Application of the D'Esopo and Lefkowitz Scoring Index Model to WCC Data and Reasoning for future use of Simulation

USF Baseball Quantitative Studies Team

WCC 2011-2018

Walks	Singles	Doubles	Triples	Homeruns
12,847	25,921	6,500	775	1,980

Introduction

One of the seemingly infinite reasons Baseball is beautiful is due to the clear and discrete nature of outcomes. There is no clock, and the states of play have very clear starts and ends. Like when a 2-2 count turns into a 3-2 count, or when no runners no outs turns into a runner on first base no outs. The same type of movement cannot be made for Football, Basketball, and Hockey. In Baseball, there are 24 potential states/combinations of baserunners and outs. Even better, a movement between base-out states is triggered by clear events such as single, double, or even just "out". The description above can possibly be simulated, if done responsibly and with clear assumptions.

The D'Esopo-Lefkowitz Scoring Index Model developed in 1960 is arguably the earliest known application of simulation to Baseball. The DLSI (D'Esopo-Lefkowitz Scoring Index) is sort of like an arcade game of Baseball. The only possible events in DLSI are BB,1B, 2B, 3B, HR, Out. In DLSI, all runners advance symmetrically by the same bases of the batter for non-walk on base events. For instance, with a runner on 3rd base, a walk will not score the runner, but a single, double, triple, or homerun will. **One exception in DLSI is that all runners from second base score on singles**. Aside from the **aggressive assumption** that all singles score runners from second base, DLSI is conservative in nature. For instance, it assumes that all outs are strikeouts and baserunners do not advance on any out, there is no way in the model to examine the likelihood of a baserunner taking an extra base, and there is no way in the model to examine double plays.

This brief research piece is an examination of a model. DLSI is an example of a very rough, conservative model. **No model perfectly reflects reality, the world is often too complex to be put into a model**. Countless examples ranging from finance/economics, natural disaster science, and even Baseball (did the Quants suggest Dave Roberts pull Rich Hill after allowing 1 hit over 7 innings because their models said to?) have shown that humans often **misinterpret** a model's output and therefore **overestimate its predictive power**.

QST is actively looking to develop a simulator like model. A model where decisions can be simulated and in game situations can be modeled before hand.

DLSI

DLSI is rooted in the **frequency of on-base outcomes/events**. For the WCC from 2011-2018, they are as follows:

$$\textit{Fraction of Walks/HBP}(f_0) = \frac{BB+HBP}{BB+HBP+1B+2B+3B+HR} = \frac{12{,}847+3{,}664}{12{,}847+3{,}664+25{,}921+6{,}500+775+1{,}980} = .319$$

$$\textit{Fraction of Singles}(f_1) = \frac{1B}{BB+HBP+1B+2B+3B+HR} = \frac{25{,}921}{12{,}847+3{,}664+25{,}921+6{,}500+775+1{,}980} = .501$$

$$\textit{Fraction of Doubles}(f_2) = \frac{2B}{BB+HBP+1B+2B+3B+HR} = \frac{6{,}500}{12{,}847+3{,}664+25{,}921+6{,}500+775+1980} = .126$$

Fraction of Triples (f_3) = $\frac{3B}{BB+HBP+1B+2B+3B+HR} = \frac{775}{12,847+3,664+25,921+6,500+775+1980} = .015$

Fraction of Homeruns (f_4) = $\frac{HR}{BB+HBP+1B+2B+3B+HR} = \frac{1,980}{12,847+3,664+25,921+6,500+775+1980} = .038$

DLSI takes these frequencies of On-Base outcomes and can answer a variety of questions.

If one were to ask, *"In the WCC, what is the probability of scoring 0 runs in an inning?"*. One could manually go through every inning of the WCC from 2011-2018, take the number of innings where 0 runs scored and divide by the number of innings played. Or they could just use DLSI principles! Let's do that:

First, look at the **number of batters reaching base where no runs can score** in DLSI:

{0 batter reaches base and 0 runs score}

OR

{1 batter reach base and 0 runs score}

OR

{2 batters reach base and 0 runs score}

OR

{3 batters reach base and 0 runs score}

Before moving ahead, the count stops at 3 batters. There is no chance, in the model, that 4 batters reach base and exactly one run scares. In the lowest form of getting on base (walks), 4 walks would result in 1 run scored. Per the model, **it is not possible for X runs to score in an inning if (X+4) batters reach base.** The set up for the calculation is as follows:

Pr( )= Probability of…

Pr (0 run) =

{Pr (0 reach base) x Pr (0 run given 0 batter reaches base)}

+

{Pr (1 reach base) x Pr (0 run given 1 batters reach base)}

+

{Pr (2 reach base) x Pr (0 run given 2 batters reach base)}

+

{Pr (3 reach base) x Pr (0 run given 3 batters reach base)}

The equation is a weighted sum of probabilities. Most importantly, the equation is entirely reliant on the characteristics of the underlying data.

The model needs to get a sense of how often batters get on base in the WCC on an inning basis. **WCC OBP comes out to .343**. OBP in the model is designated as *P*. In Short:

$$\textit{Probability that "B" batters reach base} = \frac{(B+2)(B+1)p^{B}(1-P)^{3}}{2}$$

Pr (0 reach Base) = .283

Pr (1 reach Base) = .291

Pr (2 reach Base) = .200

Pr (3 reach Base) = .114

Pr (4 reach Base) = .058

Next, need to calculate the values of scoring 1 run given X batters get on-base:

Pr (0 runs given 0 batter reaches base) = 1.0

Pr (0 runs given 1 batter reaches base) = .97

Pr (0 run given 2 batters reach base) = .82

Pr (0 run given 3 batters reach base) = .23

******Calculations are in supporting excel sheet******

We get:

Pr (0 runs) = 75.53%

Remember, DLSI is like a Baseball arcade game. In a Baseball arcade game, the possible sequences for scoring 0 runs given 2 runners reach base is as follows:

$$f_0(f_0 + f_1 + f_2) + f_1(f_0 + f_1 + f_2) + f_2(f_0) + f_3(f_0)$$

This can be translated as:

Probability of (A walk AND another walk, single, double)

+

Probability of (A single AND a walk, single, OR double)

+

Probability of (A double AND a walk)

+

Probability of (A triple AND a walk)

=

Probability of scoring 0 runs given 2 reach base

These sequences are the potential sequences of on base events in DLSI that would score 0 runs given 2 batters reach base. This exercise of examining the potential sequences where "X" runs score given "Y" runners reach base is repeated for X+4 batters reaching base.

Simulating a Player

The example above utilized the frequencies of outcomes in the WCC. The same exercise can be repeated for the frequencies of outcomes generated by players. Taking the 2018 on base outcome profiles of Riley Helland, Jonathan Allen, and Jack Winkler:

Name	Year	School	G	PA	1B	2B	3B	HR	BB
Helland, Riley	2017-18	USF	58	252	52	19	0	2	21
Allen, Jonathan	2017-18	USF	57	264	46	16	1	7	27
Winkler, Jack	2017-18	USF	58	215	34	8	0	1	17

The intuition can be framed like this: *"In DLSI, a team comprised of 2018 Riley Hellands would generate X% chance of scoring 0 runs in an inning…"* The same goes for Jonathan Allan, and Jack Winkler. So, per the DLSI model:

Pr (0 runs scored in an inning by a team of 2018 Jack Winklers) = 79.41%

Pr (0 runs scored in an inning by a team of 2018 Riley Hellands) = 66.17%

Pr (0 runs scored in an inning by a team of 2018 Jonathan Allens) = 64.96%

Per DLSI, the model uses the on base percentage and underlying on base profiles of the 3 players for the 2018 season. **All 3 players have different profiles of getting on base and the underlying outcomes, so they all play differently in the model.** The model says a team comprised of 2018 Jonathan Allens is MOST likely to experience a positive scoring inning. Remember, this is a model with its own set of assumptions that may or may not accurately display an accurate picture of reality.

Conclusion

The D'Esopo and Lefkowitz Scoring Index Model was arguably the first Baseball model developed with simulation methods in mind. A Baseball game is an incredible thing. In terms of baserunners and outs, there are only 24 potential combinations, that is it! DLSI captures the game's movements from state to state driven by underlying probabilities. For instance, it is more likely to move from the "no baserunners no outs" state to the "runner on first no outs" state than the "runner on second no outs" state. This is because the probability of walking or hitting a single is higher than hitting a double. The movement between the 24 base-out states can be simulated. For example, with 1 out and a runner on 2nd base, DLSI can answer which batters would most likely score the runner before 3 outs are made. Utilizing the assumptions, on base

frequency, and underlying profile of on base outcomes, the model can simulate chances of scoring the runner.

D'Esopo and Lefkowitz's model was built with the underlying drivers being the real distribution of outcomes. In the model, the first step is to plot the frequency of on-base outcomes. This is a good thing. **Any model that does not utilize the complete aggregate base of actual information as it's foundation should raise immediate red flags.** But, DLSI's assumptions are unfortunately not flexible enough to model a more complete version of reality (a Baseball game). The assumption that all runners from second base score on singles is overly aggressive, and not having a way to factor in double plays, steals, errors, etc… is unfortunate.

Luckily, QST is actively working to develop a similar, but more complete model. Building a simulator model of a Baseball game, in a responsible fashion, could enable decision makers to have an additional tool to make personnel or gameplay decisions. Instead of waiting for an actual in game decision, test a decision in a model days before the actual game. For instance, with a runner on first base and no outs, where in the lineup is the optimal spot to call a bunt. **The model would look at, given 1 out, and the on base profiles of all the batters involved, where in the lineup is that runner most likely to score with 2 outs to work with.** The model would simulate the situation over-and-over again. Developing a simulator model would give USF decision makers a tool that other decision makers do not have.