Moneyball Sequel

The Art of Winning an Unfair Game

Roger Qi

roger.qi@live.com

https://github.com/r-qi/Moneyball_Sequel

CONTENTS

Introduction

The Michael Lewis book *Moneyball* is a story about the Oakland Athletics baseball club's 2002 season. Being a small market team, the 2002 Athletics had one of the lowest payrolls but managed to win the most games in all of baseball through the application of empirical analytics known as sabermetric.

The book pointed out the flaws of using batting average (BA or AVG) to determine a player's offensive ability, when there are statistics such as on-base percentage (OBP) and slugging percentage (SLG) which are better indicators of offensive success. This led to the viral internet meme as shown on the right, where the only thing Bill Beane (portrayed by Brad Pitt in the movie) cared about is whether a player can get on base.

Objective

Now it is widely accepted in the baseball community that OBP and SLG are some of the most important indicators of a player's offensive production. What I wanted to do is to analyze whether a pitcher's ability in preventing batters from getting on base would determine his overall effectiveness. I will also examine the combination of preventing batters get on base and slugging a high percentage would be a better indicator. I will be using two machine learning methods to model my training data: Gradient Descent and the Normal Equation. All the coding is done in Octave.

To see how the output is generated, run the moneyball_sequel.m file in Octave and it would demonstrate how all the plots and tables are generated.

1st Training Set

Training Data

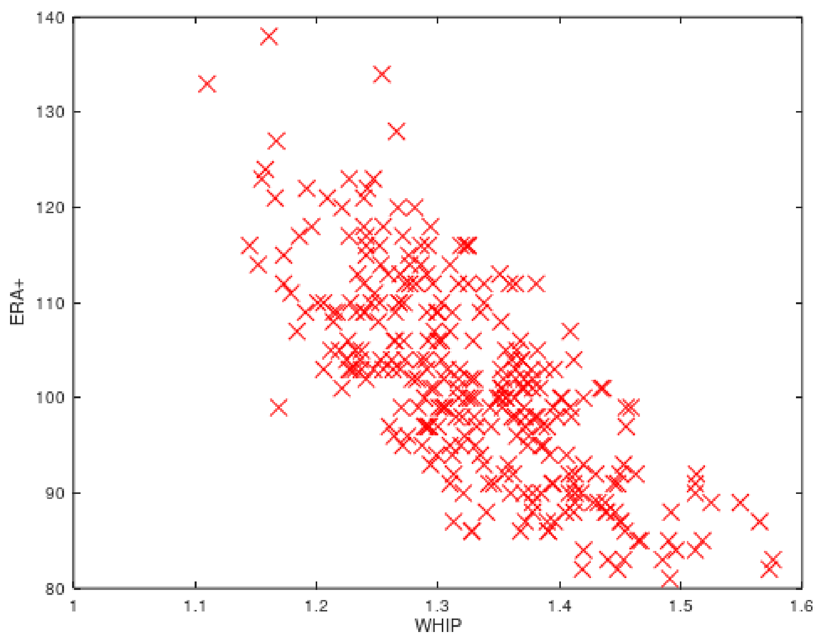
The training input or the explanatory variable will be Walks plus Hits per Innings Pitched (WHIP), and the output is ERA+ (Earned Run Average Plus) from 2008-2017 for all 30 MLB teams.

WHIP – this is a proxy for a pitcher's OBP against. I'm using WHIP over OBP because OBP includes hit-by-pitch (HBP). Although HBP is related to the pitcher's control, I think there is a high degree of randomness to it and should not be used to evaluate a pitcher's performance.

ERA+ – ERA is a measure of how many earned runs a pitcher gives up per 9 innings, the lower a pitcher's ERA, the better the pitcher is at preventing earned runs. ERA+ is a version of ERA where it is adjusted by ballpark factors and is scaled to 100. Some pitchers pitch more in bigger ballparks so they have an advantage, while some pitchers pitch more in cities with dry air like Colorado, where the ball carries further, so they are at a disadvantage. The ballpark factor puts these into consideration and adjusts the differences among ballparks. An ERA+ of 100 is league average, the higher the better.

Here is a plot for the training data on the right. We can clearly see an inverse linear relationship – and it makes sense intuitively as a team with a low WHIP is a team that allows fewer baserunners via walk or hits, which is a team that has an ERA+.

The linear model equation will be:
$y = \theta_0 + \theta_1 * x$, where y is ERA+ and x is WHIP.



Gradient Descent

Now let's put the training data through the gradient descent (gradientDescent.m) function to obtain the theta (coefficient for the independent variable or the input)

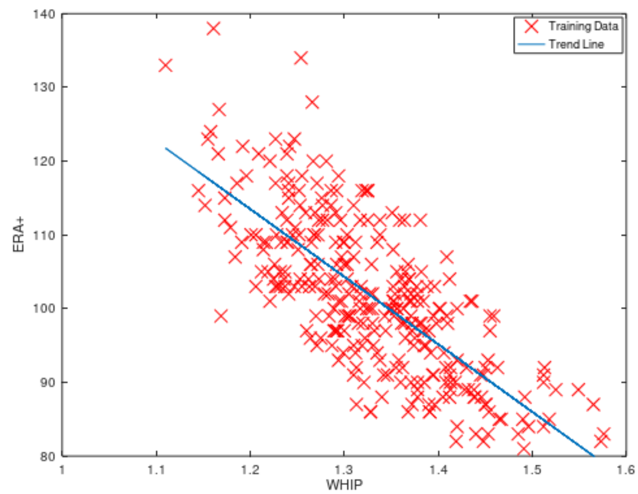
```
Theta found by gradient descent:
223.562592
-91.712376
```

So we got a y-intercept of 223.56 and -91.71 as the coefficient for our independent variable.

Our model is now **$y = 223.56 - 91.71 * x$**

This again makes sense intuitively because as WHIP increases, ERA+ decreases. But I wanted to plot the theta against my training data to see if it is the right fit.

As we can see, the trend line looks to be a good representation of the training data, so the gradient descent algorithm has done its job!

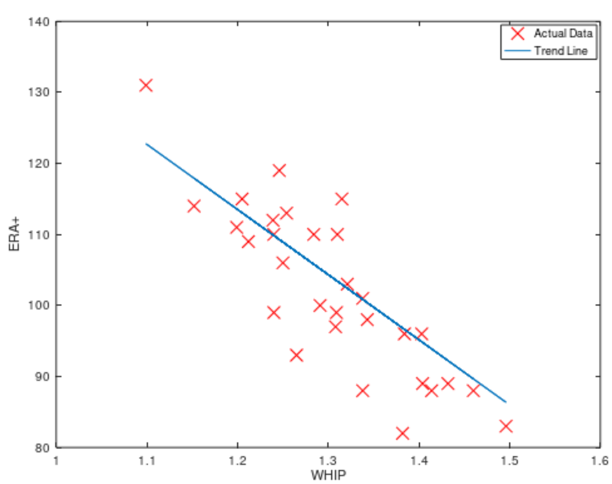


Prediction

Now we want to predict 2018's ERA+ for each team based on their WHIP using the theta obtained from gradient descent.

See the table on the right for the result and how it compares to the actual ERA+ in 2018.

Now I will plot the **y = 223.56-91.71*x** against the actual ERA+ and WHIP data for each team in 2018. We can see in the plot below that the trend line is a pretty accurate model in comparison to the actual ERA+.



Team	Predicted ERA+	Actual ERA+
ARI	108.555	113
ATL	105.804	110
BAL	86.3609	83
BOS	109.289	119
CHC	102.961	115
CHW	92.2305	89
CIN	94.7984	89
CLE	113.049	115
COL	103.419	110
DET	100.393	98
HOU	122.771	131
KCR	89.6625	88
LAA	102.411	103
LAD	117.91	114
MIA	96.8161	82
MIL	109.839	110
MIN	96.6327	96
NYM	107.546	93
NYY	109.931	112
OAK	112.407	109
PHI	105.162	100
PIT	103.511	99
SDP	100.851	88
SEA	109.839	99
SFG	103.603	97
STL	100.851	101
TBR	113.599	111
TEX	94.8901	96
TOR	93.8813	88
WSN	108.922	106

After computing the Coefficient of Determination, or R2, or that 71% off the variance in our output can be explained from the model, pretty good considering we are only using one variable.

```
Coefficient of Determination is 0.712354
```

2nd Training Set

Training Data

For the second training set, I will be introducing slugging percentage (SLG) against as another explanatory or independent variable in addition to WHIP to determine ERA+.

SLG against is simple terms is how well batters can drive the ball and get extra-base hits of a pitcher. The higher the SLG against, the pitcher is more prone to his pitches being hit hard as extra-base hits such as doubles and home runs.

We will be using the 2008-2017 team stats again, and use our model to predict the 2018 ERA+ based on each team's SLG against and WHIP. Instead of using gradient descent, we will be using the normal equation to model the theta.

The linear model equation will be: $\mathbf{y = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2}$, where y is ERA+, x1 is WHIP, and x2 is SLG.

The Normal Equation

The Normal Equation is as follows: $\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1} \cdot (\boldsymbol{X}^T\boldsymbol{y})$

After running the normal equation (normalEqn.m), we obtained the theta on the right. This gives us the model $\mathbf{y = 231.67 - 63.96 * x_1 - 110.55 * x_2}$

```
theta =

  231.672
  -63.963
 -110.554
```

Prediction

We will use the model to predict ERA+ and compare it against the actual ERA+ for teams in 2018, which gives the table on the right.

We also have an updated R2 of 73.62%, an improvement from 71.23% in the first data set. This means almost 74% of the variance in ERA+ can be explained by WHIP and SLG – not a big improvement from the 1st training set but still an enhanced model.

```
Coefficient of Determination is 0.736166
```

Team	Predicted ERA+	Actual ERA+
ARI	108.567	113
ATL	108.638	110
BAL	83.9119	83
BOS	108.858	119
CHC	105.992	115
CHW	93.6438	89
CIN	92.1181	89
CLE	108.827	115
COL	101.779	110
DET	97.236	98
HOU	121.798	131
KCR	88.9784	88
LAA	101.296	103
LAD	116.086	114
MIA	96.3997	82
MIL	109.573	110
MIN	96.3824	96
NYM	106.758	93
NYY	109.305	112
OAK	110.811	109
PHI	104.1	100
PIT	102.948	99
SDP	99.988	88
SEA	106.146	99
SFG	103.344	97
STL	104.3	101
TBR	113.301	111
TEX	90.6344	96
TOR	91.9207	88
WSN	106.28	106

Summary

The table below compares the result of the two predictions. WHIP can explain 71% of the variance in ERA+, which means it is a good indicator of a pitcher's overall effectiveness. Adding SLG into the equation improves the model but not by much, evidenced by the insignificant increase in R2.

We can conclude that WHIP is a strong indicator of pitching just like how OBP is a strong indicator of batting. One way how teams can use this information to find undervalued pitching is to look for pitchers who have a low WHIP or OBP against while other statistics might not look as good due to ballpark factors or bad luck.

Team	1st Data Set	Variance	2nd Data Set	Variance	Actual ERA+
ARI	108.555	4.44473	108.567	4.43302	113
ATL	105.804	4.1961	108.638	1.36196	110
BAL	86.3609	-3.36088	83.9119	-0.911892	83
BOS	109.289	9.71103	108.858	10.1424	119
CHC	102.961	12.0392	105.992	9.00814	115
CHW	92.2305	-3.23047	93.6438	-4.64378	89
CIN	94.7984	-5.79842	92.1181	-3.11815	89
CLE	113.049	1.95082	108.827	6.17321	115
COL	103.419	6.58062	101.779	8.22103	110
DET	100.393	-2.39287	97.236	0.763996	98
HOU	122.771	8.22931	121.798	9.20209	131
KCR	89.6625	-1.66252	88.9784	-0.978414	88
LAA	102.411	0.589458	101.296	1.70352	103
LAD	117.91	-3.90993	116.086	-2.08623	114
MIA	96.8161	-14.8161	96.3997	-14.3997	82
MIL	109.839	0.160755	109.573	0.426981	110
MIN	96.6327	-0.632663	96.3824	-0.382363	96
NYM	107.546	-14.5464	106.758	-13.7578	93
NYY	109.931	2.06904	109.305	2.69468	112
OAK	112.407	-3.40719	110.811	-1.81123	109
PHI	105.162	-5.16191	104.1	-4.09982	100
PIT	103.511	-4.51109	102.948	-3.94847	99
SDP	100.851	-12.8514	99.988	-11.988	88
SEA	109.839	-10.8392	106.146	-7.14586	99
SFG	103.603	-6.6028	103.344	-6.3441	97
STL	100.851	0.148568	104.3	-3.29959	101
TBR	113.599	-2.59945	113.301	-2.30105	111
TEX	94.8901	1.10987	90.6344	5.36564	96
TOR	93.8813	-5.88129	91.9207	-3.92073	88
WSN	108.922	-2.92212	106.28	-0.280099	106
R-Sq	0.712354		0.736166		

Sources

All the statistics are exported from baseball-reference.com