基于因子分析和逻辑回归模型预测 NBA 总冠军和常规赛 MVP

May 21, 2017

Abstract

体育赛事的定量分析已成为国际体育竞技分析的重要方法。本文正是基于这一背景，对 NBA 赛事进行数据分析，意图通过统计分析方法对 NBA 球队及球员的表现进行定量分析并进行预测。首先，本文基于探索性因子分析法从球队及球员的多项能力指标提取出三大公共因子：进攻组织能力；有效得分能力；负面影响能力。其次，本文基于多年赛事数据对 NBA 热门球队及球员表现进行定性分析。最后，本文采用逻辑回归法对 2017 年 NBA 赛季总冠军和常规赛 MVP 进行预测。

【关键词】NBA 数据分析探索性因子分析逻辑回归

Contents

1 引言

NBA，即全美职业篮球大联盟，创办于 1946 年，已有 71 年的历史，现已成为全球范围最职业化、最市场化的大联盟之一。近年来，NBA 在中国的发展受到人们越来越多的关注，其体育文化价值、商业价值等得到了充分的显现。NBA 基于科学细致的选秀制度、转会制度和限薪制度等制衡体系保障各球队的实力较平均，使得没有哪一支球队有绝对的把握能战胜另一支球队，比赛的胜负往往充满悬念，正如 NBA 的口号一样“Where Amazing Happens”。球赛越是激烈，结果悬念越大，球迷就越想预测球赛的结果。然而，球迷对球赛结果的预测基本上都是基于主观推断，有时还受个人对球队或运动员偏好的影响，预测的科学性、准确性往往较差。

如今人们对赛季总冠军和常规赛 MVP 的预测也逐渐成为一大关注点。在当今大数据时代的背景下，本文尝试着利用 R 语言软件，在前人研究的基础上，试图通过各种统计分析方法对其进行较深入地研究，目的在于通过对 NBA 球队的历史球队战绩及往年常规赛 MVP 的影响因素进行建模分析，采用逻辑回归模型进行预测，从而预测 2017 年本赛季的总冠军球队以及常规赛 MVP 得主。

本文分为五个部分，分别是 2）文献综述；3）数据处理与变量提取 4）描述统计分析 5）逻辑回归预测 5）结论。

2 文献综述

NBA 在全球的极大影响力，使国内、外有不少文献对其进行过较深入和全面的研究。国外，Chatterjee、Campbell 和 Wiseman(1994) 对 NBA 所有球队一个赛季的数据建立统计模型，对球队胜率进行回归分析，发现比赛得分、罚球、篮板和失误在统计上是显著的，并且，回归系数在各年数据之间都相对稳定；Hausman 和 Leonard(1997) 使用计量经济学方法对 NBA 赛事明星出场率与其电视收视率、门票收入等进行了相关性研究，得出了正相关的结论；Gandar、Zuber 和 Lamb(2001) 等人对 NBA 博彩市场的主客场优势进行了分析；Leeds 和 Allmen(2003) 在其著作《体育经济学》中对美国职业体育联盟的制衡机制进行了较深入的探讨；Mizak、Stair 和 Rossi(2004) 使用胜率标准差、HHI 等指标衡量了各大联盟的竞争性平衡，并指出所使用指标的优缺点；此外，国外学者还对 NBA 运动员、裁判员是否存在种族歧视、工资差异等各方面进行了研究。

国内的相关研究文献并不多，其研究特色归纳起来可分为 3 类：1) 从市场营销的角度，对 NBA 的市场价值、品牌文化传播和在中国的市场营销情况等方面进行剖析；2) 从制度经济学角度，对 NBA 的人力制衡、收益制衡及权力制衡 3 项机制的功能及相关制度的运行原理展开深入研究；3) 从 NBA 比赛本身的技术角度，如球赛中冲抢技术、不同位置的运动员、运动员的攻防能力、球赛赛程安排等方面进行分析研究，有关参考文献见刘素蓉等 (2009)、吴福珍和王晓军 (2009) 等。

另外目前国内外学者运用逻辑回归模型对 NBA 研究的文献并不多见。因此本文本着创新原则，尝试着运用逻辑回归模型对 NBA 球队战绩以及球员能力进行预测。

3 数据处理与变量提取

3.1 定义

探索性因子分析（EFA）的目标是通过发现发掘隐藏在数据下的一组较少的、更为基本的无法观测的变量，来解释一组可观测变量的相关性。这些虚拟的、无法观测的变量称为因子（每个因子被认为可解释多个观测变量间共有的方差，因此准确来说，它们应该成为公共因子）。

3.2 确定公因子数

首先判断需提取的公共因子数，因此笔者选取了 2006 年 -2017 年 30 支球队总共 360 条全数据以及 2007 年 -2017 年的全明星球员数据总共 727 条全数据，基于“场均得分”（PPG）、“进攻篮板”（OFFR）、“防守篮板”（DEFR）、“场均助攻”（APG）、“场均抢断”（SPG）、“场均盖帽”（BPG）、“场均失误”（TPG）、“场均犯规”（FPG）、“三分球命中率”（X3P.）、“罚球命中率”（FT.）以及“两分球命中率”（X2P.）这 11 个可观测变量，画出如下两个因子图形。该因子图形将会同时展示主成分和公共因子分析的结果。

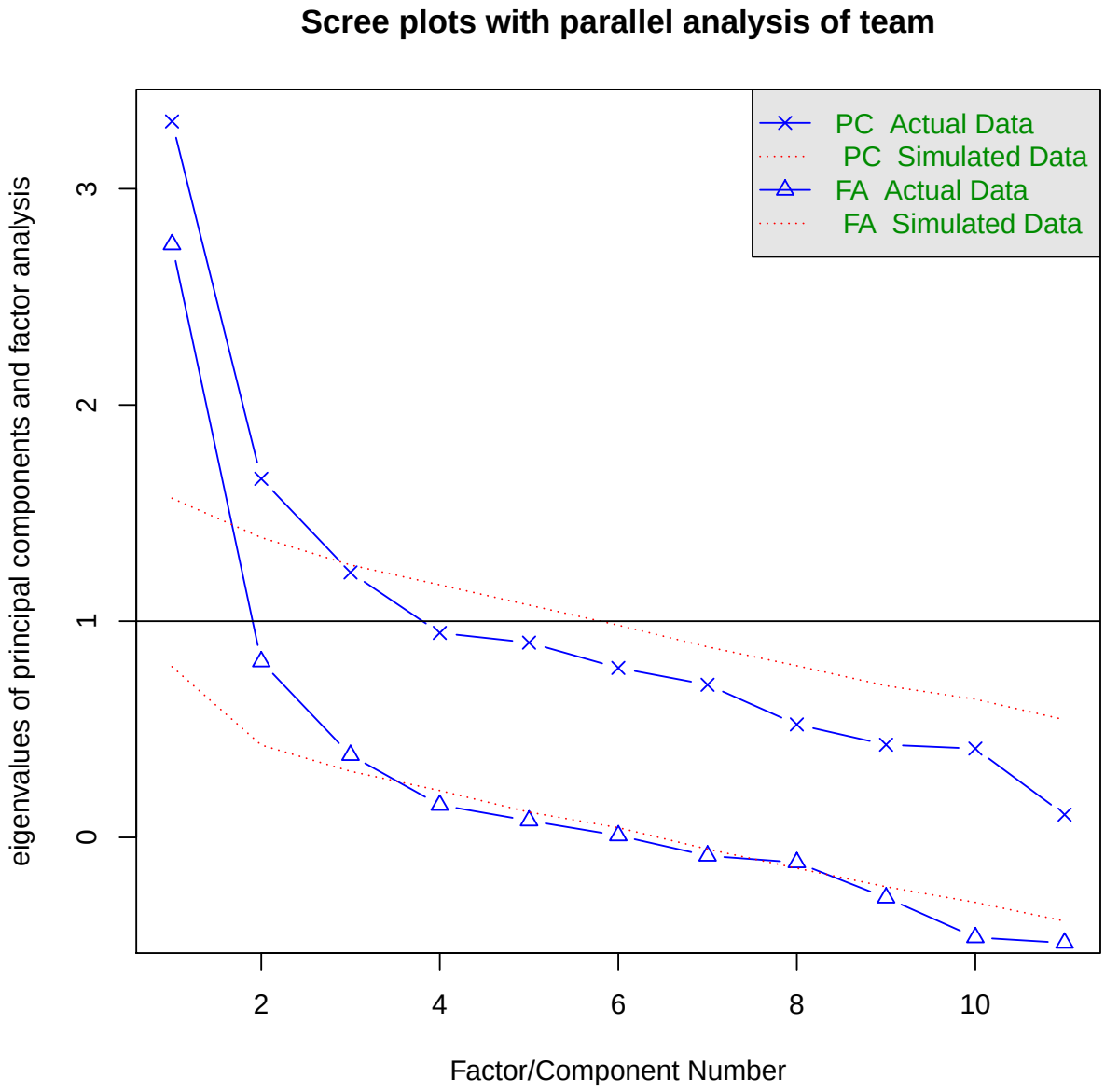
首先，数据预处理数据预处理

```
GT <- read.csv(file='file:///C:/Users/Administrator/Documents/TEAM.csv')
a<-cbind(GT[,5],GT[,6],GT[7],GT[,8],GT[,9],GT[,10],GT[,11],GT[,12],GT[,13],GT[,14],GT[,15])
a<-scale(a);colnames(a)<-names(GT)[c(-1,-2,-3,-4)]
team.cor <- cor(a) ;GP <- read.csv(file='file:///C:/Users/Administrator/Documents/PLAYER.csv')
b<-cbind(GP[,5],GP[,6],GP[,7],GP[,8],GP[,9],GP[,10],GP[,11],GP[,12],GP[,13],GP[,14],GP[,15])
b <- scale(b);colnames(b)<-names(GP)[c(-1,-2,-3,-4)]
player.cor <- cor(b)
```

判断需要提取的公因子数目，并画图

```
library(psych)
fa.parallel(team.cor,fa="both",main="Scree plots with parallel analysis of team")#Graph1

## Warning in fa.parallel(team.cor, fa = "both", main = "Scree plots with parallel
analysis of team"): It seems as if you are using a correlation matrix, but have not
specified the number of cases. The number of subjects is arbitrarily set to be 100
## The estimated weights for the factor scores are probably incorrect. Try a different
factor extraction method.
```



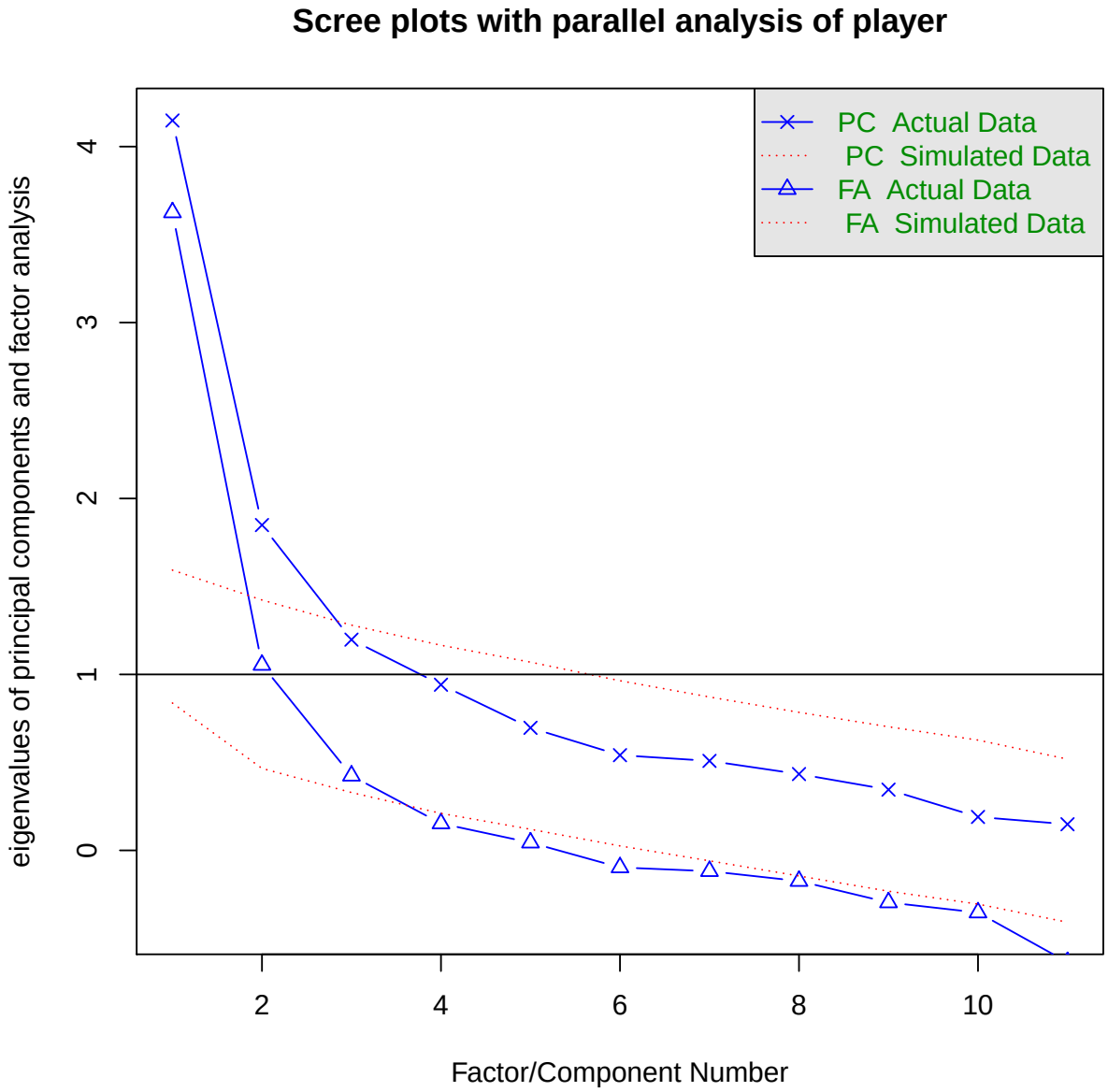

```
## Parallel analysis suggests that the number of factors =  3  and the number of components =  2
```

基于球队数据的分析，从图一可知，通过碎石检验和平行分析法得出建议的公共因子数为 3。

```
fa.parallel(player.cor,fa="both", main="Scree plots with parallel analysis of player")#G2

## Warning in fa.parallel(player.cor, fa = "both", main = "Scree plots with parallel
analysis of player"): It seems as if you are using a correlation matrix, but have not
specified the number of cases. The number of subjects is arbitrarily set to be 100
## The estimated weights for the factor scores are probably incorrect.  Try a different
factor extraction method.
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate = rotate, :  A
loading greater than abs(1) was detected.  Examine the loadings carefully.
## The estimated weights for the factor scores are probably incorrect.  Try a different
factor extraction method.
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate = rotate, : An
ultra-Heywood case was detected.  Examine the results carefully
```



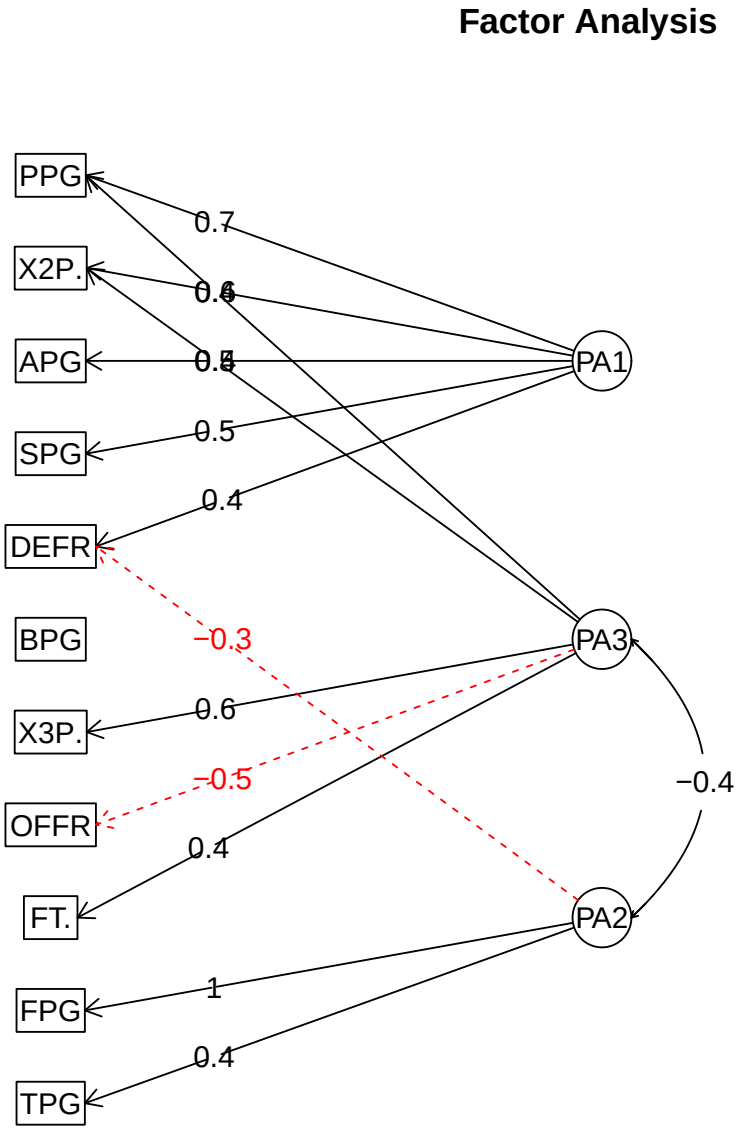

```
## Parallel analysis suggests that the number of factors =  3  and the number of components =  2
```

基于球员的数据分析，从图二可知，通过碎石检验和平行分析法得出建议的公共因子数目也同样为 3。

3.3 旋转因子提取变量

这里，笔者运用斜交旋转因子。对于斜交旋转，因子分析会考虑三个矩阵：因子结构矩阵、因子模式矩阵和因子关联矩阵。这里，笔者基于 R 语言，分别对球队和球员数据构造三因子斜交结果图，结果如下图所示。

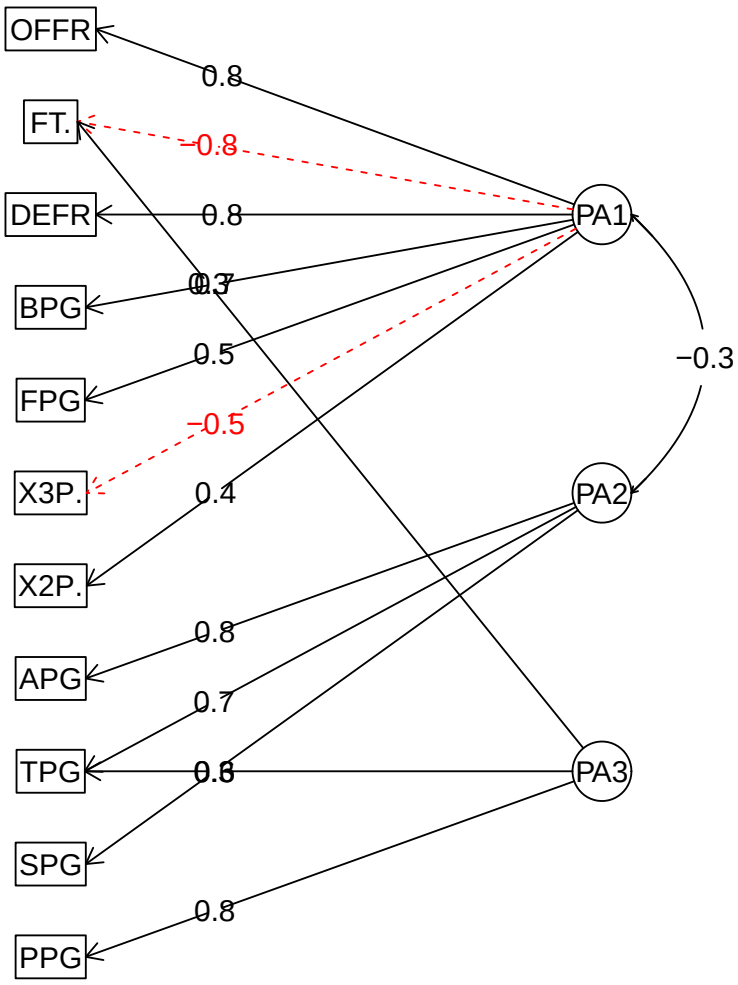
```
library(GPArotation)
fa_t.promax<-fa(team.cor,nfactors=3,rotate="promax",fm="pa")
#图像显示
fa.diagram(fa_t.promax, simple=FALSE)#G5
```



分析图三可发现，在第一个因子中，以场均得分（0.71），两分球命中率（0.62），场均助攻

（0.55），场均抢断（0.47），防守篮板（0.41）以及场均盖帽（0.28）最为显著。这些均为正指标。可见，第一个主因子是对球队进攻组织、防守反击能力的描述。在第二个因子中，三分球命中率（0.63），两分球命中率（0.42），罚球命中率（0.40），场均得分（0.39）以及进攻篮板（-0.54）最为显著。其中进攻篮板为负指标，其余为正指标。可见第二个因子是对球队有效得分能力的描述。进攻篮板为负因素可能原因为 NBA 进攻三秒规则限制了进攻球员在篮下活动及二次进攻成功率低，同时目前缺乏抢前场篮板后的对应战术。在第三个因子中，场均犯规（0.97），场均失误（0.44）及防守篮板（-0.31）最为显著。可见第三个因子是对球队负面因素的描述。其中防守篮板是稳住球队军心、重新组织进攻的重要指标，因此在此公共因子中充当负指标。

```
library(GPArotation)
fa_p.promax<-fa(player.cor,nfactors=3,rotate="promax",fm="pa")
fa.diagram(fa_p.promax, simple=FALSE)#G6
```



分析图四可发现，在第一个因子中，以进攻篮板（0.81），防守篮板（0.76），场均盖帽（0.68），场均犯规（0.53），两分球命中率（0.42），三分球命中率（-0.53）以及罚球命中率（-0.76）最为显著。其中三分球命中率、罚球命中率为负指标，其余为正指标。可见第一个因子是对中锋和大前锋能力的描述。这两个位置是内线防守的大闸，内线命中率惊人但同时三分球和罚球手感稍差。在第二个因子中，以场均助攻（0.85），场均抢断（0.61），场均失误（0.74）最为显著。这三个指标均为正指标。可见第二个因子是对控球后卫能力的描述，长时间的控球也导致了失误的增多。在第三个因子中，场均得分（0.75），罚球命中率（0.31），三分球命中率（0.20），场均失误（0.35）最为显著。这些均为正指标。可见第三个因子是对得分后卫、小前锋能力的描述。这两个位置是球队得分的基石，突破造成杀伤，同时投篮或者罚球手感柔和，但同时造成的失误也多。

4 描述性统计分析

4.1 总冠军候选球队综合实力的描述性分析

为了实现对今年总冠军球队的预测，同时出于简化分析的需要，这里笔者主要挑选了四支常规赛战绩靠前的球队，而这也是 ESPN 专家一致认为的夺冠四大热门。进一步地，我们基于 R 语言软件，对这四支球队 2017 年的球队战绩作出关于三个公共因子的雷达图，如下图所示。

球队数据处理

```
#team
pc1<-as.vector(nrow(a))
pc2<-as.vector(nrow(a))
pc3<-as.vector(nrow(a))
for(j in 1:nrow(a)){
    pc1[j] <- 0
    for(i in 1:11) pc1[j]<-pc1[j]+a[j,i]*(fa_t.promax$weights)[i,1]}
for(j in 1:nrow(a)){
    pc2[j] <- 0
    for(i in 1:11) pc2[j]<-pc2[j]+a[j,i]*(fa_t.promax$weights)[i,2]}
for(j in 1:nrow(a)){
    pc3[j] <- 0
   for(i in 1:11) pc3[j]<-pc3[j]+a[j,i]*(fa_t.promax$weights)[i,3]}
a <- cbind(a,pc1,pc2,pc3)
pca1 <- as.vector(nrow(a))
pca2 <- as.vector(nrow(a))
pca3 <- as.vector(nrow(a))
for(j in 1:nrow(a)){
    pca1[j] <- 0
 for(i in 1:11){
   pca1[j]<-pca1[j]+a[j,i]*(fa_t.promax$weights)[i,1]
}
}
for(j in 1:nrow(a)){
   pca2[j] <- 0
for(i in 1:11) {
pca2[j]<-pca2[j]+a[j,i]*(fa_t.promax$weights)[i,2]
 }
}
for(j in 1:nrow(a)){
pca3[j] <- 0
for(i in 1:11){
```

```
pca3[j]<- pca3[j]+a[j,i]*(fa_t.promax$weights)[i,3]
}
}
GT <- cbind(GT,pca1,pca2,pca3)
```

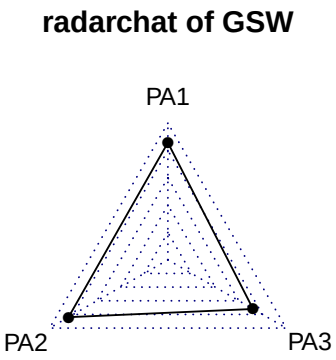
球员数据处理

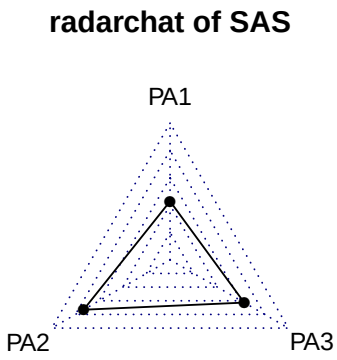
```
pcb1<-as.vector(nrow(b))
pcb2<-as.vector(nrow(b))
pcb3<-as.vector(nrow(b))
for(j in 1:nrow(b)){
   pcb1[j] <- 0
    for(i in 1:11){
    pcb1[j] <- pcb1[j]+b[j,i]*(fa_p.promax$weights)[i,1]
   }
}
for(j in 1:nrow(b)){
    pcb2[j] <- 0
    for(i in 1:11){
pcb2[j]<- pcb2[j]+b[j,i]*(fa_p.promax$weights)[i,2]
}
}
for(j in 1:nrow(b)){
   pcb3[j] <- 0
   for(i in 1:11) {
  pcb3[j]<-pcb3[j]+b[j,i]*(fa_p.promax$weights)[i,3]
  }
}
GP <- cbind(GP,pcb1,pcb2,pcb3)
```

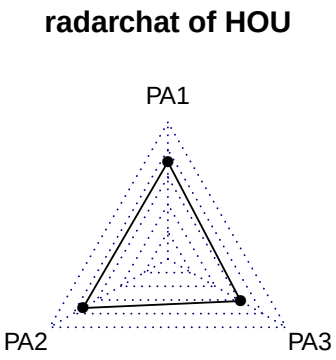
球队的雷达图

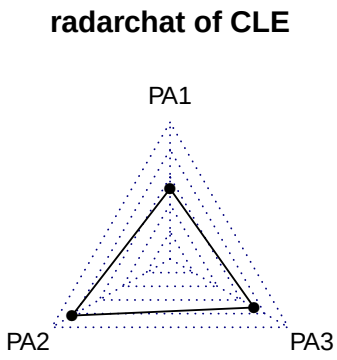
```
#1.雷达图
library(fmsb) #team
GT2<-GT[GT$Year==2017,]
opar<-par(no.readonly=TRUE)
par(mfrow=c(2,2))
maxmin<-data.frame(PA1=c(5,-2),PA2=c(3,-3),PA3=c(4,-3))
dat.A<-data.frame(PA1=GT2$pca1[GT2$TEAM=='gsw'],PA2=GT2$pca2[GT2$TEAM=='gsw'],PA3=GT2$pca2[GT2$TEA
dat.A2<-rbind(maxmin,dat.A)
radarchart(dat.A2,axistype=0,seg=5,centerzero=TRUE,title='radarchat of GSW')#Graph11
```

```
dat.A<-data.frame(PA1=GT2$pca1[GT2$TEAM=='sas'],PA2=GT2$pca2[GT2$TEAM=='sas'],PA3=GT2$pca2[GT2$TEA
dat.A2<-rbind(maxmin,dat.A)
radarchart(dat.A2,axistype=0,seg=5,centerzero=TRUE,title='radarchat of SAS')#Graph12
dat.A<-data.frame(PA1=GT2$pca1[GT2$TEAM=='hou'],PA2=GT2$pca2[GT2$TEAM=='hou'],PA3=GT2$pca2[GT2$TEA
dat.A2<-rbind(maxmin,dat.A)
radarchart(dat.A2, axistype=0, seg=5,centerzero = TRUE,title ='radarchat of HOU')#Graph13
dat.A<-data.frame(PA1=GT2$pca1[GT2$TEAM=='cle'],PA2=GT2$pca2[GT2$TEAM=='cle'],PA3=GT2$pca2[GT2$TEA
dat.A2<-rbind(maxmin,dat.A)
radarchart(dat.A2,axistype=0,seg=5,centerzero=TRUE,title ='radarchat of CLE')#Graph14
```









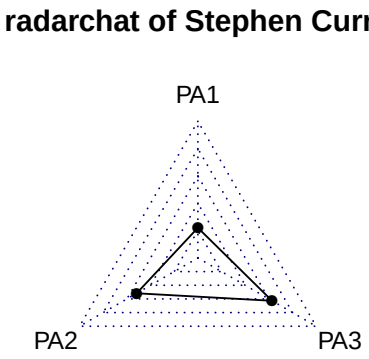
```
par(opar)
```

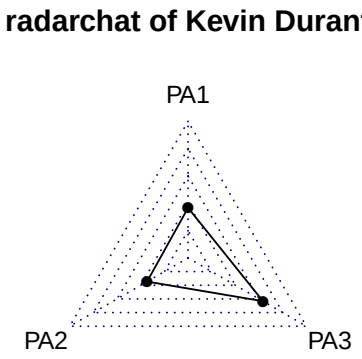
从上图我们可以看出，整体上，这四只夺冠热门球队受第一个因子和第二个因子这两个正面因素的影响很突出，而对负面因素第三个因子的控制都很突出，第三个因子对这四支球队的影响都很小。而从四支球队的内部比较方面来看，勇士队（GSW）受第一个因子和第二个因子的影响最突出，说明勇士队常规赛的球队进攻火力、防守反击转换为有效得分的能力很强，从侧面可以推断出勇士队的综合实力在这四支球队里占优。其次是火箭队（HOU），火箭队受第一个因子的影响也很强，这反映了火箭队在进攻组织方面做得不错。而从今年火箭队的常规赛表现来看，主打三分球战术，进攻火力凶猛而防守偏弱。这也正好印证了我们的推断。再者从雷达图看，东部球队骑士队（CLE）的第一因子和第二因子的表现也不错。因此我们由此预测，有极大可能，勇士队和骑士队分别杀出西部和东部重围会师总决赛，最后勇士队夺得总冠军。当然，这只是简单的根据图表描述，进一步地用数据预测见本文后面的逻辑回归预测。

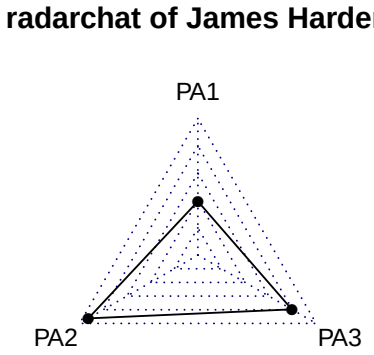
4.2 MVP 候选人综合实力的描述性统计分析

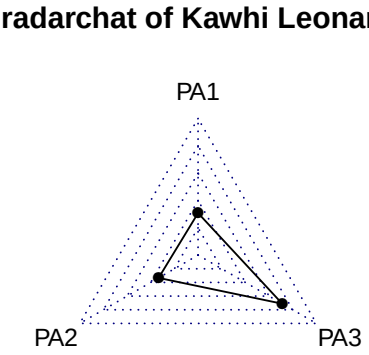
为了实现对今年常规赛 MVP 的预测，这里笔者主要挑选了今年入选全明星同时数据方面较为突出的几名球员。因为是否入选赛季全明星是衡量一个球员当下表现的一个天然的标记，同时 MVP 的决出又是媒体记者根据球员常规赛的综合表现投票得出。同样地，我们先做一些描述分析，以帮助我们更好地理解数据。因此基于 R 语言软件，对以下 7 名全明星球员的能力作出关于三个公共因子的雷达图，如下图所示。

```
#player
GP2 <-GP[GP$Year==2017,]
opar<-par(no.readonly=TRUE)
par(mfrow=c(2,2))
maxmin<-data.frame(PA1=c(3,-2),PA2=c(3,-2),PA3=c(3,-3))
dat.A<-data.frame(PA1=GP2$pcb1[GP2$Player=='Stephen Curry'],PA2=GP2$pcb2[GP2$Player=='Stephen Curry
dat.A2<-rbind(maxmin,dat.A)
radarchart(dat.A2, axistype=0, seg=5,centerzero = TRUE,title ='radarchat of Stephen Curry')#G15
dat.A<-data.frame(PA1=GP2$pcb1[GP2$Player=='Kevin Durant'],PA2=GP2$pcb2[GP2$Player=='Kevin Durant']
dat.A2 <- rbind(maxmin,dat.A)
radarchart(dat.A2, axistype=0, seg=5,centerzero = TRUE,title ='radarchat of Kevin Durant')#G16
dat.A<-data.frame(PA1=GP2$pcb1[GP2$Player=='James Harden'],PA2=GP2$pcb2[GP2$Player=='James Harden']
dat.A2 <- rbind(maxmin,dat.A)
radarchart(dat.A2,axistype=0,seg=5,centerzero = TRUE,title ='radarchat of James Harden')#G17
dat.A<-data.frame(PA1=GP2$pcb1[GP2$Player=='Kawhi Leonard'],PA2=GP2$pcb2[GP2$Player=='Kawhi Leonard
dat.A2 <- rbind(maxmin,dat.A)
radarchart(dat.A2, axistype=0, seg=5,centerzero = TRUE,title ='radarchat of Kawhi Leonard')#G18
```





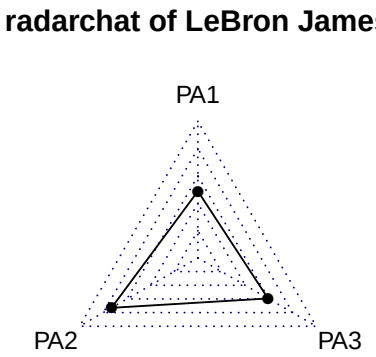


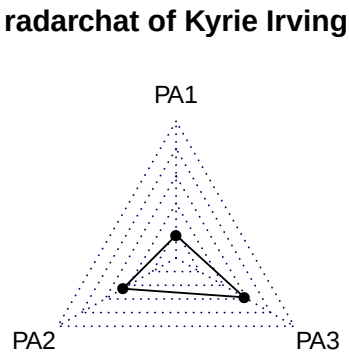


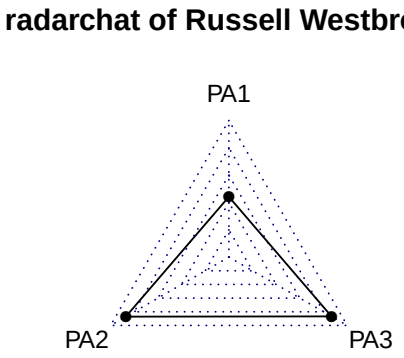

```
par(opar)
```

```
par(mfrow=c(2,2))
dat.A<-data.frame(PA1=GP2$pcb1[GP2$Player=='LeBron James'],PA2=GP2$pcb2[GP2$Player=='LeBron James']
dat.A2 <- rbind(maxmin,dat.A)
radarchart(dat.A2, axistype=0, seg=5,centerzero = TRUE,title ='radarchat of LeBron James')#G19
dat.A<-data.frame(PA1=GP2$pcb1[GP2$Player=='Kyrie Irving'],PA2=GP2$pcb2[GP2$Player=='Kyrie Irving']
dat.A2 <- rbind(maxmin,dat.A)
radarchart(dat.A2, axistype=0, seg=5,centerzero = TRUE,title ='radarchat of Kyrie Irving')#G20
dat.A<-data.frame(PA1=GP2$pcb1[GP2$Player=='Russell Westbrook'],PA2=GP2$pcb2[GP2$Player=='Russell W
```

```
dat.A2<-rbind(maxmin,dat.A)
radarchart(dat.A2, axistype=0, seg=5,centerzero = TRUE,title ='radarchat of Russell Westbrook')#G21 p
```







从以上雷达图我们可以看出，Russell Westbrook 、James Harden、LeBron James 这三名球员的能力值受这三个公共因子的影响较为突出。其中，Russell Westbrook 在前两个因子方面的表现在 7 名全明星球员中最为强势。这体现为其冲抢篮板、助攻以及得分的能力突出。本赛季 Russell Westbrook 场均三双的历史伟大表现也正很好地印证了这一点。同时，James Harden 以及 LeBron James 这两名球员也是本赛季常规赛 MVP 的有力竞争者。因此，我们可以初步预测 Russell Westbrook 最终夺得 MVP 的概率占优。具体的数据分析详见本文的逻辑回归预测部分。

5 逻辑回归预测

5.1 逻辑回归模型的构建

逻辑回归（logistic regression）是一种可以用来分类的常用统计分析方法，并且可以得到概率型的预测结果，属于一种概率型非线性回归。

考虑具有 n 个变量的向量 $x=(x_1,x_2,\cdots,x_3)$，设条件概率 $P(Y=1|X)=p$ 为根据观测量相对于某事件发生的概率，则逻辑回归模型可表示为

$$P(Y=1|X)=\pi(x)=\frac{1}{1+e^{-g(x)}}$$

其中，$g(x)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n$ ，β_n 为截距项，$\beta=(\beta_1,\beta_2,\cdots,\beta_n)$ 为自变量的回归系数。显然，$\pi(x)$ 的值域为 [0,1]，因此我们可以根据其取值来估计因变量 Y=1 时发生的概率。对逻辑回归模型的参数进行估计，通常采用极大似然函数法。设 y 是 0—1 型变量，m 个观测值为 $\{y_1,y_2,\cdots,y_n\}$ ，于是 m 个观测值的似然函数为

$$L(\beta)=\prod_{i=1}^{m}p(y_i)=\prod_{i=1}^{m}p(y_i)^{y_i}[1-p(x)]^{1-y}$$

对上式两边求自然对数，可得对数似然函数

$\ln L=\sum_{i=1}^{m}[y_i(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_nx_{in})-\ln(1+e^{\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_nx_{in}})]$

最大似然估计就是选取 $\beta_0,\beta_1,\beta_2,\cdots,\beta_n$ 的估计值，使得 lnL 的值最大化。对上式进行求导，应用牛顿—拉斐森（Newton-Raphson）方法进行迭代求解，即可获取模型的截距项和回归系数，将求得的参数代入最上式即可建立逻辑回归的预测模型。

5.2 样本数据的形成

球队数据方面，本文选取了 2006 年 -2016 年 30 支球队总共 330 条全数据形成训练样本集，2017 年 30 支球队的 30 条全数据作为测试样本集。球员数据方面，本文选取了 2007 年 -2016 年的全明星球员数据总共 702 条全数据形成训练样本集，2017 年的全明星球员数据共 25 条全数据形成测试样本集。

5.3 逻辑回归预测

5.3.1 关于球队的逻辑回归

首先，我们的目的是为了预测今年季后赛的总冠军归属。为此，笔者分为两个步骤进行。第一步，运用逻辑回归模型预测今年进入季后季后赛的前十支球队，为此将历年全联盟常规赛排名前十记为 1，其余记为 0，此为因变量。自变量选取前文得到的三个公共因子。得到的逻辑回归模型结果如下：

	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-0.7332	0.1355	-5.41	6.26e-08	***
pca1	0.5944	0.1802	3.299	0.00097	***
pca2	1.0167	0.2159	4.710	2.48e-06	***
pca3	-0.3387	0.1606	-2.109	0.03496	*

因此，将 2017 年 30 支球队的 30 条全数据代入得到的逻辑回归模型，可得：全联盟球队进入季后赛的概率排名前十位为：

因此，将 2017 年 30 支球队的 30 条全数据代入得到的逻辑回归模型，可得：全联盟球队夺得季后赛总冠军的概率排名前十位为：

rank	name	p
1	GSW	0.5534393
2	HOU	0.2746141
3	CLE	0.2274549
4	LAC	0.1701985
5	BOS	0.1619623
6	DEN	0.1534857
7	WAS	0.1226956
8	SAS	0.1144878
9	CHA	0.1102071
10	MIL	0.1043236

由此可以看出勇士队夺得最终总冠军的概率最大。

5.3.2 关于球员的逻辑回归

同样地，我们的目的是为了预测今年常规赛的 MVP 归属，同样分为两个步骤进行。第一步，运用逻辑回归模型预测今年进入全明星首发阵容的十名球员，为此将历年进入全明星首发阵容的十名球员记为 1，其余全明星球员记为 0，此为因变量。自变量选取前文得到的三个公共因子。得到的逻辑回归模型结果如下：

	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-0.3616	0.1449	-2.496	0.01255	*
pcb1	0.5061	0.1686	3.002	0.00268	**
pcb2	0.5660	0.1756	3.223	0.00127	**
pcb3	0.8479	0.1838	4.613	3.96e-06	***

因此，将 2017 年的全明星球员数据共 25 条全数据代入得到的逻辑回归模型，可得：进入全明星首发阵容概率排名前十位的为：

rank	name	p
1	Russell Westbrook	0.9501681
2	James Harden	0.9313161
3	DeMarcus Cousins	0.8459819
4	LeBron James	0.7808314
5	Anthony Davis	0.7135958
6	John Wall	0.6475794
7	Giannis Antetokounmpo	0.6329114
8	Kawhi Leonard	0.5728605
9	Stephen Curry	0.5406733
10	Kevin Durant	0.5203161

第二步，接着运用逻辑回归模型预测今年常规赛的 MVP 得主。为此将历年夺得 MVP 的球员记为 1，其余全明星球员记为 0。自变量依然为前文得到的三个公共因子。但是因为三因素模型不显著，我们就把不显著的第一个参数去掉继续做回归。得到的逻辑回归模型结果如下：

	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-4.4294	0.7126	-6.216	5.11e-10	***
pcbZ2	1.0528	0.4712	2.234	0.0255	*
pcbZ3	1.8728	0.5380	3.481	0.0005	***

因此，将 2017 年的全明星球员数据共 25 条全数据代入得到的逻辑回归模型，可得：获得常规赛 MVP 概率排名前十位的为：

rank	name	p
1	Russell Westbrook	0.898522901
2	James Harden	0.83802170
3	DeMarcus Cousins	0.26391745
4	LeBron James	0.16160967
5	John Wall	0.11394340
6	Isaiah Thomas	0.09690138
7	Stephen Curry	0.08439225
8	Kawhi Leonard	0.08303184
9	Anthony Davis	0.06095351
10	Paul George	0.05089471

由上表我们可以看出，通过逻辑回归模型预测，Russell Westbrook 获得 MVP 的概率最大，James Harden 次之，与我们前文的描述性分析基本一致。而这两个球员现实中也确实打出了现象级表现的一个赛季，可见我们的模型预测还是基本合理的。

6 结论

6.1 勇士有望夺冠，威少哈登大热双子星

基于逻辑回归预测，勇士夺得今年联盟总冠军的概率是 55%，远高于排在第二的火箭（27%）和第三的骑士（22%）。而在球员预测方面，威少和哈登获得常规赛 MVP 的概率远高于其他球员（分别是 89%，82%），这两位也正是主流媒体预测的大热门球员。

6.2 NBA 球队赢得总冠军更看重攻击和防守的协调组合

基于逻辑回归，NBA 球队赢得总冠军的影响因素里，大前锋，中锋的能力和控球后卫能力的影响是显著的，而小前锋，得分后卫的影响并不显著。说明要想获得联盟总冠军，一方面需要确保球队的进攻以获得良好的战绩得以进入季后赛，另一方面需要球队注重防守和避免失误，才可以避免在激烈的季后赛中抱憾止步。

6.3 NBA 球员赢得常规赛 MVP 主要依赖有效得分和避免失误

基于逻辑回归，NBA 球员赢得常规赛 MVP 的影响因素里，有效得分能力和避免失误能力影响非常显著，而进攻能力并非显著。这一方面说明球员的得分和失误这些可量化指标的确在 MVP 评选中占据很重要的比重，另一方面也说明一味追求进攻次数而不注重得分效率和团队合作的球员并不能受到联盟的青睐。

7 参考文献

[1] 陈建宝, 肖林, 许世杰, 林炳灿. NBA 球队战绩影响因素的统计分析 [J]. 中国体育科技. 2010(06)

[2 张志谦. 浅谈 2006-2007 赛季 NBA 总决赛各项技术统计对比赛胜负的影响 [J]. 内蒙古体育科技. 2008(02)

[3] 米勤, 李可可, 张辉, 冯杰. NBA 替补队员攻防能力对比赛结果影响的研究 [J]. 首都体育学院学报. 2008(04)

[4] S CHATTERJEE,MR CAMPBELL,F WISEMAN.Take that jam!an analysis of winning percentage for NBA teams. ManagerialDecision Economics . 1994

[5] J MGANDAR,R AZUBER,R P LAMB.The home field advantagerevisited:A search for the bias in other sports betting markets. JEconomics Business . 2001

[6] 赵利庆. 影响 CBA 各队得分的回归分析 [J]. 北京体育大学学报. 2007(02)

[7] 俞庆生. 基于云平台的逻辑回归模型构建算法的设计与实现 [J]. 科技通报. 2013(06)