Chris Dutton

Understanding BABIP:
What Determines the Outcome of a Ball in Play?

Abstract:

The primary purpose of this paper is to identify and explore the factors that influence a Major League hitter's batting average on balls in play (BABIP). While there has been significant research examining BABIP in terms of pitchers, much more limited work has been done to isolate the independent hitting qualities that determine whether a batted ball will become a hit or an out. Using simple regression analysis on a data set of batters from 2002-2007, we are able to determine the marginal effect of individual hitting abilities on a player's BABIP. Although a significant proportion can be attributed to luck, we find that certain isolated, situation-independent aspects of a player's skill can jointly explain more than 34% of the variance in BABIP. This yields a model more than three times more accurate than commonly-used formulas, and allows us to both predict future performance and examine past performance in the absence of luck. Furthermore, this model explores the potential existence of other, non-quantifiable factors that allow a hitter to consistently exceed or fall below his expected BABIP.

Introduction:

Batting average on balls in play, or BABIP, is commonly used as a measure of pitching performance. Baseball analysts such as Voros McCracken and Tom Tippett have explored this topic in depth, debating the impact of luck versus skill by exploring year-to-year trends and correlations. In 2001 McCracken published a groundbreaking article arguing that pitchers do not have an inherent ability to control the outcome of balls in play, citing the fact that strikeouts and homeruns allowed remained relatively constant from one year to the next while BABIP tended to fluctuate uncontrollably. Essentially, McCracken's conclusions suggest that hits allowed are a poor measure of pitching performance, and that BABIP is

primarily a function of defense and luck.[1] Two years later, Tippett responded by performing an analysis of all pitcher seasons dating back to 1913. After adjusting BABIP to account for park and team defense effects, Tippett revealed that certain pitchers have shown a distinct ability to prevent hits on balls in play. In fact, 12% of pitchers with at least 6,000 career balls in play prevented hits at a rate that would appear in less than 1% of random samples. Year-by-year analysis shows that pitchers such as Pedro Martinez, Charlie Hough and Walter Johnson were able to prevent hits at a significantly greater rate than the average throughout much of their careers, debunking the assertion that inherent ability is nonexistent.[2] Today, the general consensus is that pitchers do generally demonstrate some level of control over balls in play, although to a lesser degree than defense-independent factors such as walks or strikeouts.

In 2005, Dave Studeman of The Hardball Times published a pair of articles introducing the relationship between *hitting* ability and BABIP. The first emphasized the strong correlation between line drive percentage and BABIP and identified players who showed a discrepancy between these two factors in the previous season.[3] However, this article does little to investigate some of the additional factors that likely cause these discrepancies, such as park effects, contact rate, and the ability to hit to all fields. The second article is more driven towards explaining these omitted factors and arrives at a regression model using line drive rate, fly ball rate, and strikeout rate.[4] While this certainly improves upon his past models, this study still suffers from serious bias caused by omitted variables and the prohibitively small sample taken from the 2004 season alone. Extending the dataset to six years, we find that this model only explains a mere fraction of the variance in BABIP. As a result, serious improvements are necessary in order to achieve a more accurate and realistic predictive model.

Like Studeman, we aim to explore the correlation between specific performance indicators and the ability to create hits on balls in play. It is a matter of debate whether a hitter can demonstrate the ability to control the outcome of a ball in play, and anecdotal

---

[1] McCracken, Voros. "Pitching and Defense: How Much Control Do Hurlers Have?" *Baseball Prospectus* 2001.
[2] Tippet, Tom. "Can pitchers prevent hits on balls in play?" *Diamond Mind Baseball* 2003.
[3] Studeman, Dave. "If Line Drives Could Speak." *The Hardball Times 2005*.
[4] Studeman, Dave. "I'm Batty for Baseball Stats." *The Hardball Times 2005*.

evidence suggests that luck plays an extremely significant role. However, a hitter's BABIP certainly depends on a range of quantifiable factors, such as the ability to control the strike zone, make consistent contact, hit the ball hard, and keep the ball out of the air. By identifying these factors and their marginal effects, we are able to gain a deeper understanding of the role that isolated aspects of hitting skill (known as base performance indicators, or BPIs) play in determining BABIP. With sufficient data and a reliable model, we can then use this information to essentially remove luck from the equation. That is to say, we can deflate the hitting statistics of the "lucky" players who exceeded their expectations and inflate those of "unlucky" players who fell below in order to generate luck-neutral performance metrics for any given year.

Overall, our objectives were to identify and quantify the base performance indicators that influence a hitter's BABIP, design a predictive model that is unprecedented in accuracy and reliability, and explore the potential existence of non-quantifiable factors that may allow certain players to consistently exceed or fall below expectations. Our hypothesis was that performance indicators would explain roughly 25% of the variance in BABIP, and that unaccounted factors would allow a small percentage of hitters to repeatedly defy their expectations.

Methods:

In order to achieve reliable and accurate results, we focused heavily on creating a strong dataset. To do this we collected a range of raw hitting data from 2002-2007 using Baseball Prospectus' custom statistics reports as our primary source. We also gathered spray chart data provided by Bill James, as well as hitter types (lefty, righty, and switch) from Sean Lahman's publicly available database. Using this data we were able to generate more advanced metrics, which are described in detail in the selected statistics report below:

Variable	Description
BABIP	Batting average on balls in play, calculated as non-homerun hits divided by balls in play ((h-hr)/(pa-so-bb-hr)).
Hitter_Eye	A measure of plate discipline and knowledge of the strike zone, calculated as (BB rate/SO rate).
Pitches_perEBH	Pitches per extra base hit, which is a measure of how often a hitter makes solid contact (pitches/(doub+trip+hr)).
LD_per	Line drive percentage, as defined by MLB Advanced Media and provided by Baseball Prospectus.
FB_GB_ratio	Fly ball/ground ball ratio, using percentages provided by Baseball Prospectus.
Speed Score	A comprehensive measure of speed, developed by Bill James. The speed score is the average of 5 individual formulas based on SB%, SB attempts, triples, runs per time on base, and double plays.
Contact_Rate	A measure of the ability to make contact and avoid striking out, simply calculated as ((ab-so)/ab).
Spray	Measure of how well a hitter distributes balls in play to the entire field. Calculated as |1(LF%) + -1(RF%)|.
Pitches	A hitter's average number of pitches per plate appearance, to account for patience and selectiveness at the plate.
Park	A vector of binary stadium variables, to account for the influence of park effects on BABIP.
Year	A vector of year variables from 2002 through 2007, to account for potential time effects.
Lefty	A binary variable equal to "1" if the hitter is a lefty, "0" otherwise.
Switch	A binary variable equal to "1" if the player is a switch hitter, "0" otherwise.

Using this data, we generated a park-adjusted regression model to identify the factors that influence a hitter's BABIP. Our independent variables include hitter's eye, pitches per extra base hit, line drive percentage, fly ball/ground ball ratio, speed score, contact rate, spray, pitches per plate appearance, and a vector of stadium and year binaries, which collectively create a formula to determine a hitter's predicted batting average on balls in play. As a test of explanatory power, we then calculated the correlation between actual and predicted BABIP for all of the hitters in our sample.

Next, we use our regression model to identify "lucky" and "unlucky" players, as defined by those who exceeded or fell below their predicted values, respectively. By generating a variable to describe the discrepancy between predicted and actual values, we are able to rank these players accordingly. Furthermore, we are able to calculate a luck-neutral batting line (avg, obp and slg) by assuming a perfect BABIP prediction, and compare this with a player's actual line in order to provide a side-by-side comparison. Assuming regression to the mean, this can be an extremely useful tool in determining future performance.

Finally, we narrow down the sample to players with four or more years of service and observe year-by-year trends for each individual player. The reason for this is that while a player will naturally exceed or fall below his expected BABIP in any given year due to random chance, it is much more unlikely to achieve the same result over multiple years. This process identifies players who have defied the model every year observed in the sample, which suggests either a) tremendous fortune or misfortune, or b) the existence of some underlying, perhaps non-quantifiable factor that is unaccounted for in our model.

Results:

Our regression model yields an R-squared value of .3442, and all non-vector explanatory variables are significant at the 1% level. This suggests that the factors included are all highly significant, and jointly explain roughly 34% of the variance in a hitter's BABIP. As an additional test of accuracy, we find a robust 59% correlation between actual and predicted BABIP for all players in our sample. Given the tremendous uncertainty regarding the outcome of balls in play, these results are extremely promising. By contrast, commonly-used models based on line drive percentage alone explain only about 3% of the variance in BABIP when applied to the same dataset, and yield a mere 18% correlation between predicted and actual values. The regression output follows:

```
reg babip hitter_eye pitches_per_ebhit ld_per fb_gb_per speed_score contact_rate
spray pitches_per park year, r

Linear regression                                       Number of obs =    1654
                                                        F( 47,  1606) =   18.65
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.3442
                                                        Root MSE      =  .02485

------------------------------------------------------------------------------
             |               Robust
       babip |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  hitter_eye |    .009552   .0024549     3.89   0.000      .004737    .0143671
pitches_pe~t |  -.0006325   .0000554   -11.41   0.000    -.0007412   -.0005238
      ld_per |   .3395134   .0300267    11.31   0.000     .2806178    .3984091
   fb_gb_per |  -.0360601   .0035876   -10.05   0.000    -.0430971   -.0290232
 speed_score |   .0032872   .0003612     9.10   0.000     .0025789    .0039956
contact_rate |  -.1090923    .015689    -6.95   0.000    -.1398655   -.0783191
       spray |  -.0990883   .0084267   -11.76   0.000    -.1156168   -.0825597
 pitches_per |   .0085323   .0029537     2.89   0.004     .0027388    .0143258
         ana |  -.0006604   .0048539    -0.14   0.892    -.0101811    .0088602
         ari |  -.0121926   .0047733    -2.55   0.011    -.0215552     -.00283
         atl |  -.0030583    .005389    -0.57   0.570    -.0136286    .0075119
         bal |  -.0108662   .0048693    -2.23   0.026    -.0204171   -.0013152
         cha |  -.0203448    .004762    -4.27   0.000    -.0296851   -.0110045
         chn |  -.0145646   .0047335    -3.08   0.002    -.0238491   -.0052802
       cin_a |  -.0117664   .0110774    -1.06   0.288    -.0334941    .0099613
       cin_b |  -.0208366    .005036    -4.14   0.000    -.0307144   -.0109588
         cle |  -.0122645   .0053029    -2.31   0.021    -.0226658   -.0018632
         col |  -.0093492   .0048926    -1.91   0.056    -.0189458    .0002475
         det |  -.0131272   .0051457    -2.55   0.011    -.0232202   -.0030341
         flo |  -.0048644   .0049816    -0.98   0.329    -.0146355    .0049068
         hou |  -.0111054   .0048502    -2.29   0.022    -.0206189    -.001592
         kca |  -.0172583   .0044266    -3.90   0.000    -.0259407   -.0085758
         lan |  -.0155826   .0049713    -3.13   0.002    -.0253334   -.0058317
         mil |  -.0174423   .0052009    -3.35   0.001    -.0276435    -.007241
         min |  -.0077802   .0049852    -1.56   0.119    -.0175584    .0019981
         mon |  -.0207713   .0068691    -3.02   0.003    -.0342447   -.0072978
         nya |  -.0115345   .0051595    -2.24   0.026    -.0216545   -.0014145
         nyn |  -.0168474   .0048663    -3.46   0.001    -.0263924   -.0073024
         oak |  -.0179386   .0046987    -3.82   0.000    -.0271549   -.0087223
       phi_a |  -.0168998   .0087019    -1.94   0.052    -.0339681    .0001685
       phi_b |  -.0086077   .0053835    -1.60   0.110    -.0191671    .0019516
         pit |  -.0121328    .004756    -2.55   0.011    -.0214614   -.0028041
       sdn_a |  -.0146004    .006604    -2.21   0.027    -.0275538   -.0016471
       sdn_b |  -.0168768   .0047405    -3.56   0.000     -.026175   -.0075785
         sea |  -.0091226   .0050975    -1.79   0.074     -.019121    .0008758
         sfn |  -.0139924   .0049254    -2.84   0.005    -.0236533   -.0043316
       sln_a |   -.007834   .0046202    -1.70   0.090    -.0168962    .0012282
       sln_b |  -.0167127    .005451    -3.07   0.002    -.0274046   -.0060208
         tba |  -.0121543   .0047773    -2.54   0.011    -.0215247   -.0027839
         tex |  -.0164399    .004786    -3.44   0.001    -.0258273   -.0070525
         tor |   -.014793   .0044921    -3.29   0.001     -.023604   -.0059821
         was |  -.0164445   .0058431    -2.81   0.005    -.0279054   -.0049836
    year2002 |   .0347482   .0044501     7.81   0.000     .0260197    .0434768
    year2003 |   .0346184   .0044339     7.81   0.000     .0259216    .0433152
    year2004 |   .0365268   .0044753     8.16   0.000     .0277488    .0453048
    year2005 |  -.0042349    .002107    -2.01   0.045    -.0083677   -.0001021
    year2006 |  -.0005372   .0020969    -0.26   0.798    -.0046501    .0035757
       _cons |   .3569578   .0216407    16.49   0.000     .3145108    .3994047
------------------------------------------------------------------------------
```

It is important to note that park variables are all in relation to Boston, and carry negative coefficients due to the fact that Fenway Park has the strongest positive park effect on a hitter's BABIP. Team names divided into "a" and "b" reflect a stadium change at some point in the sample, such as the Padres moving from Qualcomm Stadium to Petco Park or the Phillies moving from Veteran's to Citizens Bank Park in 2004. Although the city and team remains constant, we must account for different stadiums individually since each one has its own individual effect on balls in play. Similarly, year binaries are included and reported in relation to the year 2007. F-tests of joint significance show that both the year and park binary variables are significant at the 1% level, as expected. Our initial results show a positive effect on BABIP in 2002-2004 compared to 2007, and a negative effect in 2005 and 2006.

As mentioned above, all of our key independent variables are statistically significant at the 1% level. That is to say, there is virtually no chance that the effects reflected in this model are the product of random chance. Our regression results show positive effects for hitter's eye, line drive percentage, speed score, and pitches per plate appearance, all of which conform to common sense. On the other hand, we find negative coefficients on pitches per extra-base hit, fly ball/ground ball ratio, spray and contact rate. Interpretation of these coefficients can be found in the discussion of our results.

After generating a variable to capture the difference between predicted and actual BABIP, we find that the "luckiest" player in the entire sample exceeded his expected BABIP by 113 points. This player was Matt Kemp in the 2007 season, who converted balls in play into hits at a staggering rate of .411 despite a predicted rate of .298. On the opposite end of the spectrum was Henry Blanco in 2004, who produced a measly .207 BABIP despite a prediction of .285.

Finally, our study of year-to-year differences in predicted and actual BABIP allowed us to identify the players who defied the model on a regular basis. At the top of the list we find six players who played in every year of the sample and exceeded expectations in each one. These players are Garret Anderson, Jeff Kent, Kevin Millar, Luis Castillo, Mark Loretta, and Moises Alou. Similarly, we find one player who fell below expectations for all six years,

as well four others who accomplished the same feat in five straight years. These players include Brad Ausmus, Kevin Mench, Michael Barrett, Neifi Perez and Sammy Sosa.

Discussion:

So what do these results tell us about BABIP? First of all, they tell us that park and year effects have a real influence on BABIP. The reasoning is quite obvious in terms of park effects, since stadiums with different dimensions and features have unique effects on the outcome of batted balls. Fenway Park, for example, features a shallow left field, deep center field and relatively spacious right field, which allows balls in play to become hits more often than anywhere else. As for the significance of certain years, the best explanation seems to be the overall level of offense and team defense from year to year. A season characterized by particularly poor offense and stellar defense will produce different BABIP estimates than a season with great offense but terrible defense. In our sample, there appears to be a significantly different environment in the 2002-2004 seasons compared to the 2005-2007 seasons. Other realistic explanations for this outcome are certainly possible, and open to debate.

We have also shown that all else equal, a hitter who is able to command the strike zone, hit the ball on a line, run the base paths well, or show patience and selectivity at the plate should produce a higher BABIP. We have also shown that a player who rarely hits for extra bases, tends to keep the ball in the air, or clusters balls in play to one area of the field will likely have a lower BABIP. While both of these results confirm anecdotal evidence, the negative coefficient on contact rate deserves a bit more attention. One might expect a higher contact rate to lead to a higher BABIP, but the opposite actually seems to be the case. This is likely caused by the correlation between strikeouts and power, since players who swing hard tend to either miss entirely or crush the ball for extra-base hits. If this theory is reflected in our data, it makes sense that we would expect a player with a lower contact rate to generate a higher predicted BABIP.

By summarizing our measure of luck in terms of actual and predicted performance, we isolated individual player seasons at both ends of the spectrum. Based on Kemp's

performance in 2007, we can expect that his BABIP will significantly decline as he regresses towards his predicted value. Had Kemp produced a BABIP exactly equal to his estimate (and was thus neither lucky nor unlucky), we would have expected him to have a batting line of .258/.294/.415. In reality, Kemp hit .342/.373/.521 – mostly due to the fact that over 41% of his balls in play became hits. As a result, his 2008 batting line is extremely unlikely to approach his 2007 totals. As for Henry Blanco, the "unluckiest" player in our sample, we see that his .206/.260/.368 line in 2004 improved to .242/.287/.391 the following year, as expected. A more extreme example is Brandon Inge in 2003, who had the third unluckiest season in the sample with a 72 point difference between his actual and predicted BABIP. A comparison of his 2003 and 2004 numbers show a serious rebound the following season, as he improved from .203/.265/.339 to .287/.340/.453. The last of our top three unluckiest seasons is Richie Sexson in 2007, who also suffered from a 72 point difference. That year, Sexson produced a line of .205/.295/.399 despite a predicted line of .258/.342/.468. As a result, it would be a safe bet to say that Sexson will almost certainly improve his numbers significantly in the 2008 season.

Our final area of research was to identify the players who have consistently defied the model in order to investigate potential omitted factors. To recap, the players listed below either exceeded (yellow) or fell below (blue) expectations in every year that they played:

Player	Years in sample	"Lucky" years
Garret Anderson	6	6
Jeff Kent	6	6
Kevin Millar	6	6
Luis Castillo	6	6
Mark Loretta	6	6
Moises Alou	6	6
Brad Ausmus	6	0
Kevin Mench	5	0
Michael Barrett	5	0
Neifi Perez	5	0
Sammy Sosa	5	0

So what factors are making certain players like Kent and Castillo achieve a higher than predicted BABIP, while others like Ausmus and Mench consistently fail to meet expectations? If there is some common thread between the top and bottom players on this list, it should be included in our model if possible. A preliminary analysis reveals the following hitting lines for the groups of players listed above:

Lucky: .296/.361/.455 with 15.67 HR

Average: .274/.343/.441 with 15.9 HR

Unlucky: .259/.318/.411 with 13.73 HR

Clearly, the "lucky" players appear to be ones with high average and on-base percentages, and average power. By contrast, the "unlucky" players fall significantly below the sample average in all categories described. Once again, it is important to keep in mind that although a player can be "lucky" in any given year, the ability to achieve the same result year after year is evidence of some real underlying skill or ability. With a perfect model, we would expect roughly half of the players to exceed their expectations and half to fall below in any given year, which implies a 1/64 chance due to luck alone that a player will show up in the list presented above. Since our sample here is trimmed to 220 players (with 4+ years of service), we would expect to see three or four players beat the odds. The fact that we see significantly more players make the list is strong evidence of omitted factors, as is the disparity between groups represented by the hitting lines above. This suggests that there is almost certainly some variable related to hitting performance (shared by the players shaded in yellow and lacking in the players shaded in blue) that has not yet been accounted for. If this variable can be quantified, then including that data into the model should predict higher BABIP estimates for the "lucky" players and lower estimates for the "unlucky" ones. This would improve the accuracy of the model and make it less likely for players to consistently fall above or below their predictions. As for identifying what these omitted factors may be, I will leave that question open-ended.

Overall, this study provides a comprehensive analysis of the factors that influence a hitter's BABIP. Using the model created in this study, we have shown that a deeper understanding of these factors allows for a more formal analysis of hitting performance and luck. By reporting luck-neutral predictions, this study also provides a valuable tool for player evaluation and projection. While we understand that there are also limitations and caveats associated with our research, we hope that this paper will at the very least help to serve as an inspiration for future studies.

Literature Cited

"Custom Statistics Report." Baseball Prospectus. 1996-2008. <http://baseballprospectus.com/statistics/sortable/> April 2008.

James, Bill. "Hitting Analysis Reports." *Bill James Online*. 2008. <http://www.billjamesonline.net/StatisticsReport_new.aspx?Type=01&Team=0&Player=1&men=2> 28 April 2008.

McCracken, Voros. "Pitching and Defense: How Much Control Do Hurlers Have?" Baseball Prospectus. 23 Jan 2001. <http://www.baseballprospectus.com/article.php?articleid=878> 14 Feb 2008.

Studeman, Dave. "If Line Drives Could Speak." *The Hardball Times*. 14 March 2005. <http://www.hardballtimes.com/main/article/if-line-drives-could-speak/> 30 April 2008.

Studeman, Dave. "I'm Batty for Baseball Stats." *The Hardball Times*. 10 May 2005. <http://www.hardballtimes.com/main/article/im-batty-for-baseball-stats/> 30 April 2008.

Tippett, Tom. "Can pitchers prevent hits on balls in play?" *Diamond Mind Baseball*. 21 July 2003. <http://www.diamond-mind.com/articles/ipavg2.htm> 14 Feb 2008.

Woolner, Keith. "Counterpoint: Pitching and Defense: Another Look at Pitchers Preventing Hits." Baseball Prospectus. 29 Jan 2001. <http://www.baseballprospectus.com/article.php?articleid=883> 14 Feb 2008.