关于 NBA 球员正负值的讨论

3170102187　　　孙晨旭

简介

RPM, 即 NBA 球员的正负值，表示该球员在场上时球队净胜分的情况，是衡量球员综合表现得一个指标。本次作业选取 17 年 NBA 所有球员的数据，分析 RPM 与球员出场时间以及球员薪酬的关系。首先读入数据：

```r
nba <- read_csv("~/github/Rcourse_1/project/nba_2017_nba_players_with_salary.csv")
#extract from
#https://www.kaggle.com/dhamlett/nba-player-rpm-prediction-defense-vs-offense
nba<- nba %>% select(-1)
head(nba)
```

```
## # A tibble: 6 x 38
##       Rk PLAYER POSITION   AGE    MP    FG   FGA `FG%`  `3P` `3PA` `3P%`  `2P`
##    <dbl> <chr>  <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1      1 Russe~ PG          28  34.6  10.2  24   0.425   2.5   7.2 0.343   7.7
## 2      2 James~ PG          27  36.4   8.3  18.9 0.44    3.2   9.3 0.347   5.1
## 3      3 Isaia~ PG          27  33.8   9    19.4 0.463   3.2   8.5 0.379   5.8
## 4      4 Antho~ C           23  36.1  10.3  20.3 0.505   0.5   1.8 0.299   9.7
## 5      6 DeMar~ C           26  34.2   9    19.9 0.452   1.8   5     0.361   7.2
## 6      7 Damia~ PG          26  35.9   8.8  19.8 0.444   2.9   7.7 0.37    6
## # ... with 26 more variables: `2PA` <dbl>, `2P%` <dbl>, `eFG%` <dbl>, FT <dbl>,
## #   FTA <dbl>, `FT%` <dbl>, ORB <dbl>, DRB <dbl>, TRB <dbl>, AST <dbl>,
## #   STL <dbl>, BLK <dbl>, TOV <dbl>, PF <dbl>, POINTS <dbl>, TEAM <chr>,
## #   GP <dbl>, MPG <dbl>, ORPM <dbl>, DRPM <dbl>, RPM <dbl>, WINS_RPM <dbl>,
## #   PIE <dbl>, PACE <dbl>, W <dbl>, SALARY_MILLIONS <dbl>
```

其次检查数据完整情况，其中 na 类型只出现在两项命中率数据上，说明所有数据都完整，只是某些球员整个赛季没有三分和罚球，因此对应命中率为 na。

```
rowSums(apply(nba,1,is.na))
```

```
##             Rk          PLAYER        POSITION             AGE              MP
##              0               0               0               0               0
##             FG             FGA             FG%              3P             3PA
##              0               0               0               0               0
##            3P%              2P             2PA             2P%            eFG%
##             22               0               0               0               0
##             FT             FTA             FT%             ORB             DRB
##              0               0               5               0               0
##            TRB             AST             STL             BLK             TOV
##              0               0               0               0               0
##             PF          POINTS            TEAM              GP             MPG
##              0               0               0               0               0
##           ORPM            DRPM             RPM        WINS_RPM             PIE
##              0               0               0               0               0
##           PACE               W SALARY_MILLIONS
##              0               0               0
```

RPM 与出场时间

```
temp1 <- min(nba$WINS_RPM)
nba<-nba %>% mutate(RPM1 = WINS_RPM-temp1)
temp2 <-max(nba$RPM1)
nba$RPM1 <-nba$RPM1/temp2
L1<- glm(RPM1~MP,family = binomial(),data = nba)
summary(L1)
```

```
##
```

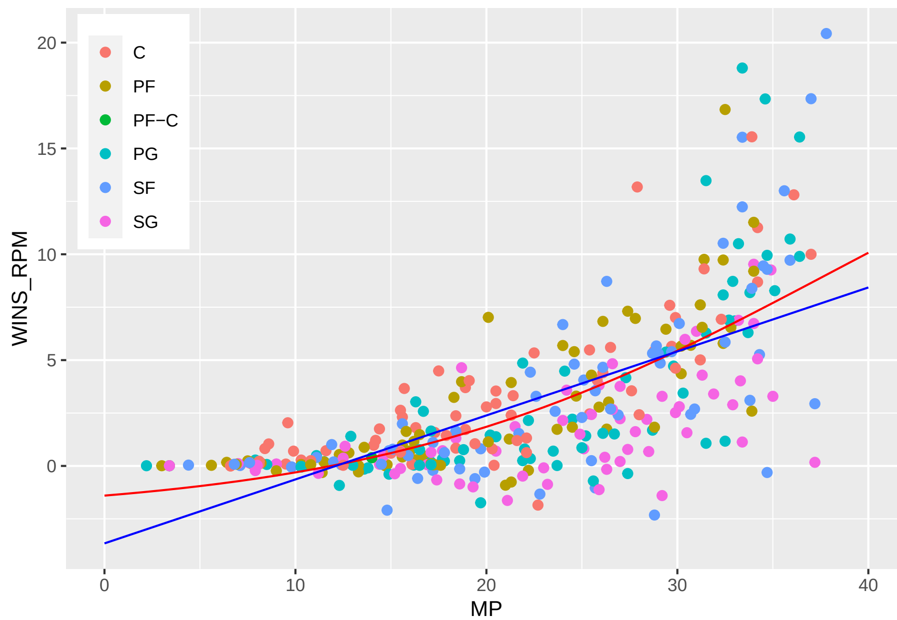
```
## Call:
## glm(formula = RPM1 ~ MP, family = binomial(), data = nba)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.87674  -0.13350   0.01699   0.13202   1.17947
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.16779    0.44091  -7.185 6.74e-13 ***
## MP           0.08368    0.01691   4.949 7.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 52.462  on 341  degrees of freedom
## Residual deviance: 24.469  on 340  degrees of freedom
## AIC: 208.11
##
## Number of Fisher Scoring iterations: 5
```

```r
L2<- lm(WINS_RPM~MP,data = nba)
summary(L2)
```

```
##
## Call:
## lm(formula = WINS_RPM ~ MP, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4157 -1.5305 -0.0712  1.3673 12.6630
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.6593     0.4053  -9.028   <2e-16 ***
## MP            0.3023     0.0174  17.373   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.829 on 340 degrees of freedom
## Multiple R-squared:  0.4702, Adjusted R-squared:  0.4687
## F-statistic: 301.8 on 1 and 340 DF,  p-value: < 2.2e-16
```

```r
k <- 0:40
Y <- exp(coef(L1)[1] + coef(L1)[2]*k)/(1+exp(coef(L1)[1] + coef(L1)[2]*k))
Y<- Y*temp2 + temp1
Y2<- coef(L2)[1] + coef(L2)[2]*k
ggplot()+
  geom_point(aes(x = nba$MP,y = nba$WINS_RPM,col = nba$POSITION),size = 2) +
  geom_line(aes(x = k,y = Y),col = "red")+
  geom_line(aes(x = k,y = Y2),col = "blue")+
  theme(legend.position = c(0.08,0.78),legend.title = element_blank())+
  labs(x = "MP",y = 'WINS_RPM')
```

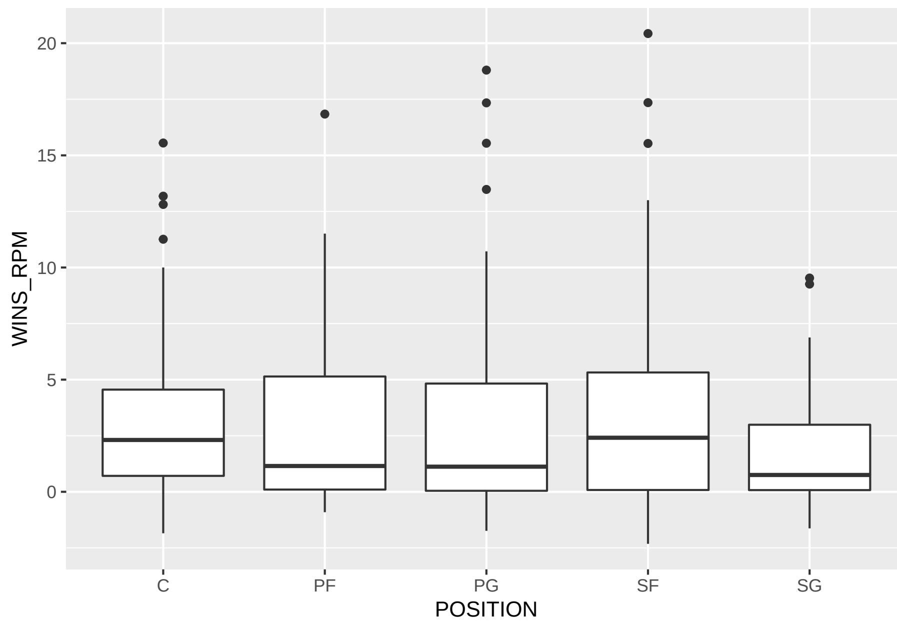


从结果来看，出场时间与 RPM 有着明显的正相关关系，这也与实际情况相符，球队往往希望表现更好的球员多打一会。另外，还可以发现，在拥有较高 RPM 的球员中，SG 位置似乎很少，说明可能现在球队更喜欢将 PG 或 SF 作为球队核心。下面的图也将说明这一关系 (因为 PF-C 分类太少，就将其去除)。

```
table(nba$POSITION)
```

```
##
##     C    PF  PF-C    PG    SF    SG
##    67    70     2    70    65    68
```

```
nba %>% filter(POSITION != 'PF-C') %>% ggplot(aes(x = POSITION,y = WINS_RPM))+
  geom_boxplot()+labs(x = "POSITION",y = 'WINS_RPM')
```



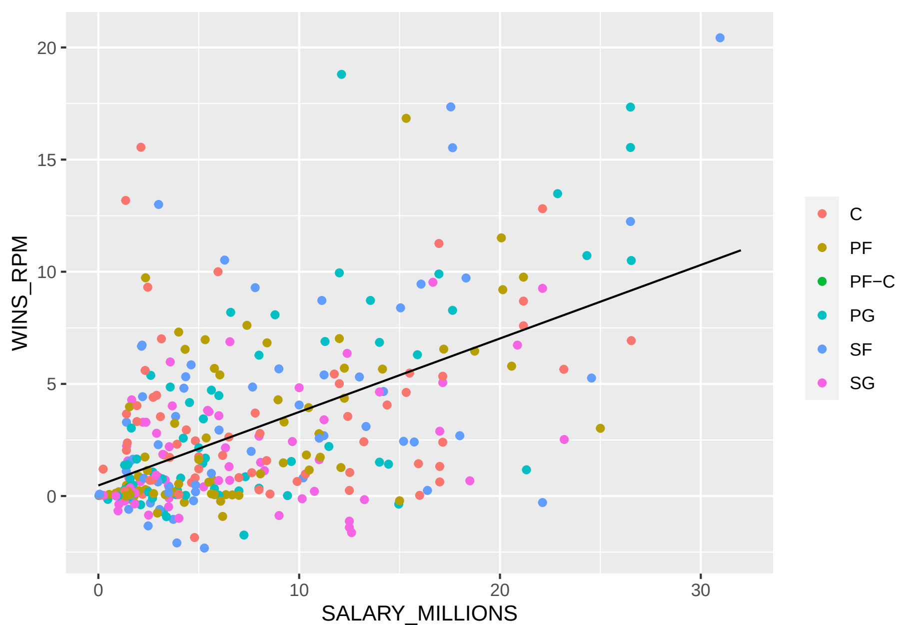
RPM 与薪酬

```r
L3<- lm(WINS_RPM~SALARY_MILLIONS,data = nba)
summary(L3)
```

```
##
## Call:
## lm(formula = WINS_RPM ~ SALARY_MILLIONS, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0102 -1.6998 -0.6954  1.2745 14.3838
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.47149    0.26362   1.789   0.0746 .
## SALARY_MILLIONS  0.32770    0.02697  12.150   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.245 on 340 degrees of freedom
## Multiple R-squared:  0.3028, Adjusted R-squared:  0.3007
## F-statistic: 147.6 on 1 and 340 DF,  p-value: < 2.2e-16
```

```r
k<-0:32
Y <- coef(L3)[1] + coef(L3)[2]*k
ggplot()+
  geom_point(aes(x = nba$SALARY_MILLIONS,y = nba$WINS_RPM,col = nba$POSITION))+
  geom_line(aes(x = k,y = Y))+
  theme(legend.title = element_blank())+
  labs(x = "SALARY_MILLIONS",y = 'WINS_RPM')
```



可见 RPM 与薪酬基本上也保持正相关，但是不如出场时间拟合的那么好。结合实际情况，有些低薪球员为了在将来获得高薪合同奋力表现自己，而有些球员拿到高额报酬后就开始“放松养生”，表现糟糕。尽管实际情况可能更加复杂，但是从 17 年的数据来看，确实有许多球员在 RPM 这一项指标上表现的与他的薪水不符。