCT FTM FTA
## 1   OKC @ FBU  W 241  95     NA  36  74  0.486    2   14   0.143  21  28
## 2   CHI @ IND  W 239  82     76  28  74  0.378    3   10   0.300  23  30
## 3 IND vs. CHI  L 240  76     82  27  82  0.329    5   17   0.294  17  23
## 4   NOP @ HOU  W 239 116    115  42  86  0.488    4   10   0.400  28  40
## 5 HOU vs. NOP  L 240 115    116  39  68  0.574    8   22   0.364  29  38
## 6   GSW @ LAL  L 241  95    104  35  88  0.398    5   21   0.238  20  28
##   FT_PCT OREB DREB REB AST STL BLK TOV PF PLUS_MINUS home not.paired
## 1  0.750   18   34  52  22   9   8  20 26         13    A        999
## 2  0.767   17   39  56  20   5  10  23 25          6    A          0
## 3  0.739   11   27  38  15  12   8  15 23         -6    H          0
## 4  0.700   12   21  33  17  12   4  15 32          1    A          0
## 5  0.763    5   30  35  24   9   4  22 27         -1    H          0
## 6  0.714   14   38  52  21  13   5  23 31         -9    A          0
```

```r
tail(NBA)
```

```
##              CITY           TEAM_NAME    TEAM_ID    GAME_ID  NICKNAME
## 18411   Cleveland Cleveland Cavaliers 1610612739 1621900005 Cavaliers
## 18412     Memphis   Memphis Grizzlies 1610612763 1621900005 Grizzlies
## 18413 San Antonio   San Antonio Spurs 1610612759 1621900006     Spurs
## 18414        Utah           Utah Jazz 1610612762 1621900006      Jazz
## 18415                                         NA         NA
## 18416                                         NA         NA
##           STATE YEAR_FOUNDED SEASON_ID TEAM_ABBREVIATION GAME_DATE
## 18411      Ohio         1970     22019               CLE  7/3/2019
## 18412 Tennessee         1995     22019               MEM  7/3/2019
## 18413     Texas         1976     22019               SAS  7/3/2019
## 18414      Utah         1974     22019               UTA  7/3/2019
## 18415                     NA        NA
## 18416                     NA        NA
##           MATCHUP WL MIN PTS PTSAGN FGM FGA FG_PCT FG3M FG3A FG3_PCT FTM
## 18411   CLE @ MEM  L 201  68     81  22  60  0.367   10   35   0.286  14
## 18412 MEM vs. CLE  W 199  81     68  31  74  0.419   10   28   0.357   9
## 18413   SAS @ UTA  L 200  81     84  29  70  0.414    7   17   0.412  16
## 18414 UTA vs. SAS  W 199  84     81  30  74  0.405   11   31   0.355  13
## 18415                 NA  NA     NA  NA  NA     NA   NA   NA      NA  NA
## 18416                 NA  NA     NA  NA  NA     NA   NA   NA      NA  NA
##       FTA FT_PCT OREB DREB REB AST STL BLK TOV PF PLUS_MINUS home
## 18411  19  0.737    3   26  29  15   9   3  17 12      -13.0    A
## 18412  13  0.692   11   36  47  19   8   5  14 19       13.0    H
## 18413  26  0.615    9   29  38  12   5   3  11 14       -7.2    A
## 18414  19  0.684   13   31  44  15   6   5  15 25        5.8    H
## 18415  NA      NA   NA   NA  NA  NA  NA  NA  NA NA         NA
## 18416  NA      NA   NA   NA  NA  NA  NA  NA  NA NA         NA
##       not.paired
## 18411          0
## 18412          0
## 18413          0
## 18414          0
## 18415         NA
## 18416        636
```

```
# The data consists of games played between 2013 and 2019.
# An important difference from the baseball and cricket data
# is that here each game appears in two rows, one for each team.
# Each pair of rows are mirror images of each other.

# The season is identified by the column SEASON_ID
# Pre-season games have the prefix "1" before the year,
# regular season games have the prefix "2"
# and postseason games have the prefix "4".

# We are going to look at the 2018 regular season and therefore
# want games with the prefix "2"
# We can use the command "summary()" to obtain descriptive statistics for our variable

NBAR18 = NBA %>% filter(SEASON_ID == 22018)
NBAR18 %>% summary()
```

```
##           CITY                      TEAM_NAME       TEAM_ID
##  Los Angeles :  92   Los Angeles Lakers    :  92   Min.   :1.611e+09
##  Memphis     :  92   Memphis Grizzlies     :  92   1st Qu.:1.611e+09
##  Atlanta     :  91   Atlanta Hawks         :  91   Median :1.611e+09
##  Golden State:  90   Golden State Warriors:  90   Mean   :1.611e+09
##  Miami       :  90   Miami Heat            :  90   3rd Qu.:1.611e+09
##  Sacramento  :  90   Sacramento Kings      :  90   Max.   :1.611e+09
##  (Other)     :2015   (Other)               :2015
##     GAME_ID              NICKNAME            STATE         YEAR_FOUNDED
##  Min.   :2.180e+07   Grizzlies:  92   California: 272   Min.   :1946
##  1st Qu.:2.180e+07   Lakers   :  92   Texas     : 264   1st Qu.:1949
##  Median :2.180e+07   Hawks    :  91   Florida   : 177   Median :1970
##  Mean   :1.280e+08   Heat     :  90   New York  : 174   Mean   :1970
##  3rd Qu.:2.180e+07   Jazz     :  90   Tennessee :  92   3rd Qu.:1980
##  Max.   :1.622e+09   Kings    :  90   Atlanta   :  91   Max.   :2002
##                      (Other)  :2015   (Other)   :1490
##    SEASON_ID     TEAM_ABBREVIATION      GAME_DATE           MATCHUP
##  Min.   :22018   LAL    :  92      11/23/2018:  27   MEM @ UTA  :   4
##  1st Qu.:22018   MEM    :  92      4/7/2019  :  27   UTA vs. MEM:   4
##  Median :22018   ATL    :  91      11/21/2018:  26   ATL @ CHI  :   3
##  Mean   :22018   GSW    :  90      4/5/2019  :  25   ATL @ IND  :   3
##  3rd Qu.:22018   MIA    :  90      12/19/2018:  24   ATL vs. NYK:   3
##  Max.   :22018   SAC    :  90      2/23/2019 :  24   BKN @ IND  :   3
##                  (Other):2015      (Other)   :2407   (Other)    :2540
##  WL            MIN           PTS            PTSAGN          FGM
##   :   0   Min.   :197   Min.   : 53.0   Min.   : 53.0   Min.   :17.00
##  L:1286   1st Qu.:239   1st Qu.:100.0   1st Qu.:100.0   1st Qu.:37.00
##  W:1274   Median :240   Median :110.0   Median :110.0   Median :40.00
##           Mean   :239   Mean   :109.2   Mean   :109.1   Mean   :40.36
##           3rd Qu.:241   3rd Qu.:118.0   3rd Qu.:118.0   3rd Qu.:44.00
##           Max.   :341   Max.   :168.0   Max.   :168.0   Max.   :61.00
##                                         NA's   :88
##       FGA             FG_PCT            FG3M            FG3A
##  Min.   : 55.00   Min.   :0.2620   Min.   : 2.00   Min.   :13.00
##  1st Qu.: 83.00   1st Qu.:0.4208   1st Qu.: 9.00   1st Qu.:27.00
##  Median : 88.00   Median :0.4580   Median :11.00   Median :31.00
##  Mean   : 88.21   Mean   :0.4580   Mean   :11.16   Mean   :31.75
##  3rd Qu.: 93.00   3rd Qu.:0.4940   3rd Qu.:13.00   3rd Qu.:36.00
##  Max.   :123.00   Max.   :0.6490   Max.   :27.00   Max.   :70.00
##
##     FG3_PCT            FTM             FTA             FT_PCT
##  Min.   :0.0800   Min.   : 2.00   Min.   : 4.00   Min.   :0.2630
##  1st Qu.:0.2920   1st Qu.:13.00   1st Qu.:18.00   1st Qu.:0.7000
##  Median :0.3480   Median :17.00   Median :22.00   Median :0.7680
##  Mean   :0.3508   Mean   :17.32   Mean   :22.72   Mean   :0.7631
##  3rd Qu.:0.4060   3rd Qu.:21.00   3rd Qu.:27.25   3rd Qu.:0.8330
##  Max.   :0.8420   Max.   :44.00   Max.   :54.00   Max.   :1.0000
##
##       OREB            DREB           REB            AST
##  Min.   : 1.00   Min.   :17.0   Min.   :22.00   Min.   : 7.00
##  1st Qu.: 8.00   1st Qu.:30.0   1st Qu.:40.00   1st Qu.:20.00
##  Median :10.00   Median :34.0   Median :44.00   Median :24.00
##  Mean   :10.34   Mean   :34.4   Mean   :44.75   Mean   :24.03
##  3rd Qu.:13.00   3rd Qu.:38.0   3rd Qu.:49.00   3rd Qu.:28.00
##  Max.   :26.00   Max.   :55.0   Max.   :71.00   Max.   :42.00
##
##       STL              BLK              TOV              PF
##  Min.   : 0.000   Min.   : 0.000   Min.   : 3.00   Min.   : 9.00
##  1st Qu.: 6.000   1st Qu.: 3.000   1st Qu.:11.00   1st Qu.:18.00
##  Median : 7.000   Median : 5.000   Median :13.00   Median :21.00
##  Mean   : 7.694   Mean   : 4.974   Mean   :13.61   Mean   :20.85
##  3rd Qu.:10.000   3rd Qu.: 6.000   3rd Qu.:16.00   3rd Qu.:24.00
##  Max.   :20.000   Max.   :19.000   Max.   :27.00   Max.   :38.00  
##                                                                    
##    PLUS_MINUS         home       not.paired    
##  Min.   :-56.00000    :   0   Min.   :  0.00  
##  1st Qu.: -9.00000   A:1279   1st Qu.:  0.00  
##  Median : -1.00000   H:1281   Median :  0.00  
##  Mean   : -0.04188            Mean   : 34.34  
##  3rd Qu.:  9.00000            3rd Qu.:  0.00  
##  Max.   : 56.00000            Max.   :999.00  
## 
```

```
# We can list all the variable names

names(NBAR18)
```

```
##  [1] "CITY"              "TEAM_NAME"         "TEAM_ID"          
##  [4] "GAME_ID"           "NICKNAME"          "STATE"            
##  [7] "YEAR_FOUNDED"      "SEASON_ID"         "TEAM_ABBREVIATION"
## [10] "GAME_DATE"         "MATCHUP"           "WL"               
## [13] "MIN"               "PTS"               "PTSAGN"           
## [16] "FGM"               "FGA"               "FG_PCT"           
## [19] "FG3M"              "FG3A"              "FG3_PCT"          
## [22] "FTM"               "FTA"               "FT_PCT"           
## [25] "OREB"              "DREB"              "REB"              
## [28] "AST"               "STL"               "BLK"              
## [31] "TOV"               "PF"                "PLUS_MINUS"       
## [34] "home"              "not.paired"
```

```
# Many datasets contain missing variables.
# Missing variables in a column will usually cause operations to fail.
# The command ".dropna()" will eliminate missing variables.
# Compare the counts of variables below after the na.omit() below to the counts in the

NBAR18 <- NBAR18 %>% na.omit()
NBAR18 %>% summary()
```

```
##          CITY                        TEAM_NAME        TEAM_ID         
##  Atlanta    :  88   Atlanta Hawks       :  88   Min.   :1.611e+09  
##  Memphis    :  88   Memphis Grizzlies   :  88   1st Qu.:1.611e+09  
##  Miami      :  88   Miami Heat          :  88   Median :1.611e+09  
##  Cleveland  :  87   Cleveland Cavaliers:  87   Mean   :1.611e+09  
##  Los Angeles:  87   Los Angeles Lakers :  87   3rd Qu.:1.611e+09  
##  Utah       :  87   Utah Jazz           :  87   Max.   :1.611e+09  
##  (Other)    :1947   (Other)             :1947                      
##     GAME_ID                NICKNAME            STATE        YEAR_FOUNDED 
##  Min.   :2.180e+07   Grizzlies:  88   California: 257   Min.   :1946  
##  1st Qu.:2.180e+07   Hawks    :  88   Texas     : 252   1st Qu.:1949
##  Median :2.180e+07   Heat     :  88   Florida   : 173   Median :1970
##  Mean   :1.281e+08   Cavaliers:  87   New York  : 170   Mean   :1970
##  3rd Qu.:2.180e+07   Jazz     :  87   Atlanta   :  88   3rd Qu.:1980
##  Max.   :1.622e+09   Lakers   :  87   Tennessee :  88   Max.   :2002
##                      (Other)  :1947   (Other)   :1444
##    SEASON_ID     TEAM_ABBREVIATION     GAME_DATE           MATCHUP
##  Min.   :22018   ATL    :  88      11/21/2018:  26   MEM @ UTA  :   4
##  1st Qu.:22018   MEM    :  88      11/23/2018:  26   UTA vs. MEM:   4
##  Median :22018   MIA    :  88      4/7/2019  :  26   ATL @ CHI  :   3
##  Mean   :22018   CLE    :  87      12/19/2018:  24   ATL @ IND  :   3
##  3rd Qu.:22018   LAL    :  87      2/23/2019 :  24   ATL vs. NYK:   3
##  Max.   :22018   UTA    :  87      4/5/2019  :  24   BKN @ IND  :   3
##                  (Other):1947      (Other)   :2322   (Other)    :2452
##  WL            MIN              PTS             PTSAGN            FGM
##   :   0   Min.   :197.0   Min.   : 53.0   Min.   : 53.0   Min.   :17.00
##  L:1236   1st Qu.:239.0   1st Qu.:100.0   1st Qu.:100.0   1st Qu.:37.00
##  W:1236   Median :240.0   Median :110.0   Median :110.0   Median :40.00
##           Mean   :238.9   Mean   :109.1   Mean   :109.1   Mean   :40.34
##           3rd Qu.:241.0   3rd Qu.:118.0   3rd Qu.:118.0   3rd Qu.:44.00
##           Max.   :341.0   Max.   :168.0   Max.   :168.0   Max.   :61.00
##
##       FGA              FG_PCT            FG3M             FG3A
##  Min.   : 55.00   Min.   :0.2620   Min.   : 2.00   Min.   :13.00
##  1st Qu.: 83.00   1st Qu.:0.4208   1st Qu.: 9.00   1st Qu.:27.00
##  Median : 88.00   Median :0.4580   Median :11.00   Median :31.50
##  Mean   : 88.14   Mean   :0.4581   Mean   :11.19   Mean   :31.83
##  3rd Qu.: 93.00   3rd Qu.:0.4940   3rd Qu.:14.00   3rd Qu.:36.00
##  Max.   :123.00   Max.   :0.6490   Max.   :27.00   Max.   :70.00
##
##     FG3_PCT            FTM              FTA            FT_PCT
##  Min.   :0.0800   Min.   : 2.00   Min.   : 4.0   Min.   :0.2630
##  1st Qu.:0.2930   1st Qu.:13.00   1st Qu.:18.0   1st Qu.:0.7000
##  Median :0.3500   Median :17.00   Median :22.0   Median :0.7670
##  Mean   :0.3511   Mean   :17.21   Mean   :22.6   Mean   :0.7626
##  3rd Qu.:0.4070   3rd Qu.:21.00   3rd Qu.:27.0   3rd Qu.:0.8330
##  Max.   :0.8420   Max.   :44.00   Max.   :54.0   Max.   :1.0000
##
##       OREB             DREB             REB              AST
##  Min.   : 1.00   Min.   :17.00   Min.   :22.00   Min.   : 7.00
##  1st Qu.: 8.00   1st Qu.:30.00   1st Qu.:40.00   1st Qu.:20.00
##  Median :10.00   Median :34.00   Median :44.00   Median :24.00
##  Mean   :10.32   Mean   :34.41   Mean   :44.73   Mean   :24.05
##  3rd Qu.:13.00   3rd Qu.:38.00   3rd Qu.:49.00   3rd Qu.:28.00
##  Max.   :26.00   Max.   :55.00   Max.   :71.00   Max.   :42.00
```

```
##
##       STL              BLK              TOV              PF
##  Min.   : 0.000   Min.   : 0.000   Min.   : 3.00   Min.   : 9.00
##  1st Qu.: 6.000   1st Qu.: 3.000   1st Qu.:11.00   1st Qu.:18.00
##  Median : 7.000   Median : 5.000   Median :13.00   Median :21.00
##  Mean   : 7.675   Mean   : 4.933   Mean   :13.64   Mean   :20.75
##  3rd Qu.:10.000   3rd Qu.: 6.000   3rd Qu.:16.00   3rd Qu.:23.00
##  Max.   :20.000   Max.   :19.000   Max.   :27.00   Max.   :37.00
##
##    PLUS_MINUS        home       not.paired
##  Min.   :-56.00000    :   0   Min.   :0
##  1st Qu.: -9.00000   A:1236   1st Qu.:0
##  Median :  0.00000   H:1236   Median :0
##  Mean   : -0.01909            Mean   :0
##  3rd Qu.:  9.00000            3rd Qu.:0
##  Max.   : 56.00000            Max.   :0
##
```

```
# The game result is the column labeled 'WL'.
# We create a variable which has a value of '1' if the team won, and zero if it lost.
# This type of variable, where a condition (here winning) is
# either true (1) or not true (0) is called a "dummy variable".
# We will encounter them frequently.

NBAR18[,'result'] = ifelse(NBAR18$WL == 'W',1,0)
NBAR18 %>% summary()
```

```
##           CITY                     TEAM_NAME       TEAM_ID
##  Atlanta    :  88   Atlanta Hawks      :  88   Min.   :1.611e+09
##  Memphis    :  88   Memphis Grizzlies  :  88   1st Qu.:1.611e+09
##  Miami      :  88   Miami Heat         :  88   Median :1.611e+09
##  Cleveland  :  87   Cleveland Cavaliers:  87   Mean   :1.611e+09
##  Los Angeles:  87   Los Angeles Lakers :  87   3rd Qu.:1.611e+09
##  Utah       :  87   Utah Jazz          :  87   Max.   :1.611e+09
##  (Other)    :1947   (Other)            :1947
##     GAME_ID               NICKNAME           STATE            YEAR_FOUNDED
##  Min.   :2.180e+07   Grizzlies:  88   California: 257   Min.   :1946
##  1st Qu.:2.180e+07   Hawks    :  88   Texas     : 252   1st Qu.:1949
##  Median :2.180e+07   Heat     :  88   Florida   : 173   Median :1970
##  Mean   :1.281e+08   Cavaliers:  87   New York  : 170   Mean   :1970
##  3rd Qu.:2.180e+07   Jazz     :  87   Atlanta   :  88   3rd Qu.:1980
##  Max.   :1.622e+09   Lakers   :  87   Tennessee :  88   Max.   :2002
##                      (Other)  :1947   (Other)   :1444
##    SEASON_ID     TEAM_ABBREVIATION       GAME_DATE             MATCHUP
##  Min.   :22018   ATL    :  88      11/21/2018:  26   MEM @ UTA  :   4
##  1st Qu.:22018   MEM    :  88      11/23/2018:  26   UTA vs. MEM:   4
##  Median :22018   MIA    :  88      4/7/2019  :  26   ATL @ CHI  :   3
##  Mean   :22018   CLE    :  87      12/19/2018:  24   ATL @ IND  :   3
##  3rd Qu.:22018   LAL    :  87      2/23/2019 :  24   ATL vs. NYK:   3
##  Max.   :22018   UTA    :  87      4/5/2019  :  24   BKN @ IND  :   3
##                  (Other):1947      (Other)   :2322   (Other)    :2452
##  WL            MIN             PTS            PTSAGN          FGM
##   :   0   Min.   :197.0   Min.   : 53.0   Min.   : 53.0   Min.   :17.00
##  L:1236   1st Qu.:239.0   1st Qu.:100.0   1st Qu.:100.0   1st Qu.:37.00
##  W:1236   Median :240.0   Median :110.0   Median :110.0   Median :40.00
##           Mean   :238.9   Mean   :109.1   Mean   :109.1   Mean   :40.34
##           3rd Qu.:241.0   3rd Qu.:118.0   3rd Qu.:118.0   3rd Qu.:44.00
##           Max.   :341.0   Max.   :168.0   Max.   :168.0   Max.   :61.00
##
##       FGA            FG_PCT           FG3M            FG3A
##  Min.   : 55.00   Min.   :0.2620   Min.   : 2.00   Min.   :13.00
##  1st Qu.: 83.00   1st Qu.:0.4208   1st Qu.: 9.00   1st Qu.:27.00
##  Median : 88.00   Median :0.4580   Median :11.00   Median :31.50
##  Mean   : 88.14   Mean   :0.4581   Mean   :11.19   Mean   :31.83
##  3rd Qu.: 93.00   3rd Qu.:0.4940   3rd Qu.:14.00   3rd Qu.:36.00
##  Max.   :123.00   Max.   :0.6490   Max.   :27.00   Max.   :70.00
##
##     FG3_PCT          FTM             FTA          FT_PCT
##  Min.   :0.0800   Min.   : 2.00   Min.   : 4.0   Min.   :0.2630
##  1st Qu.:0.2930   1st Qu.:13.00   1st Qu.:18.0   1st Qu.:0.7000
##  Median :0.3500   Median :17.00   Median :22.0   Median :0.7670
##  Mean   :0.3511   Mean   :17.21   Mean   :22.6   Mean   :0.7626
##  3rd Qu.:0.4070   3rd Qu.:21.00   3rd Qu.:27.0   3rd Qu.:0.8330
##  Max.   :0.8420   Max.   :44.00   Max.   :54.0   Max.   :1.0000
##
##       OREB            DREB            REB             AST
##  Min.   : 1.00   Min.   :17.00   Min.   :22.00   Min.   : 7.00
##  1st Qu.: 8.00   1st Qu.:30.00   1st Qu.:40.00   1st Qu.:20.00
##  Median :10.00   Median :34.00   Median :44.00   Median :24.00
##  Mean   :10.32   Mean   :34.41   Mean   :44.73   Mean   :24.05
##  3rd Qu.:13.00   3rd Qu.:38.00   3rd Qu.:49.00   3rd Qu.:28.00
##  Max.   :26.00   Max.   :55.00   Max.   :71.00   Max.   :42.00
##
##       STL             BLK              TOV              PF
##  Min.   : 0.000   Min.   : 0.000   Min.   : 3.00   Min.   : 9.00
##  1st Qu.: 6.000   1st Qu.: 3.000   1st Qu.:11.00   1st Qu.:18.00
##  Median : 7.000   Median : 5.000   Median :13.00   Median :21.00
##  Mean   : 7.675   Mean   : 4.933   Mean   :13.64   Mean   :20.75
##  3rd Qu.:10.000   3rd Qu.: 6.000   3rd Qu.:16.00   3rd Qu.:23.00
##  Max.   :20.000   Max.   :19.000   Max.   :27.00   Max.   :37.00
```

```
##
##    PLUS_MINUS        home       not.paired     result
##  Min.   :-56.00000    :   0   Min.   :0     Min.   :0.0
##  1st Qu.: -9.00000   A:1236   1st Qu.:0     1st Qu.:0.0
##  Median :  0.00000   H:1236   Median :0     Median :0.5
##  Mean   : -0.01909            Mean   :0     Mean   :0.5
##  3rd Qu.:  9.00000            3rd Qu.:0     3rd Qu.:1.0
##  Max.   : 56.00000            Max.   :0     Max.   :1.0
##
```

```r
# For the Pythagorean Expectation we need only the result, points scored (PTS) and poi

NBAteams18   <-  NBAR18 %>% group_by(TEAM_NAME)%>%
            dplyr::summarise(result = sum(result),
                             PTS = sum(PTS),
                             PTSAGN = sum(PTSAGN)
                           )%>%
                           ungroup()
head(NBAteams18)
```

```
## # A tibble: 6 x 4
##   TEAM_NAME           result   PTS PTSAGN
##   <fct>                <dbl> <int>  <int>
## 1 Atlanta Hawks           30  9742  10306
## 2 Boston Celtics          53  9489   9082
## 3 Brooklyn Nets           42  9375   9443
## 4 Charlotte Hornets       42  9290   9359
## 5 Chicago Bulls           24  8783   9467
## 6 Cleveland Cavaliers     24  8976   9697
```

```r
tail(NBAteams18)
```

```
## # A tibble: 6 x 4
##   TEAM_NAME              result   PTS PTSAGN
##   <fct>                   <dbl> <int>  <int>
## 1 Portland Trail Blazers     57  9581   9167
## 2 Sacramento Kings           42  9445   9521
## 3 San Antonio Spurs          49  9366   9305
## 4 Toronto Raptors            58  9631   9211
## 5 Utah Jazz                  52  9479   9069
## 6 Washington Wizards         32  9449   9672
```

```r
# So now we can create the value for win percentage for each team in the 82 game seaso

NBAteams18[,'wpc'] = NBAteams18[,'result']/82
NBAteams18[,'pyth'] = NBAteams18[,'PTS']**2/(NBAteams18[,'PTS']**2 + NBAteams18[,'PTSAGN
head(NBAteams18)
```

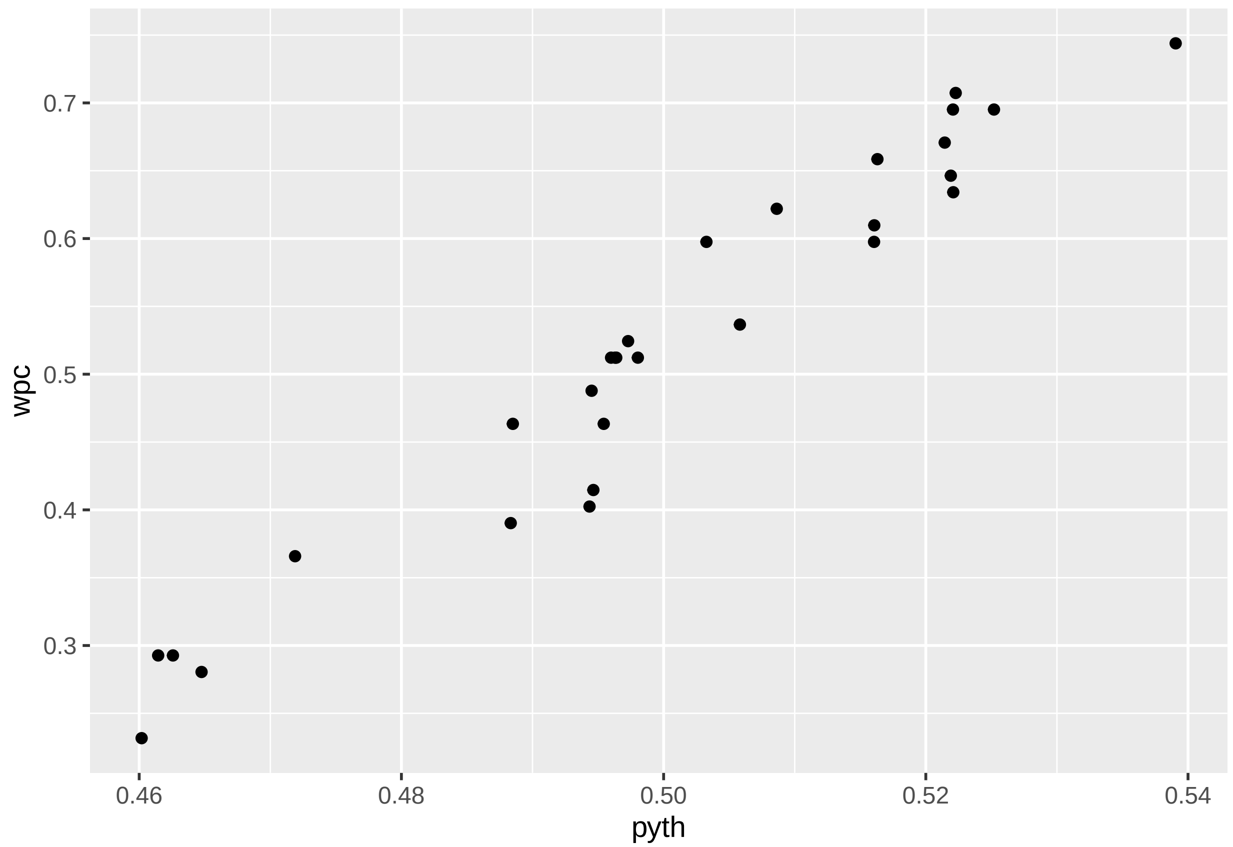
```
## # A tibble: 6 x 6
##   TEAM_NAME           result   PTS PTSAGN   wpc  pyth
##   <fct>                <dbl> <int>  <int> <dbl> <dbl>
## 1 Atlanta Hawks           30  9742  10306 0.366 0.472
## 2 Boston Celtics          53  9489   9082 0.646 0.522
## 3 Brooklyn Nets           42  9375   9443 0.512 0.496
## 4 Charlotte Hornets       42  9290   9359 0.512 0.496
## 5 Chicago Bulls           24  8783   9467 0.293 0.463
## 6 Cleveland Cavaliers     24  8976   9697 0.293 0.461
```

```
tail(NBAteams18)
```

```
## # A tibble: 6 x 6
##   TEAM_NAME              result   PTS PTSAGN   wpc  pyth
##   <fct>                   <dbl> <int>  <int> <dbl> <dbl>
## 1 Portland Trail Blazers     57  9581   9167 0.695 0.522
## 2 Sacramento Kings           42  9445   9521 0.512 0.496
## 3 San Antonio Spurs          49  9366   9305 0.598 0.503
## 4 Toronto Raptors            58  9631   9211 0.707 0.522
## 5 Utah Jazz                  52  9479   9069 0.634 0.522
## 6 Washington Wizards         32  9449   9672 0.390 0.488
```

```
# We now plot the data. Our results look very similar to the MLB case.

ggplot(data = NBAteams18,aes(x = pyth,y = wpc )) + geom_point()
```



Self test

run ggplot again, but this time write y= W instead of y= wpc. What do you find? Does it make a difference?

```
# Finally we generate a regression.

pyth_lm = lm(formula = 'wpc ~ pyth', data = NBAteams18)
pyth_lm %>% summary()
```

```
##
## Call:
## lm(formula = "wpc ~ pyth", data = NBAteams18)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08010 -0.02624  0.01040  0.02256  0.05657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.7572     0.1551  -17.77   <2e-16 ***
## pyth          6.5536     0.3100   21.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03453 on 27 degrees of freedom
## Multiple R-squared:  0.943,  Adjusted R-squared:  0.9409
## F-statistic:   447 on 1 and 27 DF,  p-value: < 2.2e-16
```

Self test

Run the regression above but instead write 'wpc ~ result' instead of 'wpc ~ result' in the line starting pyth_lm. What difference does this make?

Conclusion

We have found that the Pythagorean model fits the NBA data in roughly same way as it fits the MLB data. Let's now look at fourth example: English Premier League soccer.