Moneyball (Sports analytics)

Tool: Linear regression

The Analytics Edge,

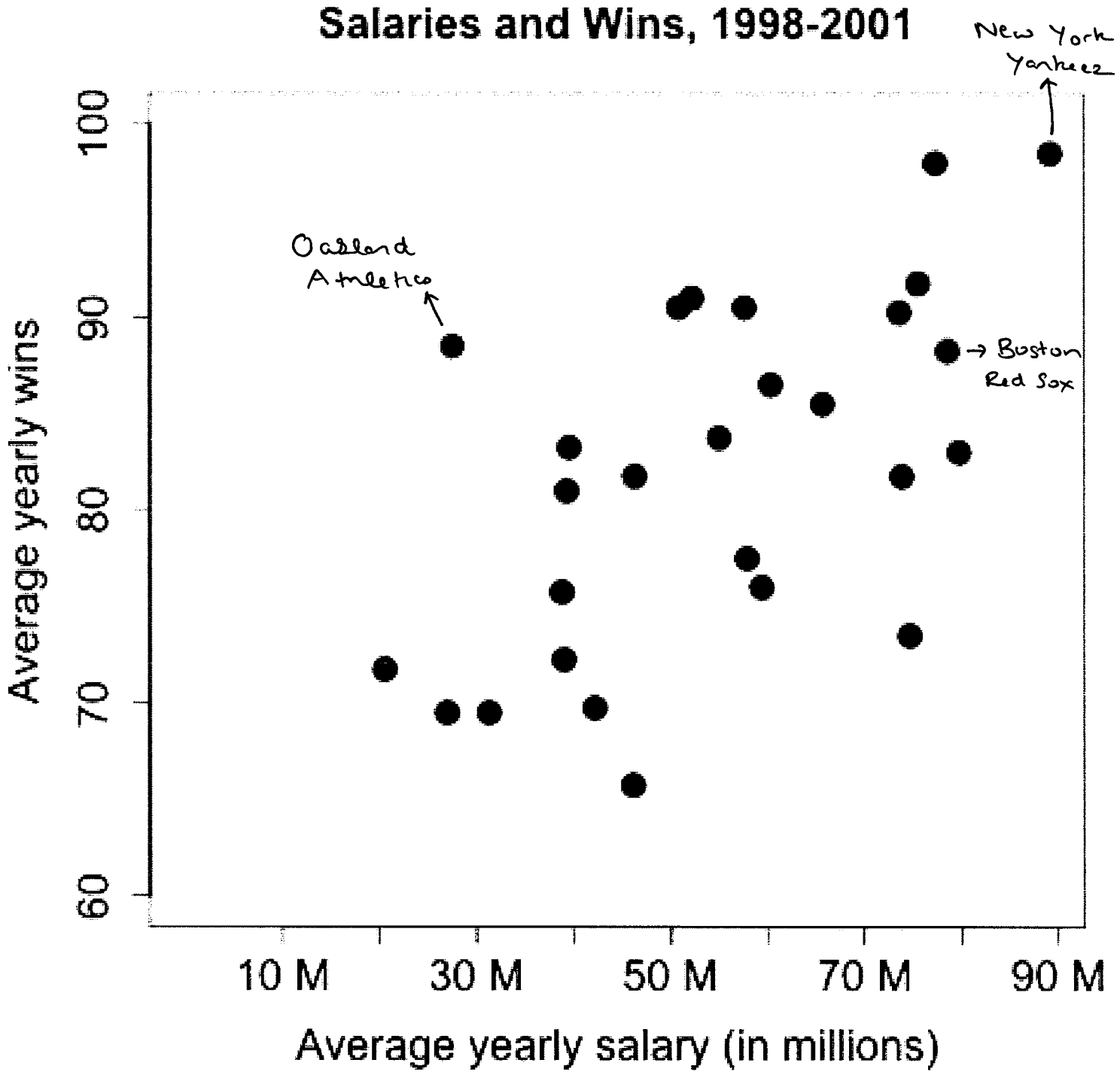
Using tools of analytics, managers can accurately value players better & minimize risk from just gut feeling evaluation of players. The hitter salaries did not reflect the contribution of batting skills to winning games which once found could be exploited to gain an advantage through having substantially lesser money to build a team. This is what Moneyball discusses.

Overview

In 2003, Michael Lewis wrote the book Moneyball which talks about how Oakland Athletics, a baseball team playing Major League Baseball in Americ identified a group of undervalued professional baseball players to turn themselves into one of the most successful franchisee. The key question that Moneyball tries to address is: How did Oakland Athletics, one of the poorest teams in Major League Baseball win so many games?

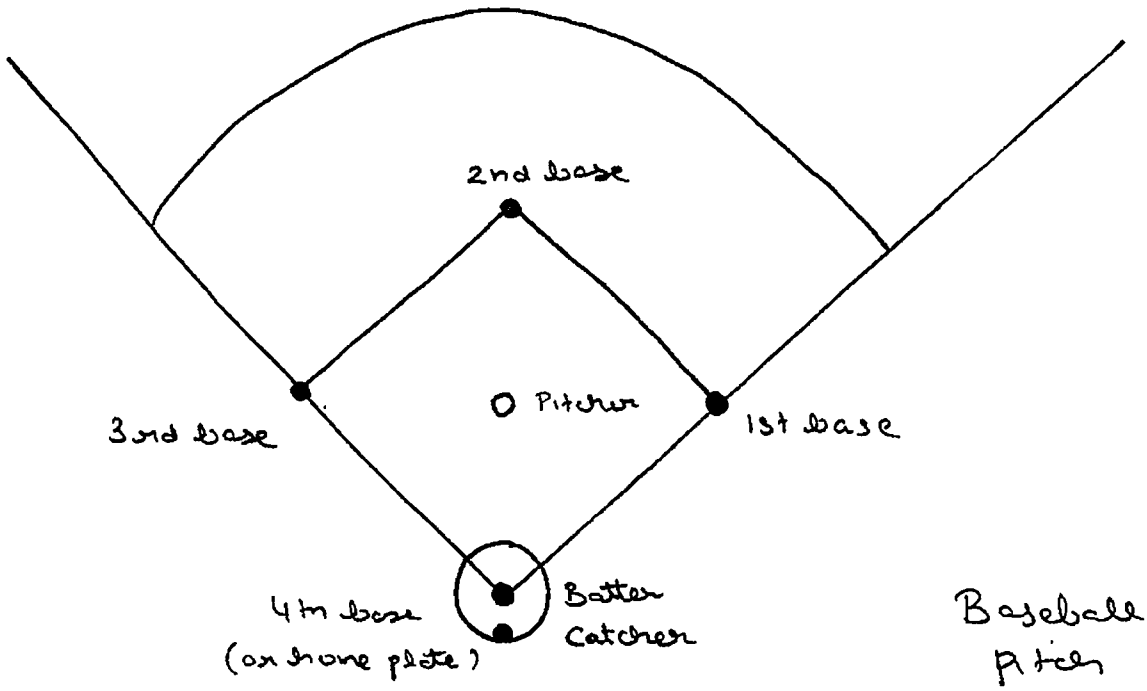
The story is about Billy Beane, the Oakland A's general manager who was willing to discard old wisdom to get an edge over big money.

In 2011 Moneyball was made into a Hollywood movie starring Brad


Salaries and Wins, 1998-2001
New York Yankees
Oakland Athletics
Boston Red Sox
Average yearly wins
100
90
80
70
60
10 M
30 M
50 M
70 M
90 M
Average yearly salary (in millions)

Baseball

Baseball is played between two teams with nine players in the field on each team.
In Major League Baseball, the game is played in nine innings in which each team gets to bat and score runs while the other team pitches and defends in the field. An innings is broken into two halves where one team bats in the first half and then the other team bats. The teams switch every time the defending team gets three of the players from the batting team out. The winner is the team with the most runs after nine innings.



Baseball Pitch

Sabermetrics : Empirical analysis of baseball especially baseball statistics that measure in-game activity (The search of objective knowledge about baseball)

Bill James : Considered as the father of Sabermetrics due to a series of books on baseball statistics in the 1970s.

Billy Beane : General Manager of Oakland Athletics who played Major league Baseball for a few years in the 1980s. He was picked by the scouts as a star but his own playing career failed to meet the expectations. As a general manager his record in the early years was :

	Won	Lost	Played
1998	74	88	162
1999	87	75	162
2000	91	71	162*
2001	102	60	162*
2002	103	59	162*
2003	96	66	162*
2004	91	71	162*

* indicates OAKLAND made it to playoffs

Paul Podesta : Helped Billy Beane use tools of analytics to merge Oakland Athletics. Currently vice president of player development & Scouting for New York Mets

1998 ⟶ 2003

Billy Beane hired as general manager of Oakland Athletics

He hires Paul De Podesta in 1999

Moneyball is published with the story of how Oakland Athletics gained a competitive advantage with data & analytics

↓ 2004

Criticism from baseball quarters & teams react by hiring analytics people to help with player management.

Use of analytics in Sports increases with games such as basketball, soccer adopting more data driven approaches ←

In this set of notes, we will try to verify the use of analytics by Oakland Athletics in the book Moneyball.

Analytics on baseball data : R

```
baseball ← read.csv("baseball.csv")
str(baseball)
summary(baseball)
```

Dataset consists of 420 observations of 17 variables

Team (name of team)

League (American or National league)

Year 1999 – 2012

Games (Number of games played. 161 to 163)

W (Number of wins)

RS (Run scored)

RA (Run agaist)

OBP (on base percentage)

SLG (slugging percentage)

BA (Batting average)

OPS (On base plus slugging)

OOBP (opposition on base percentage)

OSLG (opposition slugging percentage)

OOPS (opposition on base plus slugging)

Playoffs (1 if team made playoff, 0 otherwise)

Rank season (rank in season)

Before the 2002 season Paul De Podesta reduced the planning for the upcoming 6 months to a mathematics problem.

His goal was to help Billy Beane form an Oakland Athletics team to make the playoffs with a low budget.

```
baseball2002 <- subset(baseball, Year<2002)
```

Create a subset of data set for 1999, 2000, 2001

To understand how many games are needed to make the playoffs

```
plot(baseball2002$W, baseball2002$Team,
     col=ifelse(baseball2002$Playoffs==1,
     "red", "black"))
axis(1, at=seq(60, 120, by=5))
abline(v=95)
```

Paul De Podesta judged that it would take 95 games to make it to playoffs. If they won 95 games & could not make it to playoffs, then that was bad luck.

So how many runs needed to be scored for and against to win these many games?.

Bill James, the founder of Sabermetrics had noted over time that there was a fairly stable relationship between the seasons run total (for - against) and the number of wins.

```
plot(baseball2002$RS - baseball2002$RA,
     baseball2002$W)
```

Fairly stable linear relationship
(↑ in difference ⇒ ↑ in win)

```
baseball2002$RD ← baseball2002$RS - baseball2002$RA
```
New variable

```
model1 ← lm(baseball2002$W ~ baseball2002$RD)
```

Performs a linear regression given the fairly stable +ve relationship between RD & W.

Obtain (model 1)

$$W = 80.92 + 0.099\,RD$$

To have $W \geq 95$, $RD \geq 142$

Note Paul DePodesta estimated it to be about 135 (This could depend on how much data he used)

Though baseball is a team sport, success is a function of the achievements of individual players which can be observed easily.

To measure batting skill the most commonly used statistics are:

1) BA (Batting average) = $\frac{\text{Number of hits}}{\text{Number of at-bats}}$

This is a measure of how often a batter reaches base by hitting safely. This measure is however crude since it ignores the added productivity from hits of more than a single base (singles & home runs are counted the same)

2) SLG (slugging percentage) = $\frac{\text{Total bases}}{\text{Number of at bats}}$

This is a more refined measure that counts doubles twice as much as singles, homeruns four times as much as singles.

3) OBP (on base percentage) is the fraction of plate appearances (including at-bats and walks) in which the player reached base successfully through either a hit or a walk. Unlike batting avg it does not care about how the player gets on base

4) OPS (one base + slugging): This statistic was simply the addition of SLG and OBP.

Note that OPS added the two statistics together and it was still unclear what the relative importance was of these two statistics.

Example :

Suppose a team has OBP = 1. Every player who comes to bat, gets to first base and team would never get out. Thus it would score an infinite number of runs.

Suppose a team has SLG = 1. Then it is possible for example every player to get to first base & score infinite number of runs. Or, one player could get a home run (4 bases) while 3 players do not get on base at all. Four bases would be gained by 4 hitters & SLG = 1 but team would score only 1 run. This in this case, OBP should mean more than SLG.

Summary (baseball 2002 $ SLG)
Summary (baseball 2002 $ OBP)

} Slugging % is between 0.38 & 0.48 while OBP is between 0.31 and 0.37

Roughly similar scale,

Predicting runs scored from OBP, SLG, BA, OPS

```
m1 <- lm(RS ~ OBP, data = baseball2002)
m2 <- lm(RS ~ SLG, data = baseball2002)
m3 <- lm(RS ~ OPS, data = baseball2002)
m4 <- lm(RS ~ BA, data = baseball2002)
m5 <- lm(RS ~ OBP + SLG, data = baseball2002)
```

	Model 1	2	3	4	5
Intercept	-1046***	-442***	-846***	-778***	-1014**
OBP	3452***				3562**
SLG		2896***			1412**
OPS			2142***		
BA				5917***	
R^2	0.841	0.762	0.894	0.613	0.921
Adj R^2	0.839	0.760	0.893	0.659	0.920

The higher coefficient for OBP in comparison to SLG suggests that an extra percentage point of on base % might be more valuable than an extra % of SLG. Note that OPS weighs them equally.

Paul De Podesta decided that it was <u>three</u> times as important

m6 ← lm(RS ~ OBP + SLG + BA, data = baseball 2002)

Adding the extra BA variable kept R^2 at 0.922 and adjusted R^2 to 0.9195. Also the variable is not significant. There is also possible multiple collinearity which is why we decide to stick with m5.

m7 ⇐ lm(RA ~ OOBP + OSLG, data = baseball 2002)

Summary (m7)

Intercept Coefficient −837,
OOBP 2913, OSLG 1514
$R^2 = 0.90$, adjusted $R^2 = 0.90$
All variables are significant

The coefficients from the regression support the claim in Moneyball that Paul De Podesta believed that OBP and OOBP had a significant contribution to RS and RA.

At the start of 2002, team OBP = 0.339, SLG = 0.43 based on player statistics

Plugging into RS = −1014 + 3562 OBP + 0.43 SLG gives RS = 800.675.

Actual value for RS = 800 while Paul De Podesta believed (predicted) it to be in [800, 820].

Similarly using player statistics OOBP = 0.307

OOPS = 0.373

$$RA = -837 + 0.307(2913) + 0.373(1514)$$
$$= 622.13$$

(Paul DePodesta predicted between 650 & 670)

Prediction on number of wins

$$\hat{W} = 80.92 + 0.099(800.678 - 622.13)$$
$$= 98.59$$

(Paul De Podesta predicted they would win between 93 & 97 games)

In 2002, Oakland A won 103 games.

This was despite losing Jason Giambi to New York Yankees. He was one of the stars winning the Most valuable Player in the American league in 2000. To replace him and two other player stars Johnny Damon & Jason Isringhausen, the threw injured Scott Hatteberg, unorthodox Submarine pitcher Chad Bradford, part his prime David Justice for a smaller lesser price

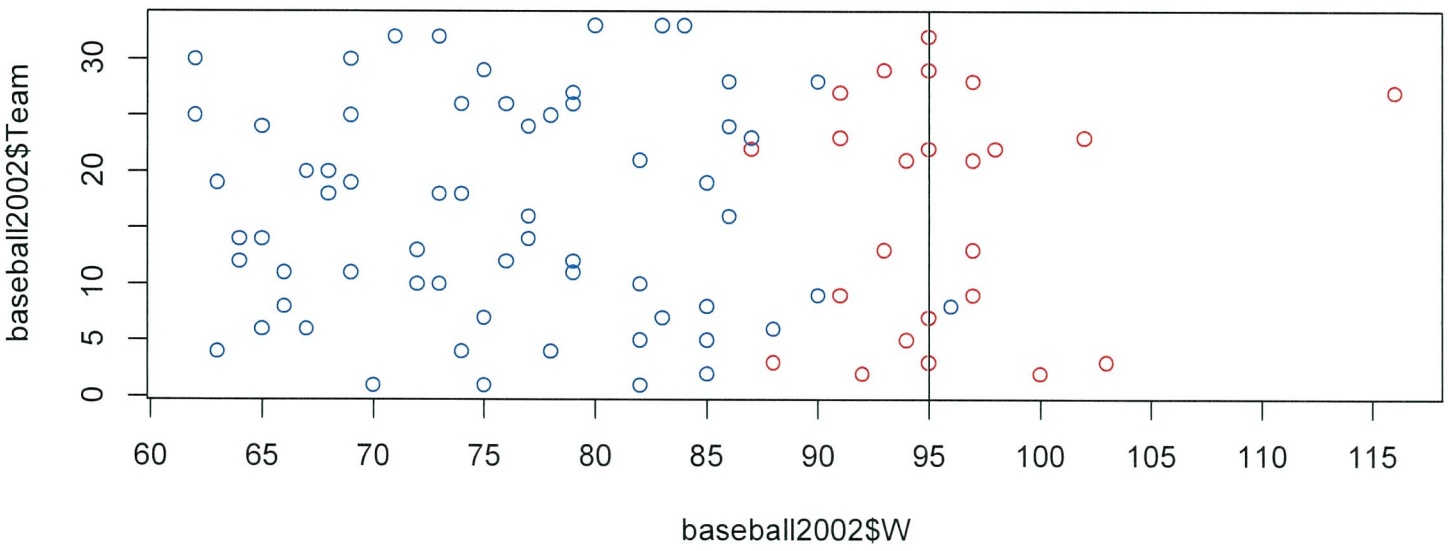
For example Scott Hatteberg has a good OBP but is valued much lesser when the Oakland Athletics sign him.

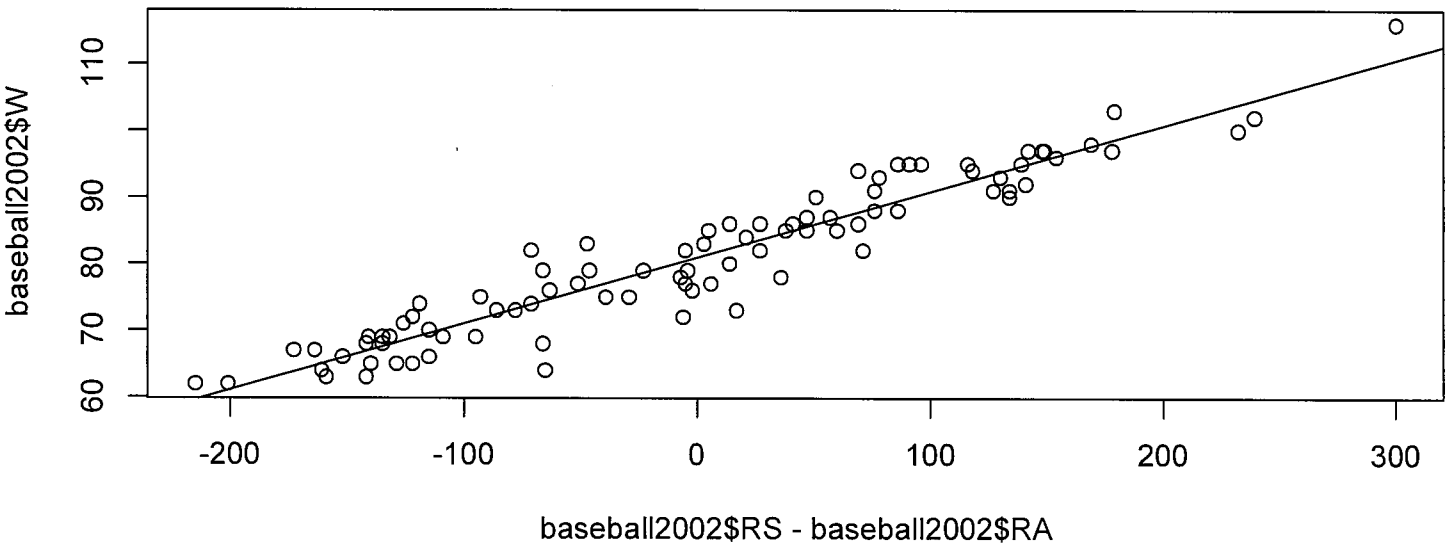
Data — Source: Data on player performance in terms of statistics on hitting (batting), fielding, pitching is increasingly available

Model — Linear regression to determine the importance of various statistics on the runs scored

Decision — Develop tools that help pick players by identifying the key statistics to look out for.

Value — Simple models provide insight to help predict team wins & develop a better team composition. Many sports have now quantitative analysts who aid in building stronger teams in efforts to win.


baseball2002$Team
0
5
10
20
30
60
65
70
75
80
85
90
95
100
105
110
115
baseball2002$W


baseball2002$W
60
70
80
90
110
-200
-100
0
100
200
300
baseball2002$RS - baseball2002$RA