HW 3 solutions

Sports and statistics

For this homework, we will be using data provided in the Lahman package.

Part 1

Our first goal is to use variables in the Teams data set to identify the model that best fits the number of runs scored for a team in a single season. We'll use all seasons since 1970. Several models are proposed.

```
library(mosaic)
library(Lahman)
library(corrplot)
data(Teams)
data(Pitching)
data(Batting)
Teams.1 <- filter(Teams, yearID >=1970)
Teams.1 <- mutate(Teams.1, X1B = H - X2B - X3B - HR)

fit.1 <- lm(R ~ X1B + X2B + X3B + HR, data = Teams.1)
fit.2 <- lm(R ~ X1B + X2B + X3B + HR + BB, data = Teams.1)
fit.3 <- lm(R ~ X1B + X2B + X3B + HR + BB + SO, data = Teams.1)
fit.4 <- lm(R ~ X1B + X2B + X3B + HR + BB + SO + CS, data = Teams.1)
fit.5 <- lm(R ~ X1B + X2B + X3B + HR + BB + SO + CS + lgID, data = Teams.1)
fit.6 <- lm(R ~ X1B + X2B + X3B + HR + BB + SO + CS + lgID + SB, data = Teams.1)

options(scipen=999)
```

Note: The `options(scipen = 999)` command disables R's scientific notation.

Question 1

Using the AIC criteria, which of the six models would you recommend for measuring runs scored on a team-wide level? From a baseball perspective, what does your choice suggest about certain measurements as far as their link to runs scored?

Model (6) has the lowest AIC, suggesting it is the strongest fit of runs scored. On a team-level, this suggests that each of the variables in model (6) are valuable as far as understanding runs scored.

Question 2

One of the coefficients in fit.5 and fit.6 is lgID. Generate a table of the lgID in your data set. What does this variable refer to?

```
tally(~lgID, data = Teams.1)
```

```
##
##   AA  AL  FL  NA  NL  PL  UA
##    0 618   0   0 616   0   0
```

There are 618 AL rows and 616 NL rows (recall: each row is a team-season). This is the league each team played in during each season.

Question 3

Using the code below, the coefficient for `league = "NL"` is negative. Interpret this coefficient. What about baseball's rules make it important to consider which league each team played in? Note: you can google the differences between the American League and the National League to guide you.

```r
summary(fit.5)
```

Teams in the National League score an estimated 6 runs fewer per season than teams in the American League, given `X1B`, `X2B`, `X3B`, `HR`, `BB`, `SO`, and `CS` in the linear regression model. This isn't surprising - since 1973, the American League has featured a designated hitter, while that spot in the National League has the pitcher hitting.

Question 4

Interpret the R-squared from `fit.5`, and produce model checks to determine if the assumptions for linear regression are appropriate.

The R-squared is 92.28. This suggests that 92.28% of the variability in a team's runs scored in a single season can be attributed to this linear fit (the one with singles, doubles, … caught stealing.)

Question 5

The first team in the data set is the Atlanta Braves, who scored 736 runs in 1970. Using `fit.5`, estimate how many runs your model expected the Braves to score. Did the Braves outperform expectations (score more runs) or underperform expectations (score fewer runs)?

There are a few ways to do this. Here's the easiest.

```r
Teams.1 <- mutate(Teams.1, predict.R = predict(fit.5, Teams.1))
Teams.1[1,]
```

```
##   yearID lgID teamID franchID divID Rank   G Ghome  W  L DivWin WCWin
## 1   1970   NL    ATL      ATL     W    5 162    81 76 86      N  <NA>
##   LgWin WSWin   R   AB    H X2B X3B  HR  BB  SO SB CS HBP SF  RA  ER  ERA
## 1     N     N 736 5546 1495 215  24 160 522 736 58 34  NA NA 772 688 4.33
##   CG SHO SV IPouts   HA HRA BBA SOA   E  DP   FP           name
## 1 45   9 24   4290 1451 185 478 960 140 118 0.97 Atlanta Braves
##                             park attendance BPF PPF teamIDBR teamIDlahman45
## 1 Atlanta-Fulton County Stadium    1078848 106 106      ATL            ATL
##   teamIDretro  X1B predict.R
## 1         ATL 1096  734.3995
```

Our model projects the Braves to have scored 734.4 runs - in scoring 736, the Braves were pretty close to our expectation (slightly underperformed.)

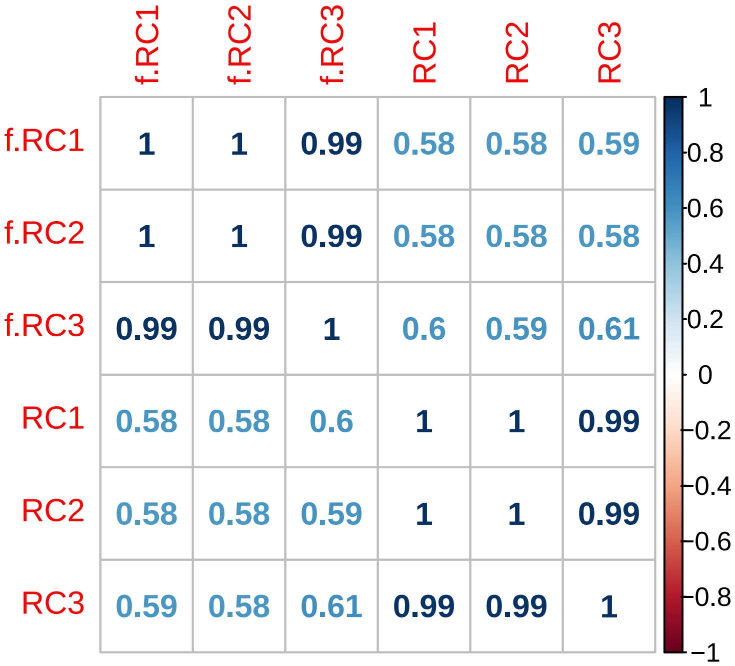
Part II

In this part, we'll use the Hitting data set, and explore the properties of runs created. In this code, we look at three formulas for runs created: RC1, RC2, and RC3.

```
Batting.1 <- Batting %>%
 filter(yearID >=1971, AB > 500) %>%
 mutate(X1B = H - X2B - X3B - HR,
        TB = X1B + 2*X2B + 3*X3B + 4*HR,
        RC1 = (H + BB)*TB/(AB + BB),
        RC2 = (H + BB - CS)*(TB + (0.55*SB))/(AB+BB),
        RC3 = ((H + BB - CS + HBP - GIDP)*(TB + (0.26*(BB - IBB + HBP))) +
          (0.52*(SH+SF+SB)))/(AB+BB+HBP+SH+SF))

Batting.2 <- Batting.1 %>%
  arrange(playerID, yearID) %>%
  group_by(playerID) %>%
  mutate(f.RC1 = lead(RC1), f.RC2 = lead(RC2), f.RC3 = lead(RC3)) %>%
  na.omit()

cor.matrix <- cor(select(ungroup(Batting.2),
    f.RC1, f.RC2, f.RC3, RC1, RC2, RC3),
    use="pairwise.complete.obs")
corrplot(cor.matrix, method = "number")
```



	f.RC1	f.RC2	f.RC3	RC1	RC2	RC3
f.RC1	1	1	0.99	0.58	0.58	0.59
f.RC2	1	1	0.99	0.58	0.58	0.58
f.RC3	0.99	0.99	1	0.6	0.59	0.61
RC1	0.58	0.58	0.6	1	1	0.99
RC2	0.58	0.58	0.59	1	1	0.99
RC3	0.59	0.58	0.61	0.99	0.99	1

Question 6

Which of the three runs created metrics more strongly correlates with its own future performance in the following year? Is that a good thing or a bad thing?

We like methods that are more repeatable: among the runs created formula, RC3 is most correlated to performance in the following year (0.61), although the other two formulas are close.

Question 7

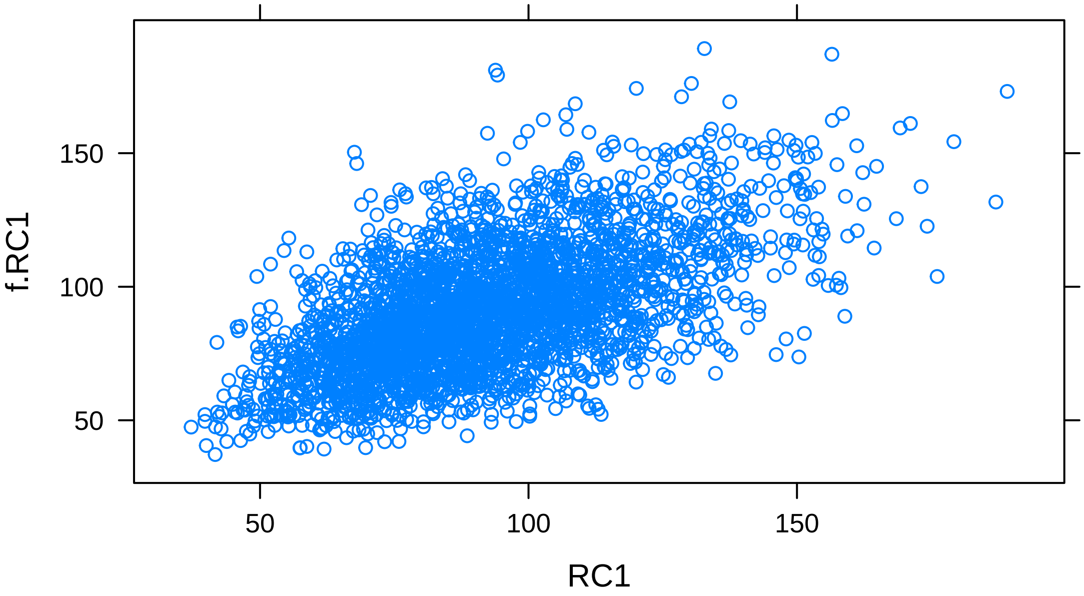
Which of the three runs created metrics more strongly correlates with `f.RC1`? What does this suggest?

Interestingly, `RC3` most strongly correlates with `RC1` in the following year (correlation = 0.59). This suggests that there's something in the formula for `RC3` (recall, its more complicated) that can generally predict offensive production. Alternatively, one could argue that because all three metrics predict `RC1` at roughly the same level, perhaps the additional complications of the runs created formula are not worth it on an individual level.

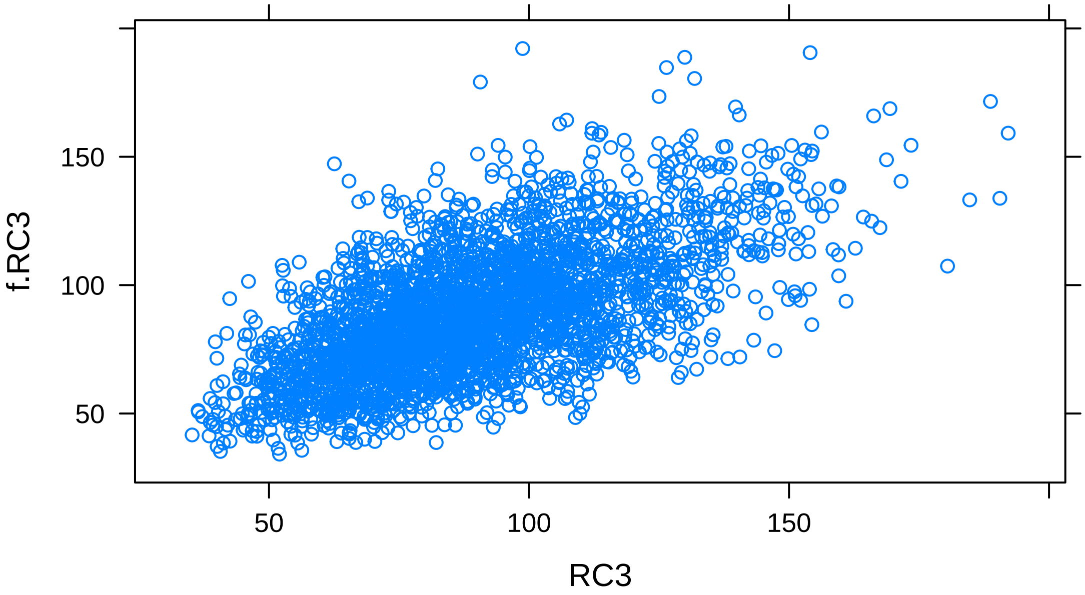
Question 8

Make a pair of scatter plots: `f.RC1` versus `RC1`, and `f.RC3` versus `RC3`. Is the difference in correlations between each pair of variables obvious in the scatter plots?

```
xyplot(f.RC1 ~ RC1, data= Batting.2)
```



```
xyplot(f.RC3 ~ RC3, data= Batting.2)
```



The strength of the association between each pair of variables looks similar in the plots.

Question 9

In place of correlation, a useful tool for measuring predictive accuracy is mean absolute error (MAE). In R, the MAE between two variables of the same length can be calculated as follows. Calculate the MAE between between each of the following three variable pairs: `RC1` and `f.RC1`, `RC2` and `f.RC2`, and `RC3` and `f.RC3`.

```
x <- Batting.2$RC1
y <- Batting.2$f.RC1
MAE <- mean(abs(x-y))
MAE
```

```
## [1] 16.57764
```

```
x <- Batting.2$RC2
y <- Batting.2$f.RC2
MAE <- mean(abs(x-y))
MAE
```

```
## [1] 16.65493
```

```
x <- Batting.2$RC3
y <- Batting.2$f.RC3
MAE <- mean(abs(x-y))
MAE
```

```
## [1] 16.57694
```

There are different ways to code this question (one is presented above). `RC3` boasts the lowest MAE, although the differences in MAE between `RC3` and the other runs created formulas are small: 16.578, 16.655, and 16.577, respectively.

Question 10

Interpret the MAE between `RC1` and `f.RC1`. Is there a noticeable difference between your MAE's found in question 9? What does this suggest about the more complicated `RC3` formula?

The average player's `RC1` in the following year will be about 16.5 runs from his `RC1` in the previous year. There does not appear to be a practical difference between the runs created formula - on an individual level, at least, suggesting the more complex formula may not be helpful as far as repeatability.