HW 2: Linear regression and prediction using MLB players

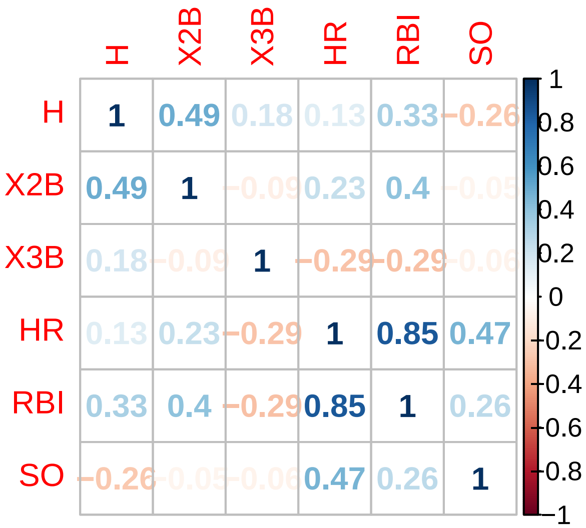
Stats and sports class

Fall 2019

3. Make a correlation matrix - both a matrix of the variables, as well as a visualization – using hits, doubles, triples, home runs, RBI, and strikeouts.

Answer

```r
library(corrplot)
library(tidyverse)
library(Lahman)
Batting_1 <- Batting %>%
  filter(yearID >= 2000) %>%
  select (playerID, yearID, AB:SO) %>%
  filter(AB >= 500)
var_cor <- Batting_1 %>%
  select(H, X2B, X3B, HR, RBI, SO)
cor_mat <- cor(var_cor)
corrplot(cor_mat, method = "number")
```



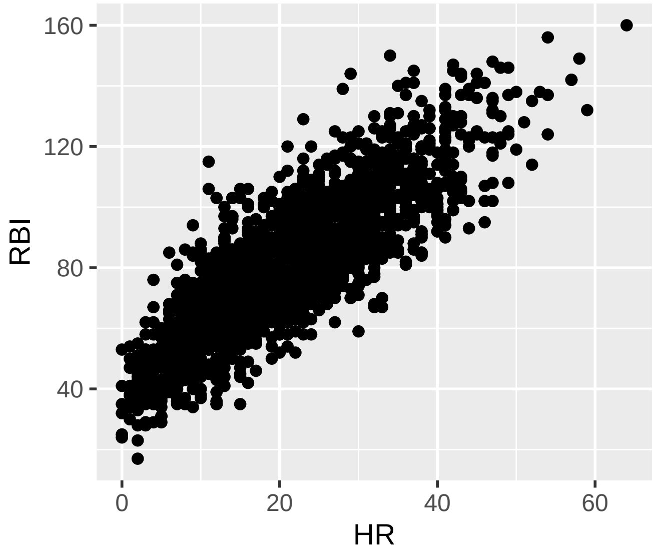
	H	X2B	X3B	HR	RBI	SO
H	1	0.49	0.18	0.13	0.33	−0.26
X2B	0.49	1	−0.09	0.23	0.4	−0.05
X3B	0.18	−0.09	1	−0.29	−0.29	−0.06
HR	0.13	0.23	−0.29	1	0.85	0.47
RBI	0.33	0.4	−0.29	0.85	1	0.26
SO	−0.26	−0.05	−0.06	0.47	0.26	1

```
## Note -- any correlation plot would be acceptable
```

4. Make a scatter plot of runs batted in (RBI, the y-variable) and home runs (HR, the x-variable). Estimate and write the regression line using the `lm` command. Finally, interpret the slope and intercept of this line.

Answer

```r
ggplot(data = Batting_1, aes(x = HR, y = RBI)) +
  geom_point()
```



```
fit_1 <- lm(RBI ~ HR, data = Batting_1)
summary(fit_1)
```

```
##
## Call:
## lm(formula = RBI ~ HR, data = Batting_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.167  -8.541  -0.716   8.177  52.352
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.66305    0.60586   70.42   <2e-16 ***
## HR           1.81679    0.02565   70.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.51 on 2001 degrees of freedom
## Multiple R-squared:  0.7148, Adjusted R-squared:  0.7147
## F-statistic:  5015 on 1 and 2001 DF,  p-value: < 2.2e-16
```

Slope of 1.81: For each additional home run, a batter is expected to have an additional 1.82 RBIs

Intercept of 42.7: A batter that hits 0 HRs would be expected to have 42.7 RBIs.

Estimated line: RBI-hat = 42.7 + 1.82*HR

5. Pete Alonso – currently with the New York Mets – has hit 47 home runs and batted in 109 runs (as of Sept 15, 2019). Given his home runs, what is his residual? That is, how many more or fewer runs batted in has he hit than we'd expect given his home runs?

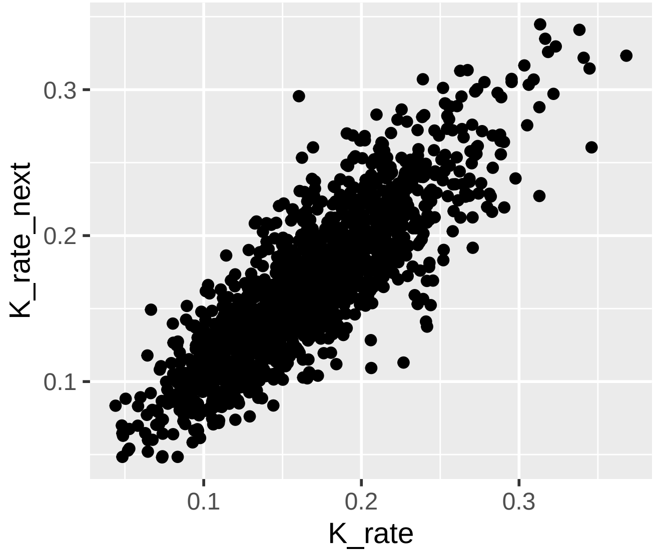
Expected RBIs: 42.7+1.82*47 = 128. We expect Alonso to have roughly 128 runs batted in. He actually has 109, which means he has fewer RBIs than we expect

Part II: Predictability of player metrics

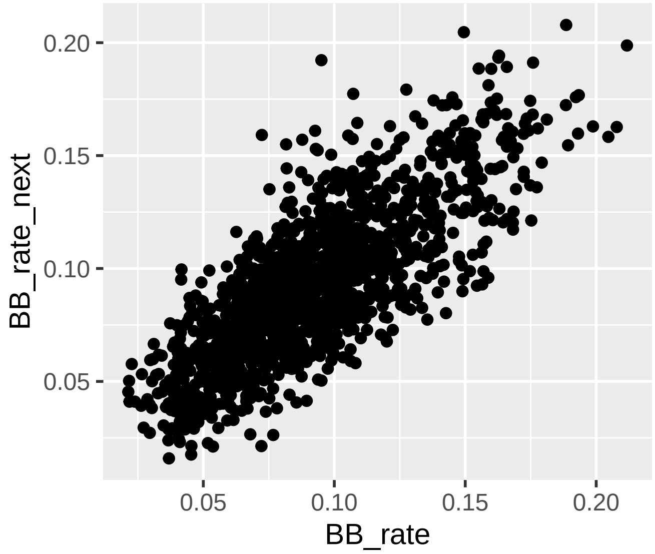
Note: The code drops the last year of a players' career – there is no future variable to look at.

10. Use (i) scatter plots and (ii) correlation coefficients to assess the year-over-year repeatability of strikeout rate, walk rate (BB_rate), HR rate, and RBI rate. That is, compare each metric in a players' curret year to the metric that he records in the following year. Which of these metrics is most repeatable? Which of these is least repeatable?

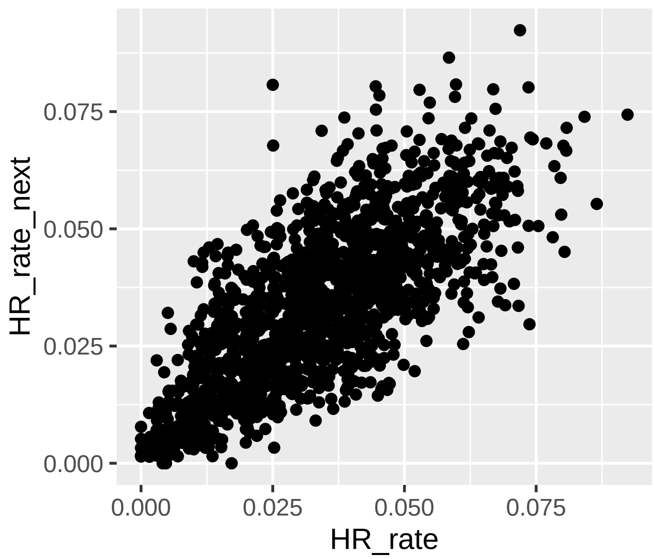
```r
#i Scatter plots
ggplot(Batting_2, aes(K_rate, K_rate_next)) +
  geom_point()
```



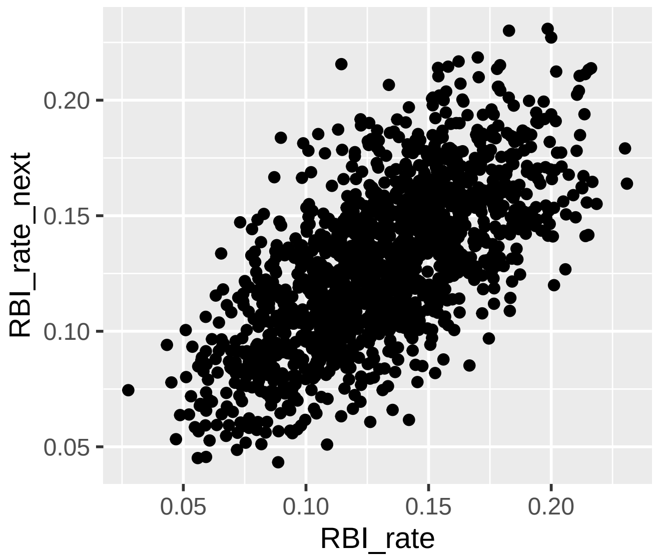
```r
ggplot(Batting_2, aes(BB_rate, BB_rate_next)) +
  geom_point()
```



```r
ggplot(Batting_2, aes(HR_rate, HR_rate_next)) +
  geom_point()
```



```
ggplot(Batting_2, aes(RBI_rate, RBI_rate_next)) +
  geom_point()
```



```
# ii correlation coefficient
Batting_2 %>%
  summarise(cor_k = cor(K_rate, K_rate_next),
            cor_bb = cor(BB_rate, BB_rate_next),
            cor_hr = cor(HR_rate, HR_rate_next),
            cor_rbi = cor(RBI_rate, RBI_rate_next))
```

```
## # A tibble: 1 x 4
##   cor_k cor_bb cor_hr cor_rbi
##   <dbl>  <dbl>  <dbl>   <dbl>
## 1 0.855  0.784  0.729   0.677
```

By both looking at the scatter plots and the correlation coefficients, strikeouts are the most repeatable from year to year, and rbi rate is the least repeatable.