An Introductory Look at the Situational Context Inherent in the RBI Using Two Modeling Approaches

Ryan Sides, Baylor University, Waco, TX

ABSTRACT

The RBI (run batted in) is a popular statistic in Major League Baseball that is extremely dependent on the situational context (i.e., which bases are occupied by runners and the number of outs in an inning) experienced by the hitter. This paper offers insight into how much this situational context affects the RBI, providing a couple of related modeling approaches that account for this information and, thus, an approach for improving a player's evaluation. The first model used to accomplish this goal is a standard multiple regression model, where the number of at-bats in each possible situation a player experiences are the independent variables. Subtracting a player's model predicted RBI count, $\widehat{RBI}$, from his actual RBI total from the season produces a performance statistic that can be described as the number of runs batted in above or below average that player achieved for the season. Prediction intervals are also calculated from the regression model to determine the "reasonable" range of RBI each player should have accrued in the season. The second model used in this study is strictly an intuitive approach based on years of experience as a player and sabermetrician on the part of the author, where league average RBI rates are multiplied by the number of plate appearances each player had for each situation. The total across these situations over the course of the season provides an "intuitive" expected RBI (ERBI), and the subtraction of a player's ERBI from his actual RBI becomes the second performance statistic utilized in this document. Both response statistics from the above models are then divided by the number of at-bats each player achieved for the season (for comparison purposes) and then converted to percentiles for ease of understanding how a player performed. Lastly, various statistical tools are utilized to check assumptions and compare the two models along with other baseball statistics.

INTRODUCTION

There is an abundance of baseball statistics in today's world, as is clearly evidenced when perusing popular baseball websites like Fangraphs or Baseball-Reference. There are traditional statistics like batting average (the proportion of hits a player gets with relation to the number of official at-bats he accumulates) and RBI, or runs batted in (an accrual of the number of players that score a run in a hitter's at-bats). There are a number of complex statistics that attempt to quantify a player's entire offensive worth into one number, despite being extremely confusing to the typical baseball fan. And there are statistics that fall in a "middle ground" between those extremes that, while not the most common of statistics in the baseball world, are still understood by many who follow the game.

One of the most traditional statistics, the RBI, has existed since nearly the beginning of the game of baseball itself, going back well into the 19th century (Thorn and Palmer 1985). And while many baseball traditionalists believe that runs batted in is an acceptable measure of a player's character and grit, baseball analysts have always been skeptical; the RBI is such a context driven statistic that even the best players can have poor totals while a mediocre hitter can amass over 100 because of the opportunities that arise (Keri 2006). And despite the relatively young movement in the baseball community to improve statistical analyses, the skepticism on the RBI actually dates back to 1879 (Thorn and Palmer 1985). In fact, the RBI was not kept as an official statistic until 1920 because for nearly half a century newspaper readers were not convinced that all players had the same opportunities to drive in runs, making it unfair to all players (Thorn and Palmer 1985).

However, sabermetricians (or baseball statisticians), have taken an extreme approach to this "fairness" issue. They have created complex statistics that most fans, coaches, and even general managers do not understand, relying solely on those and scoffing at anyone who dares to bring a statistic like runs batted in into a conversation. In fact, when Ryan Howard was recently given a 5 year, $125 million contract extension, a popular ESPN columnist stated that the deal was "obviously crazy" because best-in-baseball type money should not be paid to a good (but not great) player who only excels in the RBI category (Neyer 2010). Another sabermetrician from Baseball Prospectus, one of the highest regarded baseball organizations in existence that has published an annual baseball preview since 1996, goes on to say that the RBI is worthless because it can be inflated or deflated by a multitude of factors outside of a player's control (Perry 2004). If the whole goal of sabermetrics (the search for objective knowledge about baseball through statistics) is to

define the value of a player by what he does independent of his teammates, why should there be any focus on statistics like the RBI when there is very little valuable information contained there (Perry 2004)?

Thus, this study aims to begin to bridge the gap between the traditionalists and sabermetricians with regard to the RBI. Baseball cards, television and radio broadcasts, and ballparks across the country continue to use runs batted in because it has been around for so long and is easy to understand; even the newest of fans to the game can quickly grasp it and keep tally throughout a season. However, sabermetricians act as if this statistic does not even exist because it is too dependent on circumstances outside of the player's control. But by removing the situational context (i.e., which bases are occupied by runners and the number of outs in an inning) inherent in the RBI, a moderated form can still exist that is both understandable and practical. To take the advice of one analyst, sabermetricians need to take baby steps by making strides with their audience rather than forcing everything to be perfect immediately (Carroll 2010). And while some disagree with this approach, this study is designed solely for the purpose of finding that "middle ground" that can exist with the RBI. That way, a traditionalist who does not want to get bogged down into complex analyses and number crunching can still have and understand his RBI (albeit modified), and a sabermetrician can rest happily at night knowing that the baseball community is at least moving in the right direction towards statistics of better value.

STUDY SPECIFICS

Before the study was conducted, a judgment call needed to be made regarding which players to include in this study. Because of time limitations, only 2010 National League data was used; it was acquired from the official Major League Baseball website (mlb.com), Baseball-Reference.com, and Fangraphs.com. American League data was not included because it probably would have affected the results due to the designated hitter rule. Thus, any study done that attempts to combine the two leagues needs to take this into account. For instance, the average runs per game in the National League in 2004 was 9.34, but for the American League, the value was 10.11 (Ruane 2005); it is for this reason that it would be inappropriate to combine the two leagues into one data set.

Further, only data for players with over 250 official at-bats were included in this study. There were a couple of key reasons for choosing this cutoff. First, 250 at-bats marks roughly the halfway point needed to meet requirements that Major League Baseball has established in order to be eligible for certain hitting awards (see the Official MLB Rulebook). Also, this eliminated most players who were not essential to their team's season because they missed a large number of games due to injury or were simply not good enough to accumulate a lot of at-bats. This also removed pitchers from the study, which was a positive effect that permits the study to emphasize position players rather than inflating players' data because they were compared to pitchers. Most teams were left with about nine players at this point, enough for a starting lineup of eight position players and a good reserve hitter; in some instances, this allowed some teams to account for those situations where two players split a season's worth of at-bats. After this screening process, there were 145 players out of 16 National League teams who qualified to be a part of this study.

There are 24 different situations that the 145 batters could encounter when batting. These situations are easily enumerated as a result of eight different situations based on runners (no runners on, runner on first, runner on second, runner on third, runners on first and second, runners on first and third, runners on second and third, and bases loaded) along with three different situations based on the number of outs (none, one or two) that result in the 24 possibilities, also referred to as the 24 base/out situations. Moreover, before any model development was attempted, several of the 24 situations were combined for two reasons: first, fewer situations typically creates simpler calculations, and second, and more importantly, because many of the players in the study only had a handful of at-bats in several situations, combining situations that are known to be equal in RBI likelihood (where this explanation is provided below) should result in more reliable estimates.

Two different types of situations were combined. First, all three situations where there are no runners on base were combined because the number of outs is meaningless if there is no one (with the exception of the hitter) able to be driven in; the only way a batter gets an RBI is if he hits a home run, and players are not more or less likely to hit home runs depending on the number of outs. Second, all no outs and one out situations were combined. The logic behind this was that there should be no difference in RBI expectations unless there are two outs. In any situation where a sacrifice fly, walk, base hit, or routine out occur and generate an RBI, the fact that there are no outs or one out is irrelevant. Thus, for example, because the situation where there is a runner on first and no outs is essentially equivalent to the same situation but

with one out, both will have the same RBI likelihood and were combined into one situation. (Because it could be argued that a runner on first with two outs could provide a different result due to the runner leaving his base on contact and possibly scoring on a double, those situations were not combined.) It should be noted that if a player hits into a double play but a run is scored, he is not credited with an RBI, eliminating the possibility of the no outs and one out situations being different (see the Official MLB Rulebook).

Also, not every at-bat was used in the study. A player who walked or was hit by a pitch did not have that at-bat count towards his at-bat totals unless it generated an RBI. The thinking here is that a player should not be "penalized" (with reference to his ability to drive in runs) for being pitched around and drawing a walk (or being hit by a pitch) if there are two outs and a runner on third base (which is a common occurrence) or some other similar situation. However, if the bases are loaded, a walk or hit by pitch is not intentional because it will score a run, and in this case, a batter was credited with (for this study) an at-bat, even though this does not register as an official at-bat for Major League Baseball purposes (see the Official MLB Rulebook). Further, all sacrifice flies (defined as a fly ball out where a runner scores) were credited as at-bats in this study, despite not counting officially for Major League Baseball (see the Official MLB Rulebook), because a batter had a chance to produce an RBI and was successful. However, sacrifice bunts where a runner was not on third base were not counted as at-bats for the purposes of this study because the goal was not to score a run, but rather move a player forward a base; the batter did not actually have a chance to produce an RBI. If a sacrifice bunt occurred with a runner on third base, that at-bat was added in with the totals because a good bunt would have produced a run, and thus, an RBI.

PREVIOUS STUDIES

Historically, a similar study to this has been conducted. Tom Ruane, back in 1998 and again in 2005, compiled historic data and found modified RBI for every year (and in both leagues) from 1960 to 2004. However, the study comprising this document attempts to do several different things than were previously done. First, reliability of the data is increased by combining some of the base/out situations (thereby resulting in a more adequate sample size per base/out combination). Further, Ruane's approach only looked at the intuitive model for creating an adjusted RBI, where the multiple regression approach allows for the creation of prediction intervals because individual variability in situations can be accounted for. In addition, correlation between the newly created statistics and other baseball statistics are analyzed in an attempt to get a deeper feel for what a player does control, if anything, with regards to his number of runs batted in. However, as mentioned previously, the study comprising this document limits its use of players to only those with at least 250 at-bats, while Ruane used all players regardless of the number of at-bats accumulated. This difference should have a minimal, if any, effect on the results.

Further, it is particularly noteworthy that few, if any, analytical studies have been conducted in regards to the RBI. A 2002 paper by Gary Koop in *The Journal of the American Statistical Association* merely mentions that he is excluding statistics like runs and RBI from his study because they depend too much on the performance of other players and on a player's location in the batting order. Further, statistician Jim Albert, who has published a number of statistical papers and books on his baseball research, says in a 2010 e-mail correspondence that, "… while it would be interesting to see if particular players have a tendency to do 'better than expected' with respect to RBI, I suspect the answer is negative since this is related to the idea of clutch ability, which at best is a small effect." It appears that many in the sabermetric community have taken a similar approach to the matter.

THE MULTIPLE LINEAR REGRESSION MODEL

This first model that aims to remove situational context from the RBI is created using multiple linear regression. For this model, the number of at-bats in each of the fifteen situations were used as the independent variables and the total number of RBI was used as the dependent variable. This model is a very useful initial look into the RBI and how it is affected by situational context because reliance here is on a standard multiple regression model, allowing for the variability of the model and parameter estimates to be found (providing some insight into the reliability of the model). Moreover, its aim is directly to the point of the study; this model uses merely the number of at-bats in each of the fifteen situations and the total number of RBI accumulated on the season and attempts to predict how many RBI each player would achieve if one adjusts for what the player cannot control (i.e., the number of runners on base and the number of outs when he comes to bat).

All fifteen situations were utilized in this regression model because insight about the RBI and what it is caused by was desired based on every possible situation a player could encounter. In addition, typically a model is a better predictor when it utilizes more variables. It is noteworthy here as well that none of the independent variables had large variance inflation factors (Montgomery et. al. 2006, p. 334), and because the data is simply drawn from baseball websites, there is no additional cost or statistical penalty to utilize all fifteen independent variables. Moreover, R^2-adjusted only decreases about half a percent from its maximum value to its value when all variables are included.

Further, the intercept term was left in the model in order to provide a more optimal model estimate as opposed to "forcing" the model through the origin. More specifically, the fact that a player with no at-bats in any situation should be predicted to have no RBI (meaning no intercept), the fact that a player needed at least 250 at-bats to qualify for the study makes this situation impossible, and an attempt to use this model for players with many less at-bats is an extrapolation and should not be attempted.

Thus, the linear model in matrix/vector form is

$$\mathrm{RBI} = \mathrm{AB} * \beta + \varepsilon$$

where, in this context, RBI is a 145x1 vector comprising each player's total runs batted in during the season, AB is a 145x16 matrix comprising the players' at-bats in each of the fifteen situations (plus a leading column of ones for the intercept term), β is a 16x1 vector of coefficients to be estimated and ε is the standard 145x1 vector of error terms.

Proc Reg in SAS is utilized to compute the regression equation to find the best linear unbiased estimator of β, $\hat{\beta}$. Plugging $\hat{\beta}$ back into the equation yields

$$\widehat{\mathrm{RBI}} = \mathrm{AB} * \hat{\beta},$$

providing each player a predicted RBI value. Further, the difference between a player's actual RBI and $\widehat{\mathrm{RBI}}$, or in this case, the residual, shows the difference between how many more (or fewer) RBI each player accrued than what was "expected" with respect to the study group "average" player encountering those situations. However, at this point, a player who had more plate appearances has a larger range for his residual. If that player performs above the league average, he will have more chances to succeed, while if he underperforms, he will have more chances to fail. Thus, each player's residual was divided by his total number of at-bats so that the playing field was effectively "leveled". This statistical is called Residual/AB, and a value of 0 represents a player who performed at an average level. Further, a positive or negative Residual/AB would correspond to a player who performed above or below average, respectively. A sample of Houston Astros' players with their $\widehat{\mathrm{RBI}}$, residual and Residual/AB can be found below.

$\widehat{\mathrm{RBI}}$, Residuals and Residuals/AB for Houston Astros' Players

---------- TEAM=Astros ----------

LASTNAME	FIRSTNAME	POS	RBI	RBIHAT	RBIRESID	RBIRESIDPERAB
Johnson	Chris	CI	52	43.04	8.96	0.03
Sanchez	Angel	MI	25	20.77	4.23	0.02
Pence	Hunter	OF	91	84.19	6.81	0.01
Berkman	Lance	CI	49	47.11	1.89	0.01
Bourn	Michael	OF	38	37.02	0.98	0.00
Lee	Carlos	OF	89	88.65	0.35	0.00
Keppinger	Jeff	MI	59	63.39	-4.39	-0.01
Manzella	Tommy	MI	21	25.41	-4.41	-0.02
Quintero	Humberto	C	20	24.95	-4.95	-0.02
Feliz	Pedro	CI	40	52.90	-12.90	-0.03

From this table, it is easy to see why it is necessary to divide a player's residual by their at-bats. For instance, Lance Berkman did not get a full season's worth of at-bats. (He was injured and eventually traded to the American League.) However, comparing Berkman to Hunter Pence shows that, while Berkman had less RBI and a smaller $\widehat{RBI}$ (because of a lack of at-bats), they both performed similarly per at-bat.

Next, percentiles are found using a standardized score. Because the mean of the distribution of Residuals/AB is zero (or only off minutely due to rounding), dividing each player's Residual/AB by the standard deviation of the sample values converts each player's ratio into a z score. These ratio values are approximately normal, with four normality tests (Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling) showing no departure from normality. These values, along with a histogram and QQ plot indicating normality, were found using Proc Insight in SAS.

Thus, because Residuals/AB have an approximately normal distribution, the percentiles for each player can be found. These percentiles provide a quick and easy to understand number that shows how well a player performed (with respect to the situations he encountered) compared to the "average" National League starter. As always with percentiles, a value of 50 is average (i.e., median), while the closer a player is to the extremes of 1 and 99 demonstrates how poorly or how well, respectively, he performed over the season. Just to get an understanding of this approach, the percentiles for the Astros' players are shown below.

Residual/AB Percentiles for Houston Astros' Players

---------- TEAM=Astros ----------

LASTNAME	FIRSTNAME	POS	RBI	RESIDP
Johnson	Chris	CI	52	85.33
Sanchez	Angel	MI	25	75.02
Pence	Hunter	OF	91	67.04
Berkman	Lance	CI	49	59.79
Bourn	Michael	OF	38	52.59
Lee	Carlos	OF	89	50.57
Keppinger	Jeff	MI	59	36.14
Manzella	Tommy	MI	21	24.29
Quintero	Humberto	C	20	22.01
Feliz	Pedro	CI	40	10.07

Further, prediction intervals can now be found. These intervals give a feel for a "reasonable" range of RBI a player should have accumulated based on the situations he encountered, because $\widehat{RBI}$ has some form of variability to it with respect to the fifteen independent situations.

We can assume that $RBI_* - \widehat{RBI}_*$ (or rather, the residual resulting from the regression analysis) is distributed normally because all four normality tests (as mentioned previously), along with a histogram and QQ-Plot, support this assumption. Hence, as is easily deduced, the prediction interval for a given player's set of at-bats is given by

$$ab_*'\hat{\beta} \pm t_{\alpha/2,n-p} s\sqrt{1 + ab_*'(AB'AB)^{-1}ab_*}.$$

A player whose actual number of runs batted in falls outside of this prediction interval either performed drastically better or worse than he "should have" (with respect to his peers). A sample of 95% confidence prediction intervals were found using Proc Reg in SAS, and are shown on the following page.

Prediction Intervals for Houston Astros' Players

---------- TEAM=Astros ----------

LASTNAME	FIRSTNAME	POS	RBI	PREDICTLB	PREDICTUB	USUAL
Johnson	Chris	CI	52	19.60	66.47	OK
Sanchez	Angel	MI	25	-2.97	44.51	OK
Pence	Hunter	OF	91	59.69	108.69	OK
Berkman	Lance	CI	49	22.88	71.35	OK
Bourn	Michael	OF	38	12.64	61.40	OK
Lee	Carlos	OF	89	63.27	114.03	OK
Keppinger	Jeff	MI	59	38.77	88.02	OK
Manzella	Tommy	MI	21	1.85	48.98	OK
Quintero	Humberto	C	20	1.39	48.52	OK
Feliz	Pedro	CI	40	29.23	76.57	OK

Further, not only does this model provide prediction intervals, but also very useful information in the form of R^2. An R^2 value from this model of 0.8 implies that roughly 80% of the variability in this model can be explained by the information in the independent variables. Rather, for this data set, the total number of RBI a player accumulates over the course of a season is comprised of 80% situational context (what a player cannot control). This, in and of itself, provides statistical confirmation that a player who has a large number of RBI did so probably because he batted in a large number of favorable situations.

THE INTUITIVE MODEL

The second model for creating an estimate of how many RBI a player would accrue over the course of a season is done using an "intuitive" model. Both models utilize the same information in regard to their independent variables; however, rather than using a regression approach to estimate coefficients, the second model uses actual RBI data from the season and an intuitive algorithm, described as follows.

The total number of RBI and the total number of at-bats for the 145 players included in this study were accumulated for each of the fifteen situations. The ratio of these two numbers produces estimates of how many RBI per at-bat the average player achieved in each situation, because a player is obviously more likely to get an RBI in a bases loaded situation than one with no runners on base. The resulting RBI rates are expressed as follows:

$$\bar{r}_j = \frac{\sum_{i=1}^{145} RBI_{ij}}{\sum_{i=1}^{145} AB_{ij}}, \quad \text{for } j = 1,2,\dots,15$$

where i=1,2,…,145 represents each of the 145 players included in the study and j represents each of the fifteen situations. These fifteen RBI rates (one per situation) are found by Proc Summary in SAS and then multiplied by the number of at-bats each player had in each of the fifteen situations to find an estimate of how many RBI a player would be expected to accumulate had he been exactly average (with respect to the players in the study) for the season. The model can be described as follows:

$$ERBI_i = \sum_{j=1}^{15} AB_{ij} * \bar{r}_j, \quad \text{for } i = 1,2,\dots,145$$

The acronym (i.e., variable name) for this statistic is ERBI (derived from the model estimated "expected RBI"), because it is what is expected of the i^{th} player using the intuitive model if he were an average player from the group being studied. Each player's actual RBI is subtracted from his ERBI to create a difference that shows how many more (or less) RBI a player accrued than what was expected of him, denoted by RBI+; hence,

$$RBI +_i = RBI_i - ERBI_i, \quad \text{for } i = 1,2,\dots,145$$

where RBI_i is the total number of RBI that player i accumulated over the course of the season.

It should be noted that RBI+ compares to the residual computed from the regression model in that they are both measures of how well a player performed above or below league average. But just as before, this statistic must be divided by at-bats in order to "level the playing field" (i.e., to permit comparability). As before, an RBI+/AB of 0 represents a player who performed at an average league level, while a positive or negative value of this statistic indicates a player who performed above or below average, respectively. A sample of Astros' players with their ERBI, RBI+ and RBI+/AB can be found below.

ERBI, RBI+ and RBI+/AB for Houston Astros' Players

---------- TEAM=Astros ----------

LASTNAME	FIRSTNAME	POS	RBI	ERBI	RBIPLUS	RBIPLUSPERAB
Berkman	Lance	CI	49	39.61	9.39	0.03
Johnson	Chris	CI	52	44.08	7.92	0.02
Pence	Hunter	OF	91	77.30	13.70	0.02
Lee	Carlos	OF	89	76.45	12.55	0.02
Keppinger	Jeff	MI	59	61.56	-2.56	0.00
Sanchez	Angel	MI	25	31.17	-6.17	-0.02
Bourn	Michael	OF	38	53.69	-15.69	-0.03
Feliz	Pedro	CI	40	57.00	-17.00	-0.04
Manzella	Tommy	MI	21	33.34	-12.34	-0.05
Quintero	Humberto	C	20	32.97	-12.97	-0.05

Also, as with the regression model, percentiles can be computed on the residual statistics resulting from this model. The technique is the same as before, with RBI+/AB being distributed approximately normal with a mean of approximately zero. Also, all four normality tests (as mentioned previously) show no departure from normality; these values, along with a histogram and QQ plot showing normality, were again computed using Proc Insight in SAS. The percentiles for the Astros' players are shown below.

RBI+/AB Percentiles for Houston Astros' Players

---------- TEAM=Astros ----------

LASTNAME	FIRSTNAME	POS	RBI	PLUSP
Berkman	Lance	CI	49	87.72
Johnson	Chris	CI	52	80.75
Pence	Hunter	OF	91	80.03
Lee	Carlos	OF	89	78.46
Keppinger	Jeff	MI	59	46.20
Sanchez	Angel	MI	25	22.03
Bourn	Michael	OF	38	17.51
Feliz	Pedro	CI	40	9.00
Manzella	Tommy	MI	21	6.02
Quintero	Humberto	C	20	5.30

However, this model fails to allow for analysis in a key area of interest. Namely, a covariance matrix detailing the variability of each of the fifteen situations would be ideal in order to create prediction intervals for each player, as was done for the regression model. However, because of the nature of its form, it cannot be determined analytically for the intuitive model.

The other useful information that the regression model was able to provide was in the form of R^2. Proc Reg in SAS was able to compute this value with ease, but a detailed look into sums of squares will illustrate why this value is not as easy to come by with the intuitive model. Notice that the total sum of squares,

$$\sum(y-\bar{y})^2 = \sum(y-\hat{y}+\hat{y}-\bar{y})^2 = \sum(y-\hat{y})^2 + \sum(\hat{y}-\bar{y})^2 + 2\sum(y-\hat{y})(\hat{y}-\bar{y}).$$

In the case of least-squares regression (utilized in this first model), the last term reduces to 0, allowing for the total sum of squares to equal the sum of squares due to regression plus the sum of squares due to error. Thus,

$$R^2 = 1 - \frac{SS_{Err}}{SS_{Tot}} = \frac{SS_{Reg}}{SS_{Tot}}.$$

This simplification is what allows R^2 to be interpreted as the proportion of variation in the response variable that is attributable to the covariates. However, the estimates used in the intuitive model do not allow for the last term in the above equation to reduce to 0, so there remains three parts to the total sum of squares breakdown instead of just two.

Thus, there are multiple approaches to estimating R^2 given this situation. The first is to compute R^2 as is traditionally done. For the intuitive model, the calculations would be

$$\widehat{R^2} = 1 - \frac{SS_{Err}}{SS_{Tot}} = 1 - \frac{25{,}750.76}{84{,}111.49} = 0.694.$$

Another estimate of R^2 could be computed as

$$\widehat{R^2} = \frac{SS_{Reg}}{SS_{Tot}} = \frac{36{,}658.38}{84{,}111.49} = 0.436.$$

Lastly, R^2 could be estimated by examining the ratio of only the sum of squares due to regression and the sum of squares due to error, so that

$$\widehat{R^2} = \frac{SS_{Reg}}{SS_{Reg} + SS_{Err}} = \frac{36{,}658.38}{36{,}658.38 + 25{,}750.76} = 0.587.$$

For whichever calculation is used, it is clear that less information in the intuitive model is explained by the situations that a hitter experiences when compared to the regression model. However, it seems safe to assume that somewhere between 50% and 60% of the information contained in the RBI can be explained by these situations when utilizing the intuitive model, though a deeper look into this phenomenon is warranted.

RESULTS

First, it should be noted that this study is merely an introductory look into the RBI and how much the situational context inherent in batting (i.e., the runners on base and the number of outs in an inning) plays into it. Second, all results listed in this section are preliminary findings and are based solely on this group of players for this one season; more data should be collected and analyzed in order to enhance and improve these findings. Unfortunately, time limitations did not permit a more extensive analysis here; in particular, more questions were generated than answered.

Despite these limitations, there are several interesting things to take note of. First, two different methods were utilized in this study in order to answer the question: "Do the top RBI hitters amass a large number of runs batted in solely because of the situations they are presented with?" Further, is it more skill or luck that influences a player's RBI totals from year to year? Utilizing various correlations, the following discussion will attempt to answer both of these questions.

Recall from the regression model and the data utilized in this study that roughly 80 percent of the RBI can be explained by the situations a player experiences. Also recall that it was concluded that roughly 50 to 60 percent of the RBI can be explained by these situations when looking at the intuitive model. Thus, combining these estimates gives a relatively safe assumption that approximately two-thirds of the RBI is attributable to the situational context provided to a hitter.

Thus, it would appear that most of what comprises the RBI is a result of the situational context that a hitter experiences. (This is not to say that a hitter only has a high RBI total because of the situations he

encounters. Rather, most managers understand who their best players are and purposely arrange their batting lineup in order to maximize their top hitter's RBI opportunities.) However, what about the remaining part of the RBI that cannot be explained by the situations a player encounters when batting? Is this attributable to skill or random variation? We will look at some noteworthy baseball statistics to help us find a conclusion here, detailing them briefly. Specifically, RE24 is a "baseline" for this study in that it attempts to estimate how many runs above or below average a player contributed to his team given the game situation when he comes to bat (Skoog 1987). (Note: this is divided by a player's total at-bats for comparison purposes.) Further, weighted on-base average, or wOBA, is an overall measure of how good a player is, created by placing value for each outcome of an at-bat (like a double) rather than situational contexts like the RBI (Tango et. al. 2007); batting average on balls in play, or BABIP, is a statistic that has been shown to be mostly random (and thus, having little relation to the actual ability of the player). Also, both slugging percentage, or SLG, and isolated slugging percentage, or ISO, are essentially a measure of a player's power. Thus, if weighted on-base average has a high correlation with the newly created statistics, Residual/AB (stemming from the multiple regression model) and RBI+/AB (stemming from the intuitive model), it would seem that a large part of the unexplained variation in the RBI is actually within a player's control and can be credited as skill. However, if batting average on balls in play has a high correlation to Residual/AB and RBI+/AB, a large part of that unexplained variation is more likely luck-based. Further, if slugging percentage or isolated slugging percentage compares favorably to these newly created statistics, it would appear that power hitters do better than speed players with respect to scoring runners. (This result would be expected considering that batters who hit for more doubles, for example, will score runners from first base more often whereas primarily singles hitters are more likely to only drive in runners in specific favorable situations.) All of these same comparisons can also be made regarding RE24/AB because of the similarity of that statistic to those generated in this paper. A correlation matrix using Proc Corr in SAS is computed and presented below in order to gain a better understanding of the variation in the RBI not attributable to situational context.

Correlation Coefficients Table

	WOBA	BABIP	SLG	ISO	RBI
RE24PERAB	0.90513	0.45934	0.82782	0.71351	0.64628
RBIRESIDPERAB	0.55796	0.17724	0.63245	0.67718	0.39644
RBIPLUSPERAB	0.78065	0.21969	0.86835	0.88902	0.77117
RBIPERAB	0.67260	0.13897	0.79736	0.85304	0.79019

As seen in the table, it would appear that (once situational context is removed), a large majority of what goes into a run batted in is actually skill-based. Both of the newly created statistics, Residual/AB and RBI+/AB, have significantly larger correlations with wOBA than BABIP. Further, as expected, both also have a large correlation with slugging percentage and isolated slugging percentage, providing more evidence that these newly created statistics do indeed measure skill because this expectation is met.

However, of these two created statistics utilized in this study, which is actually better? Instinctively speaking, the intuitive model should be an improved model over the regression model from a reliability standpoint since it utilizes more information (i.e., all of the achieved situational RBI totals) in its development. Recall that the main benefits to the regression model were to permit reporting of more analytical information like prediction intervals and R^2; however, it was not able to allow for the inclusion of the information included in the intuitive model because the situational RBI totals incorporated in the intuitive model only served to estimate the model coefficients. (Had that information been included as part of the independent variables in the regression model, not only would the models not be comparable because they would have different response variables, but the regression model would then only serve to provide a comparison of each player against himself rather than the "average" player in the study group.) Further, as is evidenced, the intuitive model statistic, RBI+/AB, has correlations that align (for the most part) with the correlations of RE24/AB, the baseline for the study. Note as an example that both RE24/AB and RBI+/AB have similar correlations with slugging percentage, while the regression statistic, Residual/AB, differs.

Further, an in-depth breakdown of RBI+ (by looking at each of the fifteen components that comprise it) appears to show that the regression model does not punish players for a lack of power and instead

attempts to estimate how many RBI they "should" accumulate by weighing situations more evenly; contrast this with the intuitive model which is very heavily weighted towards power hitters, potentially helping or hurting a player with regard to his ability to hit home runs (which is especially noticeable with nobody on base), fluctuating rankings more for those hitters who are at the extremes with respect to strength. This is evidenced in the correlations table with the fact that the correlation between RBI+/AB and both slugging percentages are amongst the highest in the correlation matrix. If this intuitive model, however, is indeed a correct way to analyze runs batted in, no more analysis needs to be done on the RBI in favor of simply using slugging percentage (or isolated slugging percentage) as a baseline for which batters should be placed in the middle of a batting lineup. It should be noted here that many in the sabermetric community are already in favor of this approach.

CONCLUSION

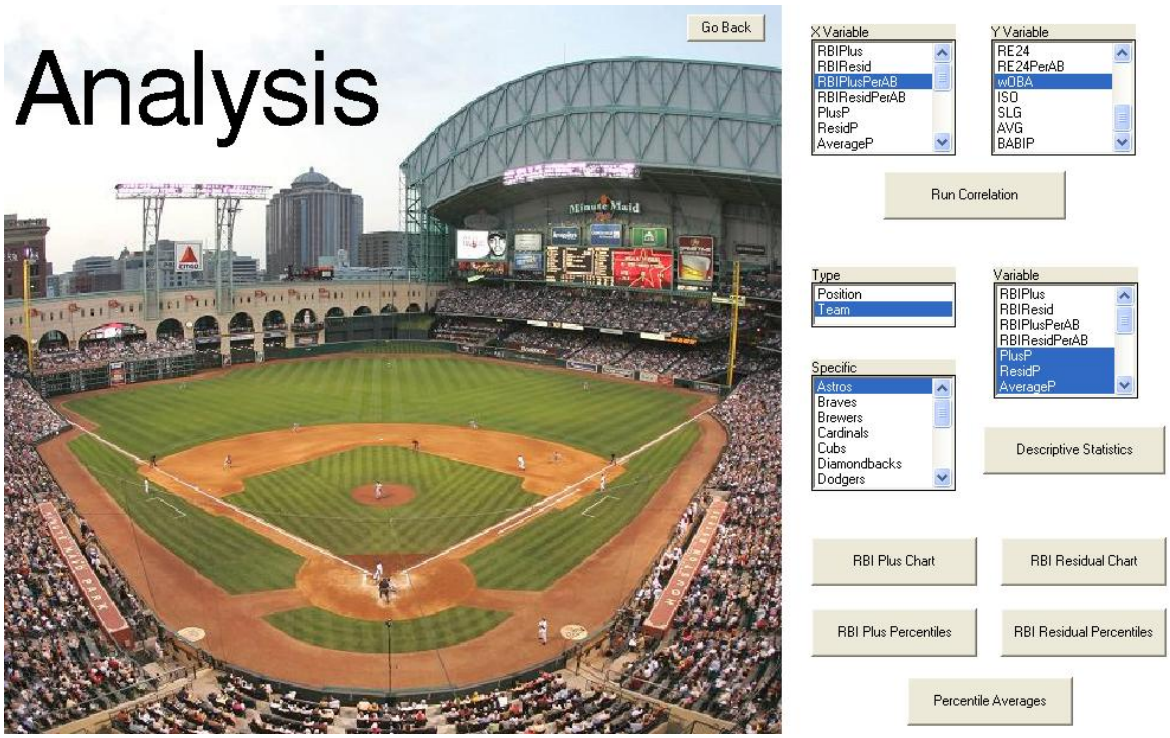
Both the regression and intuitive models seem to agree with the notion that the majority of the information in the run batted in statistic is dependent on the situations that a hitter is presented. However, of the information remaining once this situational context is removed, a large portion of it does seem to relate to skill rather than luck or randomness. Thus, the use of one the statistics created in this paper would be recommended in order to correctly reward players for "coming through in the big moments", but at the same time, not over-reward them simply for being placed in the middle of a potent batting lineup.

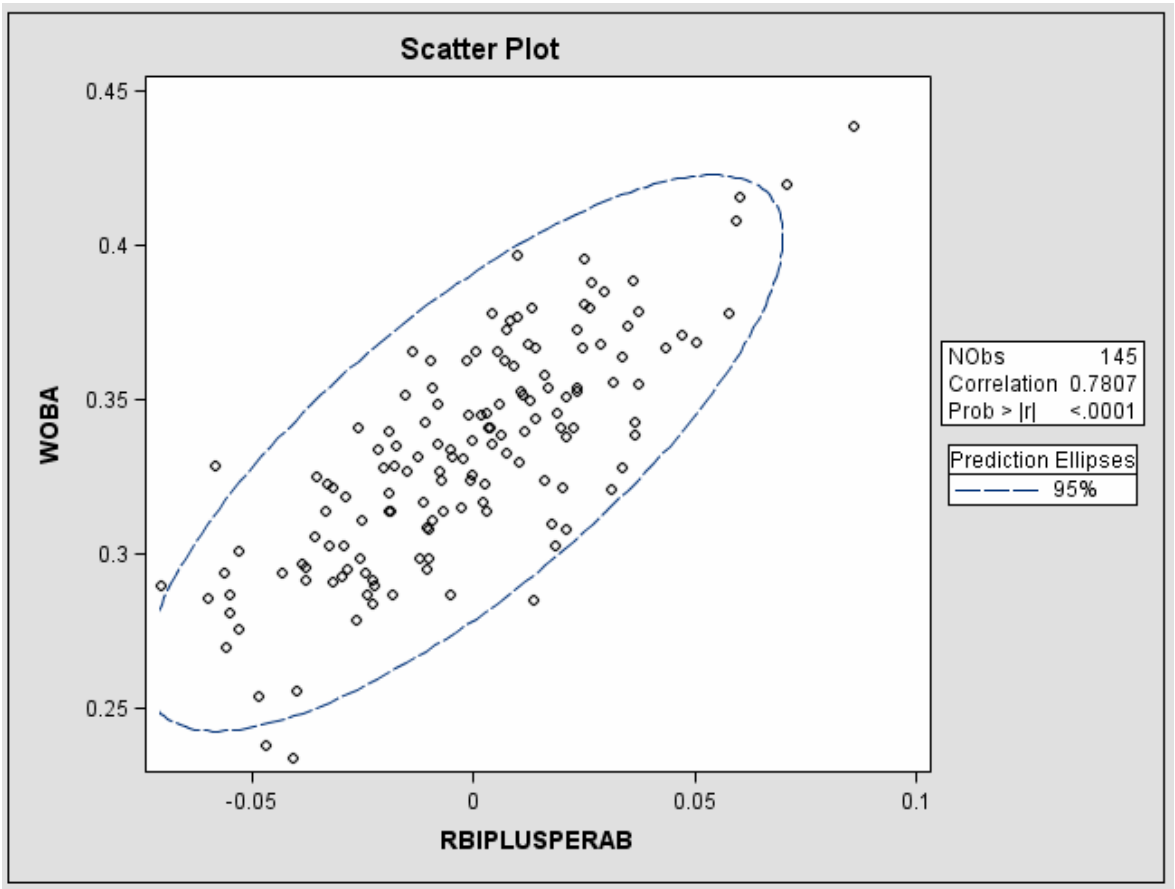
As to which one of the approaches in this paper is a "better" approach, more research should be done before a conclusion can be safely made. While the intuitive model utilizes more information and should, thus, provide better results, it also places great emphasis on the ability to hit for power. While this skill is no doubt important in accruing RBI over the course of a season, it is impossible to tell at this point whether this model is correctly rewarding this ability or not. Further analysis will hopefully show whether this model or the regression one which does not place as much weight on power is more appropriate.

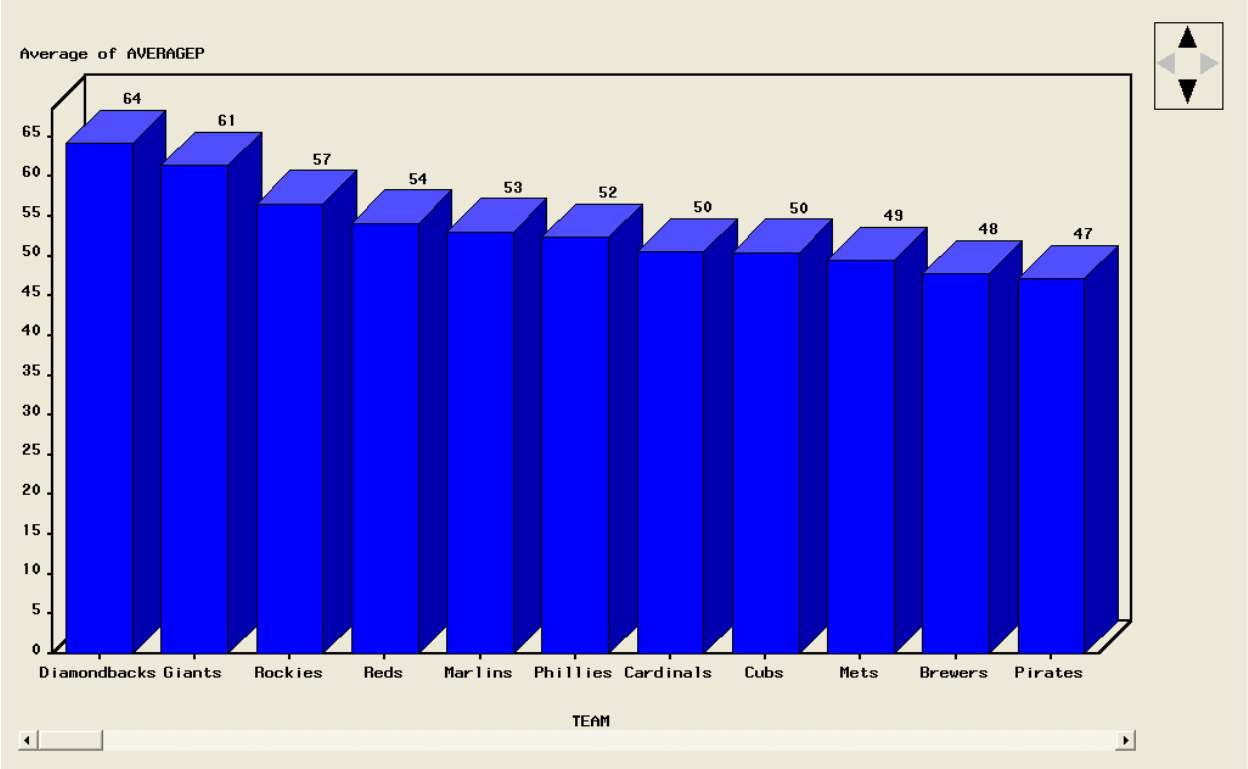
As for future recommendations, the highest priority at this point is to simply collect more data and see if those results reported herein hold "true to form". Ideally, data from another recent National League season would be collected and analyzed; if the analysis on this data set provides similar results, it would be reasonable to say that the validity of the results reported herein are further reinforced. However, if another year provides different results, like drastically dissimilar coefficient estimates, it is quite possible that the data utilized in this sample did not contain enough at-bats to provide reliable estimates, and that a multi-year study is necessary. It should be noted, though, that any data set combing multiple years should have similar properties in runs scored. It is for this reason that combining years should be done cautiously; Major League Baseball has had a history of run totals that change from era to era, so one should be sure that the data being combined comes from a similar time period and that the number of years combined does not exceed a reasonable number like four. Ideally, once a certain number of at-bats are reached by a majority of the players in the study, the intuitive model results and the regression model results would not differ drastically. Further, the fact that each model in this study had considerably different coefficients from the other may be indicative that the number of at-bats included in this study is insufficient.

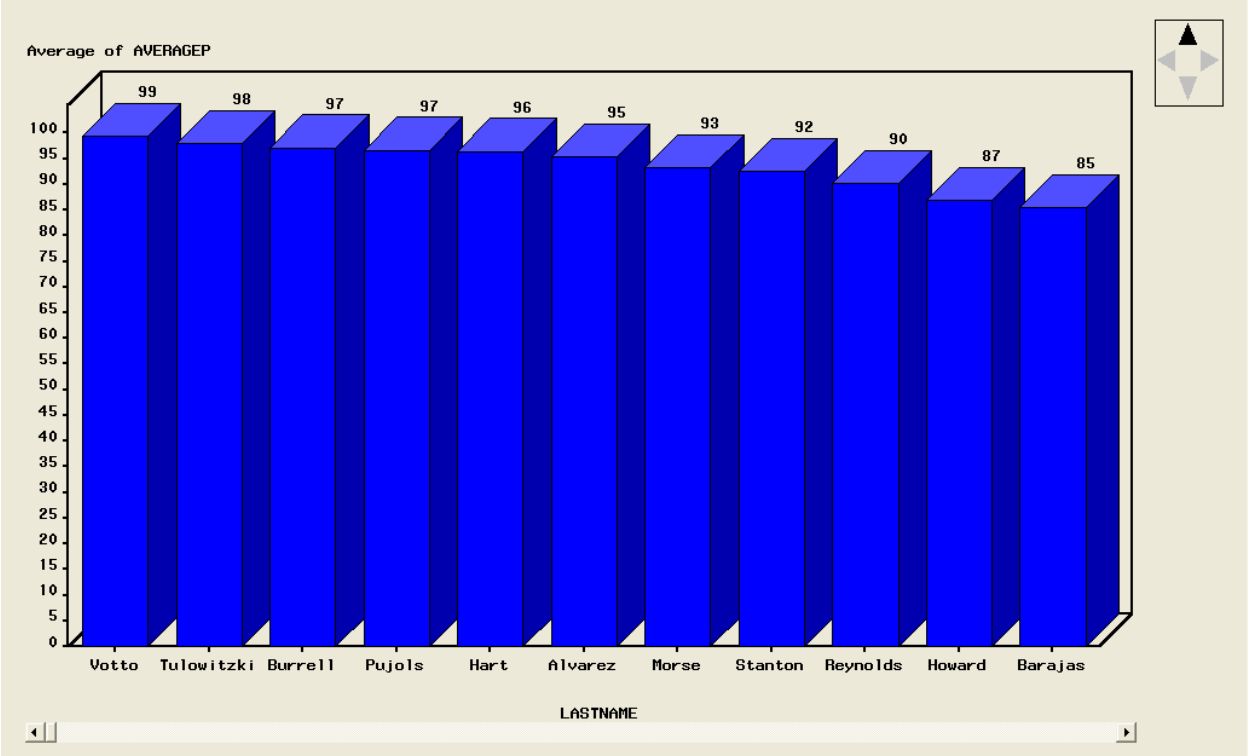
In summary, while the results from this study provide considerable insight into the run batted in statistic and create several interesting conclusions, a more expansive study across years would allow for a few more outcomes of interest along with hopefully confirming the initial results shown herein. As a final wrap up, presented next are some of the GUI (Graphical User Interface) screens created from SAS for this project. These sample screens show graphical highlights of the data and allow for conclusions to be made about individual players or teams much easier.

SAS GUI SCREENSHOTS





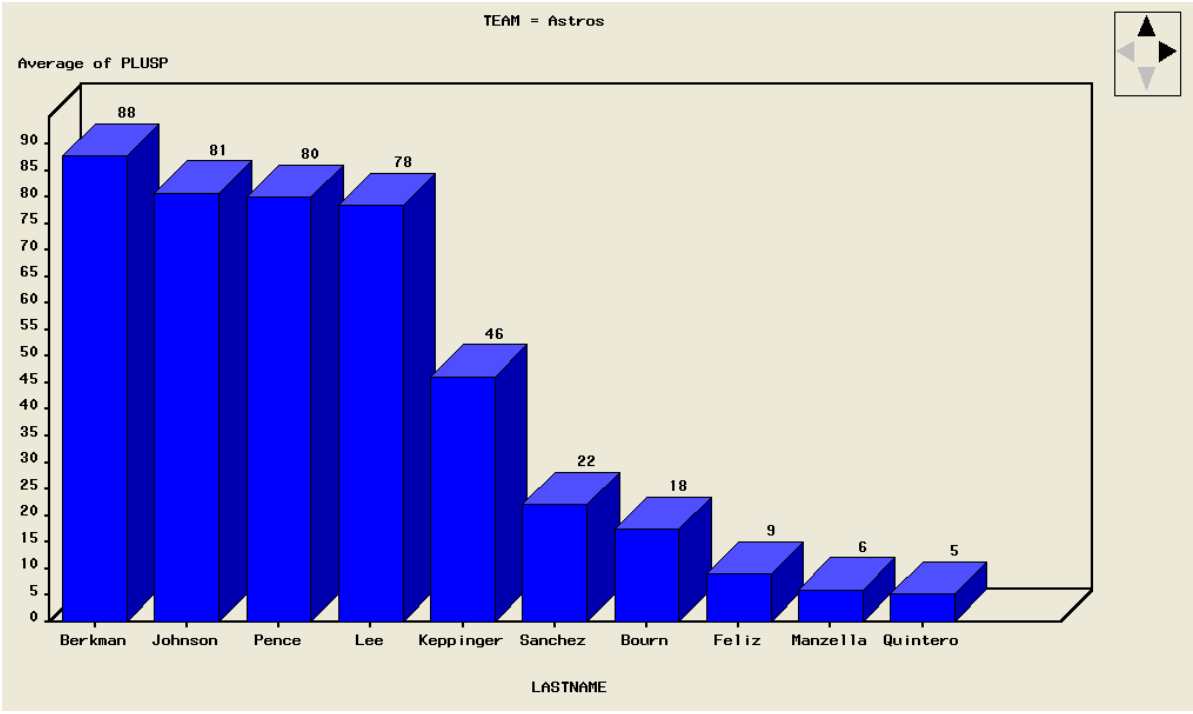

Average of AVERAGEP
64
61
57
54
53
52
50
50
49
48
47
0
5
10
15
20
25
30
35
40
45
50
55
60
65
Diamondbacks
Giants
Rockies
Reds
Marlins
Phillies
Cardinals
Cubs
Mets
Brewers
Pirates
TEAM

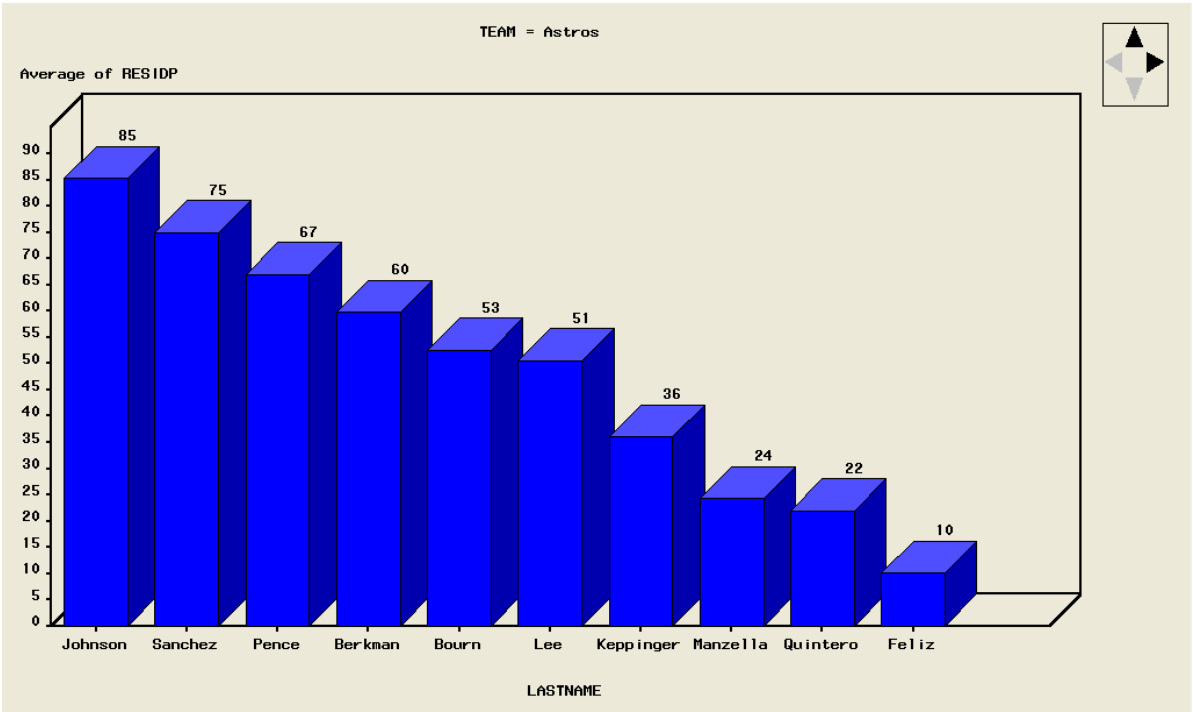

Average of AVERAGEP
99
98
97
97
96
95
93
92
90
87
85
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Votto
Tulowitzki
Burrell
Pujols
Hart
Alvarez
Morse
Stanton
Reynolds
Howard
Barajas
LASTNAME

Team = Astros

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
PLUSP	10	43.3030235	35.1156728	5.2991780	87.7228049
RESIDP	10	48.2842096	24.6494224	10.0722789	85.3275081
AVERAGEP	10	45.7936166	27.3248442	9.5349344	83.0401538





ACKNOWLEDGEMENTS

The author would first like to thank his advisor, Dr. Cecil Hallum, for going outside of his comfort zone of basketball into baseball, and hanging with him through endless discussions on the subject. Also, the author would like to thank his father for helping him with the baseball side of this project; many of the judgment calls made in this paper are a credit to countless phone conversations with him.

REFERENCES

Carroll, Will (2010), "Stop Making Sense: Baseball Advanced Stats Thru the iTunes Catalog of Stars," http://presscoverage.us/soapbox/stop-making-sense-baseball-advanced-stats-thru-the-itunes-catalog-of-stars.

Keri, Johan (2006), "What's the Matter with RBI? … and Other Traditional Statistics," *Baseball Between the Numbers*, New York: Basic Books.

Koop, Gary (2002), "Comparing the Performance of Baseball Players: A Multiple-Output Approach," *Journal of the American Statistical Association*, 97, 459, 710-720.

Montgomery, Douglass C., Peck, Elizabeth A. and Vining, G. Geoffrey (2006), *Introduction to Linear Regression Analysis*, New Jersey: Wiley.

Myers, Raymond H. and Milton, Janet S. (1998), *A First Course in the Theory of Linear Statistical Models*, New York: McGraw-Hill.

Neyer, Rob (2010), "Phillies pay top dollar for RBIs," http://espn.go.com/blog/sweetspot/post/_/id/3345/phillies-pay-top-dollar-for-rbis.

Perry, Dayn (2004), "Baseball Prospectus Basics," http://www.baseballprospectus.com/article.php?articleid=2562.

Ruane, Tom (1998), "RBI Production – A New Look at an Old Stat," http://www.baseballthinkfactory.org/btf/scholars/ruane/articles/rbi_production.htm.

Ruane, Tom (2005), "RBI Production – A New Look at an Old Stat," http://www.retrosheet.org/Research/RuaneT/rbipro_art.htm.

Skoog, Gary R. (1987), "Measuring Runs Created: The Value Added Approach," *The 1987 Bill James Abstract*, New York: Ballantine Books.

Tango, Tom M., Lichtman, Mitchel and Dolphin, Andrew (2007), *The Book: Playing the Percentages in Baseball*, Dulles, Virginia: Potomac Books.

Thorn, John and Palmer, Pete (1985), *The Hidden Game of Baseball*, New York: Doubleday.

CONTACT INFORMATION

Your comments and questions are encouraged. The author can be reached at ryan_sides@baylor.edu.