MA388 Sabermetrics: Lesson 10

Value of Plays - Base Hits and Stealing

LTC Jim Pleuss

Review

Last class, we discussed the Run Expectancy Matrix and how to obtain it from the Retrosheet play-by-play data.

```
library(tidyverse)
library(knitr)
library(Lahman)

erm_2011_wide <- read_csv(file = "erm_2011_wide.csv")

kable(erm_2011_wide)
```

bases	Outs=0	Outs=1	Outs=2
000	0.4711649	0.2546956	0.0971845
001	1.4543568	0.9374359	0.3173913
010	1.0582804	0.6501976	0.3091673
011	1.9304897	1.3388641	0.5407407
100	0.8350992	0.4960492	0.2179536
101	1.7526998	1.1496169	0.4882597
110	1.4144549	0.8739176	0.4222569
111	2.1718266	1.4745146	0.7610094

Interpret the value in the first row, third column.

True or False: The probability of scoring at least one run in the remainder of the inning is 0.25 when there is no one on base with one out.

Let's say there is a runner on second base (and no other baserunners). Stealing third base is the most valuable when there are how many outs?

Let's say there is a runner on second base with one out. Calculate the expected value of a stolen base attempt if the probability of successfully stealing third base is 70%. In addition, calculate the minimum probability of successfully stealing third base required to make it a better strategy in terms of expected runs.

Measuring the Success of a Batting Play

We can estimate the value of a plate appearance as the difference in run expectancy between the new and old states plus the number of runs scored on the play.

[RUN VALUE = E[RUNS]*new state* - *E[RUNS]*old state + RUNS_scored on play]

Calculate the RUN VALUE of the following plays:

(1) There are runners on first and second with no outs. The batter successfully sacrifice bunts, resulting in runners on second and third with one out.

(2) There is a runner on second base with no outs. The batter hits a single that scores the runner from second. The batter stops at first base.

(3) What is the most valuable play in baseball? Explain.

Measuring Batting Success Over a Season: David Ortiz

We'll start with all batting events in the 2011 season.

```
# Load the data.
site = "https://raw.githubusercontent.com/maxtoki/baseball_R/"
fields <- read_csv(file = paste(site, "master/data/fields.csv", sep =""))
retro2011 <- read_csv(file = paste(site, "master/data/all2011.csv", sep = ""),
                      col_names = pull(fields, Header),
                      na = character())
colnames(retro2011) <- tolower(colnames(retro2011))

# Add states and run values.
# (Note the data set has to be called retro2011.)
source("./RunExpectancyMatrix.R")
```

Now let's look specifically at David Ortiz.

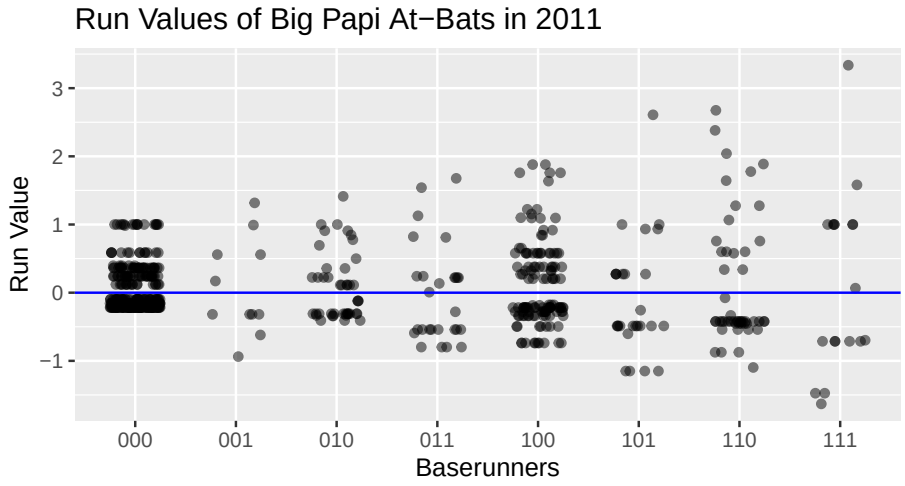
```
# Get Ortiz's playerID.
ortizID <- People |>
  filter(nameFirst == "David", nameLast == "Ortiz") |>
  pull(retroID)

ortiz_df <- retro2011 |>
  filter(bat_id == ortizID,
         bat_event_fl == TRUE)

ortiz_df |>
  select(game_id, inn_ct, state, new_state, run_value) |>
  head(10) |>
  kable()
```

game_id	inn_ct	state	new_state	run_value
ANA201104210	2	000 0	010 0	0.5871155

game_id	inn_ct	state	new_state	run_value
ANA201104210	4	000 0	000 1	-0.2164694
ANA201104210	6	000 0	100 0	0.3639342
ANA201104210	7	110 2	110 3	-0.4222569
ANA201104210	9	000 2	100 2	0.1207691
ANA201104220	2	000 0	000 1	-0.2164694
ANA201104220	4	000 1	000 2	-0.1575110
ANA201104220	5	101 2	101 3	-0.4882597
ANA201104220	7	010 2	010 3	-0.3091673
ANA201104230	2	000 0	000 1	-0.2164694

Here's a plot of Ortiz's run values by runner state.

```
ortiz_df |>
  ggplot(aes(x = bases, y = run_value)) +
  geom_jitter(width = 0.25, alpha = 0.5) +
  geom_hline(yintercept = 0, color = "blue") +
  labs(x = "Baserunners",
       y = "Run Value",
       title = "Run Values of Big Papi At-Bats in 2011")
```



Next, let's look at total runs in different baserunner situations.

```
runs_ortiz <- ortiz_df |>
  group_by(bases) |>
  summarize(runs = sum(run_value) |> round(1),
            PA = n())
```

```
runs_ortiz |> kable()
```

bases	runs	PA
000	13.1	301
001	0.8	11
010	3.4	44
011	-0.9	27
100	14.9	127
101	-2.2	27
110	4.4	50
111	2.1	18

Lastly, we can calculate a statistic called *RE24* which sums runs over the different baserunner combinations.

```
runs_ortiz |>
  summarize(RE24 = sum(runs))
```

```
# A tibble: 1 x 1
   RE24
  <dbl>
1  35.6
```

How do we interpret RE24?

Come up with a modification for RE24 that could be used to rank batters according to what they produced given the opportunity they had. How would you do it?

Part II - Measuring the Success of a Stolen Base Attempt

Your book dedicates roughly three pages to evaluating base stealing. We're going to discuss it briefly and return to our evaluation of batting performance. Your task: Develop an RE24-inspired measure to rank MLB's top base stealers.

Let's start with the same approach to valuing an individual event.

[RUN VALUE = RUNS*new state* - *RUNS*old state + RUNS_scored on play]

where $\text{RUNS}_{\text{new state}}$ and $\text{RUNS}_{\text{old state}}$ are their respective entries in the run expectancy matrix and $\text{RUNS}_{\text{scored on play}}$ is the number of runs that scored on the play.

```
kable(erm_2011_wide)
```

bases	Outs=0	Outs=1	Outs=2
000	0.4711649	0.2546956	0.0971845
001	1.4543568	0.9374359	0.3173913
010	1.0582804	0.6501976	0.3091673
011	1.9304897	1.3388641	0.5407407
100	0.8350992	0.4960492	0.2179536
101	1.7526998	1.1496169	0.4882597
110	1.4144549	0.8739176	0.4222569
111	2.1718266	1.4745146	0.7610094

Let's say there is a runner on first base with no outs and you are considering bunting. Historically, a bunt in this situation would result in a runner on second with one out 75% of the time, a runner on first and second with no outs 10% of the time, a runner on first with one out 10% of the time, and runners on second and third with no outs 5% of the time. If you have an average hitter at bat, does bunting or hitting away have the highest expected run value?

For more on bunting, see our textbook author's blog.

RE24

Earlier, we investigated the run value of David Ortiz's 2011 plate appearances. We can aggregate these run values for each player to a single statistic called "RE24".

If we sum all players RE24 together, what should be the result?

In baseball, there are context-dependent and context-independent statistics. *Name some examples of each. Which type of statistic is RE24?*

Calculate the RE24, plate appearances, and total starting run expectancy for each player in 2011.

```
re24_2011 <- retro2011 |>
  group_by(bat_id) |>
  summarize(re24 = sum(run_value),
            PA = length(run_value),
            runs_start = sum(rv_start)) |>
  arrange(-re24)

# Add names from the People data frame (note it's not playerID, but retroID).
re24_2011 <- re24_2011 |>
  left_join(select(People, retroID, nameLast, nameFirst),
            by = c("bat_id" = "retroID")) |>
  mutate(name = paste(nameFirst, nameLast, sep = " ")) |>
  select(-nameLast, -nameFirst)

re24_2011 |>
  head(15) |>
  kable(digits = 1)
```

bat_id	re24	PA	runs_start	name
cabrm001	73.9	701	332.9	Miguel Cabrera
bautj002	73.6	693	332.8	Jose Bautista
fielp001	65.6	734	363.6	Prince Fielder
vottj001	64.5	743	343.3	Joey Votto
braur002	61.9	663	316.3	Ryan Braun
kempm001	60.1	704	349.1	Matt Kemp
ellsj001	58.1	748	347.4	Jacoby Ellsbury
gonza003	57.3	755	367.3	Adrian Gonzalez
berkl001	55.4	600	289.4	Lance Berkman
martv001	49.3	606	318.2	Victor Martinez
canor001	49.1	695	335.9	Robinson Cano
napom001	46.6	444	215.0	Mike Napoli
granc001	44.1	727	362.9	Curtis Granderson

bat_id	re24	PA	runs_start	name
teixm001	41.2	718	347.9	Mark Teixeira
gorda001	41.1	727	344.3	Alex Gordon

First, let's see if the average is about zero.

```
re24_2011 |>
  summarize(mean_RE24 = mean(re24))
```

```
# A tibble: 1 x 1
  mean_RE24
      <dbl>
1     0.411
```

Now, let's plot RE24 for each player versus starting runs.

```
library(ggrepel)

# Create data frame of points to label.
labels <- re24_2011 |>
  filter(PA >= 502) |>
  filter(re24 > 50 | re24 < -30)


re24_2011 |>
  filter(PA >= 502) |>
  ggplot(aes(x = runs_start, y = re24)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  geom_text_repel(data = labels, aes(x = runs_start,
                                     y = re24,
                                     label = name)) +
  labs(title = "Player RE24 vs. Opportunity",
       x = "Cumulative Starting Run Value for All At-Bats",
       y = "RE24")
```

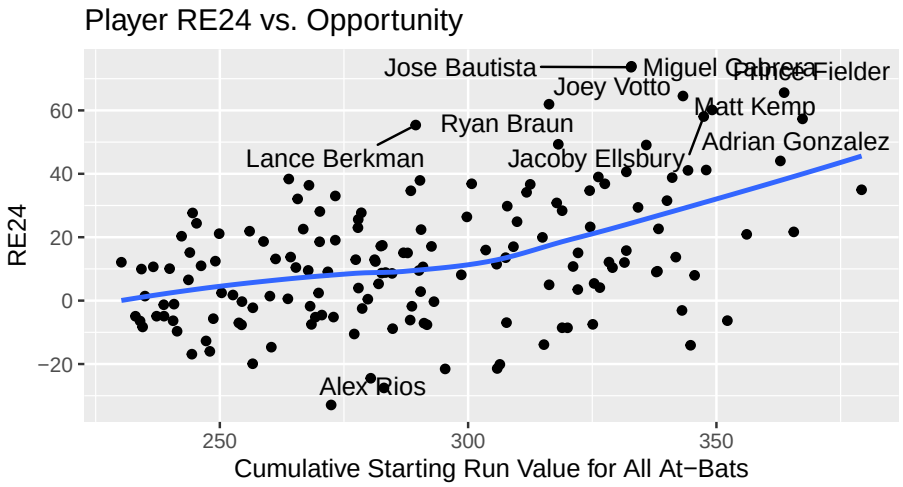


Figure 1: RE24 versus starting runs for all players with at least 502 plate appearances (2011)

Briefly summarize what you would conclude from this chart.

Value of Hits

Batting Average - Traditionally, baseball analysts and fans used batting average to assess players.

$$AVG = \frac{H}{AB}$$

where H is the number of hits and AB is the number of at bats.

Slugging Percentage - Alternatively, slugging percentage weights the different types of hits by the number of bases the hitter gets.

$$SLG = \frac{X_{1B} + 2 \times X_{2B} + 3 \times X_{3B} + 4 \times HR}{AB}$$

Discuss strengths and limitations of batting average and slugging percentage in measuring batting performance.

How can we use run values to improve these statistics?

```
# Event Code (variable: EVENT_CD)
# Single: 20
# Double: 21
# Triple: 22
# Home Run: 23

# Create a look-up table to add event names (single, etc) to the data set.
codes <- data.frame(event_cd = 20:23,
```

```
                    hit_type = factor(c("single", "double", "triple", "home run"),
                                      levels = c("single", "double", "triple", "home run")))

codes |>
  kable(caption = "Event codes for each hit type.")
```

Table 6: Event codes for each hit type.

event_cd	hit_type
20	single
21	double
22	triple
23	home run

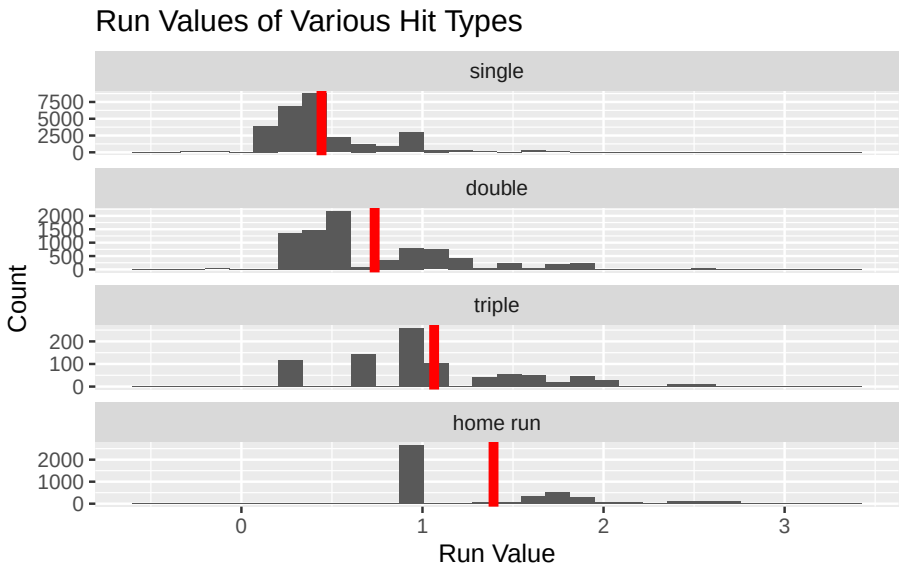
```
# Calculate mean hit values.
mean_hit_values <- retro2011 |>
  filter(event_cd %in% 20:23) |>
  left_join(codes) |>
  group_by(hit_type) |>
  summarize(mean_value = mean(run_value))

mean_hit_values |>
  kable(digits = 2, caption = "Mean run value for each hit type (2011).")
```

Table 7: Mean run value for each hit type (2011).

hit_type	mean_value
single	0.44
double	0.74
triple	1.06
home run	1.39

```
# Plot run values.
retro2011 |>
  filter(event_cd %in% 20:23) |>
  left_join(codes) |>
  ggplot(aes(x = run_value)) +
  geom_histogram() +
  geom_vline(data = mean_hit_values, aes(xintercept = mean_value), col = "red", size = 2) +
  facet_wrap(~ hit_type, ncol = 1, scales = "free_y") +
  labs(title = "Run Values of Various Hit Types",
```

```
       x = "Run Value",
       y = "Count")
```



Based on these results, what would you conclude about AVG and SLG?

Let's see if we can find players whose slugging percentage is *not* a good predictor of RE24.

```
# Add slugging percentage to the RE24 data frame.
re24_2011 <- Batting |>
  filter(yearID == 2011) |>
  group_by(playerID) |>
  summarize(H = sum(H),
            X2B = sum(X2B),
            X3B = sum(X3B),
            HR = sum(HR),
            AB = sum(AB)) |>
  mutate(X1B = H - X2B - X3B - HR,
         SLG = (X1B + 2*X2B + 3*X3B + 4*HR)/AB) |>
  left_join(select(People, playerID, retroID)) |>
  right_join(re24_2011, by = c("retroID" = "bat_id"))
```

(Random aside: As an alternative to a figure title, consider labeling the figure in a caption. Professional publications often prefer a caption that provides a title and description for a well-labeled figure.)

```
re24_2011 |>
  filter(AB > 400) |>
  ggplot(aes(x = SLG, y = re24)) +
  geom_point() +
  labs(x = "Slugging Percentage",
       y = "RE24")
```

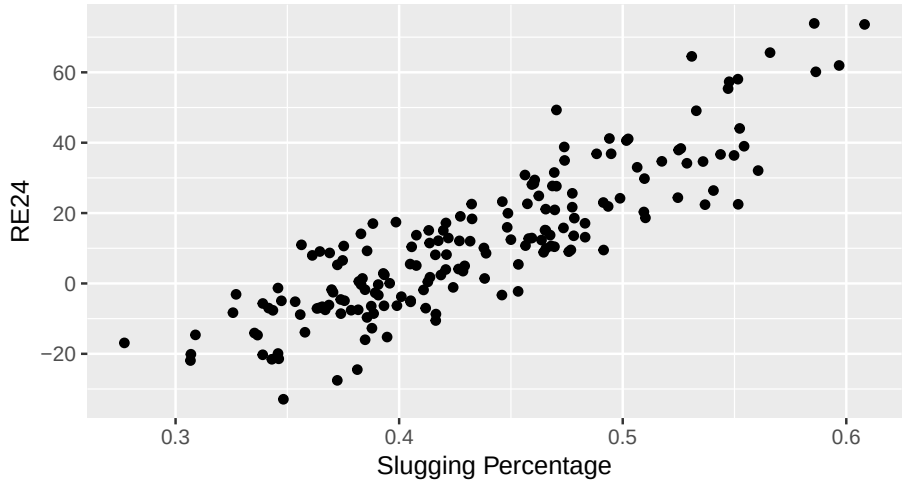


Figure 2: RE24 versus slugging percentage

```
re24_2011 |>
  filter(AB > 400) |>
  arrange(-SLG) |>
  select(name, H, AB, PA, X1B, X2B, X3B, HR, SLG, re24, runs_start) |>
  head(10) |>
  kable()
```

name	H	AB	PA	X1B	X2B	X3B	HR	SLG	re24	runs_start
Jose Bautista	155	513	693	86	24	2	43	0.608187	173.6170	2332.8156
Ryan Braun	187	563	663	110	38	6	33	0.596802	861.9475	8316.2900
Matt Kemp	195	602	704	119	33	4	39	0.586378	760.1466	0349.1314

name	H	AB	PA	X1B	X2B	X3B	HR	SLG	re24	runs_start
Miguel Cabrera	197	572	701	119	48	0	30	0.585664	73.90067	332.8927
Prince Fielder	170	569	734	95	36	1	38	0.565905	165.6006	363.6424
Adrian Beltre	144	487	543	79	33	0	32	0.560574	93.08008	265.6730
David Ortiz	162	525	627	92	40	1	29	0.554285	38.99476	326.2239
Curtis Granderson	153	583	727	76	26	10	41	0.552315	144.0581	362.8833
Pablo Sandoval	134	426	476	82	26	3	23	0.551643	22.49106	217.8158
Jacoby Ellsbury	212	660	748	129	46	5	32	0.551515	258.0531	347.4389