Lecture 3: Baseball stats & Multivariate regression

Skidmore College, MA 276

Goals

- Player specific measures, baseball pitching statistics
- Example: Fielding independent pitching
- Extensions: WAR, deserved run average
- Tools: multivariate regression, player prediction

Review: multivariate regression

Model:

$$y_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \ldots + \beta_{p-1} * x_{i,p-1} + \epsilon_i$$

Assumptions:

- $\epsilon_i \sim N(0, \sigma^2)$
- ϵ_i,$\epsilon_{i'}$ independent for all i, i'
- Linear relationship between y and x

Review: multivariate regression

Estimated model:

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_{i1} + \hat{\beta}_2 * x_{i2} + \ldots + \hat{\beta}_{p-1} * x_{i,p-1}$

Interpretations:

- $\hat{\beta}_0$:
- $\hat{\beta}_1$:

Back to baseball

Fielding Independent Pitching (FIP)

1. Is it important to success?
2. How well does it measure a player's contribution?
3. Is it repeatable?

Ex: FIP

"measurement of a pitcher's performance that strips out the role of **defense, luck, and sequencing**"- FanGraphs

- defense
- luck
- sequencing

$$FIP = \frac{(13*HR)+(3*(BB+HBP))-(2*K)}{IP} + constant$$

Ex: FIP

```
library(Lahman)
library(mosaic)
data(Teams)

#filter: only look at years starting with 1970
Teams.1 <- filter(Teams, yearID >= 1970)

#mutate: create new variable FIP
Teams.1 <- mutate(Teams.1,
    FIP = ((13*HRA) + 3*(BBA) - 2*SOA)/IPouts)
```

Ex: FIP

```
fit.pitcher <- lm(RA ~ HRA + BBA + SOA, data = Teams.1)
```

Write the multiple regression model:

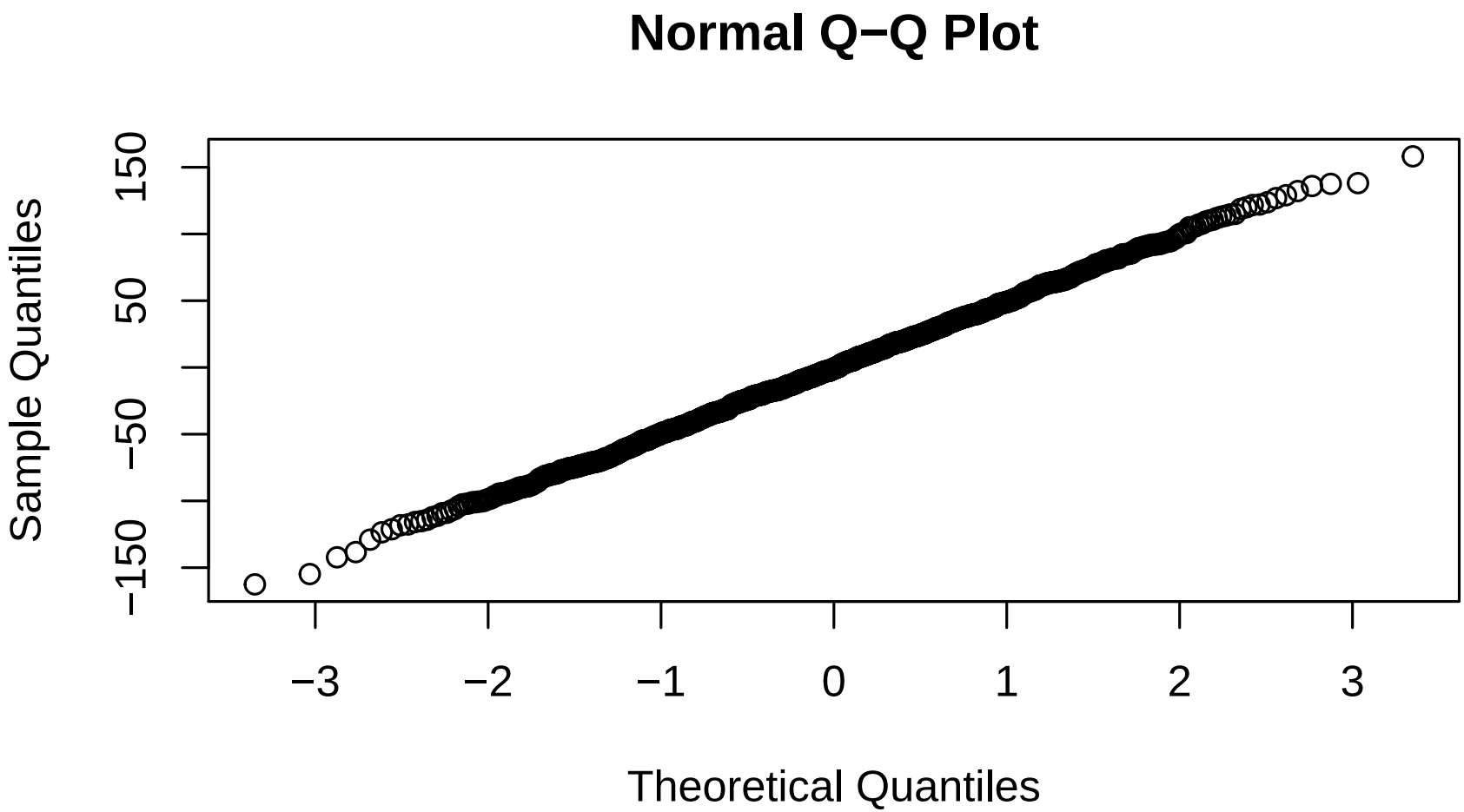
Ex: FIP

```
msummary(fit.pitcher)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 205.176535  11.736668   17.48   <2e-16 ***
## HRA           2.007380   0.047354   42.39   <2e-16 ***
## BBA           0.571729   0.020551   27.82   <2e-16 ***
## SOA          -0.089678   0.008745  -10.26   <2e-16 ***
##
## Residual standard error: 49.74 on 1230 degrees of freedom
## Multiple R-squared:  0.7612, Adjusted R-squared:  0.7607
## F-statistic:  1307 on 3 and 1230 DF,  p-value: < 2.2e-16
```

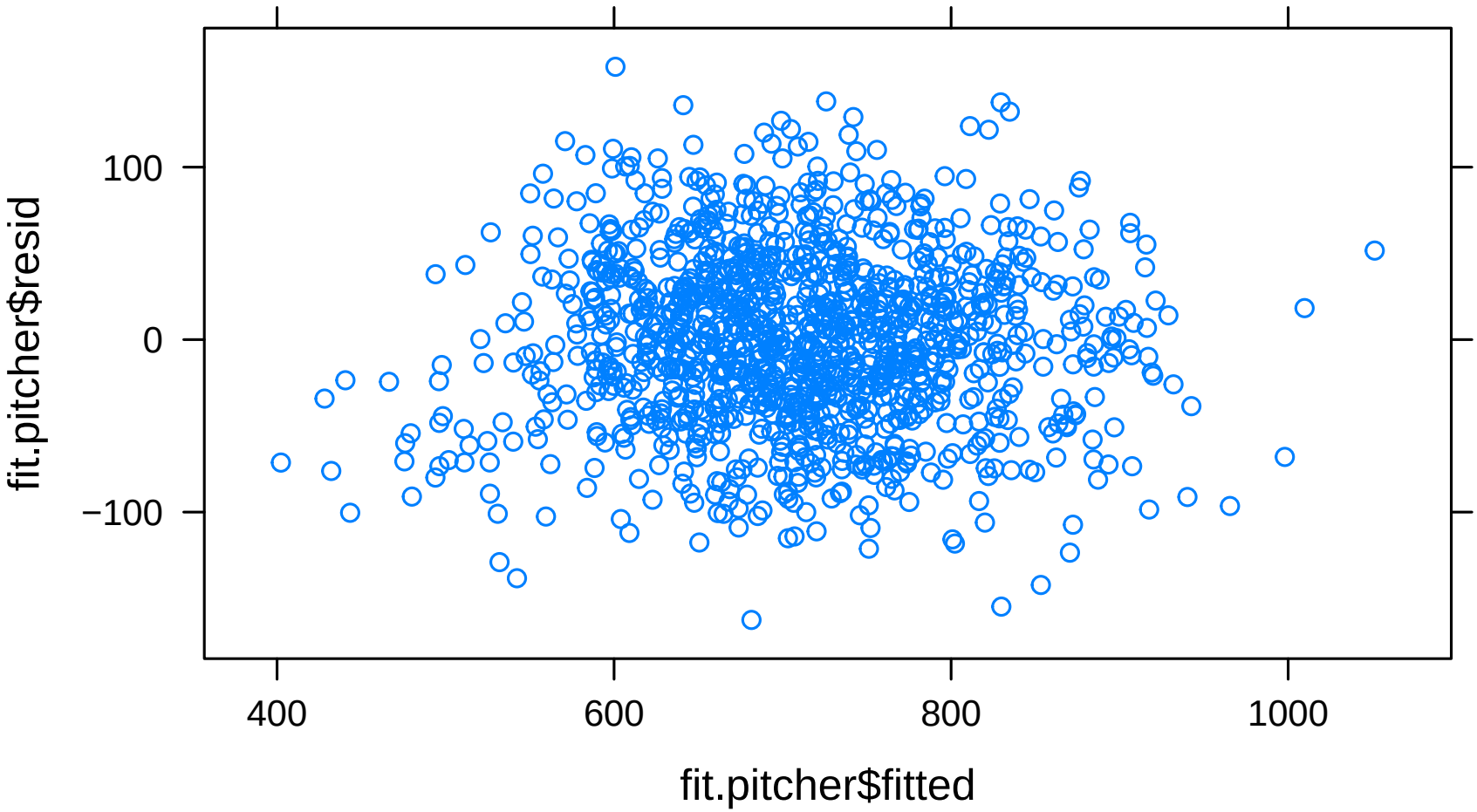
Ex: FIP

```
qqnorm(fit.pitcher$resid)
```



Ex: FIP

```
xyplot(fit.pitcher$resid ~ fit.pitcher$fitted)
```



Ex: FIP

Write the estimated regression model, interpret slope for HRA

```
msummary(fit.pitcher)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 205.176535  11.736668   17.48   <2e-16 ***
## HRA           2.007380   0.047354   42.39   <2e-16 ***
## BBA           0.571729   0.020551   27.82   <2e-16 ***
## SOA          -0.089678   0.008745  -10.26   <2e-16 ***
##
## Residual standard error: 49.74 on 1230 degrees of freedom
## Multiple R-squared:  0.7612, Adjusted R-squared:  0.7607
## F-statistic:  1307 on 3 and 1230 DF,  p-value: < 2.2e-16
```

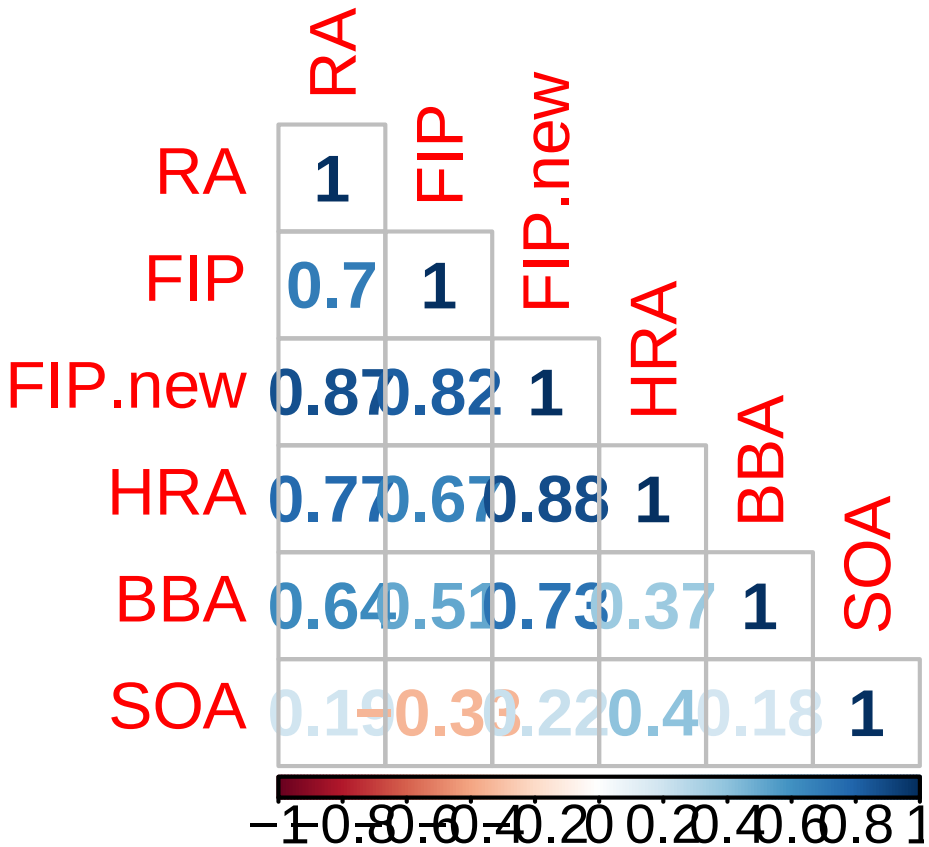
Ex: FIP

```
Teams.1 <- mutate(Teams.1,
   FIP.new = predict(fit.pitcher, Teams.1))
mat <- select(Teams.1, RA, FIP, FIP.new, HRA, BBA, SOA)
cor.matrix <- cor(mat)
round(cor.matrix, 2)
```

```
##              RA   FIP FIP.new  HRA  BBA   SOA
## RA         1.00  0.70    0.87 0.77 0.64  0.19
## FIP        0.70  1.00    0.82 0.67 0.51 -0.33
## FIP.new    0.87  0.82    1.00 0.88 0.73  0.22
## HRA        0.77  0.67    0.88 1.00 0.37  0.40
## BBA        0.64  0.51    0.73 0.37 1.00  0.18
## SOA        0.19 -0.33    0.22 0.40 0.18  1.00
```

Ex: FIP

```
library(corrplot)
corrplot(cor.matrix, method="number", type = "lower")
```



Ex: FIP

Conclusions:

- Multiple regression model estimated FIP (`FIP.new`) closely approximates actual FIP formula
- High link (team-wide) between FIP and RA

To the players!

```
data(Pitching)
head(Pitching)
```

```
##    playerID yearID stint teamID lgID  W  L  G GS CG SHO SV IP
## 1 bechtge01   1871     1    PH1   NA  1  2  3  3  2   0  0
## 2 brainas01   1871     1    WS3   NA 12 15 30 30 30   0  0
## 3 fergubo01   1871     1    NY2   NA  0  0  1  0  0   0  0
## 4 fishech01   1871     1    RC1   NA  4 16 24 24 22   1  0
## 5 fleetfr01   1871     1    NY2   NA  0  1  1  1  1   0  0
## 6 flowedi01   1871     1    TRO   NA  0  0  1  0  0   0  0
##   HR BB SO BAOpp   ERA IBB WP HBP BK BFP GF   R SH SF GIDP
## 1  0 11  1    NA  7.96  NA NA  NA  0  NA NA  42 NA NA   NA
## 2  4 37 13    NA  4.50  NA NA  NA  0  NA NA 292 NA NA   NA
## 3  0  0  0    NA 27.00  NA NA  NA  0  NA NA   9 NA NA   NA
## 4  3 31 15    NA  4.35  NA NA  NA  0  NA NA 257 NA NA   NA
## 5  0  3  0    NA 10.00  NA NA  NA  0  NA NA  21 NA NA   NA
## 6  0  0  0    NA  0.00  NA NA  NA  0  NA NA   0 NA NA   NA
```

To the players!

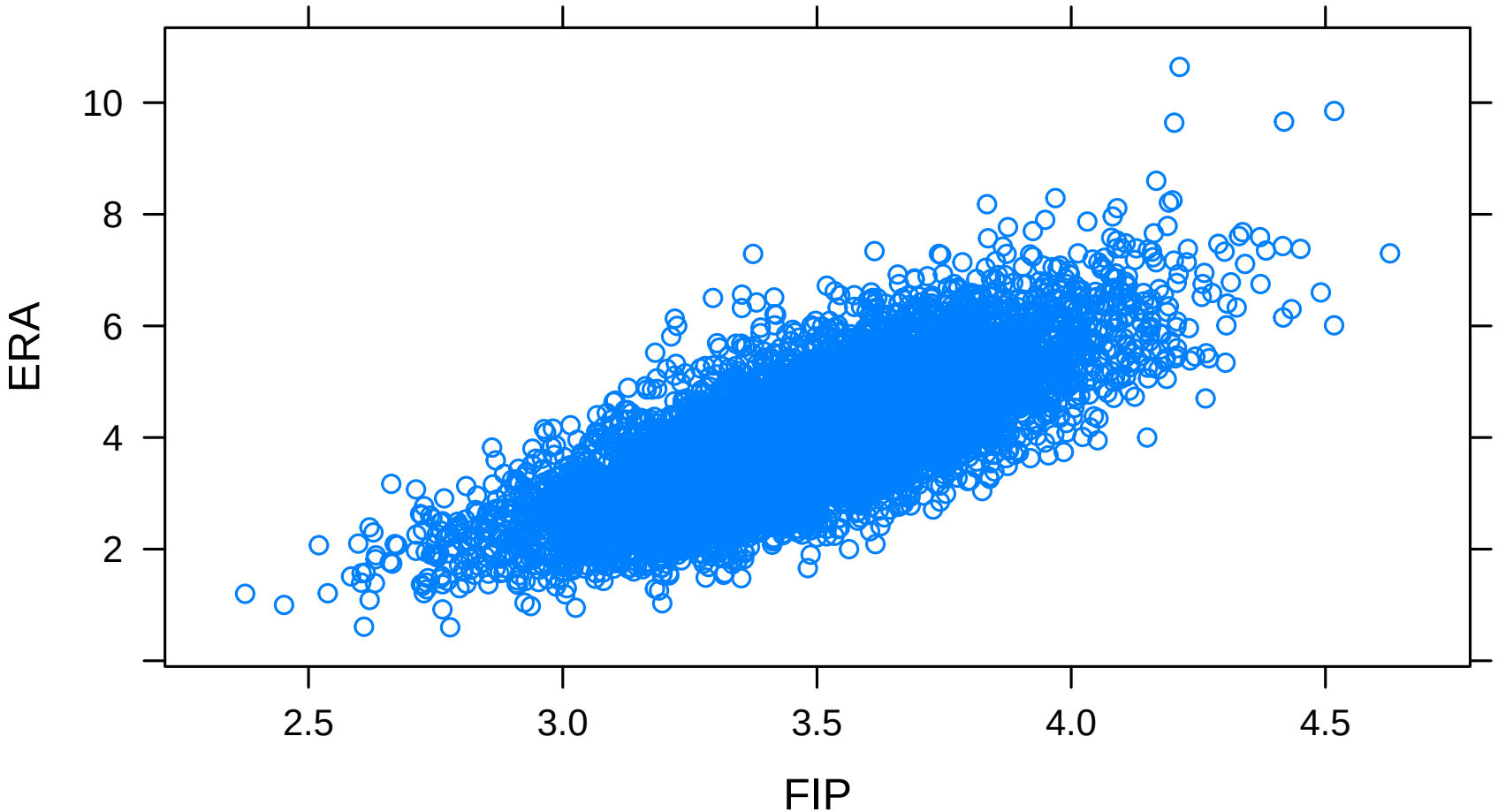
```
Pitchers.1 <- filter(Pitching, yearID >= 1970, IPouts > 200)
Pitchers.1 <- mutate(Pitchers.1,
   FIP = ((13*HR) + 3*(BB + HBP) - 2*SO)/IPouts + 3.1)
Pitchers.1[3,]
```

```
##    playerID yearID stint teamID lgID  W  L  G GS CG SHO SV IP
## 3 bahnsst01   1970     1    NYA   AL 14 11 36 35  6   2  0
##   HR BB  SO BAOpp  ERA IBB WP HBP BK BFP GF   R SH SF GIDP
## 3 23 75 116  0.25 3.33   4  3   2  0 977  0 100 NA NA   NA 3.
```

Link between ERA, FIP

What accounts for uncertainty?

```
xyplot(ERA ~ FIP, data = Pitchers.1)
```



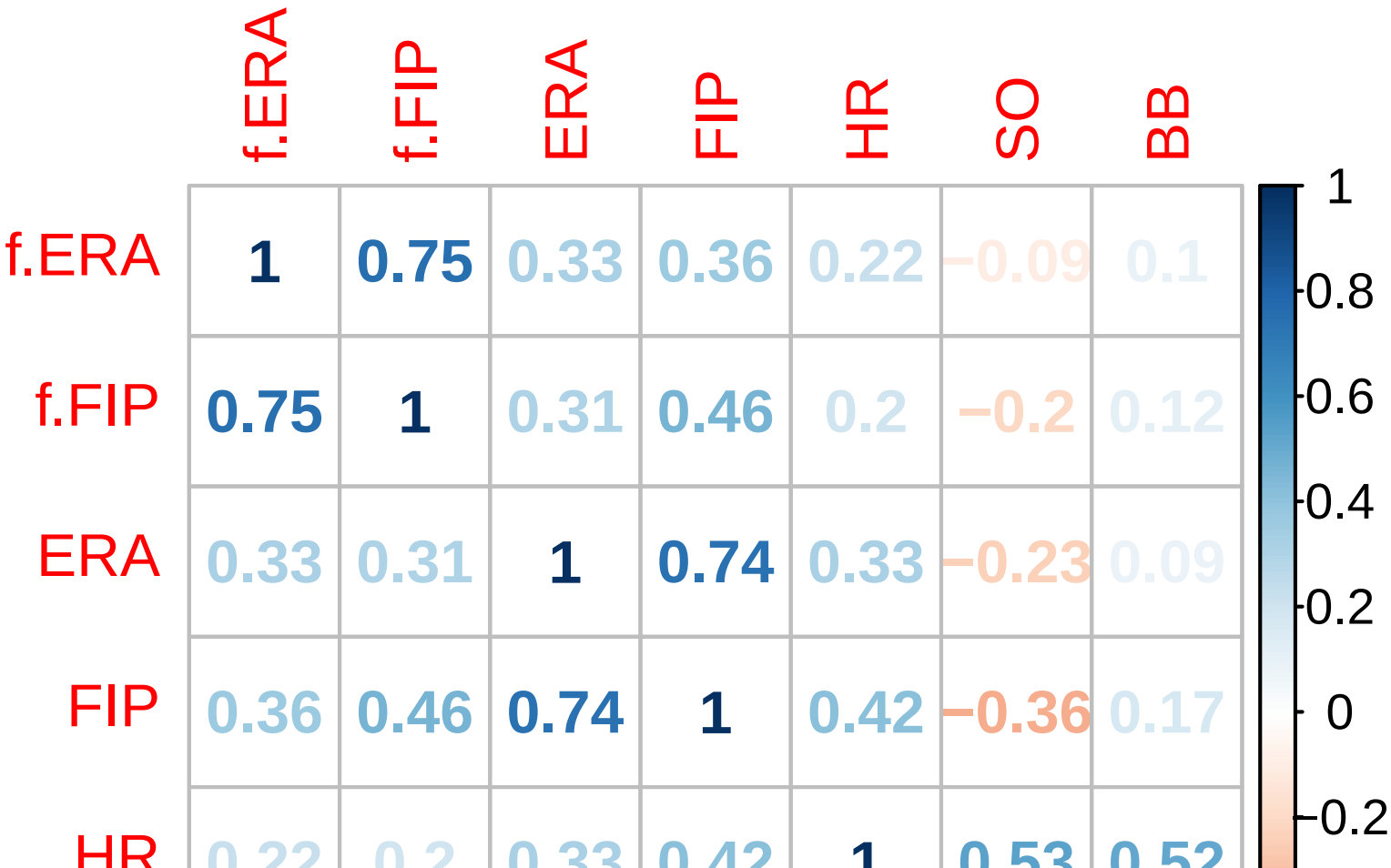
Link to the future

All of the above is within a single year. How does it project to future years?

```
Pitchers.2 <- Pitchers.1 %>%
  arrange(playerID, yearID) %>%
  group_by(playerID) %>%
  mutate(f.ERA = lead(ERA), f.FIP = lead(FIP))
```

Link to the future

```
cor.matrix <- cor(select(ungroup(Pitchers.2),
    f.ERA, f.FIP,
    ERA, FIP, HR, SO, BB),
    use="pairwise.complete.obs")
corrplot(cor.matrix, method = "number")
```



Conclusions

Fielding Independent Pitching (FIP)

1. Is it important to success?

- Yes. Why?

2. How well does it measure a player's contribution?

- Preferred to ERA - Why?

3. Is it repeatable?

- Moreso than other metrics - Why?

Just for fun

```
Pitchers.2 %>%
  group_by(playerID) %>%
  tally(G) %>%
  top_n(5)
```

```
## Source: local data frame [5 x 2]
##
##    playerID     n
##       (chr) (int)
## 1 fingero01   803
## 2 houghch01   771
## 3 riverma01   752
## 4 rogerke01   751
## 5 tekulke01   923
```

Just for fun

```
Pitchers.2 %>%
  filter(IPouts > 600) %>%
  group_by(yearID) %>%
  slice(which.min(FIP)) %>%
  tail() %>%
  select(playerID, yearID, ERA, FIP, W, L)
```

```
## Source: local data frame [6 x 6]
## Groups: yearID [6]
##
##    playerID yearID   ERA      FIP     W     L
##       (chr)  (int) (dbl)    (dbl) (int) (int)
## 1 greinza01   2009  2.16 2.844186    16     8
## 2 wainwad01   2010  2.42 3.026194    20    11
## 3 hallaro01   2011  2.35 2.824679    19     6
## 4 hernafe02   2012  3.06 3.013793    13     9
## 5 kershcl01   2013  1.83 2.879661    16     9
## 6 klubeco01   2014  2.44 2.838331    18     9
```