Simple Linear Regression and Run Differential

Ryan Jung

2/03/2023

```
library(tidyverse)
library(Lahman)
library(broom)
```

Summary: In these examples I fit two simple linear regression models to demonstrate baseball is a game of offense, pitching, and defense. Data is from the Lahman database, which contains statistics for Major League Baseball from 1871 through 2021.

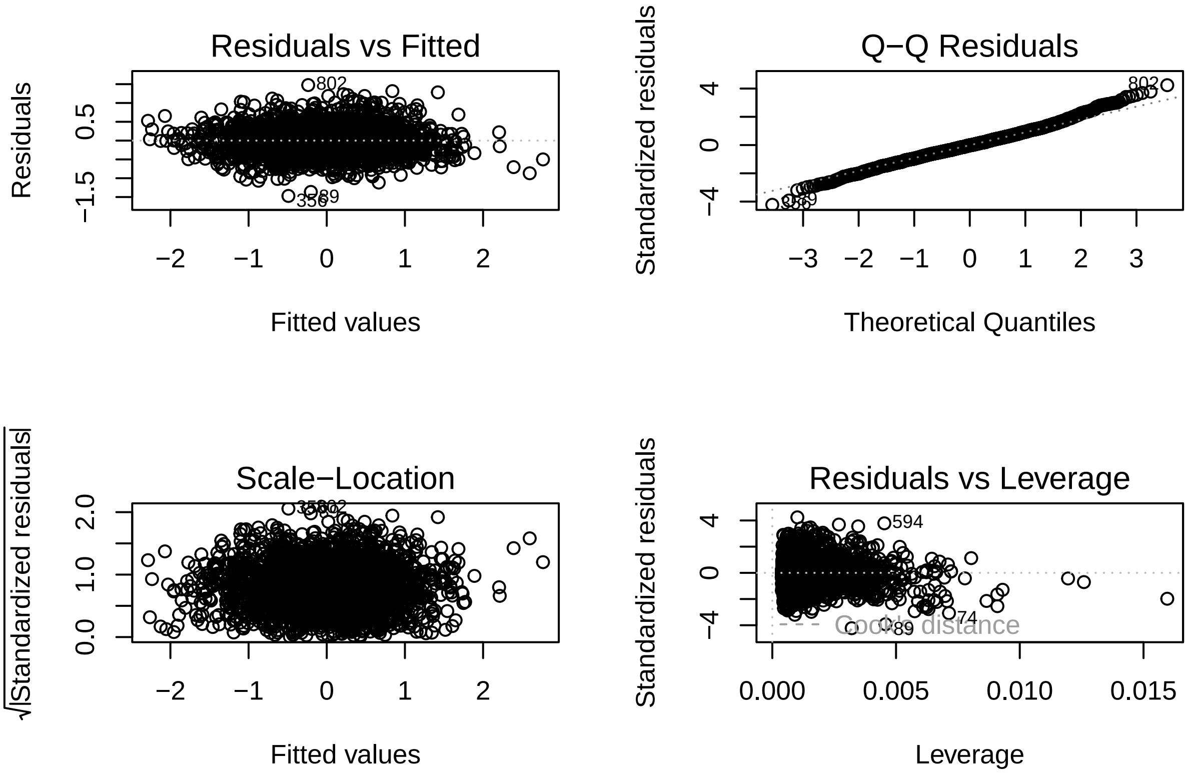
First, I fit a model with RD as the response and OPS, WHIP, and FP as our explanatory variables. Prior to that, I have to create some variables that are not explicitly included in the Lahman database, namely rate stats such as OBP and SLG.

```
dat <- Teams %>%
  select(yearID, franchID, G, W, L, AB, H, X2B, X3B, HR, BB, HBP, SF,HA, HRA,
         BBA, SOA, IPouts, FP, R, RA) %>%
  filter(yearID >= 1900) %>%
  replace_na(list(HBP = 0, SF = 0)) %>%
  mutate(X1B = H - (X2B + X3B + HR)) %>%
  mutate(OBP = (H + BB + HBP)/(AB + BB + HBP + SF)) %>%
  mutate(SLG = (X1B + 2*X2B + 3*X3B + 4*HR)/AB) %>%
  mutate(OPS = OBP + SLG) %>%
  mutate(WHIP = 3*(HA + BBA)/IPouts) %>%
  mutate(RD = (R - RA)/G)
```

```
m <- lm(RD ~ OPS + WHIP + FP, data = dat)
summary(m)
```

```
##
## Call:
## lm(formula = RD ~ OPS + WHIP + FP, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4707 -0.2169 -0.0061  0.2103  1.4758
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.71282    0.78206   17.53   <2e-16 ***
## OPS          11.79734    0.15538   75.92   <2e-16 ***
## WHIP         -5.40521    0.06403  -84.42   <2e-16 ***
## FP          -15.19191    0.83208  -18.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3484 on 2636 degrees of freedom
## Multiple R-squared:  0.7891, Adjusted R-squared:  0.7889
## F-statistic:  3288 on 3 and 2636 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c (2,2))
plot(m, add.smooth = FALSE)
```



The model performs reasonably well, and doesn't have any glaring issues.

This model can be improved through a principled rescaling of OPS, WHIP, and FP. To do this, I first have to create league average variables for each season, then use those league averages to scale each individual team's performance.

```
scaled = dat %>%
  group_by(yearID) %>%
  mutate(avgOBP = sum(H + BB + HBP)/sum(AB + BB + HBP + SF)) %>%
  mutate(avgSLG = sum(X1B + 2*X2B + 3*X3B + 4*HR)/sum(AB)) %>%
  mutate(avgOPS = avgOBP + avgSLG) %>%
  mutate(avgWHIP = 3*sum(HA + BBA)/sum(IPouts)) %>%
  mutate(OPSscaled = OPS/avgOPS) %>%
  mutate(WHIPscaled = avgWHIP/WHIP) %>%
  mutate(FPscaled = mean(FP)/FP)
```

```r
m3 = lm(RD ~ OPSscaled + WHIPscaled + FPscaled, data = scaled)
summary(m3)
```

```
##
## Call:
## lm(formula = RD ~ OPSscaled + WHIPscaled + FPscaled, data = scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06459 -0.17447  0.00401  0.17717  0.96594
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.80111    1.66971   4.073 4.77e-05 ***
## OPSscaled     8.97239    0.10604  84.612  < 2e-16 ***
## WHIPscaled    7.01191    0.08715  80.455  < 2e-16 ***
## FPscaled    -22.80770    1.61081 -14.159  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2666 on 2636 degrees of freedom
## Multiple R-squared:  0.8766, Adjusted R-squared:  0.8764
## F-statistic:  6239 on 3 and 2636 DF,  p-value: < 2.2e-16
```

This model performs much better because of different league environments. Some eras of baseball had more or less run scoring than others, so predicting RD based on raw rate stats over the history of baseball can over/underestimate run differential depending on the balance between pitching and hitting in any given season. Additionally, by scaling each team’s performance by the league average, we gain context about how a team performed relative to other teams in the same season.