

# STAT 428 Final

Ziqin Xiong, *zxiong8*

21 April, 2019

## 1. MLB game logs cleaning

```
setwd('~/.courserel/STAT 428/final')
# gl18 <- read.csv("GL2018.TXT",header = F)
# head(gl18)
# unique(gl18$V4)
# unique(gl18[,c(5,8)])
# nrow(subset(gl18,V5=='AL'&V8=='NL'))
# nrow(subset(gl18,V161!='Y'))

# data files were downloaded from
# https://www.retrosheet.org/gamelogs/
# The information used here was obtained free of
# charge from and is copyrighted by Retrosheet. Interested
# parties may contact Retrosheet at "www.retrosheet.org".
gl16 <- read.csv("GL2016.TXT",header = F)[,c(4,5,10,7,8,11)]
gl17 <- read.csv("GL2017.TXT",header = F)[,c(4,5,10,7,8,11)]
gl18 <- read.csv("GL2018.TXT",header = F)[,c(4,5,10,7,8,11)]
#simplified game logs from 2016 to 2018
gl <- rbind(gl16,gl17,gl18)
colnames(gl) <- c('vteam','vleague','vscore','hteam','hleague','hscore')

win <- lose <- draw <- matrix(0,30,30,dimnames =
  list(paste0(levels(gl$vteam),'v'),paste0(levels(gl$hteam),'h')))
for(i in 1:nrow(gl)){
  if(gl[i,3]>gl[i,6])
    win[gl[i,3],gl[i,6]] = win[gl[i,3],gl[i,6]]+1
  else if(gl[i,3]<gl[i,6])
    lose[gl[i,3],gl[i,6]] = lose[gl[i,3],gl[i,6]]+1
  else
    draw[gl[i,3],gl[i,6]] = draw[gl[i,3],gl[i,6]]+1
}
total <- win+lose+draw
#winning probability of visiting teams, where rows are visiting teams and
#cols are home teams.
#For example (ANAv,CHAh) means the avg probability of ANA winning CHA as a
#visiting team. This also means the avg probability of CHA losing or drawing
#ANA as a home team (P(lose/draw) = 1-P(win))
#NaN means there's no game records for 2 teams
winprob.v <- win/total
winprob.v[1:8,1:8]
```

	ANAh	ARiH	ATLh	BALh	BOSh	CHAh	CHNh	CINh
ANAv	NaN	0.00000	NaN	0.55556	0.33333	0.54545	0.00000	NaN
ARiV	0.50000	NaN	0.55556	0.00000	0.00000	NaN	0.50000	0.44444
ATLv	0.33333	0.54545	NaN	NaN	0.40000	0.66667	0.50000	0.50000

```
## BALv 0.33333      NaN 0.66667      NaN 0.50000 0.45455      NaN 0.66667
## BOSv 0.60000      NaN 1.00000 0.75862      NaN 0.60000      NaN 1.00000
## CHAv 0.20000 0.00000      NaN 0.40000 0.45455      NaN 0.28571 0.33333
## CHNv 1.00000 0.70000 0.77778 1.00000 0.33333 0.57143      NaN 0.57143
## CINv 0.00000 0.44444 0.60000      NaN      NaN      NaN 0.24138      NaN

# avg winning prob not considering visiting & home
alltotal <- total*upper.tri(total) + t(total)*upper.tri(total)
alltotal <- alltotal+t(alltotal)
allwin <- win*upper.tri(win) + t(lose)*upper.tri(win) +
  win*lower.tri(win) + t(lose)*lower.tri(lose)
winprob <- allwin/alltotal
dimnames(alltotal) <- dimnames(winprob) <- list(levels(gl$vteam),levels(gl$vteam))
#total matches between 2 teams
alltotal[1:10,1:10]
```

```
##      ANA  ARI  ATL  BAL  BOS  CHA  CHN  CIN  CLE  COL
## ANA    0   4   3  18  19  21   4   3  19   4
## ARI    4   0  20   3   3   3  20  18   3  57
## ATL    3  20   0   3  10   3  19  20   3  21
## BAL   18   3   3   0  57  21   3   3  20   3
## BOS   19   3  10  57   0  21   3   3  20   3
## CHA   21   3   3  21  21   0  14   3  57   3
## CHN    4  20  19   3   3  14   0  57   4  19
## CIN    3  18  20   3   3   3  57   0  14  20
## CLE   19   3   3  20  20  57   4  14   0   4
## COL    4  57  21   3   3   3  19  20   4   0
```

```
#avg winning prob not considering visiting & home
#For example (ANA,CHA) means the avg probability of ANA winning CHA
#This matrix could be used directly in sim_tournament_initial.R
winprob[1:9,1:9]
```

```
##      ANA  ARI    ATL    BAL    BOS    CHA    CHN    CIN    CLE
## ANA      NaN 0.25 0.66667 0.61111 0.36842 0.66667 0.00000 1.00000 0.21053
## ARI 0.75000  NaN 0.50000 0.00000 0.00000 1.00000 0.40000 0.50000 1.00000
## ATL 0.33333 0.50    NaN 0.33333 0.20000 0.66667 0.36842 0.45000 0.00000
## BAL 0.38889 1.00 0.66667    NaN 0.36842 0.52381 0.00000 0.66667 0.40000
## BOS 0.63158 1.00 0.80000 0.63158    NaN 0.57143 0.66667 1.00000 0.55000
## CHA 0.33333 0.00 0.33333 0.47619 0.42857    NaN 0.35714 0.33333 0.33333
## CHN 1.00000 0.60 0.63158 1.00000 0.33333 0.64286    NaN 0.66667 0.25000
## CIN 0.00000 0.50 0.55000 0.33333 0.00000 0.66667 0.33333    NaN 0.28571
## CLE 0.78947 0.00 1.00000 0.60000 0.45000 0.66667 0.75000 0.71429    NaN
```

```
#names of teams
levels(gl$vteam)
```

```
## [1] "ANA" "ARI" "ATL" "BAL" "BOS" "CHA" "CHN" "CIN" "CLE" "COL" "DET" "HOU"
## [13] "KCA" "LAN" "MIA" "MIL" "MIN" "NYA" "NYN" "OAK" "PHI" "PIT" "SDN" "SEA"
## [25] "SFN" "SLN" "TBA" "TEX" "TOR" "WAS"
```