Testing Efficiency in the Major League of Baseball Sports Betting Market.

Jelle Lock 328626, Erasmus University

July 1, 2013

Abstract

This paper describes how for a range of betting tactics the sports betting market of the major league of baseball is quite efficient and that on the long term almost no abnormal returns can be made. It shows that there is no underestimation present of the home field or underdog chances as proven in papers on other sports betting markets. New methods, using minimum margins and weighting methods also do not result in abnormal returns. Also when using a probit model using readily available data no abnormal returns are found. Except when the Money Lines of bookmakers are included in the model. Using this a small profit can be made.

1 Introduction

Efficiency in the market is a source of much research and debate. Due to emotions and bad use of available data prices in stock markets often do not represent the real value of a firm. Although it has been believed that the Efficient Market Hypothesis must be true due to the forces of arbitrage. It has however turned out that the market can be inefficient without giving the opportunity for arbitrage. Because of this more research is required on the subject.

It was already pointed out by Jaffe and Winkler (1976) that the (American) football sports betting markets are analogous to securities markets. But where the range of returns on the stock exchanges are pretty much endless, the possible returns in the sports betting markets are known before hand and limited. Gray and Gray (1997) argue that this creates a relatively simple test of market efficiency. In this paper old test for efficiency will be revisited and we try some new methods which have proven successful on financial markets.

To begin a small introduction on how prices at the bookmakers are reached. As the betting period for a certain game is opened the bookmaker gives every team a return, representing the change the bookmaker gives both teams on winning. A small margin is built in to allow the bookmaker to make money (Just as it is for market makers). These returns for both teams are called Money Lines (MLs). For example as am writing this you can bet on the Braves vs Reds game, where the Braves have a ML of a 100 (meaning you can make a 100 euro by betting a 100 euro) and the Reds have a ML of -108 (meaning you have to bet a 100 euro to make a 107 euro). A more detailed explanation of Money Lines is found in Section 2.

As the game comes closer and more people start betting the Bookmaker updates the Money Lines to make sure he always makes a profit. This means that when the market for a game closes, the prices represent the opinion of the gamblers on the game. If the market is efficient this would mean that on average the MLs are an unbiased representation of a teams chance on winning. But research has shown that, for the football market at least, this is not always the case. It can be expected for the sports-betting market to be inefficient because sports-fans can be very emotional and biased towards a certain team. Also they may overestimate the effect of certain statistics and in reaction to these biases the bookmakers would set incorrect MLs. There has been found proof that in the NFL market, home teams and underdogs are often underestimated (Golec and Tamarkin (1992) & Gray and Gray (1997)). But the bias is not very big and seems to be declining.

In this paper data of the Major League Baseball sports betting market will be analysed. This has some advantages over the football and soccer market for example. Most important all teams in the MLB play about 160 games a season, which gives a much larger data-set, which will hopefully give a better opportunity for analysis. This paper will also focus on Money Lines instead of spreads. With spreads you bet if the actual outcome of a game is above or under a certain outcome, the spread. You can for instance bet for a ML of -108 that the Reds will beat the Braves by more than 1.5 points. Most papers make the assumption that the returns for betting on the score being below or above this spread are the same. But this is actually seldom the case with actual bookmakers.
First we try betting on home teams and underdogs to see if there is an inefficiency in this regard. It is shown that no extraordinary profits can be made this way. Also when using minimal or maximal Money Lines it results in barely any positive returns. Using weights based on a teams scores standard deviation also does not improve on these results. Lastly when using Probit and Logit models to make predictions, no positive results are made. Pointing to an efficient use of available information. However when using money lines as a variable it turns out small profit can be made, pointing to a small inefficiency.

In section 2 it is introduced what data is used. Section 2.1 introduces how the existence of a home team or underdog bias is tested. Section 2.2 describes how the models where constructed and which variables are used. In section 3.1 the results and some first analysis are shown on the home team and underdog bias. In section 3.2 the results of the different probit and logit models are shown. Section 4 gives more analysis and a general conclusion of the results.

2 Methods

The data used in this thesis is provided by covers.com. Here average Money Lines of big USA bookmakers and the results of the games are listed for every team in the competition for a large amount of seasons. In this thesis data from the last 5 seasons will be used. These are the seasons from 2008 untill 2012. The website makes use of American money lines which work as follows: If team A has a ML of X it means you have to bet X to win a 100, where $X > 100$. In this case team A is the favorite and most likely to win. Team B on the other hand has a ML of Y, which means you have to bet a 100 to win Y, where $Y > 100$. Team B is thus the underdog. Due to the margins of the bookmakers it might happen that, when the odds for winning are quite close, both teams have a negative ML.

Using these money lines we can calculate the implied probabilities for both teams to win using a formula suggested by Gandar, Zubar and Lamb (2000). Assuming p is the implied chance the favorite team wins and c the margin of the bookmakers, we get the following formulas:

$$-c = -100 * (1-p) + p * 100\frac{100}{-ML_f} \tag{1}$$

$$-c = -100 * p + (1-p) * ML_u \tag{2}$$

The first formula describes the average return of the favorite and the second one describes the average return of the underdog. Putting these formulas together gives the following:

$$p = \frac{ML_u * ML_f + 100 * ML_f}{-10^4 + ML_u * ML_f + 200 * ML_f} \tag{3}$$

Using this data we will try two different methods to make profits on the MLB sports betting market. First of all we test the hypothesis that home playing teams and the underdogs chances on winning are undervalued and secondly we will try using the data to create a probit and logit model to predict the outcome of the games.

2.1 Homefield and Underdog

First the home field and underdog advantage will be revisited. We distinguish four different positions a team can be in. A team can be a home playing favorite (HF), a home playing underdog (HU), an away playing favorite (AF), and an away playing underdog (AU). To test if the bookmakers are under or over estimating any of these positions we will bet for every game for five seasons a 100 on the different positions and see what the average return is for every strategy.

Furthermore we bet for every state on teams only when their chances on winning are higher or lower than a given value. For the underdog/favorite we will bet on these teams only when the money line is higher/lower than: 125/-125; 166/-166$\frac{2}{3}$; 200/-200. Also we will try this tactic the other way around, by betting on these teams only when the ML for the underdog is lower/higher than the previously given values. The reasoning behind this strategy is that people tend to underestimate very small chances and overestimate very big chances. Using the margins it might be possible to filter this out.

We will also try using a weighted betting method, where we bet more on teams with a low volatility and less on teams with a high volatility. In the stock market it is shown that using this method can result in higher average returns (Chaves, Hsu, Li, Shakernia (2011), among others). Depending on the distribution of the standard deviations different weighing methods will be tried.

2.2 Probit and logit model

Next we'll try building a probit and logit model to beat the predictions made by the bookmakers. We will try both models because the distribution of the results is most likely not distributed normally and one model might come to better results. We will test different variables:

Average number of wins over the whole season Shows the overall performance of a team.

Average number of games won at home Shows how well the team performs at home.

Average number of games won away Shows how well the team performs away.

Consecutive games won Shows the form the team is in.

Games won in the last ten games Other way of showing the form of a team.

Results of the last three games Shows some of the rivalry between teams and other factors specific for these teams.

This information is all readily available to the public and easily accessible. Using these variables a model will be made to try and predict the outcome of the games. If the model gives an higher chance of winning than given by formula (3) a bet will be placed on the favorite, if it is lower, a bet will be made on the underdog. Also results will be shown when using a margin, meaning we only place a bet when the chance of winning giving by the model exceeds the chance given by the bookmakers by a certain margin. We will try a moving and expanding window with a logit and probit model. First all the games will be analyzed as a whole and after that the teams will be analyzed separately. Modelling the teams separately might lead to better results because every team might be influenced differently by certain statistics and it seems unlikely the these effects are equal for the competition as a whole. Because most of the variables are not very informative at the beginning of the season the results of the first 20 games of every team will be omitted.
Lastly to take look a look at how these statistics are incorporated into the bookmakers odds a regression is done including these odds as variable. Statistics which are well incorporated into the odds should turn out insignificant. This should show if certain attributes are over or under valued in the betting market. This model will also be used to make bets following the same method as described above.

The observations of the first four, 7th and 8th variables are between 0 and 1. The observations of the 5th and 6th variable are -n for a n games losing streak and +n for an n games winning streak.

3 Results

3.1 Homefield and Underdog

In table 1 an overview of the stats of the results of the different groups can be found. As mentioned before they have been sorted in: HF, HU, AF and AU. The mean shows the average margin by which these teams win (or lose). It can be seen that the home field advantage does exist and home teams on average score more than their opponents. This margin is even bigger when looking at favorites. A little surprising is that away playing favorites/underdogs outperform home playing favorites/underdogs. A possible explanation is the existence of a negative bias towards away playing teams. Some away playing underdog teams should perhaps actually be marked as favorite and some home favorites actually as underdogs, causing this small difference in scores.

In table 2 the results of betting on these different groups are shown. The table shows what you would make on average by betting a 100 on the teams in this group. For reference it is also shown what percentage of games the teams in this group won. The implied chance shows the chance of the border of this group that follows from formula (3). No significance tests are shown because of the distribution that arises when

Table 1: *Overview of stats of the results of the different games. The mean gives the average margin by which the teams in this group wins or loses.*

	Home	Away	Favorite	Underdog
Mean	0.197	-0.187	0.619	-0.684
Observations	12055	12059	12757	11357
	Home Favorite	Home Underdog	Away Favorite	Away Underdog
Mean	0.591	-0.800	0.677	-0.634
St. dev.	4.262	4.133	4.167	4.248
Jarque Bera	88.43 (0.000)	85.50 (0.000)	86.89 (0.000)	84.61 (0.000)
Observations	8644	3411	4113	7946

looking at the wins. There are a large amounts of observations at -100 and the rest of the observations lies between 0 and a 100 (in the case of betting on favorites) or at a 100 and above (in the case of betting on underdogs). Any significance test would therefore hold no information whatsoever.
It immediately shows that most of these tactics give no winnings on average. There are only two tactics giving a positive results. The first one is betting on the home playing favorite only when the ML is lower than -200. Which gives an average winning of 6.292. But only 70 games during the five analysed seasons meet this requirement. So even if the result is a significant one, it would hardly qualify as good betting strategy, due to the low amount of bets that could be placed. The second positive returns arises when betting on the away playing underdogs with a ML higher than $166\frac{2}{3}$, an 1.634 average return is made here. Although it would qualify as valid strategy, because of the many observations, the return is not very big and probably not worth it on the long run.
We do however see a somewhat better performance by the bets on the home playing teams, when looking only at the groups with more than a 1000 observations. This might suggest a slight underestimation of the home field advantage. But it is hard to say if these are significant due to reasons mentioned above.

Next it is shown in table 3 what happens when using the volatility of the win margins to weigh the bets. Two different weights were used. It can be seen that this does not perform better than the results of table 2. This method is probably not very effective because the volatilities of the win margins are very close. Most likely, with a mean of 4.244 and a standard deviation of 0.184, the effect, if present, is to small to make any difference. Using a moving window does seems to improve on our returns for the different groups except for AU. So the theory of better average returns for teams with lower volatility might be valid, but if it exists it is not a very big effect. Also these results also seem to support the existence of a small underestimation of the home field advantage. Because these, especially the results of the moving window, seem to give the best results.

3.2 Logit and Probit model

We start off by testing the significance of the different variables separately. For every variable we look at the one of the home and away playing team together. For these regressions the first 20 observations of every team of every season were omitted. This is done because most variables do not contain much information at the beginning of the season. This resulted in 9146 unique games, the observation were then regressed using

Table 2: *Overview of betting on the different groups: HF, HU, AF, AU. The results show the average winnings when betting a 100 on every team in this group. It is also shown what happens when a certain min. or max. ML is taken into account.*

	max. ML	min. ML	Implied min./max. Chance	Average return	Percentage won	Observations
HF						
	-100		0.51	-0.758	0.584	8644
	-125		0.55	-0.870	0.611	5846
	$-166\frac{2}{3}$		0.62	-2.500	0.648	2164
	-200		0.66	-0.582	0.6947	845
		-125	0.55	-1.037	0.5228	2697
		$-166\frac{2}{3}$	0.62	-0.407	0.561	6481
		-200	0.66	-0.9711	0.571	7764
HU						
		100	0.49	-0.788	0.444	3411
		125	0.44	-1.523	0.447	1320
		$166\frac{2}{3}$	0.37	-4.672	0.339	180
		200	0.33	-72.30	0.087	23
	125		0.44	-0.229	0.4747	2037
	$166\frac{2}{3}$		0.37	-0.480	0.4525	3233
	200		0.33	-0.240	0.449	3385
AF						
	-100		0.51	-2.998	0.5410	4113
	-125		0.55	-1.239	0.589	1837
	$-166\frac{2}{3}$		0.62	-2.721	0.630	319
	-200		0.66	6.292	0.7267	70
		-125	0.55	-3.658	0.506	2255
		$-166\frac{2}{3}$	0.62	-2.938	0.533	3233
		-200	0.66	-3.076	0.5374	4040
AU						
		100	0.49	-1.609	0.411	7946
		125	0.44	-1.564	0.382	5013
		$166\frac{2}{3}$	0.37	1.634	0.346	1607
		200	0.33	-6.053	0.285	547
	125		0.44	-1.893	0.465	2803
	$166\frac{2}{3}$		0.37	-2.560	0.429	6335
	200		0.33	-1.422	0.4215	7373

a probit model, where we model the chance for the home playing team to win. First we regression different variables in groups. Were we put every separate variable for every team in a group and past games together. The results of these regressions are shown in table 4.

Table 3: *Overview of betting on the different groups: HF, HU, AF, AU; when weighted using the volatility of the results of this team. These results show the profit when betting an average of 100 on every game. The volatility was calculated using an expanding window and a moving window. The moving window has a window of 20 games.*

Expanding	Weigth	HF	HU	AF	AU
Return	$\frac{1}{Volatility}$	-0.875	-1.071	-2.556	-1.7945
Return	$\frac{1}{Volatility^2}$	-0.875	-1.194	-2.382	-1.7448
Observations		8418	3341	4010	7745
Moving					
Return	$\frac{1}{Volatility}$	-0.545	-0.960	-2.282	-2.184
Return	$\frac{1}{Volatility^2}$	-0.202	-1.073	-1.832	-2.570

We can see that the average wins (the first four variables) are all very significant. Which is no surprise because it shows the overall performance of a team during a season. It is harder to explain the difference between significance of the winning streaks. Although it is expected for these variables to be somewhat significant, mostly because well performing teams will have higher average winning streaks, it is strange that this is only the case for the away playing team. A possible explanation might be that an away playing team with a long winning streak is more motivated to keep the streak going than is the case for a home playing team.

The average results of last 10 games seem to be a much better indicator for the the form a team is in. But these variables might also gain much of their significance from the fact that once again well performing teams also perform better on this variable on average. Lastly, looking at past results, only the last game seems to influence the chance of winning.

Table 4: *Result of the different probit models. The dependent variable is the chance of winning for the home team. Different regressions are separated by a line.*

Variable	Beta	Z-value	p-value
Home Team Average Wins	1.601	12.44	0.000
Away Team Average Wins	-1.37	-10.67	0.000
Home Team Average wins at home	0.874	8.822	0.000
Away Team Average wins away	-0.685	-6.691	0.000
Winning streak Home team	0.006	1.089	0.276
Winning streak Away team	-0.017	-2.871	0.004
Average wins of last 10 games for home team	0.476	9.048	0.000
Average wins of last 10 games for away team	-0.250	-4.859	0.000
Result 1 game ago	0.011	3.482	0.001
Result 2 games ago	0.003	0.989	0.323
Result 3 games ago	0.005	1.333	0.182

Using the variables which turned out to be significant in these regressions a probit model is created. The results of which are shown in table 5. It can be seen that most variables are no longer significant and that the average winnings of both teams contain most of the information already. It is interesting to see though

Table 5: *Overview of the results of a regression using a probit model with the significant variables shown in table 4.*

Variable	Beta	Z-value	p-value
Home Team Average Wins	1.789	5.759	0.000
Away Team Average Wins	-1.828	-6.875	0.000
Home Team Average wins at home	-0.246	-1.010	0.312
Away Team Average wins away	-0.589	2.923	0.004
Winning streak Away team	-0.004	0.588	0.557
Average wins of last 10 games for home team	0.049	0.539	0.590
Average wins of last 10 games for away team	-0.127	-1.254	0.210
Result 1 game ago	0.003	0.836	0.403

Table 6: *Results of betting using chance generated by the use of logit and probit models with a moving and expanding window. The chances were compared to the chances found using the ML and formula 3. Furthermore it is shown what happens when betting when given a certain margin.*

Probit	Moving		Expanding	
Margin	Result	Observations	Result	Observation
0.00	-3.545	8746	-2.388	8746
0.05	-2.904	5443	-2.381	5718
0.10	-1.844	3147	-2.591	3198
Logit				
0.00	-3.983	8746	-3.536	8746
0.05	-3.069	5535	-3.003	6040
0.10	-1.608	3057	-1.768	3751

that average wins of the away team at away games is still very significant. This suggest that certain teams are better at away games independent of how they perform on average and that this should be taken in account. It seems that this is not the case for how a team performs at home.

In table 5 it is shown what happens when these variables are used for different models to make predictions. The predictions are compared with the probability found using the MLs and formula 3. First a probit model is tried and after that a logit model. Both trying a moving window and an expanding window. For the moving window a window of 400 observations is used. It is also shown what happens when taking a minimum margin in account when comparing the predictions with the chances of the bookmakers.

It is immediately clear that no positive returns can be made using this method. The results of the different models do not seem to differ very much, except for the expanding window probit model. This is also the only model where the returns do not go up when a margin is used. For the other models the use of margins improve the performance of our betting strategy. Also none of these methods outperform simple tactics as betting on the home favorite.

Next we try using these models making regression per team. The results of this strategy are shown in table 3.2. Because of the smaller amount of observations per team only an expanding window is used. In this case

Table 7: *Results of betting using chances generated by the use of logit and probit models. Regressions were made for individual teams using an expanding window. The used variables are the same as 5, except for winning streak away team which was excluded. The chances were compared to the chances found using the MLs and formula 3. Furthermore it is shown what happens when betting given a certain margin.*

	Logit		Probit	
Margin	Result	Observations	Result	Observation
0.00	-1.177	19385	-1.175	19385
0.05	-2.141	11644	-1.940	11641
0.10	-2.797	5941	-2.946	5917

the results get worse when the margins get bigger. Which is strange because it suggests that when the more certain the model is about a result the worse it performs, which contradicts the results of the models used in 5.

In table 8 the results are shown of probit model including all variables plus the odds of the bookmakers which were generated using formula (3). It can be seen that the average wins of the home team become very insignificant, but surprisingly enough this is not the case for the average wins of the away team. This suggests a overvaluing of the home field advantage by the bookmakers odds. The next variable that stays significant is the average wins of the away team at away games. Which has a positive Beta, this most likely compensates for the high Beta for the average wins of the away teams. Meaning that teams with a high average overall winning-rate but a low average winning-rate at away games should have a lower chance on winning.(This is supported by the fact that the Beta turns negative when we do a regression with only the bookmakers-odds and the average winning-rate at away games.)

Now using these significant variables once again a probit model is created with a moving and expanding window. The results of which can be found in table 3.2. We see that this model gets positive results for both the moving and expanding window. For the expanding window the amount of observations is pretty small and because of this the result might not be significant. For the expanding window however the amount of observations is quite big and this could qualify as good betting strategy. However further research might be required to to test this, because without any good significant tests we can not attach much value to this outcome.

Table 8: *Results of a probit model including the odds generated using the Money Lines and formula (3).*

Variable	Beta	Z-value	p-value
Bookmaker odds	1.666	10.235	0.000
Home Team Average Wins	0.133	0.381	0.703
Away Team Average Wins	-1.825	-10.92	0.000
Home Team Average wins at home	-0.3601	-1.466	0.143
Away Team Average wins away	0.543	2.680	0.074
Winning streak Home team	0.004	-0.591	0.554
Winning streak Away team	-0.007	0.957	0.338
Average wins of last 10 games for home team	0.093	0.893	0.372
Average wins of last 10 games for away team	-0.130	-1.278	0.201
Result 1 game ago	0.005	1.316	0.188

Table 9: *Results of betting using chance generated by the use of probit model with the significant variables of 8. Once again the estimated chances are compared to the odds of the bookmakers and 100 is bet on every game according to this and given margins.*

	Expanding		Moving	
Margin	Result	Observations	Result	Observation
0.00	-1.959	9146	-2.488	9146
0.05	-0.139	1865	-0.868	3111
0.10	2.077	134	2.6965	724

4 Conclusion

When looking at all the different methods used to make bets it shows that there are very few which actually beat the odds of the bookmaker. On of the methods that does make a profit is using a simple strategy as betting on the away playing underdog while keeping in mind a certain minimal or maximal chance of winning. These methods would however, most likely not be very profitable tactics because of a low amount of games or because the made profit is fairly small. It does show that the odds of the bookmaker are not completely correct and that there is a small amount of inefficiency. But this inefficiency is very small and can not be used tot make abnormal profits. Also the use of weights does not seem to improve the results enough but does indeed improve them. The methods used here are fairly simple and more sophisticated methods might give even better results. But it seems unlikely that this method will every become very profitable.
When looking at the probit models used, almost no profits can be made. Using different tactics can improve the results of the models but not enough and not by much. The results of the regressions show that there are some statistics of teams that have a significant effect on the results but there are not many. Also looking at the betting results they seem to be incorporated very well into the odds of the bookmaker. Based on this gamblers do not seem to have a bias towards the effect of a certain statistic and use the readily available information efficient. However making use of the bookmakers odds as a variable shows that not all the data is incorporated correctly into the Money Lines, and using the bookmakers odds in combination with these variables in a probit model does lead to some interesting results and some profit. Perhaps in future research using better models even better results could be found. It also does give some proof for a not completely

efficient MLB betting market.
Putting all these results together it can be said that the MLB betting market is quite but not completely efficient, but that the gamblers in this market make on average a very good prediction of the results of a game. This is not surprising as the Major League of Baseball is a very well monitored and renown league. Every day teams, players and match-ups are discussed on blogs, news-sites, tv-shows and in magazines. Sports are very important in the USA and almost every fan is very knowledgable about it. The bias towards ones own favorite team does not seem to bring an unbalance to the odds, or they even each other out. Also it seems that fans are not over- or underestimating the influence of things as winning streaks and the home-field advantage.

It is often said by more fanatic or professional gamblers that profits can be made in competitions with much less monitoring than is the case for MLB. There are several reasons why this might be the case. First of all because of less and/or less professional talk by analysts and coverage by the media, gamblers might be less informed on the form of a team and its players. This could lead to biases on certain teams which would not be possible in the MLB. Second, because the gamble market of these competitions is much smaller than is the case for the MLB, bookmakers will put less focus on these competitions and this might lead to worse starting money lines and slower reactions to news and such. Therefor further research could point out that in these market there are indeed inefficiencies.
When applying the results of this paper to the financial market it would support the Efficient Market Hypothesis. All though research has shown that the stock exchange can be inefficient, as shown by a paper of Foot and Dabora's (1999) for example, and that a stock can be mispriced for a long period of time; these results do not have to contradict this, because as is not the case for stocks the actual value of the bet gets known at a certain date whereas the true value and payoff of a stock might never be known for certain.

References

[1] John M. Gandar, Richard A. Zuber, Reinhold P. Lamb *The homefield advantage revisited: a search for the bias in other sports betting markets.*:Journal of Economic and Busniess 53 (2001) 439-453.

[2] Joseph Golec and Maurry Tamarkin *The degree of inefficiency in the football betting market.* : Journal of financial economic 30 (1991) 311-323

[3] Philip K. Gray and Stephen F Gray *Testing Market Efficiency: Evidence from the NFL sports betting market.* : The journal of Finance - VOL. LII, NO. 4 (1997) 1725-1735.

[4] William H. Dare & A. Steven Holland *Efficiency in the NFL betting market: modifying and consolidating research methods.* : Applied Economics, 36:1, (2004) 9-15

[5] Linda M. Woodland and Bill M. Woodland *Market Efficiency and the Favorite-Longshot Bias: The Baseball Betting Market*: The journal of Finance - VOL. XLIX, NO. 1 (1994) 269-279

[6] Florian Peters *Lecture notes: Lecture 2 Behavorial Finance*: DSF & UvA November 15 (2010)

[7] N. Barberis and R. Thaler *A survey of behavorial finance, Ch18* (2002)

[8] K.A. Froot and E.M. Dabora *How are stock prices affected by the location of trade?*: Journal of Financial Economics, 53 (1999), 189-216