chap03. Correlation by using batting indices

Peter

2018-11-13

Pre-Processing

Make the basic set merged together into a table

```
library(Lahman)
library(plyr)
a = subset(Batting, yearID>2014)
a$teamID = as.numeric(as.factor(a$teamID))
head(a)
```

```
##        playerID yearID stint teamID lgID   G  AB  R   H X2B X3B HR RBI SB
## 99848 aardsda01   2015     1      4   NL  33   1  0   0   0   0  0   0  0
## 99849  abadfe01   2015     1     96   AL  62   0  0   0   0   0  0   0  0
## 99850 abreujo02   2015     1     33   AL 154 613 88 178  34   3 30 101  0
## 99851 achteaj01   2015     1     79   AL  11   0  0   0   0   0  0   0  0
## 99852 ackledu01   2015     1    116   AL  85 186 22  40   8   1  6  19  2
## 99853 ackledu01   2015     2     93   AL  23  52  6  15   3   2  4  11  0
##       CS BB  SO IBB HBP SH SF GIDP
## 99848  0  0   1   0   0  0  0    0
## 99849  0  0   0   0   0  0  0    0
## 99850  0 39 140  11  15  0  1   16
## 99851  0  0   0   0   0  0  0    0
## 99852  2 14  38   0   1  3  3    3
## 99853  0  4   7   0   0  0  1    0
```

```
b = function(a){return(data.frame(
                team=ifelse(mean(a$teamID) == a$teamID, 0, 1),
                a$playerID, a$lgID, a$SF, a$SH, a$H,
                a$yearID, a$teamID, a$RBI, a$AB))}
d = ddply(a,.(playerID), b)
head(d)
```

```
##    playerID team a.playerID a.lgID a.SF a.SH a.H a.yearID a.teamID a.RBI
## 1 aardsda01    0  aardsda01     NL    0    0   0     2015        4     0
## 2  abadfe01    1   abadfe01     AL    0    0   0     2015       96     0
## 3  abadfe01    1   abadfe01     AL    0    0   0     2016       79     0
## 4  abadfe01    1   abadfe01     AL    0    0   0     2016       16     0
## 5 abreujo02    0  abreujo02     AL    1    0 178     2015       33   101
## 6 abreujo02    0  abreujo02     AL    9    0 183     2016       33   100
##   a.AB
## 1    1
## 2    0
## 3    1
## 4    0
## 5  613
## 6  624
```

Get lag_ variables for previouse years data

```
d$lag_team = as.numeric(sapply(1:nrow(d), function(x){d$a.teamID[x-1]}))
head(d)
```

```
##    playerID team a.playerID a.lgID a.SF a.SH a.H a.yearID a.teamID a.RBI
## 1 aardsda01    0  aardsda01     NL    0    0   0     2015        4     0
## 2  abadfe01    1   abadfe01     AL    0    0   0     2015       96     0
## 3  abadfe01    1   abadfe01     AL    0    0   0     2016       79     0
## 4  abadfe01    1   abadfe01     AL    0    0   0     2016       16     0
## 5 abreujo02    0  abreujo02     AL    1    0 178     2015       33   101
## 6 abreujo02    0  abreujo02     AL    9    0 183     2016       33   100
##   a.AB lag_team
## 1    1       NA
## 2    0        4
## 3    1       96
## 4    0       79
## 5  613       16
## 6  624       33
```

```
d$lag_RBI = as.numeric(sapply(1:nrow(d), function(x){d$a.RBI[x-1]}))
d$lag_AB = as.numeric(sapply(1:nrow(d), function(x){d$a.AB[x-1]}))
d$lag_SF = as.numeric(sapply(1:nrow(d), function(x){d$a.SF[x-1]}))
d$lag_SH = as.numeric(sapply(1:nrow(d), function(x){d$a.SH[x-1]}))
d$lag_H = as.numeric(sapply(1:nrow(d), function(x){d$a.H[x-1]}))
d$lag_playerID = as.character(sapply(1:nrow(d), function(x){d$playerID[x-1]}))
head(d)
```

```
##    playerID team a.playerID a.lgID a.SF a.SH a.H a.yearID a.teamID a.RBI
## 1 aardsda01    0  aardsda01     NL    0    0   0     2015        4     0
## 2  abadfe01    1   abadfe01     AL    0    0   0     2015       96     0
## 3  abadfe01    1   abadfe01     AL    0    0   0     2016       79     0
## 4  abadfe01    1   abadfe01     AL    0    0   0     2016       16     0
## 5 abreujo02    0  abreujo02     AL    1    0 178     2015       33   101
## 6 abreujo02    0  abreujo02     AL    9    0 183     2016       33   100
##   a.AB lag_team lag_RBI lag_AB lag_SF lag_SH lag_H lag_playerID
## 1    1       NA      NA     NA     NA     NA    NA character(0)
## 2    0        4       0      1      0      0     0    aardsda01
## 3    1       96       0      0      0      0     0     abadfe01
## 4    0       79       0      1      0      0     0     abadfe01
## 5  613       16       0      0      0      0     0     abadfe01
## 6  624       33     101    613      1      0   178    abreujo02
```

Filtering the data

```
d$lag_team = ifelse(d$playerID==d$lag_playerID, d$lag_team, 'NA')
head(d)
```

```
##    playerID team a.playerID a.lgID a.SF a.SH a.H a.yearID a.teamID a.RBI
## 1 aardsda01    0  aardsda01     NL    0    0   0     2015        4     0
## 2  abadfe01    1   abadfe01     AL    0    0   0     2015       96     0
## 3  abadfe01    1   abadfe01     AL    0    0   0     2016       79     0
## 4  abadfe01    1   abadfe01     AL    0    0   0     2016       16     0
## 5 abreujo02    0  abreujo02     AL    1    0 178     2015       33   101
## 6 abreujo02    0  abreujo02     AL    9    0 183     2016       33   100
```

```
##   a.AB lag_team lag_RBI lag_AB lag_SF lag_SH lag_H lag_playerID
## 1    1       NA      NA     NA     NA     NA    NA character(0)
## 2    0       NA       0      1      0      0     0    aardsda01
## 3    1       96       0      0      0      0     0     abadfe01
## 4    0       79       0      1      0      0     0     abadfe01
## 5  613       NA       0      0      0      0     0     abadfe01
## 6  624       33     101    613      1      0   178    abreujo02
```

```
d$lag_avg = d$lag_H / d$lag_AB
d$sac = d$lag_SF + d$lag_SH
d = subset(d, d$a.AB > 400 & d$lag_AB > 400)
d$change_rbi = d$a.RBI / d$lag_RBI
d = subset(d, !(d$lag_team == 'NA' | d$a.teamID == d$lag_team))
head(d)
```

```
##      playerID team a.playerID a.lgID a.SF a.SH a.H a.yearID a.teamID a.RBI
## 366 cabreas01    1  cabreas01     NL    2    0 146     2016       94    62
## 444 castrst01    1  castrst01     AL    5    1 156     2016       93    70
## 675 desmoia01    1  desmoia01     AL    3    0 178     2016      131    86
## 796 escobyu01    1  escobyu01     AL    4    3 157     2016       71    39
## 885 frazito01    1  frazito01     AL    7    1 133     2016       33    98
## 893 freesda01    1  freesda01     NL    0    0 118     2016      106    55
##     a.AB lag_team lag_RBI lag_AB lag_SF lag_SH lag_H lag_playerID
## 366  521      130      58    505      6      1   134    cabreas01
## 444  577       35      69    547      4      1   145    castrst01
## 675  625      137      62    583      4      6   136    desmoia01
## 796  517      137      56    535      2      1   168    escobyu01
## 885  590       38      89    619      7      1   158    frazito01
## 893  437       71      56    424      3      0   109    freesda01
##       lag_avg sac change_rbi
## 366 0.2653465   7  1.0689655
## 444 0.2650823   5  1.0144928
## 675 0.2332762  10  1.3870968
## 796 0.3140187   3  0.6964286
## 885 0.2552504   8  1.1011236
## 893 0.2570755   3  0.9821429
```

add some additional variables for plotting

```
d$lg_col = ifelse(d$a.lgID =='NL', 'gray', 'black')
d$lg_shape = ifelse(d$a.lgID == 'NL', 2, 15)
```

Calculate the Corvariance between

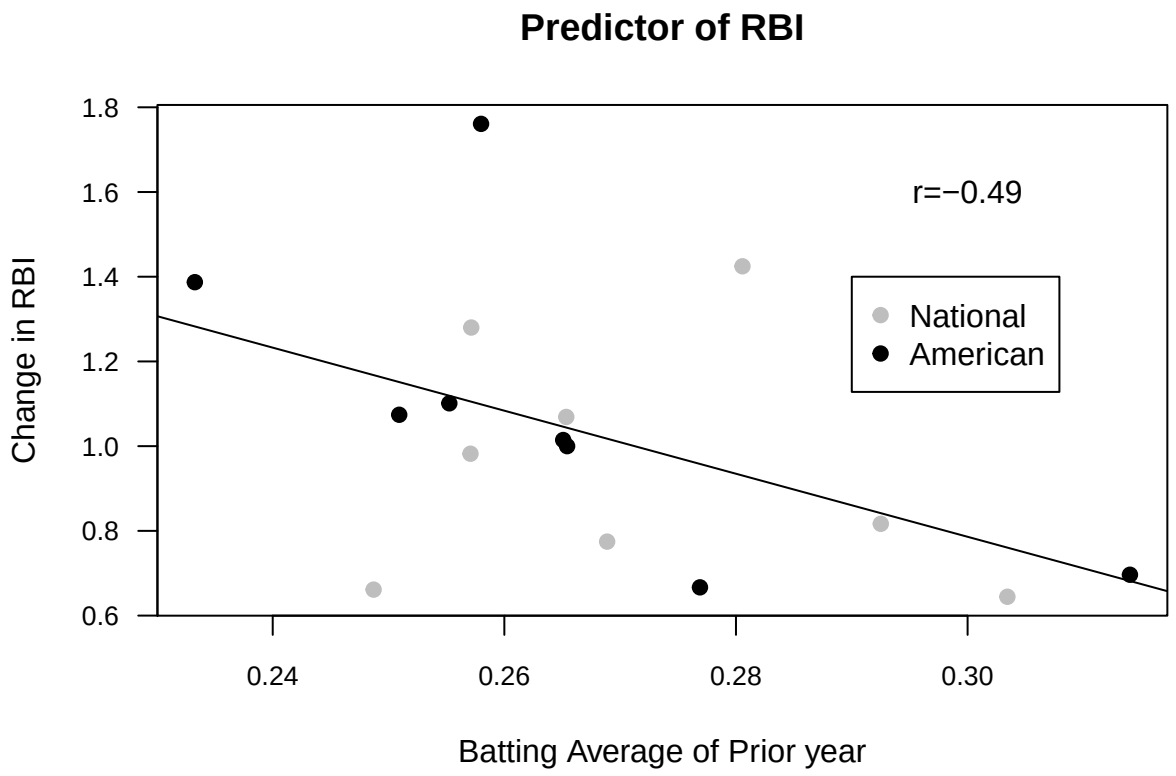
```
cor(d$lag_avg, d$change_rbi)
```

```
## [1] -0.4890067
```

Plot

```
plot(d$lag_avg, d$change_rbi, main='Predictor of RBI', xlab=
       'Batting Average of Prior year', ylab='Change in RBI', las = 1,
       cex.axis = 0.8, pch=19, col=d$lg_col)
text(x=0.3, y=1.6, label='r=-0.49')
abline(lm(d$change_rbi~d$lag_avg, d))
```

```
legend(x=0.29, y=1.4, c('National', 'American'), col=c('gray', 'black'), pch=c(19, 19))
```



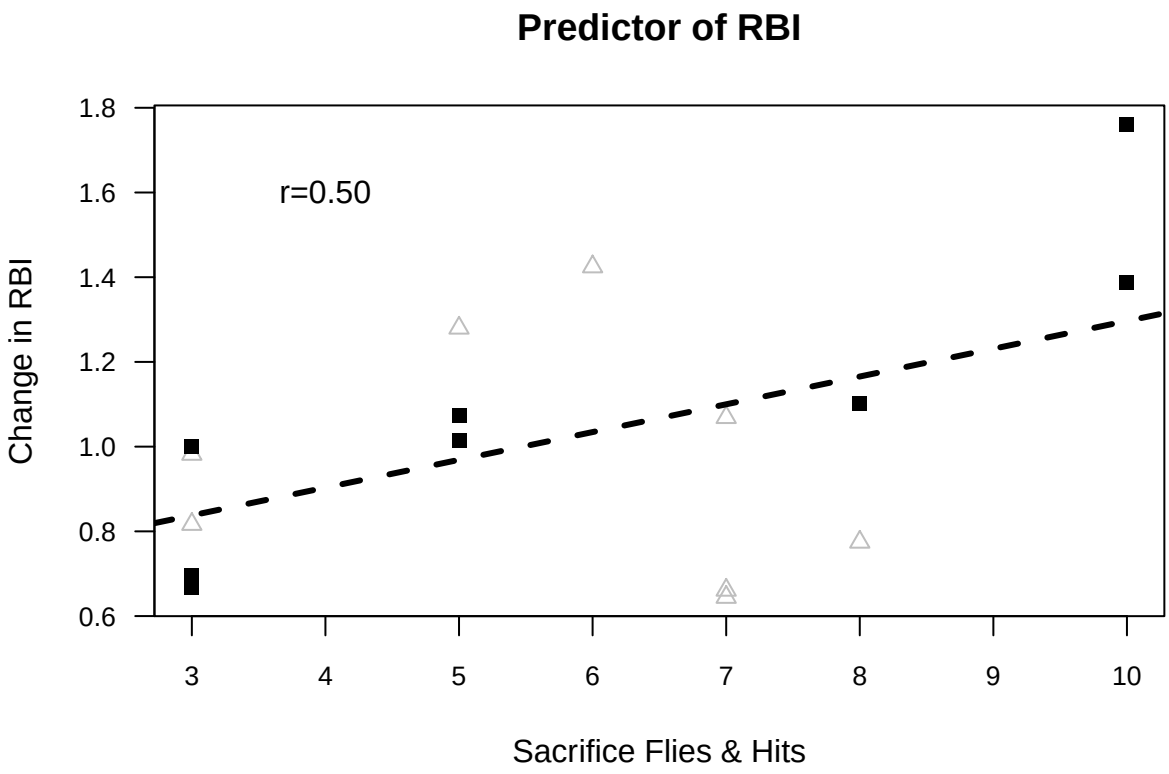
sac vs change_rbi

```
cor(d$sac, d$change_rbi)
```

```
## [1] 0.5042151
```

plot

```
plot(d$sac, d$change_rbi, main='Predictor of RBI', xlab=
       'Sacrifice Flies & Hits', ylab='Change in RBI', las = 1,
       cex.axis = 0.8, pch=d$lg_shape, col=d$lg_col)
text(x=4, y=1.6, label='r=0.50')
abline(lm(d$change_rbi~d$sac, d), lty=2, lwd=3)
```



Get Correlation Table by Using tableHTML

```
#install.packages('tableHTML')
library(tableHTML)
e = with(d, data.frame(change_rbi, sac, lag_avg))
colnames(e) = c('c_RBI', 'Sacrifice', 'AVG')
cor(e)
```

```
##                c_RBI  Sacrifice        AVG
## c_RBI      1.0000000  0.5042151 -0.4890067
## Sacrifice  0.5042151  1.0000000 -0.4597213
## AVG       -0.4890067 -0.4597213  1.0000000
```

```
tableHTML(round(cor(e), 3))
```

c_RBI

Sacrifice

AVG

c_RBI

1

0.504

-0.489

Sacrifice

0.504

1

-0.46

AVG

-0.489

-0.46

1