The Relation Between Runs and Wins

Kevin Cummiskey

January 28, 2020

Introduction

In 2002, the Oakland Athletics were losing three star players (Johnny Damon, Jason Giambi, and Jason Isringhausen) to free agency. Their general manager, Billy Beane, had a limited budget to work with. He wanted to find players who were effective at producing runs but were undervalued in the market. At the time, teams used mostly traditional statistics like batting average (BA), home runs (HR), and runs batted in (RBI) to value players. However, evidence from sabermetrics suggested there are better statistics for determining players' contributions to run scoring. The book *Moneyball* by Michael Lewis and the film by the same name follow the 2002 Oakland A's.

Put yourself in the position of Billy Beane. How would you approach this problem?

How many wins to make the playoffs?

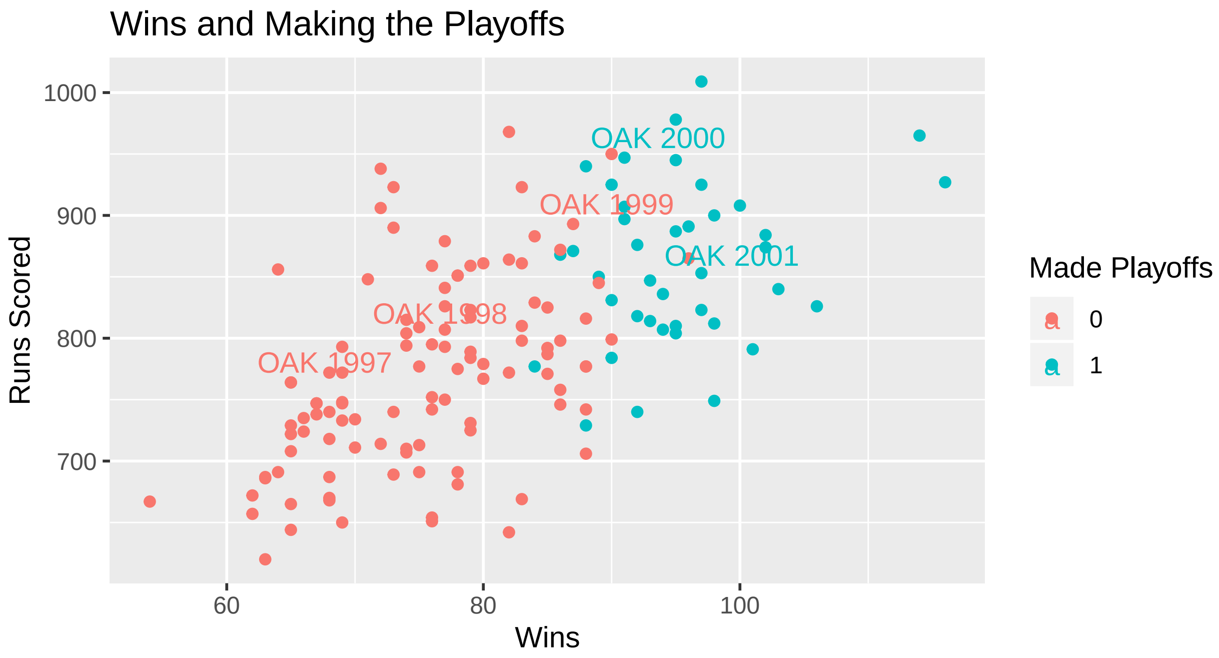
Let's see historically how many wins it takes to make the playoffs. Below, we will look at the five seasons (1997-2001) preceeding the 2002 season. First, we need to add a variable `playoff` to the Teams database which is an indicator of whether the team made the playoffs (1) or did not make the playoffs (2).

```r
library(Lahman)
library(tidyverse)
library(ggrepel)

#get teams in the 1997-2001 season
my_teams <- Teams %>%
  filter(yearID >= 1997, yearID <= 2001)

#determine which teams made the playoffs each year
#Use the SeriesPost data
my_series <- SeriesPost %>%
  filter(yearID >= 1997, yearID <= 2001)%>%
  select(yearID, teamIDwinner, teamIDloser) %>%
  gather(key = "result", value = "teamID", -yearID) %>%
  select(-result) %>%
  mutate(playoffs = 1) %>%
  unique()

#Merge with my_teams and replace NAs
my_teams = my_teams %>%
  left_join(my_series, by = c("yearID", "teamID")) %>%
  replace_na(list(playoffs = 0))

#let's check if we did this correctly
# there should be eight playoff teams each year
my_teams %>%
  group_by(yearID) %>%
  summarize(teams.playoffs = sum(playoffs))
```

```
## # A tibble: 5 x 2
##   yearID teams.playoffs
##    <int>          <dbl>
## 1   1997              8
## 2   1998              8
## 3   1999              8
## 4   2000              8
## 5   2001              8
```

Next, let's look at the data.

```
my_teams %>%
  ggplot(aes(x = W, y = R, color = factor(playoffs))) +
  geom_point() +
  labs(x = "Wins", y = "Runs Scored", color = "Made Playoffs",
       title = "Wins and Making the Playoffs") +
  geom_text_repel(data = filter(my_teams, teamID == "OAK"),
                  aes(x = W, y = R,label = paste(teamID,yearID)))
```



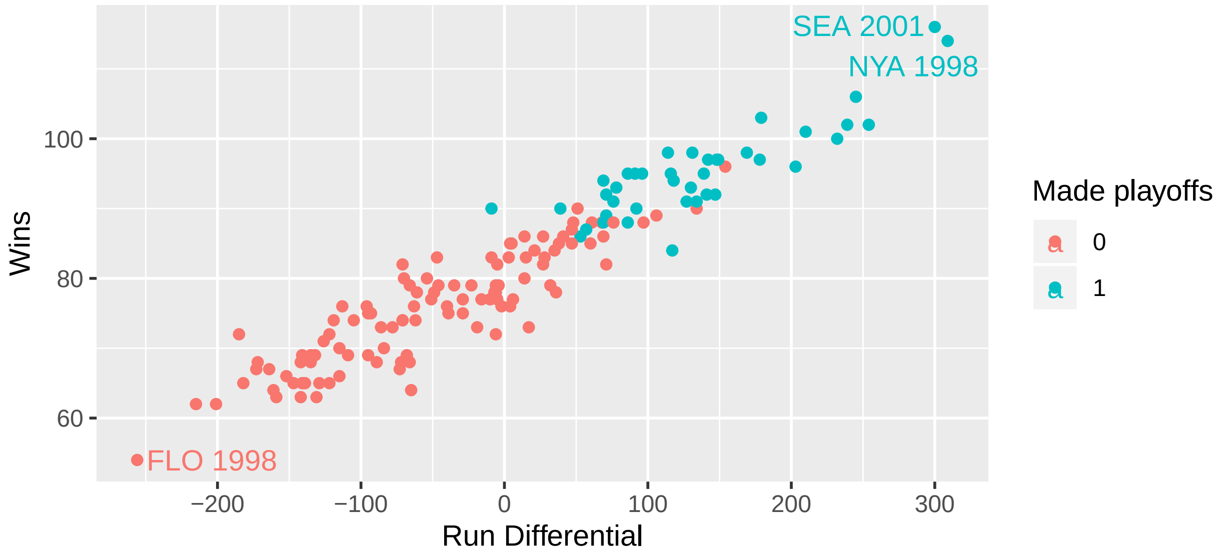
Based on the figure above, if you were Billy Beane, how many wins should you shoot for in order to have a good chance of making the playoffs?

In the previous question, you came up with a number of wins by inspecting the figure. What other methods have you learned that you could use?

Run differential and wins

Now that we have a goal for the number of wins for our season, we will look at how much we have to outscore our opponents over the course of a season to achieve that number of wins. Towards this end, we will calculate the run differential and use linear regression.

```
#calculate the run differential
my_teams = my_teams %>%
  mutate(RD = R - RA)

my_teams %>%
  ggplot(aes(x= RD,y = W, color = factor(playoffs))) +
  geom_point() +
  geom_text_repel(data = filter(my_teams, W > 110 | W < 60),
                  aes(label = paste(teamID,yearID))) +
  labs(x = "Run Differential", y = "Wins",
       color = "Made playoffs")
```



What run differential would you recommend to feel confident of getting the number of wins you want?

Let's look further at the relationship between run difference and wins with linear regression. Here is the model we will fit:

$$Wins_i = \beta_0 + \beta_1 RD_i + \epsilon_i \quad \epsilon_i \sim \text{Normal}(0, \sigma^2)$$

where $Wins_i$ and RD_i are the wins and run differential for team i.

Interpret the coefficients in this model.

```
#fit a linear regression model
lin.fit = lm(W ~ RD, data = my_teams)
summary(lin.fit)
```

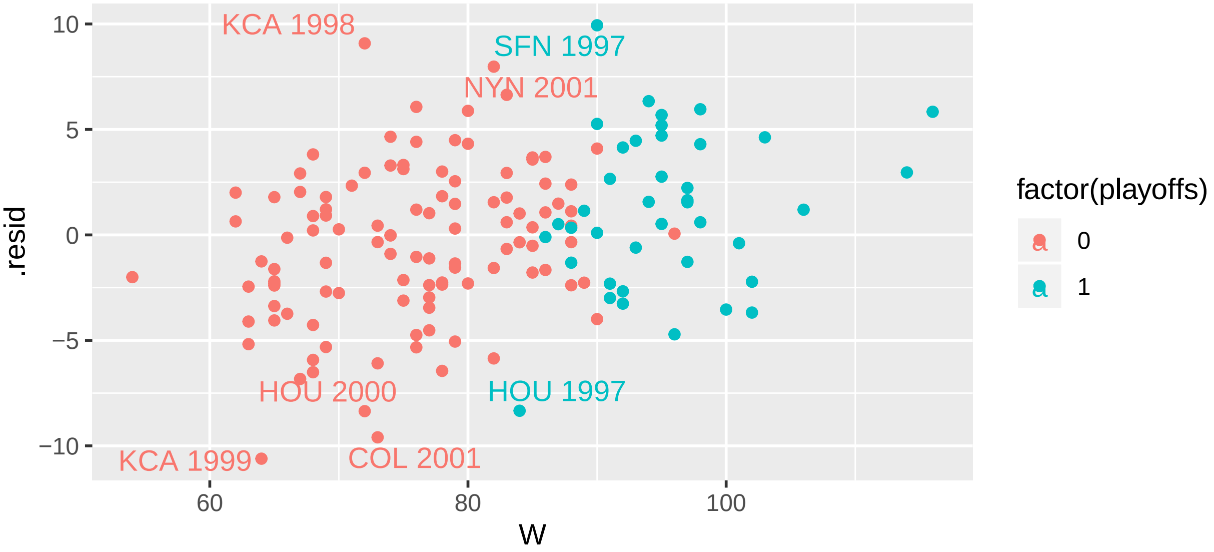
```
##
## Call:
## lm(formula = W ~ RD, data = my_teams)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6075  -2.3612   0.2371   2.4585   9.9375
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 80.939189   0.305841  264.64   <2e-16 ***
## RD           0.097411   0.002693   36.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.721 on 146 degrees of freedom
## Multiple R-squared:  0.8996, Adjusted R-squared:  0.899
## F-statistic:  1309 on 1 and 146 DF,  p-value: < 2.2e-16
```

Calculate the run differential that would typically result in one additional win.

Based on the regression model, do you feel confident in your recommended run differential? Explain how you decided.

Let's look at the residuals for this regression. The broom package helps get model results.

```
library(broom)

my_teams = augment(lin.fit, data = my_teams)

my_teams %>%
  ggplot(aes(x = W, y = .resid, color = factor(playoffs))) +
  geom_point() +
  geom_text_repel(data = filter(my_teams, abs(.resid) > 7.5),
                  aes(label = paste(teamID,yearID)))
```



What units are the residuals in?

What do the residuals tell us about our model?

Above, we used run differential to predict number of wins. Typically, we use win percentage instead of wins. Why?

Repeat the above analysis using win percentage instead of wins. Report the run differential that typically results in one additional win.

How should we proceed next with this analysis?