THE WALL STREET JOURNAL.
WSJ.com

July 18, 2014, 1:38 PM ET

How We Did It: Investigating Whether There's a DH Advantage in Baseball

By Rob Barry and Tom McGinty

Agence France-Presse/Getty Images

To complete the analysis of interleague Major League Baseball games highlighted in this week's "The Numbers" column, we needed detailed data for decades of games, and we found it–for free!–on the website of Retrosheet, an amazing resource for baseball/data junkies.

Retrosheet has compiled database-ready files with play-by-play descriptions of most games from 1940 through 2013, and many other games from prior years. We loaded all plays and all available data points for each play from 1950 through 2013 into a SQL Server table that has 9,772,477 records and takes up about 4.8 gigabytes of disk space, including indexes.

We also loaded Retrosheet's game logs going back to 1940 (136,459 games). The logs have a summary of each game with every statistic imaginable.

Once the files are imported into a database manager, it's a snap to create statistics for any scenario a nerd can envision. Curious about Alex Rodriquez's career batting and slugging averages while batting with various counts? Here you go:

"Plate Appearances" counts the number of times A-Rod had a complete at bat, no matter how it ended (walk, strikeout, hit, etc.). What we've called "Statistical At Bats" here are plate appearances that ended with a hit, the batter making an out (not counting sacrifice flies and sacrifice hits), or an error by the fielding team. Statistical at bats are used to calculate batting average (hits divided by statistical at bats) and slugging percentage. Slugging percentage not only credits batters for hitting often, but also adds value for extra-base hits, with batters getting 1 point for each single, two for each double, three for each triple and four for each home run. The total is then divided by statistical at bats to derive the slugging percentage.

So using the Retrosheet data to calculate simple statistics is easy. Using the data to figure out whether the American League has an advantage in interleague play due to the designated hitter turned out to be much more difficult.

Our investigation was motivated by a stark fact: the American League's extraordinarily high win rate during interleague play. The AL wins 57.5% of its home interleague games — much higher than the 53.7% win rate for the home team in AL/AL games.

This difference is statistically significant, according to Fisher's Exact test (for those of you dying to know, the p-value is 0.0007682).

By contrast, there isn't a statistically significant difference between home games in National League parks where the league is playing itself (54.0%) and when it's hosting the American League (52.7%). If the entire advantage held by the AL was that they were simply a better team, why wouldn't they also win a statistically significant number of games in NL parks?

Regular Season Play 1997-2013		Home Team			
		AL		NL	
		Wins	Losses	Wins	Losses
Away Team	AL	9,237	7,980	1,123	1,009
	NL	1,226	906	10,602	9,046

(Nerdy factoid footnote: there are eight ties in this period, though none in interleague games.)

To answer this question, we decided to take two paths: a top-down method (a regression model), and a bottom-up approach (tallying up the actual impact of the DH in each game).

For the regression model, which we created with guidance from Serge Sverdlov, a statistician at the University of Washington, we decided to see if we could measure what we termed an "AL host effect": is there something different about the outcomes of interleague games played in AL parks?

We incorporated several standard baseball statistics into our model, including each teams' batting averages, win rates, starting pitcher's ERAs (for both the year and his entire career), average errors, assists and runs per game, and the teams' on base and slugging percentages for the prior 162 games. We also included variables for home team advantage and the relative strength of the two different leagues.

In addition, we added another variable: an Elo rating, based on the method described by Nate Silver, intended to capture the differences in teams' relative strengths. Elo ratings help our model differentiate between teams with tougher schedules and, although interleague play is fairly rare, allow for some comparison of the strength of teams between the two leagues.

Our model predicts that by removing the special factor inherent to interleague games in American League

parks, the AL's win rate falls more than two percentage points, to 55.1% from 57.5%.

However, that rate is still higher than the average home team win rate in both leagues (53.8%) during regular season non-interleague play. One possible reason: the American League is simply better.

In fact, our model can offer a prediction about this. If the interleague games played in American League parks had been moved to a neutral location (removing home team advantage) with no designated hitter rule, the league would still have won about 51.3% of its games.

For our ground-up analysis, the goal was to measure how important the designated hitters were in the outcomes of interleague games played in AL parks.

To do this, we subtracted the DH's RBIs from the box score and added the number of RBIs an average pitcher on that team would have produced with the same count of at bats.

Why? Because that's who would have been at the plate had the designated hitter rule not been in effect.

Obviously, this works out poorly for the American League, whose pitchers only go to bat in National League games. That's also why we had to use average pitcher performance: individual American League pitchers bat so infrequently, their skill at the plate is hard to estimate.

Indeed, we found the American League's win rate plunged to 54.3% from 57.5% when we performed this replacement, a slightly lower win rate than our other model predicts.

But we knew that would happen, so we tried taking the experiment a step further: we ran through the batting order, replacing each position's RBIs with the pitchers'.

Our finding: the biggest effect came from the designated hitter. (The position with the least impact was second base, which, when swapped for the pitcher, caused the win rate to drop to just 56.2%.)