The secret behind “Moneyball”

Nabhan Kazi

Introduction

Baseball has always revolved around numbers and statistics. In recent years, the statistics part of baseball has become more emphasized and recognized, although it was still present back in the day to some extent.

A crucial part of baseball, nowadays, is sabermetrics. The term Sabermetrics was invented by Bill James. It was derived from SABR, which is an acronym for Society for American Baseball Research. Sabermetrics is known as "the science of learning about baseball through objective evidence" (Albert, 2010). Sabermetrics combines some old and new methods of evaluating players into an accurate way of determining player performance.

"Moneyball", by Michael Lewis, is the story of studying and scouting baseball players using a new data driven method that was first implemented in major league baseball by the Oakland Athletics' general manager, Billy Beane. In 2002 the Oakland Athletics was one of the poorest baseball teams in Major League Baseball. Compared to some of the richest teams, like the New York Yankees with a payroll of $132 million, the Oakland A's had only one-third of that, with a payroll of only $41 million. Billy realized that he was playing in an unfair game, where having the most money meant you had the most successful team. He knew that he could not succeed using the traditional methods of finding good players for his team. He realized that he had to figure out new and improved ways to stay alive in the game with his limited resources. So, with the help of his assistant, Paul DePodesta, Billy found a way to go around and beat the unfair system.

Billy Beane and Paul DePodesta

It dawned on Billy in 2001 that he could not compete with the other teams if he abided by the conventional ways of scouting players. He noticed that the older scouts used traditional methods to find players, which relied heavily on a player's prominence, the way he looked physically and back dated statistics that had been used for over a century. "Measuring employee talent in baseball has been and probably continues to be a subjective assessment based on limited, biased, qualitative data" (Wathen). Many of the scouts also strongly believed that they could foresee a player's future in professional baseball simply by the way he looked. Lewis writes, "they had a phrase they used: 'the Good Face'" (Lewis, 2011, p. 7). In conventional methods the scouts looked for specific attributes within a player. They called these attributes 'tools'. "There were five tools: the abilities to run, throw, field, hit, and hit with power. A guy who could run had "wheels"; a guy with a strong arm had "a hose"" (Lewis, 2011, p. 3).

When the scouts recruited Billy Beane in his senior year, they saw that he had all the 5 tools and he had "the Good Face." They heavily relied on what they witnessed in the tryouts and how Billy surpassed everyone in all the criteria. After he started playing pro ball, even then the scouts and the general manager failed to notice that his batting average dropped to 0.300 in senior year compared to 0.500 in his junior year (Lewis, 2011, p. 9). They also chose to ignore the fact that Billy had a very bad tendency of dealing with failure. Instead of learning from his failures, he would lash out in anger and lose control, and as a result lose his confidence on the field.

2002 was Beane's 5th year as the Oakland A's general manager. He decided that he has had enough and that it was time for a change. He could not compete with the other teams with the

limited resources that he had. He could not afford the players he wanted, lost the good ones he had and was stuck with creating a major league team with 1/3 of the money that richer teams had. Billy was fed up of the approach that the scouts took to recruit players. He believed that their approach was flawed and would not help them out in any way during the current crisis. Michael Lewis summarizes the approach of the scouts in three points:

> "there was, for starters, the tendency of everyone who actually played the game to generalize wildly from his own experience. People always thought that their own experience was typical when it wasn't. there was also a tendency to be overly influenced by a guy's most recent performance: what he did last was not necessarily what he would do next. Thirdly- but not lastly- there was the bias toward what people saw with their own eyes, or thought they had seen" (Lewis, 2011).

He sought new and more efficient ways of recruiting players, and that is when he met Paul DePodesta. If it had not been for Paul, Billy would have been stuck dealing with the old scouts all by himself. So why was Paul a crucial part of Billy's scouting team and why was Billy listening to Paul in the first place?

Paul DePodesta graduated from Harvard University with a degree in Economics. His first job was with the Cleveland Indians for three years, before he was hired as the assistant to the general manager of the Oakland Athletics in 1999. Lewis tells us that unlike major league baseball, Paul really pays attention to how frequently a college player walks. Over the years, Paul has researched the factors that led amateur players to become professional players and the reasons that caused others to not be able to reach the big leagues. From his studies he learnt that there

are some skills that were overvalued and sought after, even though they were, in fact, not that substantial. For example, scouts and managers put a lot of emphasis on foot speed, fielding ability and even raw power. But Paul figured out that the factor that played a crucial role in determining a player's accomplishment in the big leagues was his ability to control the strike zone. And a player's ability to control the strike zone could be determined from the number of walks that he drew (Lewis, 2011, p. 33). He also used statistical analysis on a large sample of players to figure out that on-base percentage and pitches-per-plate are very important indicators of a strong offense. Paul was strictly instructed by Billy not to bother explaining the reasonings behind their new and innovative ways of recruiting to the older scouts. He said that they would not be able to comprehend it at all and that "when you try to explain probability theory to baseball guys, you just end up confusing them" (Lewis, 2011, p. 34). Paul's research and analyses also made him favor college players over high school players, because college players were more experienced, and were used to playing against hard hitters and more intense competition. He favored college players also because they presented a bigger sample that Paul could work on and produce accurate results with minimal error or uncertainty.

Thus, using all of Paul's research and findings, Billy and Paul started the process of "performance scouting". Meaning that instead of going after a player's looks or his anticipated success, what should be focused on is his stats on the scoreboard and achievements. By hiring Paul and watching the numbers that his laptop spat out and the players it suggested, Billy had successfully taken an economic approach to building his team. More importantly, he had eliminated the bias of personal baseball experience and unnecessary factors.

Where it all started: James and Alderson

One might ask, which the older scouts definitely did, that why was Billy putting all this trust in a Harvard graduate and risking taking this new approach of recruiting players? The plain and simple answer is that it was not exactly the first time someone had tried to do something like this.

In the 1970's, Bill James, an American baseball article writer, was a firm believer in numbers and implementing statistics into baseball. He despised the notion of judging and evaluating players just by watching them play. He strictly believed that numbers spoke louder than fame and experience. To back up his argument he gives us an example, "One absolutely cannot tell, by watching, the difference between a 0.300 hitter and a 0.275 hitter" (Lewis, 2011, p. 68). He adds, "the difference between a good hitter and an average hitter is simply not visible- it is a matter of record" (Lewis, 2011, p. 68). He then came up with a way to be able to distinguish the abilities of one player from another using the number of successful plays he displayed in each game. He called this the "range factor". His persistent attempts to figure out the nitty gritty in baseball triggered something remarkable. Soon many other people started following his methods and joined his cause to hunt for new and undiscovered baseball insight. People, especially analysts and mathematicians, realized that "baseball, more than other sports, gave you meaningful things to count, and that by counting them you could determine the value of the people who played the game" (Lewis, 2011, p. 69).

Bill James decided to come up with a model that would project the number of runs a team would score in a game. He used the number of walks, hits, stolen bases, etc. as variables for his model. He played around with several different models until he finally figured out a formula that

correctly predicted the total run a team scored given number of hits, walks, bases, at bats. He called it the “runs created” formula:

```
Runs created= (hits + walks)*total bases / (at bats + walks)
```

This equation proved to be quite accurate for the most part. So basically, if the number of hits, walks, total bases and at bats are entered for a team from previous years, the projected runs created proved to be very similar to the runs that they actually scored during that year. This meant that Bill James was definitely on the right track and had possibly struck gold. This also meant that what he has been preaching all that time was true: that baseball experts “didn’t place enough value on walks and extra base hits, which featured prominently in the “runs Created” model and placed too much value on batting average and stolen bases, which James didn’t even bother to include” (Lewis, 2011, p. 77). James stated that the significant correlation between his variables, the projected runs and the actual run scored meant that these variables are not to be taken lightly. These variables are an essential part of projecting teams’ stats, and therein, baseball itself. In 1977, Bill James started publishing the details of his research and findings to his annual baseball journals and abstracts. At first, his abstracts were not popular at all. But few years down the line, when people started realizing that he was onto something extraordinary, his abstracts and journals became very popular and well known.

Bill James was definitely headed in the right direction. But the main reason that his ideas and systematic search to discover the mysteries of baseball did not pan out as successfully as he had hoped for was because he lacked a proper sample size. Statistical analysis heavily relies on determining statistically significant results from large sample sizes. The bigger the sample size of the experiment the significant the outcomes would be, and the more reliable the experiment will

be. Reliability, replicability and validity are three very important aspects of statistical analysis. If there is low reliability and validity in an experiment, then that experiment would be deemed insignificant and untrue.

Now we fast forward to 1983, when Sandy Alderson was hired as the Oakland Athletics' general manager. He had no previous baseball experience. He studied Law at Harvard University and later worked as a Marine Corps Officer. When he became general manager, he started reading some of Bill James' journals, and as it turns out, he became a huge fan and a strong believer of James' ways right away. To better understand the game, Alderson decided to use scientific analysis to evaluate players and determine their on-field strategies. He used past baseball data to conduct hypothesis tests and statistical analyses, instead of relying on the judgements of the baseball scouts. He believed, "by analyzing baseball statistics you could see through a lot of baseball nonsense" (Lewis, 2011, p. 57). Alderson found out from his analyses that contrary to what the scouts and other general managers believed, the number of runs scored had very little to do with batting average, and instead had a much better correlation with on-base and slugging percentages. Alderson realized that "a lot of the offensive tactics that made baseball managers famous- the bunt, the steal, the hit and run- could be proven to have been, in most situations, either pointless or self-defeating" (Lewis, 2011, p. 57). Alderson had no prior math or statistical experience, but he said that he knew what a regression analysis was, and the results that it spat out made sense to him.

A decade later, in 1993, Billy became Alderson's assistant. He learned a lot from Alderson, and the methods that Alderson used to evaluate players really opened his eyes. He had finally found someone whose views on baseball were objectively based on numbers and stats instead of

subjective opinions based on personal experience. Billy Beane later learnt that Alderson's ideas and views of baseball had derived from Bill James, the father of sabermetrics. Bill James' journals and abstracts were an eye opener for Alderson, and in turn an eye opener for Billy Beane.

Therefore, in 2001 and 2002, Billy Beane and his protégé, Paul DePodesta set out to follow in the footsteps of Alderson and Bill James.

Moneyball: the movie

Moneyball was depicted into a movie in 2011, starring Brad Pitt as Billy Beane and Jonah Hill playing the fictional character of Paul DePodesta as Peter Brand. Paul's character's physical attributions were slightly changed in the movie because he did not want to be affiliated with the movie, but his contributions to Moneyball and to major league baseball in general remained unchanged throughout the movie. Hence, we do not really miss out on any critical information. The movie was critically acclaimed worldwide, received several awards, and was even Oscar nominated for best picture. In the movie, when Peter (Paul's fictional character) first encounters Billy, he explains to him that baseball is not as it used to be back in the day. He tells him that there is a crucial aspect that people are not acknowledging and accounting for. He says, "there is an epidemic failure within the game to understand what is really happening. And this leads people who run major league baseball teams to misjudge their players and mismanage their teams" (Moneyball, 2011). He tells Billy that the conventional ways of scouting players are flawed and that the scouts are focusing on the wrong things. He adds, "people who run ball cubs, they think in terms of buying players. Your goal shouldn't be to buy players, your goal should be to buy wins. And in order to buy wins, you need to buy runs" (Moneyball, 2011).

Statistical Analysis

The following processes and analyses have been replicated from *The Analytics Edge-Unit 2: Moneyball* by Giovanni Fossati. He used Major League Baseball (MLB) Team Stats data from espn.com and compiled it into a csv file called *"baseball".*

I used some of my own names and variables, but I followed the same steps and guidelines as Giovanni to come up with similar results.

At first, I set the proper directory and I installed the necessary packages.

```
> setwd("C:/Users/Nabz/Dropbox/moneyball data/data")
> install.packages("ggplot2")
> library(ggplot2)
```

Then I read in the csv file containing all the teams' stats. I displayed all the variables and their types and displayed the first few lines of the dataset.

```
> baseball_data= read.csv("baseball.csv",header=T)
> str(baseball_data)
```

```
'data.frame':    1232 obs. of  15 variables:
 $ Team        : Factor w/ 39 levels "ANA","ARI","ATL",..: 2 3 4 5 7 8 9 10 11 12 ...
 $ League      : Factor w/ 2 levels "AL","NL": 2 2 1 1 2 1 2 1 2 1 ...
 $ Year        : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
 $ RS          : int  734 700 712 734 613 748 669 667 758 726 ...
 $ RA          : int  688 600 705 806 759 676 588 845 890 670 ...
 $ W           : int  81 94 93 69 61 85 97 68 64 88 ...
 $ OBP         : num  0.328 0.32 0.311 0.315 0.302 0.318 0.315 0.324 0.33 0.335 ...
 $ SLG         : num  0.418 0.389 0.417 0.415 0.378 0.422 0.411 0.381 0.436 0.422 ...
 $ BA          : num  0.259 0.247 0.247 0.26 0.24 0.255 0.251 0.251 0.274 0.268 ...
 $ Playoffs    : int  0 1 1 0 0 0 1 0 0 1 ...
 $ RankSeason  : int  NA 4 5 NA NA NA 2 NA NA 6 ...
 $ RankPlayoffs: int  NA 5 4 NA NA NA 4 NA NA 2 ...
 $ G           : int  162 162 162 162 162 162 162 162 162 162 ...
 $ OOBP        : num  0.317 0.306 0.315 0.331 0.335 0.319 0.305 0.336 0.357 0.314 ...
 $ OSLG        : num  0.415 0.378 0.403 0.428 0.424 0.405 0.39 0.43 0.47 0.402 ...
```

```
> head(baseball_data)
```

```
Team League Year  RS  RA  W   OBP   SLG    BA Playoffs RankSeason RankPlayoffs   G  OOBP  OSLG
1  ARI     NL 2012 734 688 81 0.328 0.418 0.259        0         NA           NA 162 0.317 0.415
2  ATL     NL 2012 700 600 94 0.320 0.389 0.247        1          4            5 162 0.306 0.378
3  BAL     AL 2012 712 705 93 0.311 0.417 0.247        1          5            4 162 0.315 0.403
4  BOS     AL 2012 734 806 69 0.315 0.415 0.260        0         NA           NA 162 0.331 0.428
5  CHC     NL 2012 613 759 61 0.302 0.378 0.240        0         NA           NA 162 0.335 0.424
6  CHW     AL 2012 748 676 85 0.318 0.422 0.255        0         NA           NA 162 0.319 0.405
```

The basic gist of winning a game in baseball is to score more runs than the opponent. The Oakland Athletics, in 2002, used a formula to calculate how many games they needed to win to be able to make it to the postseason, how many runs needed to score in order to win those games, and how many runs they can afford to allow the opponents to score. The formula was called the win percentage formula:

$$Win\% = (Runs\ scored)^2 / (Runs\ scored^2 + Runs\ allowed^2)$$

Accordingly, they figured out that they needed to win 95 games to be able to proceed to the next season, and that they needed to score at least 135 runs more than they allowed to be able to accomplish that.

To check if their formula and predictions held true, I will perform a linear regression using the data from the years before Moneyball. First, I created a new variable called *moneyball*, and subset all the data before 2002 into it.

```
> moneyball = subset(baseball_data, Year < 2002)
> str(moneyball)

'data.frame':    902 obs. of  15 variables:
 $ Team        : Factor w/ 39 levels "ANA","ARI","ATL",..: 1 2 3 4 5 7 8 9 10 11 ...
 $ League      : Factor w/ 2 levels "AL","NL": 1 2 2 1 1 2 1 2 1 2 ...
 $ Year        : int  2001 2001 2001 2001 2001 2001 2001 2001 2001 2001 ...
 $ RS          : int  691 818 729 687 772 777 798 735 897 923 ...
 $ RA          : int  730 677 643 829 745 701 795 850 821 906 ...
 $ W           : int  75 92 88 63 82 88 83 66 91 73 ...
 $ OBP         : num  0.327 0.341 0.324 0.319 0.334 0.336 0.334 0.324 0.35 0.354 ...
 $ SLG         : num  0.405 0.442 0.412 0.38 0.439 0.43 0.451 0.419 0.458 0.483 ...
 $ BA          : num  0.261 0.267 0.26 0.248 0.266 0.261 0.268 0.262 0.278 0.292 ...
 $ Playoffs    : int  0 1 1 0 0 0 0 0 1 0 ...
 $ RankSeason  : int  NA 5 7 NA NA NA NA NA 6 NA ...
```

```
 $ RankPlayoffs: int  NA 1 3 NA NA NA NA NA 4 NA ...
 $ G           : int  162 162 162 162 161 162 162 162 162 162 ...
 $ OOBP        : num  0.331 0.311 0.314 0.337 0.329 0.321 0.334 0.341 0.341 0.35 ...
 $ OSLG        : num  0.412 0.404 0.384 0.439 0.393 0.398 0.427 0.455 0.417 0.48 ...
```

I also created a new variable within the *moneyball* dataset, called *RD*, which calculates the run difference between runs scored and runs allowed.

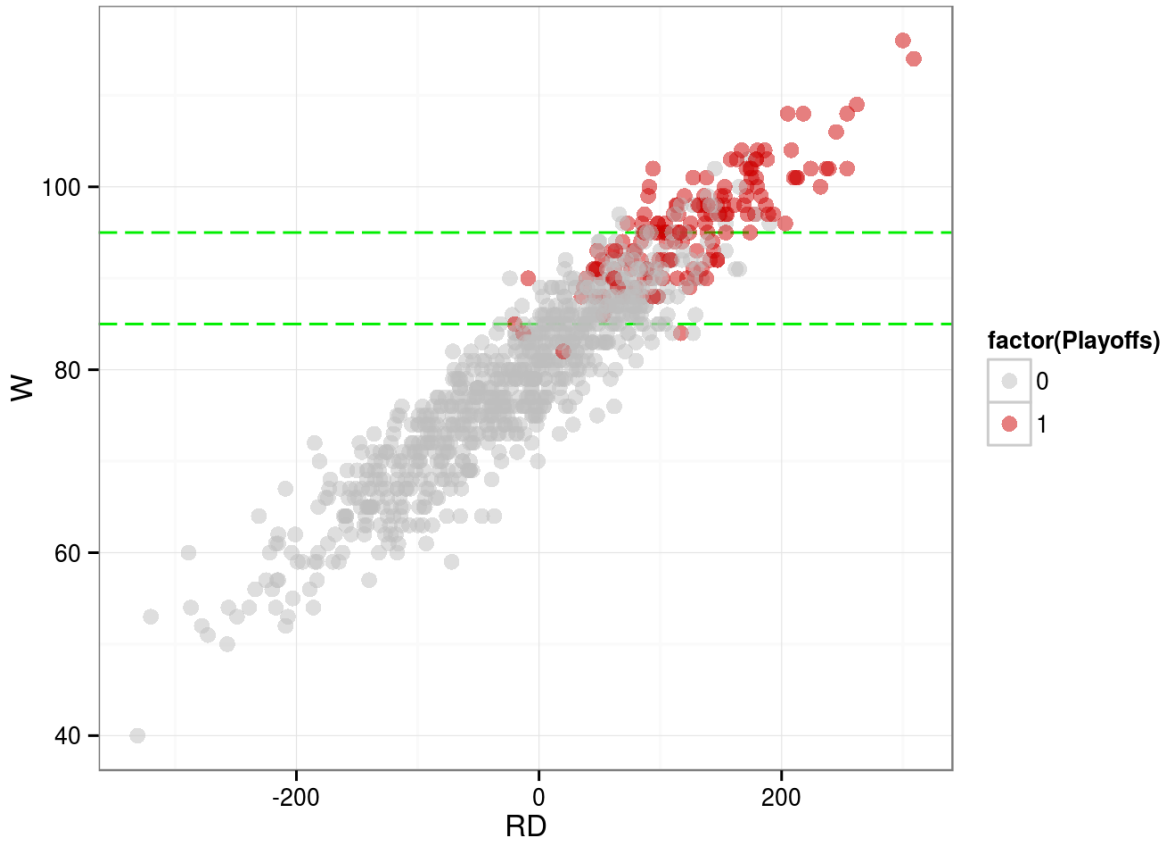
```
> moneyball$RD = moneyball$RS - moneyball$RA
> str(moneyball)

'data.frame':    902 obs. of  16 variables:
 $ Team        : Factor w/ 39 levels "ANA","ARI","ATL",..: 1 2 3 4 5 7 8 9 10 11 ...
 $ League      : Factor w/ 2 levels "AL","NL": 1 2 2 1 1 2 1 2 1 2 ...
 $ Year        : int  2001 2001 2001 2001 2001 2001 2001 2001 2001 2001 ...
 $ RS          : int  691 818 729 687 772 777 798 735 897 923 ...
 $ RA          : int  730 677 643 829 745 701 795 850 821 906 ...
 $ W           : int  75 92 88 63 82 88 83 66 91 73 ...
 $ OBP         : num  0.327 0.341 0.324 0.319 0.334 0.336 0.334 0.324 0.35 0.354 ...
 $ SLG         : num  0.405 0.442 0.412 0.38 0.439 0.43 0.451 0.419 0.458 0.483 ...
 $ BA          : num  0.261 0.267 0.26 0.248 0.266 0.261 0.268 0.262 0.278 0.292 ...
 $ Playoffs    : int  0 1 1 0 0 0 0 0 1 0 ...
 $ RankSeason  : int  NA 5 7 NA NA NA NA NA 6 NA ...
 $ RankPlayoffs: int  NA 1 3 NA NA NA NA NA 4 NA ...
 $ G           : int  162 162 162 162 161 162 162 162 162 162 ...
 $ OOBP        : num  0.331 0.311 0.314 0.337 0.329 0.321 0.334 0.341 0.341 0.35 ...
 $ OSLG        : num  0.412 0.404 0.384 0.439 0.393 0.398 0.427 0.455 0.417 0.48 ...
 $ RD          : int  -39 141 86 -142 27 76 3 -115 76 17 ...
```

As we can see, a 16th variable, *RD*, has been added.

Then I carried out a scatterplot to check the relationship between the run difference and the number of wins.

```
> ggplot(data = moneyball, aes(x = RD, y = W)) + theme_bw() + scale_color_man
ual(values = c("grey", "red3")) + geom_hline(yintercept = c(85.0, 95.0), col
= "green2", linetype = "longdash") +geom_point(aes(color = factor(Playoffs)),
alpha = 0.5, pch = 16, size = 3.0)
```



There seemed to be a very strong positive correlation between run difference and wins, so I carried on to run a linear regression. I used *RD* as my predictor first.

```
> wins_rd = lm(W ~ RD, data = moneyball)
> summary(wins_rd)

Call:
lm(formula = W ~ RD, data = moneyball)

Residuals:
     Min       1Q   Median       3Q      Max
-14.2662  -2.6509   0.1234   2.9364  11.6570

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 80.881375   0.131157  616.67   <2e-16 ***
RD           0.105766   0.001297   81.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.939 on 900 degrees of freedom
Multiple R-squared:  0.8808,   Adjusted R-squared:  0.8807
F-statistic:  6651 on 1 and 900 DF,  p-value: < 2.2e-16
```

The p-value was much less than 0.05 with a strong adjusted R-squared value of 88%. Hence, the formula to predict wins from given Run differences can be written as follows:

```
W = 80.881 + (0.106 * RD)
```

In this case, since we know that Paul calculated they needed to win 95 games, our *W* is >= 95. Using the above formula, I calculated the *RD* to be:

```
W >= 95
80.881 + (0.106 * RD) >= 95
RD >= (95 - 80.881) / 0.106
RD >= 133.5
```

Which can be rounded up to the nearest integer to be ~134. Hence, yes, Paul's math checks out to be correct, that however many runs they allowed, they needed to score about 135 more than that.

The Oakland A's used Alderson's findings that *runs scored* had a much better correlation with on-base percentage *(OBP)* and slugging percentage *(SLG)* than with batting average *(BA).* On-base percentage is the percentage of time a player gets on base (including walks). Slugging percentage is how far a player gets around the base on his turn (measures power). Batting average is when a player gets on base by hitting the ball.

Paul and Billy believed that on-base and slugging percentages were the most important stats in determining performance, and that batting average was just an overestimated stat. Ergo, I will try to verify using linear regression whether their notion was correct.

First, I will run a regression using all three variables to check which variables have a significant effect on runs scored.

```
> rs_reg1 = lm(RS ~ OBP + SLG + BA, data = moneyball)
> summary(rs_reg1)

Call:
lm(formula = RS ~ OBP + SLG + BA, data = moneyball)

Residuals:
    Min      1Q  Median      3Q     Max
-70.941 -17.247  -0.621  16.754  90.998

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -788.46      19.70 -40.029  < 2e-16 ***
OBP          2917.42     110.47  26.410  < 2e-16 ***
SLG          1637.93      45.99  35.612  < 2e-16 ***
BA           -368.97     130.58  -2.826  0.00482 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.69 on 898 degrees of freedom
Multiple R-squared:  0.9302,   Adjusted R-squared:   0.93
F-statistic:  3989 on 3 and 898 DF,  p-value: < 2.2e-16
```

From the regression table, it can be seen that both *OBP* and *SLG* have p-values of much less than 0.05 and adjusted R-squared value of 93%. However, *BA* has a p-value of 0.236, which implies insignificance. Moreover, the coefficient of *BA* is negative, which means runs scored is inversely proportional to batting average, which does not make sense at all. This is most likely due to multicollinearity between the predictor variables. So, I will run the regression again, but this time without the batting average as one of my predictors.

```
> rs_reg2 = lm(RS ~ OBP + SLG, data = moneyball)
> summary(rs_reg2)

Call:
lm(formula = RS ~ OBP + SLG, data = moneyball)

Residuals:
    Min      1Q  Median      3Q     Max
-70.838 -17.174  -1.108  16.770  90.036

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -804.63      18.92  -42.53   <2e-16 ***
OBP          2737.77      90.68   30.19   <2e-16 ***
SLG          1584.91      42.16   37.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 24.79 on 899 degrees of freedom
Multiple R-squared:  0.9296,   Adjusted R-squared:  0.9294
F-statistic:  5934 on 2 and 899 DF,  p-value: < 2.2e-16
```

Again, both *OBP* and *SLG* are statistically significant with positive coefficients, p-values of much less than 0.05 and adjusted R-squared of 92.94%. This means that *OBP* and *SLG* are in fact statistically significant predictors of runs scored while *BA* is overestimated, meaning that Billy and Paul's approach was accurate. The formula is:

```
RS = -804.63 + (2737.77*OBP) + (1584.91*SLG)
```

We can also use the data to calculate the number of runs that could be allowed using the stats for opponents on-base percentage *(OOBP)* and opponents slugging percentage *(OSLG).* I ran a linear regression using *OOBP* and *OSLG* as my predictor variables to predict runs allowed.

```
> ra_reg = lm(RA ~ OOBP + OSLG, data = moneyball)
> summary(ra_reg)

Call:
lm(formula = RA ~ OOBP + OSLG, data = moneyball)

Residuals:
    Min      1Q  Median      3Q     Max
-82.397 -15.178  -0.129  17.679  60.955

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -837.38      60.26 -13.897  < 2e-16 ***
OOBP         2913.60     291.97   9.979 4.46e-16 ***
OSLG         1514.29     175.43   8.632 2.55e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.67 on 87 degrees of freedom
  (812 observations deleted due to missingness)
Multiple R-squared:  0.9073,   Adjusted R-squared:  0.9052
F-statistic: 425.8 on 2 and 87 DF,  p-value: < 2.2e-16
```

From the regression model we can see that *OOBP* and *OSLG* are very significant predictors of runs allowed with p-values less than 0.05 and adjusted r-squared of 90.52%. The formula is shown as follows:

```
RA = -837.38 + (2913.60*OOBP) + (1514.29*OSLG)
```

Since I have the formulas for predicted wins, runs scored and run allowed figured out, I can now use the data from 2001 to predict how many games the Oakland Athletics would have won in 2002. Then we can compare my predicted values, with Paul's predicted values for the A's and the A's actual number of wins.

These are the stats the Oakland A's had for the 2001 season.

```
OBP = 0.345, SLG = 0.439, OOBP = 0.308, OSLG = 0.38
```

Using these, we can calculate *RS* and *RA* as follows:

```
RS = -804.63 + (2737.77*OBP) + (1584.91*SLG)
   = -804.63 + (2737.77*0.345) + (1584.91*0.439)
   = 835.676
RA = -837.38 + (2913.60*OOBP) + (1514.29*OSLG)
   = -837.38 + 2913.60*0.308 + 1514.29*0.38
   = 635.439
```

Now that we have runs scored and runs allowed, we can use these to calculated the run difference. Then we can sub in the run difference value into the predicted wins formula to get the number of predicted wins for the 2002 season.

```
RD = RS - RA
   = 200.237
W = 80.881 + (0.106 * RD)
  = 80.881 + (0.106 * 200.237)
  = 102.106
```

The table below shows the predictions that I made for the Oakland A's 2002 season using the stats for the 2001 season and all the above formulas. The table also compares my predicted values with Paul's predicted values and the Oakland A's actual values for the 2002 season.

	My prediction	Paul's prediction	Actual
Runs Scored	835.676	800 – 820	800
Runs Allowed	635.439	650 - 670	653
Wins	102.106	93 - 97	103

As we can see, my predicted runs scored and runs allowed was slightly off compared to their actual runs, but my predicted wins were just one-win shy off their actual number of wins. Their actual numbers were within the range that Paul predicted, and they ended up winning more games than Paul predicted, which worked out for the best for them. That year, they ended up setting a League record by winning 20 games in a row.

Conclusion

Billy Beane and Paul DePodesta, with the help of Bill James and Sandy Alderson, showed the world in 2002 that there is a lot more to baseball than meets the eye. Baseball is not just about random and boring numbers. There is an art to it, and with the correct tools and proper interpretation, these numbers and stats can take you to the championships. Billy and Paul, using statistical analysis and the relevant stats of a player, figured out how to build a Major League team with limited funding and undervalued players. They changed the game. Now people all over the world and in other sports are using statistical analysis to determine the value of their players. Billy is still waiting for the Oakland A's to win the Championship one day.

References

- Lewis, M. (2011). *Moneyball: The Art of Winning an Unfair Game*. S.l.: WW Norton & Company.
- Fossati, G. (2018). *The Analytics Edge - Unit 2 : Moneyball*. [online] Rstudio-pubs-static.s3.amazonaws.com. Available at: https://rstudio-pubs-static.s3.amazonaws.com/71693_88f6f5dea1ea4d9aa4808dd526d38429.html [Accessed 7 Apr. 2018].
- Coachup.com. (2018). *The Magic Formula Behind Moneyball*. [online] Available at: https://www.coachup.com/nation/articles/the-magic-baseball-formula-behind-mo [Accessed 7 Apr. 2018].
- Albert, J. (2010). Sabermetrics: The Past, the Present, and the Future. In J. Gallian (Ed.), *Mathematics and Sports* (pp. 3-14). Mathematical Association of America. doi:10.5948/UPO9781614442004.002 [Accessed 7 Apr. 2018].
- Wassermann, E. (2005). *An Examination of the Moneyball Theory: A Baseball Statistical Analysis*. [online] The Sport Journal. Available at: http://thesportjournal.org/article/an-examination-of-the-moneyball-theory-a-baseball-statistical-analysis/ [Accessed 7 Apr. 2018].
- *2002 MLB Team Batting Stats - Major League Baseball - ESPN*. [online] ESPN.com. Available at: http://www.espn.com/mlb/stats/team/_/stat/batting/year/2002 [Accessed 7 Apr. 2018].

- A.espncdn.com. (2018). *ESPN.com - MLB Playoffs 2002 - Baseball's 2002 payrolls*. [online] Available at: http://a.espncdn.com/mlb/playoffs2002/s/2002/1011/1444560.html [Accessed 7 Apr. 2018].
- Wright, R. (2018). *Moneyball: A Look Inside Major League Baseball and the Oakland A's*. [online] Bleacher Report. Available at: http://bleacherreport.com/articles/858470-moneyball-a-look-inside-major-league-baseball-and-the-oakland-as [Accessed 7 Apr. 2018].
- espnW. (2018). *Amanda Rykoff: Inside the stats that created 'Moneyball'*. [online] Available at: http://www.espn.com/espnw/news-commentary/article/7577771/stats-created-moneyball [Accessed 7 Apr. 2018].
- Wathen, S. (2018). *An Application of Statistics: Using the "Moneyball" Story in a Basic Statistics Course*. [online] Southeastinforms.org. Available at: http://southeastinforms.org/Proceedings/2013/proc/p130520003.pdf [Accessed 7 Apr. 2018].
- Oracle.com. (2018). *The Moneyball Method of Field Service Management*. [online] Available at: http://www.oracle.com/us/products/applications/moneyball-method-2412955.pdf [Accessed 7 Apr. 2018].
- SwingSmarter.com. (2018). *Oakland A's: Modern Day David & Goliath & What This Means to YOU*. [online] Available at: http://www.swing-smarter-baseball-hitting-drills.com/oakland-as.html [Accessed 7 Apr. 2018].
- Mathgoespop.com. (2011). *Moneyball*. [online] Available at: http://www.mathgoespop.com/2011/09/moneyball.html [Accessed 7 Apr. 2018].
- *Moneyball*. (2011). [DVD] Directed by B. Miller.