RUNNING A REGRESSION

The regression shown in Figure 3-1 predicts team runs scored from a team's singles, doubles, triples, HRs, BBs + HBPs, SBs, and CSs.

To run the regression, first go to the sheet Data of the workbook Ch3Data.xlsx. In Excel, bring up the Analysis Toolpak by selecting the Data Tab and then Choosing Data Analysis from the right-hand portion of the tab.

Now select the Regression option and fill in the dialog box as shown in Figure 3-7.

This tells Excel we want to predict the team runs scored (in cell range C2:C211) using the independent variables in cell range D2:J211 (singles, doubles, triples, HRs, BBs + HBPs, SBs, and CSs). We checked the Labels box so that our column labels shown in row 1 will be included in the regression output. The output (as shown in Figure 3-1) will be placed in the worksheet MLR.

CHAPTER 4

EVALUATING HITTERS BY MONTE CARLO SIMULATION

In Chapters 2 and 3 we showed how to use Runs Created and Linear Weights to evaluate a hitter's effectiveness. These concepts were primarily developed to "fit" the relationship between Runs Scored by a team during a season and team statistics such as BB, singles, doubles, triples, and HRs. We pointed out that for players whose event frequencies differ greatly from typical team frequencies these metrics might do a poor job of evaluating a hitter's effectiveness.

A simple example (described by famed *USA Today* sports statistician Jeff Sagarin) will show how Runs Created and Linear Weights can be very inaccurate. Consider a player (let's call him Joe Hardy after the hero of the wonderful movie and play *Damn Yankees!*) who hits an HR during 50% of his plate appearances and makes an out during the other 50% of his plate appearances. Since Joe hits as many HRs as he makes outs, you would expect Joe "on average" to alternate HR, out, HR, out, HR, out and average three runs per inning. In the Appendix to Chapter 6 we will use the principle of conditional expectation to give a mathematical proof of this result.

	K	L	M	N	O	P
3	Method	At Bats	Home Runs	Outs	Runs Created	Runs Created Per Game
4	Bill James	8748	4374	4374	8748	54
5	Linear Weights	8748	4374	4374	5957.26	36.77321

FIGURE 4.1 Runs Created and Linear Weights Predicted Runs per Game for Joe Hardy.

In 162 nine-inning games our Joe Hardy will make on average $162 * 27 = 4{,}374$ outs and hit 4,374 home runs. As shown in Figure 4-1 (see fileSimulationmotivation.xlsx), we find that runs created predicts that Joe Hardy would generate 54 runs per game (or six per inning) and linear weights predicts Joe Hardy would generate 36.77 runs per game (or 4.01 runs per inning). Both these estimates are far away from the true value of 27 runs per game!

INTRODUCTION TO MONTE CARLO SIMULATION

How can we show that our player generates three runs per inning or 27 runs per game? By programming the computer to play out many innings and averaging the number of runs scored per inning. Developing a computer model to repeatedly play out an uncertain situation is called Monte Carlo simulation.

Physicists and astronomers use Monte Carlo simulation to simulate the evolution of the universe. Biologists use Monte Carlo simulation to simulate the evolution of life on earth. Corporate financial analysts use Monte Carlo simulation to evaluate the likelihood that a new GM car model or a new Proctor and Gamble shampoo will be profitable. Wall Street "rocket scientists" use Monte Carlo simulation to price exotic or complex financial derivatives. The term *Monte Carlo simulation* was coined by the Polish born physicist Stanislaw Ulam, who used Monte Carlo simulation in the 1930s to determine the chance of success of the chain reaction needed for an atom bomb to successfully detonate. Ulam's simulation was given the military

code name Monte Carlo, and the name Monte Carlo simulation has been used ever since.

How can we play out an inning? Simply flip a coin and assign a toss of heads to an out and a toss of tails to a home run. Or we could draw a card from a deck of cards and assign a red card to an out and a black card to an HR. Either the coin tossing or the card drawing method will assign a .5 chance to a home run and a .5 chance to an out. We keep flipping the coin or drawing a card (with replacement) until we obtain three outs. Then we record the number of HRs. We repeat this procedure 1,000 or so times and average the number of runs scored per inning. This average should closely approximate the average runs per inning scored by our hypothetical player. We will get very close to 3,000 total runs, which yields an estimate of three runs per inning. We choose to implement our simple Monte Carlo simulation using Microsoft Excel. See Figure 4-2 and file Simulationmotivator.xls. Excel contains a function RAND( ). If you type = *RAND( )* in any cell and hit the F9 key the number in the cell will change. The RAND( ) function yields any number between 0 and 1 with equal probability. This means, for example, that half the time RAND( ) yields a number between 0 and .5 and half the time RAND( ) yields a number between .5 and 1. The results generated by the RAND( ) function are called random numbers. Therefore, we can simulate an inning for our player by assigning an outcome of an HR to a random number less than or equal to .5 and assigning an outcome of an out to a random number between .5 and 1. By hitting F9 in spreadsheet Simulationmotivator.xls we can see the results of a simulated inning. See Figure 4-2. For our simulated inning, each random number $\leq .5$ yielded an HR and any other random number yielded an out. For our simulated inning five runs were scored.

Cells J6:J1005 contain the results of 1,000 simulated innings while cell J3 contains the average runs per inning generated during our 1,000 hypothetical innings. (Note that rows 17 to 1,002 are hidden.) The chapter appendix explains how we used Excel's Data Table feature to perform our simulation 1,000 times. Whenever you hit F9

	B	C	D	E	F	G	H
2	Batter	Random Number	Result	Outs	Runs	Over?	Total Runs
3	1	0.9732	out	1	0	no	5
4	2	0.2423	HR	1	1	no	
5	3	0.7489	out	2	1	no	
6	4	0.3429	HR	2	2	no	
7	5	0.4219	HR	2	3	no	
8	6	0.0333	HR	2	4	no	
9	7	0.0767	HR	2	5	no	
10	8	0.9828	out	3	5	yes	
11	9					yes	
12	10					yes	
13	11					yes	
14	12					yes	
15	13					yes	
16	14					yes	

FIGURE 4.2 Simulating One Inning for Joe Hardy.

you will see cell J3 is very close to 3, sometimes a little lower and sometimes a little higher, indicating that our player will generate around three runs per inning or 27 runs per game (not 54 runs per game as Runs Created predicts).

SIMULATING RUNS SCORED BY A TEAM OF NINE TROUTS

Buoyed by the success of our simple simulation model, how could we simulate the number of runs that would be scored by a team of, say, nine Mike Trout 2016s? We need to follow through the progress of an inning and track the runners on base, runs scored, and number of outs. In our model the events that can occur on each plate appearance are displayed in Figure 4-3.

- We assume each error advances all base runners a single base.
- A long single advances each base runner two bases.
- A medium single scores a runner from second base but advances a runner on first one base.

	C	D
13		Event
14	1	Strikeout
15	2	Walk
16	3	HBP
17	4	Error
18	5	Long Single
19	6	Medium Single
20	7	Short single
21	8	Short double
22	9	Long double
23	10	Triple
24	11	Home run
25	12	Ground into Double Play
26	13	Normal Ground ball
27	14	Line drive or Infield fly
28	15	Long Fly
29	16	Medium Fly
30	17	Short Fly

FIGURE 4.3 Possible Batter Outcomes for Baseball Simulation.

- A short single advances all runners by one base.
- A short double advances each base runner two bases.
- A long double scores a runner from first.
- A GIDP (ground into double play) is a ground ball double play if there is a runner on first, first and second, first and third, or bases loaded. In other situations, the batter is out and the other runners stay where they are.
- A normal GO is a ground out that results in a force out with a runner on first, first and second, first and third, or bases loaded. We assume that with runners on second and third, the runners stay put; with a runner on third, the runner scores; and a runner on second advances to third.
- A long fly ball advances (if there are fewer than two outs) a runner on second or third base.
- A medium fly ball (if there are fewer than two outs) scores a runner from third.
- A short fly or a line drive or an infield fly does not advance any runners.

Our next step is to assign probabilities to each of these events. During recent seasons around 1.8% of all at bats result in an error. For a given player we input the information in the highlighted cells. Let's input Mike Trout's 2016 statistics. See Figure 4-4 and file Trout2016.xlsm. We note that our simulation omits relatively infrequent baseball events such as steals, caught stealings, passed balls, wild pitches, balks, etc.

	D	E	F
1	Outcome	Number	Probability
2	Plate Appearances	681	
3	At Bats+ Sacrifice Hits + Sacrifice Bunts	554	
4	Errors	10	0.0146843
5	Outs (in Play)	234	0.3436123
6	Strikeouts	137	0.2011747
7	Walks	116	0.1703377
8	Hit by Pitch	11	0.0161527
9	Singles	107	0.1571219
10	Doubles	32	0.0469897
11	Triples	5	0.0073421
12	Home Runs	29	0.0425844

FIGURE 4.4 Inputs to Trout Simulation.

For Trout, At bats + SH + SB = 554. He walked 116 times, hit 107 singles, etc.

Outs (in play) are plate appearances that result in non-strikeout outs.

Outs (in play) = (At Bats + SH + SB) − Hits − Errors − Strikeouts.

Historically, errors are 1.8% of At Bats + SH + SB, so we compute

Errors = .018 * (At Bats + SH + SB).

Also,

Total plate appearances = BB + HBP + (At bats + SH + SB)
= 554 + 116 + 11 = 681.

We now compute the probability of various events as (Frequency of Event)/(Total Plate Appearances). For example, we estimate the probability of a Trout single as 107/681 = .157.

We need to also estimate probabilities for all possible types of singles, doubles, and outs in play. For example, what fraction of outs in play are GIDP? Using data from Earnshaw Cook's *Percentage Baseball* (1966) and discussions with *USA Today*'s Jeff Sagarin (who has built many accurate baseball simulation models) we estimated these fractions as follows.

- 30% of singles are long singles, 50% are medium singles, and 20% are short singles.
- 80% of doubles are short doubles and 20% are long doubles.
- 53.8% of outs in play are ground balls, 15.3% are infield flies or line drives, and 30.9% are fly balls.
- 50% of ground outs are GIDP and 50% are normal GOs.
- 20% of all fly balls are long fly balls, 50% are medium fly balls, and 30% are short fly balls.

To verify that these parameters provide an accurate representation of baseball, we simulated 50,000 innings using the composite major league statistics for the 2016 season. We found that our simulated runs per game were within 1% of the actual runs per game.

We now use the Excel simulation add-in @RISK to "play out" an inning thousands (or millions!) of times. You can download a free 15-day trial version of @RISK from Palisade.com. Basically, @RISK generates the event for each plate appearance based on the probabilities that we input (of course, these probabilities are based on the player we wish to evaluate). Essentially, for each plate appearance, @RISK generates a random number between 0 and 1. For example, for Trout a random number .157 would yield a single. This will cause 15.7% of Trout's plate appearances (just as in the actual 2016 season!) to result in a single. In a similar fashion, each other possible batter outcome will occur in the simulation with approximately the same probability as the outcome actually occurred!

One sample inning of our Trout 2016 simulation in which two runs were scored is shown in Figure 4-5.

	D	E	F	G	H	I	J
56	Outcome	State #	Outcome #	Runs	Outs Made	Outs	Done?
57	Strikeout	1	1	0	1	1	no
58	Short double	1	8	0	0	1	no
59	Medium Fly	6	16	0	1	2	no
60	Long Single (advance 2 bases)	6	5	1	0	2	no
61	Walk	2	2	0	0	2	no
62	Long Single (advance 2 bases)	3	5	1	0	2	no
63	Line drive or Infield fly	4	14	0	1	3	yes

FIGURE 4.5 Example of Simulating a Single Inning.

The Entering State column tracks the runners on base; for example, 101 means runner on first and third while 100 means runner on first. The Outcome column tracks the outcome of each plate appearance using the codes shown in Figure 4-3. For example, event code 6 represents a medium single.

In the inning shown in Figure 4-5 our team of nine Trouts scored two runs. Playing out thousands of innings with @RISK enables us to estimate the average number of runs scored per inning by a team of nine Trouts. Then we multiply the average number of innings a team bats during a game (26.72/3) to estimate the number of runs created per game by Trout. Since we are playing out each inning using the actual probabilities corresponding to a given player, our Monte Carlo estimate of the runs per inning produced by nine Trouts (or any other player) should be a far better estimate than runs created or linear weights. Again, the Monte Carlo estimate of runs per game should be accurate for **any player, no matter how good or bad he is.** As we have shown with our Joe Hardy example, the accuracy of runs created and linear weights as a measure of hitting effectiveness breaks down for extreme cases.

SIMULATION RESULTS FOR TROUT, BRYANT, AND CABRERA (AND BONDS04)

For Trout16, Bryant16, Cabrera13, and Bonds04 our simulation yields the following estimates for Runs Created per game:

- Trout16: 9.38 runs per game.
- Bryant16: 7.95 runs per game.
- Cabrera13: 10.24 runs per game.
- Bonds04 : 21.02 runs per game!!!

There is a problem with our Bonds04 result. Barry Bonds received 232 walks during 2004. However, 120 of these walks were intentional. Of course, Bonds received intentional walks because pitchers would rather pitch to the other players (who were not as good at hitting as Bonds). For a team consisting of nine Bonds04 players there would be no point in issuing an intentional walk. Therefore, we reran our simulation after eliminating the intentional walks from Bonds's statistics. We found that the no Intentional Walks Bonds created 15.98 runs per game, which is still quite high in comparison to the other players.

HOW MANY RUNS DID TROUT ADD TO THE 2016 ANGELS?

To be honest, there will never be a team of nine Mike Trouts or nine Barry Bonds. What we really want to know is how many runs a player adds to his team. Let's try and determine how many runs Mike Trout 2016 added to the 2016 LA Angels. The hitting statistics for the 2016 LA Angels (excluding Mike Trout) are shown in Figure 4-6.

	D	E	F
1	Outcome	Number	Probability
2	Plate Appearances	5360	
3	At Bats+Sacrifice Hits + Sacrifice Bunts	4962	
4	Errors	89	0.0166045
5	Outs (in play)	2782	0.5190299
6	Strikeouts	854	0.1593284
7	Walks	355	0.0662313
8	HBP	40	0.0074627
9	Singles	848	0.158209
10	Doubles	247	0.0460821
11	Triples	15	0.0027985
12	Home Runs	127	0.023694

FIGURE 4.6 2016 LA Angels Statistics (sans Trout).

Note, for example, that 4.3% of Trout's plate appearances resulted in HRs but for the rest of the 2016 Angels only 2.4% of all plate appearances resulted in home runs. We can now estimate how many runs Trout added to the LA Angels. For the Angels without Trout we assume that each hitter's probabilities are governed by the data in Figure 4-6. Playing out 25,000 innings we found that, based on the runs per inning from our simulation, the Angels were projected to score an average of 626 runs without Trout. With Trout, the Angels actually scored 717 runs. So, compared to an average LA Angels hitter, how many wins can we estimate that Trout added? Enter the Pythagorean Theorem from Chapter 1. During 2016 the Angels gave up 727 runs. This yields a scoring ratio of $717/727 = 0.986$. The Pythagorean Theorem predicts they should have won $\frac{162 * .986^2}{.986^2 + 1} = 79.88$ games.

Without Mike Trout our simulation yielded a scoring ratio of $626/727 = .861$. Therefore without Mike Trout the Pythagorean Theorem predicts that the Angels would have won $\frac{161 * .861^2}{.861^2 + 1} = 68.97$ games.

Thus our model estimates that Trout added $79.88 - 68.97 = 10.91$ wins for the Angels (assuming that Trout plate appearances were replaced by an average non-Trout Angels hitter).

TROUT VS. THE AVERAGE MAJOR LEAGUER

In his *Historical Baseball Abstract* (2001), Bill James advocated comparing a player to an "average major leaguer." Let's try and determine how many extra runs an "average 2016" team would score if we replaced 681 of the average team's plate appearances by Trout's statistics (shown in Figure 4-4). The file Troutoveraverage.xlsx allows us to input two sets of player statistics. We input Trout's 2016 statistics in cells B2:B12. Then we input the average 2016 major league team's statistics in H2:I12. See Figure 4-7.

	H	I	J
1	Outcome	Number	Probability
2	Plate Appearances	6153	
3	At Bats+Sacrifice Hits + Sacrifice Bunts	5593	
4	Errors	101	0.01641476
5	Outs (in play)	2784	0.45246221
6	Strikeouts	1299	0.21111653
7	Walks	503	0.08174874
8	HBP	55	0.00893873
9	Singles	918	0.14919551
10	Doubles	275	0.04469365
11	Triples	29	0.00471315
12	Home Runs	187	0.03039168

FIGURE 4.7 Average 2016 Team Statistics.

We can see that Trout hit more HRs and had many more walks per plate appearance than the average 2016 batter. When simulating an inning, each batter's probabilities will be generated using either Trout's data from column D or the team data in column J. Since Trout had 681 plate appearances and the average team had 6,153 plate appearances, we choose each batter to be Trout (column D data) with probability $681/6{,}153 = .102$, and choose each batter to be an "average batter" (column J data) with probability $1 - .111 = .889$. After running 50,000 innings for the average team, and the team replacing 11.1% of the average team's at bats by Trout, we find the marginal impact is that Trout would increase the number of runs scored over an average team from 727 to 804. How many wins is that worth? With Trout, our scoring ratio is $804/727 = 1.106$. Using the Pythagorean Theorem of Chapter 1 we predict that the team with Trout would win $\frac{162 * (1.106)^2}{(1 + 1.106^2)} = 89.12$. Therefore, we estimate that adding Trout to an average team would lead to $89.12 - 81 = 8.12$ wins. We will see in Chapter 9 that an alternative analysis of Trout's 2016 batting record indicates that he added around 6.64 wins more than an average player.

CHAPTER 4 APPENDIX: USING A DATA TABLE TO PERFORM A SIMULATION IN EXCEL

In the cell range B2:H22 of the file Simulationmotivator.xlsx we have programmed Excel to "play out" an inning for a team in which each hitter who has a 50% chance of striking out or hitting an HR. Hit F9 and the number of runs scored by the team is recorded in cell H3. Note that whenever the Excel RAND( ) function returns a value <.5 the batter hits an HR, and otherwise the batter strikes out. To record the number of runs scored during many (say, 1,000) innings we enter the numbers 1 through 1,000 in the cell range I6:I1005. An easy way to accomplish this is to enter a 1 in cell I6 and, after choosing Fill from the Editing group on the Home tab, choose Series . . . and fill in the dialog box as shown in Figure 4-8.

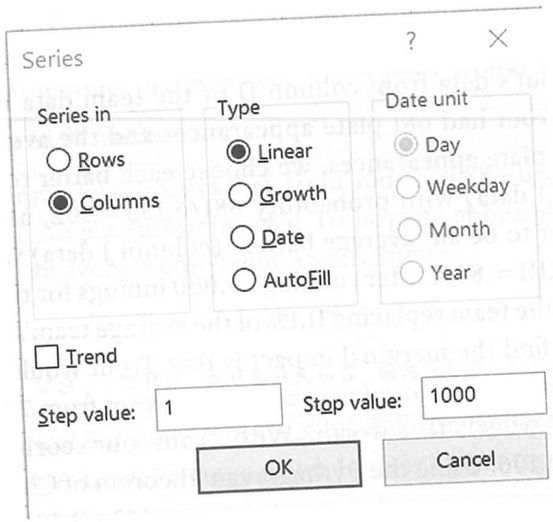

FIGURE 4.8 Entering the Numbers 1–1000 in a Column.

Next, we enter in cell J3 the formula (=H3) that we want to play out or simulate 1,000 times. Now we select the cell range I3:J1005 (this is called the Table Range). We select the Data tab and then

	I	J
2		Average Runs per Inning
3		3.01
4		
5	Iteration Number	1
6	1	6
7	2	6
8	3	5
9	4	4
10	5	0
11	6	0
12	7	7
13	8	1
14	9	2
15	10	6
16	11	1
17	998	1
18	999	8
19	1000	4

FIGURE 4.9 Simulating 1,000 Innings of Joe Hardy Hitting.

choose the What-If icon (the one with a question mark) and choose Data Table.

Next, we leave the row input cell blank and **choose any blank cell as your column input cell.** Then Excel puts the numbers 1, 2, . . . , 1000 successively in our selected blank cell. Each time cell H3 (runs in the inning) is recalculated as the RAND( ) functions in column C recalculate. Entering the formula = AVERAGE(I6:I1005) in cell J3 calculates the average number of runs scored per inning during our 1,000 simulated innings. For the 1,000 innings simulated in Figure 4-9, the mean number of runs scored per inning was 3.01.