
# Live Betting Markets Efficiency: the NBA case

**André Cardoso Dias**

Dissertation submitted as partial requirement for the conferral of

**Master in Economics**

Supervisor:

Professor José Joaquim Dias Curto, Associate Professor

ISCTE Business School, Quantitative Methods Department

September 2016

# Live Betting Markets Efficiency: the NBA case

André Cardoso Dias

September 2016

## Abstract

Several studies have been assessing the efficiency of sports betting markets, by comparing the pre-game prices with the actual outcomes of each event. While some have documented particular forms of inefficiency, as the favourite/longshot bias, an important part has been unable to reject the efficiency hypothesis, while identifying the betting volume and the event's notoriety as key factors for a market to be efficient.

In this study, we seek to bridge a gap in the literature, by assessing the efficiency of betting markets as the inherent sports events are taking place. To this extent, we tested the in-play (live) betting markets efficiency of 4 NBA Finals games, by comparing, on a near second by second basis, the winning probability of the home team, implicitly expressed in its betting odds – the price element –, with a theoretical estimation of its winning probabilities – the information element –, generated through a logit regression based on a sample considering all plays and pre-game odds for the NBA seasons between 2007/2008 and 2014/2015.

Our results show that, for the 4 games considered, the in-play betting markets are not efficient, as we reject the hypothesis that the difference between the price and the information element is zero. Although the testing framework and the limited set of considered games prevents us from validating the cause and persistence of our findings, we identify as possible mechanisms inducing these results the asymmetric valuation of game-related events by bettors and market rigidities preventing agents from reacting instantaneously to important events.

**Keywords**: Betting Markets Efficiency; NBA; Live efficiency; Logistic Regression;

**JEL codes:** G14; Z29;

# A eficiência dos mercados de apostas em tempo real: o caso da NBA

André Cardoso Dias

Setembro de 2016

## Resumo

Vários estudos têm avaliado a eficiência de mercados de apostas desportivas, comparando os preços adstritos a cada interveniente antes do jogo começar com os respetivos resultados do evento. Embora alguns documentem formas pontuais de ineficiência, como o enviesamento favorito/não-favorito, uma parte importante da literatura não rejeita a hipótese de mercados eficientes, apontando como principais factores de promoção dessa eficiência o volume de apostas e a notoriedade do evento.

Este estudo procura preencher um vazio nesta literatura, avaliando a eficiência dos mercados de apostas desportivas enquanto o jogo decorre. Nesse sentido, a eficiência dos mercados de apostas em tempo real foi testada, para 4 jogos das finais da NBA, através da comparação, ao quasi-segundo, da probabilidade de vitória da equipa da casa, implícita na sua cota – o elemento preço –, com uma estimativa (teórica) da probabilidade de vitória da mesma equipa – o elemento informação –, gerada através de uma regressão logística alicerçada numa amostra que contem todos as jogadas e cotas pré-jogo das épocas compreendidas entre 2007/2008 e 2014/2015.

Os resultados demonstram que, para os 4 jogos considerados, os mercados de apostas são ineficientes em tempo real, uma vez que é rejeitada a hipótese de que a diferença entre os elementos preço e informação seja zero. Apesar da metodologia de teste e do conjunto de jogos considerados impedir a validação das causas e da persistência destes resultados, identificaram-se como potenciais factores a valorização assimétrica dos eventos pelos apostadores e rigidezes de mercado que impeçam a reação imediata dos apostadores.

**Palavras-Chave**: Eficiência de mercados de apostas; NBA; Eficiência de mercados em tempo real; Regressões Logísticas;

**Códigos JEL:** G14; Z29;

**Acknowledgements**[1]

*To my supervisor, who took on the challenge of researching a completely new topic and has supported me in every critical moment of this long journey.*

*To my co-workers and coordinator at Banco de Portugal, whose personalities I have learned tremendously from and whose personal opinions proved to be particularly insightful.*

*To my girlfriend, whose endless patience, trust and motivation have kept me focused on completing this ambitious project.*

*To my family, whose limitless support and encouragement have led me to become the man I am today.*

[1]The opinions and arguments contained in this report are the sole responsibility of the author and not of ISCTE-IUL or of the author's employer.

# Contents

# List of Figures

# List of Tables

*"Choose a job you love and you will never have to work a day in your life"*

Confucius

# 1 Introduction

Fama (1970)'s work on market efficiency set a cornerstone contribute, in economic and financial literature, by brilliantly setting forth a solid theoretical framework on the pre-conditions and analysis of financial markets' efficiency. Although this author has initially focused primarily on stock prices, many researchers have been applying this theoretical framework to very distinct realities. Indeed, several studies, published in recent times, have been testing for the efficiency of betting markets pertaining to a wide range of sports, with applications spanning from basketball to horseracing. Some of these studies have been documenting particular forms of consistent betting markets inefficiencies. The most notorious case is, undoubtedly, the favourite/longshot bias in horseracing, which pertains to the empirically observed overbetting on the favourite horse, thus pressuring its prices upwards and inversely effecting their returns towards non-efficient levels. Notwithstanding, a very important part of these studies, spanning throughout many different sports, have been unable to reject the efficient markets hypothesis for their betting markets. In this sense, recent studies have investigated the factors which promote betting markets efficiency, in the context of the aforementioned sports, and concluded that the number of participants, the betting volume and the event's notoriety are among the key set of factors identified.

Although these perspectives are somewhat different, the vast majority of their supporting works are conducted following the same *ex-post* perspective, that is, in a nutshell, they compare the pre-game prices for betting on a given event, in a given sport, with the actual outcomes of such events. In this sense, under Fama (1970)'s setting, the price for such event should be adjusted such that it is impossible to extract long-term profits from such market.

In light of the contributes identified, it seems that the sports betting markets efficiency literature is leaving behind the most important part of sports: the actual moments during which the underlying events take place. Indeed, in recent times, betting markets have evolved up to a point where it is also possible to bet during the course of a sports event. Naturally, the progressive dynamics of such event will, inevitably, change the pre-event prices. But is this in-game fluctuation also market efficient? Or are there any consistent biases to be explored? So far, the literature has not followed up to this point, hence leaving these important questions

unanswered.

Following this glaring gap in the literature, this study seeks to perform a market efficiency test to the home team betting prices, for a reduced set of NBA games, as the underlying game events are taking place. This first approach at in-game (live) betting markets efficiency testing strives to set forth a methodological framework and, ideally, groundbreaking insights on the efficiency of betting odds as the events underlying the selected basketball games are occurring. To this extent, in Fama (1970)'s spirit, we have gathered two very detailed elements for the testing framework: a price and an information element.

The price element comprises, among other important information, the betting prices for the home team, on a near second by second basis, for 4 games of the NBA finals of a season at our choice. The information element includes the estimated winning probabilities of the home team, on a near second by second basis, which are generated through a logistic regression of the binary victory/defeat of the home team on a set of key game-related variables (time elapsed, pre-game odds and winning margin) and of event-related variables (*e.g.* home team rebound, away team turnover). The parameters of the logistic regression were calculated, on a play by play basis, considering a sample of 10 255 NBA games, from the seasons spanning between 2007/2008 and 2014/2015, which includes a grand total of 4 108 439 plays.

To accomplish the task we have ventured for, we will test whether the difference between the aforementioned elements is not statistically different from zero. If that hypothesis is not rejected, then we are able to conclude that the in-play betting markets are efficient, for the games considered. Contrarily, if we reject this hypothesis, then we would prove the inefficiency of these markets. In this sense, this study is structured as follows: section 2 provides a theoretical overview on the concept of efficiency and on the findings of betting markets efficiency studies pertaining to basketball and other sports; section 3 describes with detail the construction steps of the price and information elements and of the testing framework itself; section 4 assesses the properties of the testing equation parameters and encompasses the findings of the application of the testing framework, alongside the respective comments on its results; finally, section 5 concludes by reviewing the path undertaken during this study and presents future avenues of study that can be drawn from it.

***"A market in which prices always "fully reflect" available information is called "efficient""***

Fama (1970, p. 383)

# 2 Literature review

The starting point of this essay returns to the Marshallian definition[1] of *what* a market is: the "place" where, in *equilibrium*, supply and demand equal each other, allowing for the definition of *equilibrium* prices and quantities. In betting markets - the main focus of this essay - the core concept underlying market phenomena is not different, whether one is dealing with "bookmaker" markets, pari-mutuel markets or betting exchange markets.[2] Notwithstanding, as explained in the previous section, the goal of this essay is determining the efficiency of live betting markets. By definition, this is only possible when one takes into consideration betting exchanges, as the conceptual framework underlying other types of markets leaves little to no room for odd fluctuation during the event.[3] Moreover, there are many arguments that support the hypothesis that bettors are at least as accurate (if not more) as bookmakers in forecasting the outcome of events. Smith et al. (2009) devoted a paper exactly on this topic: understanding if bookmakers evidence superior skills in forecasting outcomes relatively to betting exchange markets. He identifies sources claiming that bookmakers have superior information processing power, inducing a better forecast of the probabilities associated to each event,[4] while also earning an additional return for exploring bettor biases, as, for example,

[1] As in Marshall (1890).

[2] In a nutshell, a bookmaker market is one in which the bookie - the price setter - sets the prices and returns for each event based on an *ad hoc* evaluation of the probability associated to each outcome and on the bets placed on each outcome, prior to the event's start.

A pari-mutuel market is one in which the prices and returns of each outcome in the event are determined by the amount placed on that outcome and on the total amount bet on the event. In this case, the return for betting on outcome $j$ for all $i$ possible outcomes in Z is given by $(\frac{\sum^{Z} bets_j}{\sum^{Z} bets_i})^{-1}, \forall i \in Z$.

Finally, a betting exchange market is one in which prices and returns are set by bettors, with a major caveat: one can bet *for* (back) and bet *against* (lay) a determined event. In this case, one is allowed to trade bets as if they were a commodity, buying and selling at determined prices, yielding profits/taking losses from the event's outcome and the odds' fluctuation during the event.

[3] Note that we are only interested in what happens *during* the event, as opposed to what happens between the opening of betting pools and the events' start. In that case, one should consider all three types of markets, as they share some similarities between them *pre-live*.

[4] See Levitt (2004).

the overround,[5] which further increases the incentive on bookmakers to enhance their predictive power. Additionally, he stresses that the framework under which betting exchanges are built uppon are themselves contributors towards a better predictive power of market prices.[6] Despite these apparently contradictory perspectives, the results obtained in this study show that, considering an estimation corrected for insider activity via Shin (1993)'s $z$ estimator, exchange odds are marginally superior as forecasting elements of each event's outcome when compared to bookmaker forecasts.

In this sense, working with exchange markets seems to be a viable solution in analyzing live betting markets efficiency. However, while at this moment the betting market type that is most pertinent for this purpose is determined, the meaning of the attribute *efficient* is still shrouded in mist. In the next subsection, we shed light into this topic, building on the major contributes of Fama (1970).

## 2.1 The concept of efficiency

The quotation that precedes this section is a central element in any market efficiency study and is exactly the baseline definition we take forward in this discussion, given that it is widely used in betting markets efficiency studies.[7] In his innovative approach, Fama (1970, p. 383) not only defined market efficiency as the reflection of "*all available information*", but also established a clear distinction between three forms of efficiency, based on the underlying information sets: weak form, semi-strong form and strong form efficiency. In a nutshell, for Fama (1970, p. 383), a market is weak form efficient if its prices reflect all available historical information, semi-strong form efficient if the prices reflect all publicly available information and strong form efficient if they also reflect privately held information (in the event that a group of users has monopolistic access to relevant information for price setting).

Following this rationale, Fama (1970) elaborated three models of efficient markets that

[5] For a broader discussion on the impact of the overround and on the factors causing it, see Bruce and Marginson (2013).

[6] Namely the possibility of backing and laying bets, *i.e.*, the existence of possible trading strategies that can mitigate the risk involved in betting. Moreover, the margin between both sides of the market tends to decrease, converging towards the commission's take on winning bets as the betting volume increases - which ultimately is documented to increase the market's efficiency (Gramm and Owens, 2005).

[7] *e.g.*in Schnytzer and Weinberg (2004), Smith et al. (2009) and Paul et al. (2004) the methodology undertaken to test for market efficiency is inspired by the definition stated in Fama (1970), *i.e.* the explanation of the actual event's outcome through the observable market fluctuations up until the event's start.

reflected his view on efficient markets theory: the expected return (or "*fair game*") model, the submartingale model[8] and the random walk model[9]. For the sake of this essay, we are only interested in the fundamentals attached to the expected return model, since as Fama (1970, p. 386-387) himself stated, the submartingale model is "*a special case of the fair game model*" and the random walk model can be interpreted as an extension of the fair game.

In that sense, the expected return theory is nothing more that the logical formalization of the following principle: if prices do reflect the available information, then, in *equilibrium*, there should be no long-term profitable strategies. Fama (1970) formalizes this statement, in expected return terms, such that a market is in *equilibrium* and is efficient if, and only if, the excess market value of any security in the coming periods is zero, conditional on the information sets that are available in the current period.[10] In effect, this formalization translates exactly the underlying concept: if one is conditioned to make decisions based on today's information sets and the excess market value of a given security is zero, then the impossibility of an expected positive net return in the next period is implied (Fama, 1970, p. 384).

Although these definitions seem quite pristine, formalized and empirically testable,[11] they are not exempt of criticism. In fact, they assume very strong assumptions for individuals and questionable market conditions for efficiency.[12] In this spirit, Beaver (1981) postulated that the aforementioned definition was ill constructed, on the basis that it focused to much

[8]In a nutshell, the submartingale model, as envisioned by Fama (1970), implies that the assumption of non-negativity on returns leads to trading decisions based on the information set $\theta_t$ that ultimately will have greater return than just buying and holding the security. In this sense, for a market to be efficient, prices should not follow a martingale sequence as explained in Fama (1970, p.386).

[9]The random walk model can be easily interpreted as a natural extension of the expected return theory: the assumption that prices *fully reflect all available information* implies that they follow an independent and identically distributed process with zero mean, leaving no space for long-term profitable strategies. Apart from Fama (1970) see also Malkiel (1973) for more implications and testing on the random walk theory.

[10]Formalized, Fama (1970)'s expected return model imposed that, for price of security j at $t+1$ "$P_{j,t+1}$", the information set at time t "$\theta_t$" and excess market value of security j at time t+1 "$X_{j,t+1}$", equation 1 and 2 are met in an efficient market:

$$X_{j,t+1} = P_{j,t+1} - E[P_{j,t+1}|\theta_t] \tag{1}$$

$$E[\tilde{X}_{j,t+1}|\theta_t] = 0 \tag{2}$$

[11]Fama (1970, p. 414) himself provides numerous testing frameworks into these hypothesis, while concluding that there is evidence in support of the weak-form and semi-strong form efficiency, corroborating the perspective of his "*fair game*" model.

[12]Namely that market participants are perfectly rational, generate no information asymmetries and are perfectly capable of processing all information accurately. Simultaneously, the condition that there are no transaction costs in the market, no costs of information and that all individuals agreed on the implication of current prices and distribution of returns is also assumed by Fama (1970, p. 387).

on empirical testing rather than clearly bounding the concept in hand. Furthermore, after classifying it as a vague, nonoperational and tautological definition (Beaver, 1981, p. 27), the same author stresses the importance of distinguishing between system and signal efficiency[13] and the role of the information systems underlying market decisions which, in his opinion, were being overlooked. Consequently, he proposes a new twist into Fama (1970)'s definition: a market is efficient if, and only if, prices act *as if* all agents know the information and the market is system efficient (Beaver, 1981, p. 35), which further reflects the concern to complement the baseline concept.

Notwithstanding, the core definition in study is far from being consensual, even when adjusted for such caveats. As Boettke (2010) denotes, there is extensive market imperfection literature arguing for the shortcomings of individuals in market context. Boettke (2010, p. 368) himself defends that Stiglitz's work on information asymmetry would have the "*upper hand*" today, as the huge volumes of information available severely constrain the individual's ability to process the famous *all available information* hypothesis. Moreover, there are significant contributions in the economic literature sustaining that individuals are themselves, by nature, instigators of market non-efficiency due to their non-perfect rationality: Kahneman and Tversky (1979) formulated a prospect theory spotlighting the asymmetric value that individuals tend to confer to their gains and losses, which ultimately affects their investment decisions, while Shiller (2000) alerted to the bounds of individual rationality and the role of heuristics in causing non-rational movements in the markets and deviations from the "efficient" level.[14]

In light of the aforementioned arguments, it appears consensual that Fama (1970)'s work was instrumental in drawing attention to the market efficiency topic and laid the foundations for further work on the conceptual definition of efficiency and its empirical testing. Despite this, one shall not overlook several suggestions of individual bounded rationality that may jeopardize the pristine efficiency pre-conditions stated by Fama (1970) and ultimately shuffle

[13] In a nutshell, a market is to be considered signal efficient with respect to signal $y'_t$ if, and only if, in *equilibrium*, the security price $P_{j,t}$ is the same as in a theoretically identical economy. Moreover, according to Beaver (1981), a market is said to be system efficient if, and only if, all signals in the economy prove to be signal efficient. For more details on these formulations and their implications, see Beaver (1981, p. 28).

[14] In this line of argument, Shiller (2000) elaborates on the role of herding behavior and heuristics (such as quantitative and moral anchorage) which lead to non-maximizing behaviors that might induce deviations from the true fundamental value of the market.

the determinants and concept of market efficiency.

Therefore, it is now pertinent to analyze concrete non-NBA literature on betting market efficiency, seeking valuable input in known inefficiencies, the application of the efficiency concept, the testing framework endured and the conclusions yielded, in order to bridge the theoretical efficiency literature to empirical studies on sports betting markets.

## 2.2 Non-NBA betting literature

Following last subsection's input, we know at this point that typically efficiency is regarded, *grosso modo*, as the existence of null long-term returns on investment, for any market or security. Furthermore, we also know that this might be threatened by, *inter allia*, herding behavior, moral and quantitative anchorage and asymmetric valuation of market fluctuations (Shiller, 2000; Kahneman and Tversky, 1979) leading to possibly profitable market exploits. Before delving into market efficiency studies applied specifically to the NBA, it is important to briefly discuss how efficiency is tested in other betting markets and its respective results.

Horseracing is, by far, the sport with the most extensive literature on betting efficiency. Johnson et al. (2010) approach this sports' betting market efficiency *à la* Fama (1970), *i.e.*, assessing how *pre-live* market odds are able to explain the races' eventual outcomes, along with the possibility of extracting net positive returns. They test for ordinal efficiency of the first three finishers of each race using an exploded logit approach, concluding that, as the race class[15] improves, the ordinal efficiency implied also improved significantly due to the higher notoriety these races attract, hence decreasing the chance to benefit from insider information and from the discrepancies in the incentives to the participants.

In the same stream of thought, Gramm and Owens (2005), following a similar logical testing, conduct an empirical review on the determinants of horseracing betting market efficiency.[16] Based on Vaughan Williams and Patton (1988), they test for parametric evidence [17] of a widely documented form of inefficiency - the favourite-longshot bias - and for statistical

[15]Note that, in this study, each race is classified qualitatively from A (highest) through H (lowest) according to, among others, the betting volume, the quality of the field, the media attention, the division and the location (Johnson et al., 2010).

[16]The factors influencing efficiency under review were: breakage (as in Busche and Walls (2001)), the track's commission, the track volume, the quality of the field, if the race was the first or last race of the day and if it was run on a weekend or weekday.

[17]Namely the negativeness of parameter $\beta_1$ in equation 3, where $NR_ij$ and $Odds_ij$ are the actual returns

significance in explaining net returns on bets. Using a tobit regression for data censored at -1, they conclude that the factors influencing net returns are the odds (as expectable), the betting pool size, the number of race participants, the race class and the maiden,[18] which, apart from odds, all show a negative relation with net returns.[19] Notwithstanding, the authors denote that the estimated parameters point for the existence of the aforementioned inefficiency in horse racing betting markets, which is itself mitigated as the efficiency of the market increases via the referred factors (Gramm and Owens, 2005, p. 184).

In fact, as pointed out above, this inefficiency is not new in the betting literature.[20] The favourite-longshot bias is, as Busche and Walls (2000) pointed, the empirically observable evidence that individuals tend to overbet on favourites and underbet on longshots, which ultimatly increases the prices (decreasing the implied returns) of the first and inversely effecting the second. This creates the opportunity for the existence of positive net excessive returns on horse racing markets if one adopts the strategy of constantly betting on the underbet horse, which ultimately jeopardizes the markets' efficiency. In that study, Busche and Walls (2000), test the ability of a pari-mutuel betting market in generating positive per-dollar returns and also conclude that this bias is inversely related to the size of the betting pool and that the deviations from market efficiency are mainly due to this phenomena.

At this point, two arguments seem to be important in the framing of the problem in hand: there is a documented inefficiency in horse racing markets which allows the possibility of the generation of positive net returns and that this is mitigated by the factors that promote betting efficiency: betting volume, number of participants, race class, among others. (Busche and Walls, 2000; Gramm and Owens, 2005; Johnson et al., 2010; Ali, 1977; Thaller and Ziemba, 1988; Asch and Quandt, 1990). Finally, regarding aditional transferable contributes

and odds on horse $i$ in race $j$, respectively, which would constitute an evidence of a form of inefficiency in horse racing: the favourite-longshot bias.

$$NR_{ij} = \beta_0 + \beta_1 Odds_{ij} + \beta_2 Odds_{ij}^2 + \beta_k Odds_{ij} * RaceFactors_j + \epsilon \tag{3}$$

[18] A maiden race is a sub-race for horses that have not won the main race. It can be interpreted in line with Johnson et al. (2010)'s study of ordinal efficiency: the higher the ordinal efficiency, the higher the likelihood of encountering an efficient betting market.

[19] That is, for a higher number of race participants, higher race class or bigger betting volume, the net return on the respective betting market decreases towards zero, hence increasing the efficiency of the market.

[20] In all truthfulness, the standard favourite-longshot bias has been identified in many studies as in, *e.g.* Asch and Quandt (1990) or Ali (1977). Moreover, Thaller and Ziemba (1988, p. 163) even described it as "*the most robust anomalous empirical regularity*".

of horse racing betting literature, the contributions by Gramm and McKinney (2009) are also noteworthy. Using Fama (1970)'s notion of efficiency, the authors compare the closing odds to last-minute odds,[21] to assess the ability of last-minute money in promoting market efficiency. Using a sample of 1 644 races, in which 40 % of the money is classified as "*last-minute*" (late money), they conclude that late money pressures odds movements towards more efficient levels, [22] while also reducing the standard favourite-longshot bias. The argument behind this conclusions is that late-money is more informed money, most often triggered by informed bettors and insiders who compete to take advantage of non-adjusted odds for a profit, which, under that process, pressures odds to their "*true*" efficient level (Gramm and McKinney, 2009, p. 370-371).

Finally, before shifting the scope of analysis towards NBA betting literature, it is also pertinent to consider Dare and Holland (2004)'s work on NFL betting markets efficiency. Taking, yet again, Fama (1970)'s notion on efficiency as the ability to yield positive returns, they employ a probit regression to the explanation of a binary victory/defeat of the team in which a bet is placed.[23] The parametrical results show that there is evidence of a bias against visiting favourites, favouring home underdogs, which ultimately taints the efficient markets hypothesis for the handicaps market during the NFL seasons considered (1976-1994) (Dare and Holland, 2004). This conclusion resembles those that were found for horseracing but, simultaneously, Dare and Holland (2004, p. 14) defend that it is a very risky strategy to endure (always betting the home underdog), since its profits may be too small to be exploited.

## 2.3 NBA betting literature

The previous subsection highlighted that, in other sports, the betting volume and the notoriety of the event, usually proxied by the event class, are typically two of the most relevant determinants of betting market efficiency encountered. In this subsection, we review

[21] Note that the closing odds refer to those that are settled when the markets close just before the event's start, while last minute odds are those placed inside the last 60 seconds.

[22] That is, towards the actual probability of success of each horse, progressively eliminating the possibility of consistent profits.

[23] As in equation 4,with W as a win/lose dummy variable, $\alpha^{HF}$ as the parameter associated with playing as the home favourite, $\alpha^{VF}$ likewise as the visiting favourite and CL as the closing line of each game:

$$W = \alpha^{HF} HF + \alpha^{VF} VF + (\beta - 1)CL + \epsilon \tag{4}$$

the major modelization techniques for framing the game of basketball and the conclusions of the efficiency studies surveyed.

One of the possible forms of modeling basketball is by studying which regression is the one which fits better possession based data, estimating and simulating the outcome of the games through statistical techniques. Following this logic, Parker (2010) seeks to provide an alternative to conventional college basketball teams' ranking, approaching the game of basketball as a sequence of possessions that each team has "available" during the course of the game, by studying which model is the one that fits better the distribution of points scored per possession of a given team, for a determined sample of games.[24] His conclusions show that, despite of the fact that the multinomial logistic regression was the one which fitted better the data, the estimates of the winning probabilities for the NCAA[25] men's basketball tournament games derived render that none of the considered models emerges as statistically "better" in predicting the eventual outcomes of NCAA games, for the estimated parameters and sample considered.

Another possible way of modeling and deriving the winning probabilities on a play by play basis is following Kenter (2015)'s combinatorial game approach. In his study, basketball is modeled as a sequential game between the defending and the attacking teams, in which the former decides to defend or foul and the latter chooses between shooting or passing the ball. Solving the game through the probabilistic minimax theorem, and conditioning the time remaining on the game to 60 seconds,[26] the winning probabilities are derived as a result of the combinatorial game, capturing the short and long-run behavior of win probabilities in a basketball game (Kenter, 2015).

Complimentary to the exposed approaches, there are also two other methods of framing the game of basketball which can prove to be useful in determining the efficiency of in-play betting markets. Firstly, Štrumbelj and Vračar (2012) derive each teams' winning probabilities using a Markov possession-led model, estimating the transition matrix from NBA play by play data and from box score statistics (Štrumbelj and Vračar, 2012, p. 533-

[24]The tested models were the simple linear regression model, the Poisson regression, the negative binomial regression, the zero altered Poisson regression and the multinomial logistic regression (Parker, 2010).

[25]The "*National Collegiate Athletic Association*".

[26]Otherwise, since each observation is locally exponential (more than 100 potential outcomes), the game would be computationally very demanding to solve (Kenter, 2015, p. 3).

535). When the performance of this model is compared against other common forecasting methods, no evidence was found on the inferior statistical quality of the produced results in explaining the eventual outcomes of the games. Finally, the last methodology worth noting is the one Michael Beuoy has been developing in his thematic project[27] as in Beuoy (2015). In a nutshell, this big data driven approach currently models each team's winning probability, on a per-minute, per-possession basis, by running a locally weighted logistic regression of a binary win/lose variable, explained by the game time, possession, point differential and the Vegas Line point spreads,[28] based on play by play data of NBA games since the 1996-1997 season.

In essence, all of the previous models seek to predict the same thing: the theoretical winning probability of an NBA team during the course of the game and/or before the game, despite doing it through very different methods. However, none of these studies seeks to compare these theoretical results with the empirically observable odds fluctuation during the game, which would constitute the efficiency test *à la* Fama (1970) we have described in the previous two subsections.

By contrast, there are several studies which address the power of bookmaker odds in predicting the actual outcome of the events underlying, with a major caveat: the actual odds fluctuations during the course of the game, caused by the events taking place, are left off the analysis, relinquishing it to a pre-event basis as in the studies surveyd in the previous subsection. In this sense, Baryla Jr et al. (2007) seek to understand how odds prices are formed, prior to the event's start, for the market considering the total points scored by both teams in a game (*i.e.*, the totals market), while also testing for the efficiency of the resulting prices throughout the course of the season. Using a definition of efficiency *à la* Fama (1970), the authors prove that, for the NBA seasons spanning from 1985-1986 through 2004-2005, the fluctuations of opening and closing lines for the totals market, obtained from Stardust Casino, show that there is a systematic inefficiency during the first four games of the season, given that the totals lines are upwardly biased, enabling the existence of strategies with net positive returns (Baryla Jr et al., 2007, p. 160-163). According to the same authors, the

[27] See *http://www.inpredictable.com/*.

[28] The Vegas Line point spread controls for the favourite/underdog status of a team, as defined by the Las Vegas bookmakers and respecting the 11-for-10 rule inherent to handicap markets.

reason for this phenomena is intrinsically related to information processing deficiencies, since bookmakers are not able to *fully* reflect in prices the changes occurred during the off-season, [29] hence being unable to immediately reflect them in the first games of the season. Despite this, the authors also conclude that this problem is corrected swiftly by the bookmakers (market learning effect), completely fading after the 10th game of each team, evidencing after that threshold an efficient behavior (Baryla Jr et al., 2007, p.163).

Using the same type of market and logic, Paul et al. (2004) test for the efficiency of NBA totals markets from the seasons spanning between 1995-1996 to 2002-2003, following the regression based methodologies set forth in Gandar et al. (1988) and in Sauer et al. (1988). Their results show that, for the overall sample considered, the hypothesis of efficiency cannot be rejected for the totals market, hence rendering no possible profitable strategies, despite that the fair bet hypothesis was rejected for the 202, 204, 206, 207 and 208 points total. This rejection was the result of a phenomena similar to the favourite-longshot bias we have identified above for horseracing: the overbetting on favourite led overs on selected totals, which ultimately induce deviations from the efficient level of unders[30] of the respective totals market (Paul et al., 2004, p. 626-30).

Finally, the contributions by Schnytzer and Weinberg (2004) are also noteworthy regarding possible efficiency tests to be applied to NBA betting markets. In effect, in their study, after highlighting the bennefits of applying efficiency theory to the game of basketball[31] and adopting strictly the efficiency definitions and forms laid by Fama (1970), Schnytzer and Weinberg (2004) test for weak and semi-strong efficiency of NBA handicap markets for the seasons spanning from 1999-2000 to 2003-2004. By comparing the point spread of the actual games with those predicted by the Las Vegas bookmakers,[32] the parametrical results show that there is clear evidence of weak-form efficiency, given that the Vegas lines are able to reflect the actual event's outcomes, translating the infamous *all available information* hy-

[29] As an example, the author relates this deficiency to the difficulty that economic agents tend to evidence when evaluating the true price of a company set for an initial public offer (IPO). Since much of the information is new to the market, it may not be immediately processed accordingly, which can ultimately bias its respective prices (Baryla Jr et al., 2007, p. 156).

[30] In a nutshell, in the totals markets, agents choose whether to bet that the total will be exceeded (the *over*) or that it will not be exceeded (the *under*), with prices following an 11-for-10 rule.

[31] Namely that, given that it is a very repetitive game, it is good for statistical inference and that it is not influenced by weather or severe injuries (Schnytzer and Weinberg, 2004, p. 2).

[32] Via the data obtained from *covers.com*.

pothesis set forth in Fama (1970) (Schnytzer and Weinberg, 2004, p. 6-10). In spite of this result, when a matrix of descriptive statistics was added to test for semi-strong efficiency, the derived parameters rejected the hypothesis that the market was semi-strong efficient, while at the same time no trading strategies proved to be profitable since the residual between the vegas line and the "true" point spread was too small to be profitable to explore (Schnytzer and Weinberg, 2004, p. 9-10).

In essence, from the literature surveyed in this section, one major idea stands out from the rest in the context of this empirical study of live NBA market efficiency: it is a common trend to approach the efficiency question *à la* Fama (1970), *i.e.* seeking to understand if, and how, bookmaker odds are able to reflect the actual event's outcome. Notwithstanding, the true value of this study lies in performing this evaluation, following the same conceptual framework, *during* the course of the game, something that none of the surveyed studies ventured for and constitutes an interesting research hypothesis to be explored. However, the literature surveyed hinted several strict requirements that must be met in order to avoid biasing the study at birth. In the next section, we discuss the methodologies to adopt for the empirical study and identify the relevant data considered.

# 3 Data and methodology

As digressed in the previous section, the application of the concept of market efficiency in betting markets efficiency studies is roughly consistent for all of the sports analysed. Following this consistent theoretical framework for the efficiency problem, we will assume, in this study, an efficiency approach *à la* Fama (1970), *i.e.*, we will try to prove whether the fluctuation of in-game odds for NBA games reflects, or not, the infamous *all available information* hypothesis, which would ultimately render the markets to be efficient (or not).

To attain this goal, we need two crucial elements, in order to construct the market efficiency testing framework:

- The in-game fluctuations of the odds settled through a betting exchange market which, in practice, represent the prices at which bettors are willing to pay to bet *for* or *against* a given team; and
- The *actual* probabilities inherent to the events underlying the bets;

This approach would enable one to conduct a rigorous efficiency test *à la* Fama (1970), as one is effectively incorporating in the testing framework both the prices element under scrutiny – *i.e.* the betting exchange odds – and the *all available information* element, which should be reflected in the *actual* winning probabilities inherent to the events under analysis.

Although the latter element of the testing framework is easily interpreted and incorporated in the model, that is not the case for the former, since, from a statistical perspective, the *true* probability of a team winning a basketball game is not directly observable. For that reason, we incorporate in the testing framework a model which allows us to estimate the winning probability of the home team, on a play by play basis, which enables the derivation of a near second-by-second estimate of this phenomena and permits the matching of the theoretical winning probabilities to their current market prices, hence completing the efficiency testing framework.

In that sense, in this section we describe the data concerning the in-game odds fluctuations, the data used to fuel the play by play estimate of the home team's winning probability, the model's construction steps and the efficiency testing methodology we have endured.

## 3.1 The price element – In-game fluctuation of betting odds

When one considers constructing a market efficiency test *à la* Fama (1970), the choice of the prices to be tested is quintessential in fostering the quality of the study, since if one chooses particular realities where the preconditions for market efficiency, as laid down by Fama (1970), are not verified, then one might be biasing the study at birth. Taking this matter into careful consideration, we thoroughly analysed all possible hypothesis for the prices element.

In that sense, one could consider using the in-game fluctuation of the home team winning odd for all games of a given NBA season, including both the regular season and playoffs. Although this approach could yield a very high number of games and observations to be scrutinized, which could improve the quality of the test, it presents three problems worth noting. Firstly, it mixes different types of games, in the sense that one is not controlling for regular season versus playoff games, the calendar or for heavily favoured teams versus unfavoured teams, which could induce eventual biases into the testing framework – in fact, as explained in the previous section, Baryla Jr et al. (2007) demonstrated the existence of an early season bias in NBA betting markets, which would be included if one encompasses all the games in one season. Secondly, by including all games in a season, one would eventually be considering games with very different liquidity levels, which could scramble the efficiency testing framework, given that, as argued before, the betting pool's volume is widely regarded as an instigator of market efficiency of betting odds. Finally, this approach would also create a very cumbersome – if not impossible – task in matching the market data to the theoretical model we are creating, as there is no possible direct link between one and the other.[33]

Naturally, the judgement on this approach is based on the premise that the data we are looking for – the in-game fluctuations of betting odds for NBA games for a given selection of games/seasons – is actually available and possible to be worked upon and published. To that effect, we have contacted 6 betting houses and betting exchanges operating in Portugal in the summer of 2015, from which we have been granted access to betting data for a limited set of games from one of these entities. Due to data protection policies or other unnamed issues,

[33]This is due to the fact that the NBA's application program interface (API) only allows to extract play by play information with time stamps including the hour and the minute during which the play as occurred. In section 3.3, we address this question with greater detail and propose a way to overcome it.

all other entities have denied access to the requested information, hence we will maintain throughout this study both the providing and the denying entities under anonymity.

Taking into account both the limitations of the inclusion of all games identified previously and the access to data constraints, we have decided to select the first 4 games of the NBA Finals of a season at our choice as the games in which we will perform the market efficiency test to the home team winning odds. As decided for the entities mentioned in the previous paragraph, we also opted not to disclose the particular season we have spotlighted. In addition, the number of chosen games was also a consequence of this protection decision.[34] However, note that the choice of the NBA Finals games also seems to be the most logical one in avoiding the problems encountered in the previous approach. In fact, by choosing these 4 games, one is concentrating the spotlight on games with the highest "class", visibility and dimension of NBA games, since the NBA Finals are the culmination of the NBA season, when the Western and Eastern Conferences champions play each other for the league title, and, for that reason, it is pertinent to hypothesize that these are the games with the most attention of the season. On that note, if one proxies the attention factor with the number of TV viewers, it is clearly noticeable the heterogeneity of viewership within the regular season and between itself and the NBA Finals. For example, when one considers the 2015-2016 season, the average number of regular season TV viewers on ABC, ESPN, TNT and NBA TV were 3.9, 1.7, 1.7 and 0.3 millions, respectively (Karp, 2016), whereas game 7 of the 2016 Finals had over 30 million viewers (Pallota, 2016). Therefore, it seems plausible to assume, in line with the literature surveyed, that the betting volume would accompany the increase in the event's notoriety (Johnson et al., 2010) which is crucial in promoting betting markets efficiency (Johnson et al., 2010; Gramm and Owens, 2005). Moreover, this assumption is also fostered by the repetitiveness inherent to the game of basketball (Schnytzer and Weinberg, 2004) and the NBA Finals, thus solidifying the validity and pertinence of the price element choice and pre-empting a bias at birth of the testing framework.

* * *

With the price element of our testing framework chosen and duly justified, it is now

[34]This is due to the fact that all NBA Finals games are played through a best of 7 series, which guarantees that for each season there are, at least, 4 Finals games. For further clarification on the dataset involved, please contact the author.

pertinent to describe the raw data we have received from our provider. Note that we will only present the variables and observations from the raw dataset which are relevant for our study, since the database provided comprehended many auxiliary markets (*e.g.* spreads and totals) and identification variables (*e.g.* event identification) not relevant for our purpose. Moreover, for simplicity, we have opted to include only the odds corresponding to the home team, while the "inplay" flag is signalled positively. These decisions have narrowed the total number of observations from the original 150 022 to 27 639.

Table 1: Summary of the *market* dataset

| **Variable** | **Description** | **Number of Observations** | **Min** | **Max** |
|---|---|---|---|---|
| Name of market | Indicates the date, the season, the teams playing and the respective betting market of the game under scrutiny | 27 639 | - | - |
| Time stamp | Indicates the date, hour, minute, second and tenth of a second at which the record took place. | 27 639 | - | - |
| Inplay flag | Binary variable indicating if the game is in play or not. | 27 639 | - | - |
| Market status | Binary variable indicating if the market is active or suspended. | 27 639 | - | - |
| selection | Indicates if the observation shows betting information for the home team or for the visiting team | 27 639 | - | - |
| total matched[35] | Sum of all the back and lay stakes on the market in analysis | 27 639 | 139.096 | 1.705.818 |
| last price matched | Indicates the latest price matched on the market | 27 639 | 1,01 | 1 000 |
| back price1 | Indicates the first available price to back on the selection chosen | 27 639 | 1,01 | 1 000 |
| lay price1 | Indicates the first available price to lay the selection chosen | 27 639 | 1,01 | 1 000 |

In table 1, we describe the aforementioned variables considered from the raw dataset, henceforth referred to as *market* dataset, and provide their minimum and maximum values, when pertinent. Note that we did not compute additional descriptive statistics – *e.g.* mean, variance, standard deviation – as they would have no interpretation in these cases, since in variables *"last price matched"*, *"back price1"* and *"lay price 1"* the dataset simply accompanies their fluctuation during the timespan of each respective game and the *"total matched"* variable is an accumulation variable, which renders additional descriptive statistics meaningless. Moreover, it is also noteworthy that the frequency of the *"time stamp"* variable is

[35] As for the data provider and the NBA Finals under scrutiny, we decided not to disclose the currency of this variable. Please consider the values presented in monetary units and enquire the author for any additional detail.

irregular, which bears additional consideration since the granularity conferred through this variable is crucial in allowing the near second by second analysis we have ventured for. To test the granularity of the dataset, we have computed the intra-game average of the difference, in seconds, between each observation and its previous one. The result – 1,342811727 seconds – clearly shows the high granularity of our *market* dataset and warrants the near second by second logic we seek, which is instrumental in promoting the quality of this study.

Now that the variables considered in the price element of our testing framework are clearly defined, we conclude this subsection by presenting below the in-game fluctuations of the variables *"last price matched"* and *"total matched"*. Note that we show the latter in its modified form, thus representing the implied probability in the respective price and not the raw decimal form.[36] This is done by simply calculating its inverse,[37] and shown in figures 1, 2, 3 and 4.

$$Implied\ probability\ home\ team\ win_t = \frac{1}{last\ price\ matched_t} \tag{5}$$

[36]The purpose of this transformation is simply fostering a clearer visualisation of the fluctuations of the data, since when the home team lost, the "*last price matched*" variable quickly converged to 1 000 and rendered the graphic meaningless.

[37]In line with Cortis (2015).

Figure 1: Game 1 – *"implied winning probability of the home team"* and *"total matched"*


Figure 2: Game 2 – *"implied winning probability of the home team"* and *"total matched"*

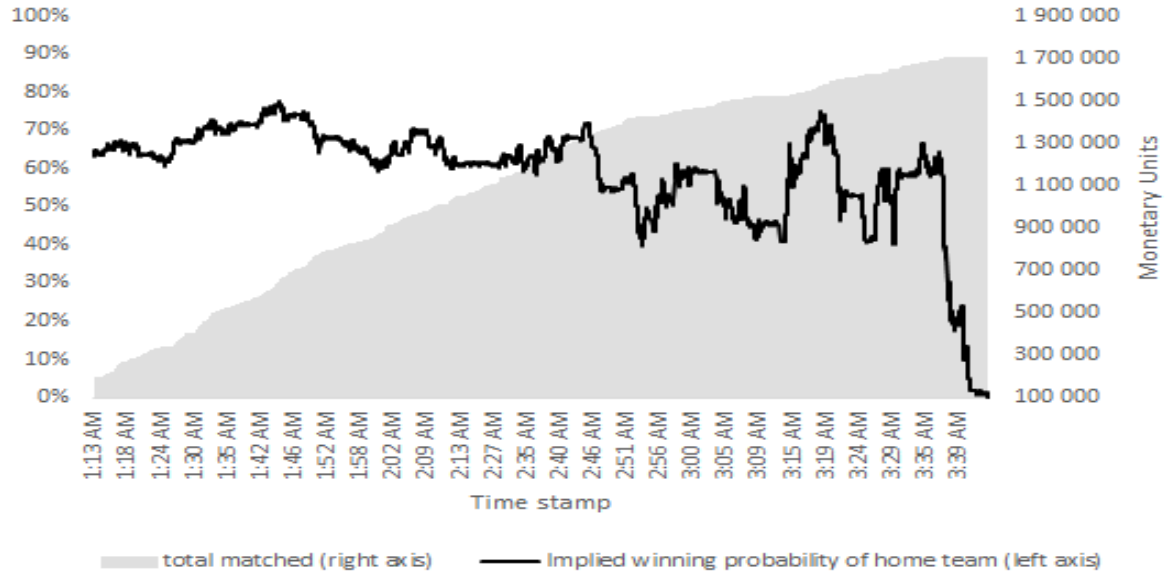


Figure 3: Game 3 – *"implied winning probability of the home team"* and *"total matched"*

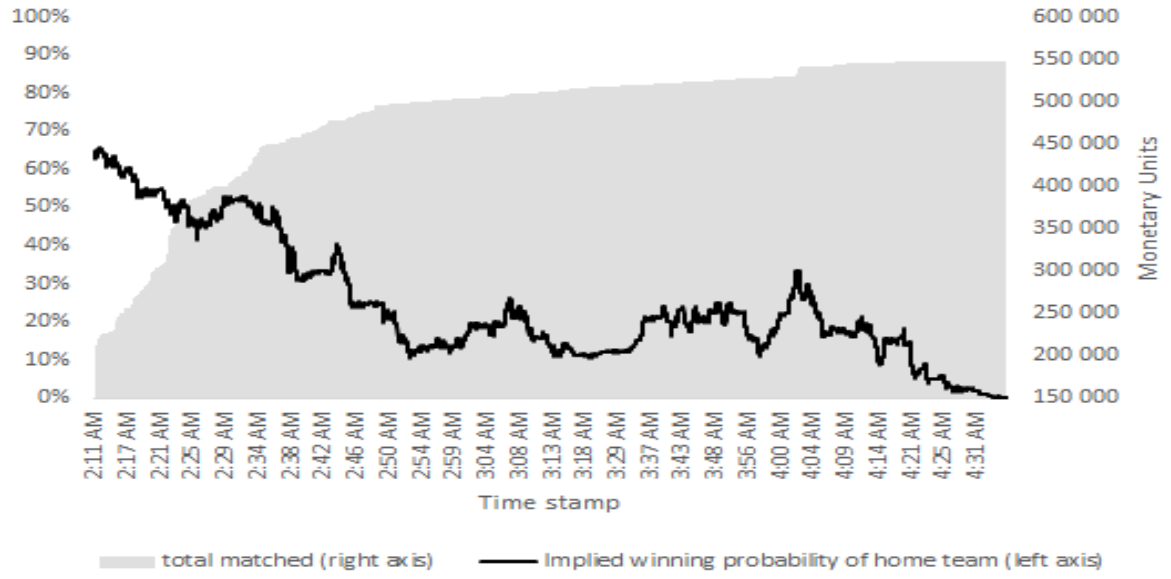
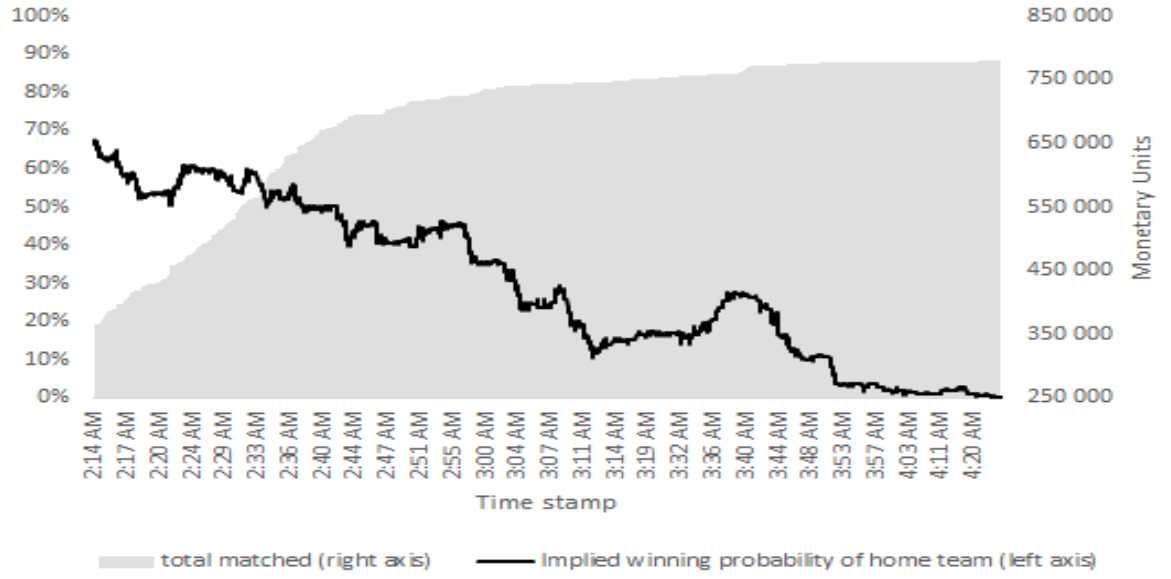



Figure 4: Game 4 – *"implied winning probability of the home team"* and *"total matched"*



## 3.2 The information element – Play by play estimation of the home team's winning probability

Contrary to the price element, for which we chose to narrow the set of the games under analysis to foster the pertinence of the testing framework, we opted for a comprehensive approach to the information element. Taking Fama (1970)'s *all available information* concept as a key element in our testing framework, we strived to include *all* of the games that were possible to be extracted through a time-efficient process on a play-by-play basis. To that extent, we have investigated several methods[38] through which this extraction could be operated, having opted to extract all possible games, on a play by play basis, since the 1996-1997 season. This was done by compiling, through a web scrapping programme written in Python, a list of each game's identification code and then extracting their respective plays by querying the NBA's application program interface (API). This Python programme, which also converted the .json input to a user-friendly .csv format, allowed us to achieve a *quasi*-population coverage[39] of NBA games – regular season and playoffs – occurred between the 1996-1997 and the 2014-2015 seasons. In annex 1, we share the extraction programme used in this study.

This *Big Data*[40] driven approach has allowed us to extract detailed play by play information from 23 857 games – regular season and the playoffs – for the aforementioned seasons. These games are in turn scattered through 10 732 035 observations[41] and comprise a very rich, detailed and granular information dataset which promotes the quality of our estimation of the theoretical winning probability of the home team on a play by play basis. In the next subsection, we describe this dataset and refer to it as the *baseline* dataset.

[38] The first method we tried was manually extracting plays from the game logs held at the NBA's website. Naturally, for the number of games we wanted to cover, this approach was utterly ineffective. In that sense, we investigated thoroughly the applicability of machine driven extraction techniques, which eventually led us to the solution applied. Nevertheless, during this process, we created and corrected several web scrapping programmes that iteratively showed different problems as, for example, the inability to extract games prior to 2010 and the vulnerability against any changes that occurred in the game codes (*e.g.* when the games were rescheduled).

[39] Punctual extraction errors, due to incoherences in the raw data, have prevented us from covering the population of games occurred between the 1996-1997 season and the 2014-2015 season.

[40] Given that the data we used stems from the statistical system in place at the NBA and is framed within the V's of the Big Data concept – Velocity, Variety, Volume, Value – (Vorhies, 2014), it falls in line with one of UNECE's types of Big Data (Vale, 2013).

[41] Note that we are only considering the observations which are valid for our model, since the initial extraction also yielded additional non-relevant observations (*e.g.* substitutions) and some punctual extraction errors (*e.g.* observations without content).

### 3.2.1 Baseline dataset

As mentioned above, the extraction process we have designed has allowed the construction of a detailed dataset comprising nearly all NBA games dating back to the 1996-1997 season. In table 2, we describe the variables we have used from the raw input and compute their average, minimum and maximum when pertinent.

Table 2: Summary of the *baseline* dataset

| Variable | Description | Number of Observations | Min | Max | Average |
|---|---|---|---|---|---|
| GAME_ID | Indicates the code of the game under analysis. This identification is done through an 8-digit number, in which the first number reflects if the game is a regular season or playoff game,[42] the second and third numbers reflect the last two numbers of the year during which the respective season has started[43] and the last 5 numbers reflect the ordering of the games for the combination of the previous two inputs in the identification code.[44] | 10 732 035 | 29600001 | 41400406 | - |
| EVENTNUM | Indicates the ordering of the events registered for each game | 10 732 035 | 0 | 827 | 450[45] |
| EVENTMSGTYPE | Indicates the type of play registered in the observation.<br>This code comprises the following possibilities:<br>1 – Field Goal Made<br>2 – Field Goal Missed<br>3 – Free Throw Made/Missed<br>4 – Offensive/Defensive Rebound<br>5 – Turnover/Steal<br>6 – Foul<br>7 – Violations (*e.g.* kicked ball)<br>8 – Substitutions<br>9 – Full/Short Time-Out<br>10 – Jump-Ball<br>12 – Start of Period<br>13 – End of Period | 10 732 035 | 0 | 13 | - |
| PERIOD | Indicates the game period during which the play registered in the observation took place | 10 732 035 | 1 | 8[46] | - |
| WCTIMESTRING | Indicates the hour and minute, in Eastern Standard Time (EST), at which the play took place. | 10 732 035 | - | - | - |
| PCTIMESTRING | Indicates the time remaining on the game period when the play took place | 10 732 035 | 00:00 | 12:00 | - |
| HOMEDESCRIPTION | Indicates if the acting team was the home team and briefly describes the registered play and indicates the player/s involved. | 10 732 035 | - | - | - |
| VISITORDESCRIPTION | Indicates if the acting team was the visiting team and briefly describes the registered play and indicates the player/s involved. | 10 732 035 | - | - | - |
| SCORE | Indicates the score of the game through an alphanumeric string such as: [away team score] - [home team score] | 10 732 035 | - | - | - |
| SCOREMARGIN | Indicates the winning/losing margin of the team at the moment when the play registered in the observation took place | 10 732 035 | -60 | 65 | - |

This dataset has also allowed to understand both how the home team winning frequency has fluctuated over the years and the differences observed for this reality in the regular season and the playoffs. In table 3 and in figure 5 it is clear that there exists in the NBA a home-court advantage, given that, for the games in our database, the home team wins consistently more frequently than the away team – on a simple average, the home team wins 60% of the time –, which is concurrent to the existing literature on this issue.[47] Moreover, figure 5 also undoubtedly shows the different dynamics between the regular season and playoffs home team winning frequency, which is mainly due to two factors: *(i)* firstly, the playoff sample is much smaller[48] than that of the regular season, which renders the observed increase in the volatility of the series across time and *(ii)* secondly, it is due to the very nature of the playoffs, where each team plays its opponent on a best of 7 series throughout a sequence of elimination rounds, conferring extra importance to the games played at home during the playoffs versus when they are played in the regular season.

Table 3: Home team winning frequency

| | **Frequency** | **Percent** | **Min** | **Max** |
|---|---|---|---|---|
| Regular Season – Home Team Lost | 8 952 | 40,08% | 37,34% | 42,52% |
| Regular Season – Home Team Won | 13 384 | 59,92% | 57,48% | 62,66% |
| Playoffs – Home Team Lost | 536 | 35,24% | 25,58% | 44,05% |
| Playoff – Home Team Won | 985 | 64,76% | 55,95% | 74,42% |
| Total – Home Team Lost | 9 488 | 39,77% | 37,43% | 42,41% |
| Total – Home Team Won | 14 369 | 60,23% | 57,59% | 62,57% |

This empirically observable home court advantage effect will be one of the first methods of

[42] 2 if the game was played during the regular season or 4 if it was during the playoffs.

[43] For example, if the observation reports to a game occurred during the regular season of the 2012-2013 season, then the code will start by 212.

[44] This means that, for example, the 234th game of the 2007-2008 regular season will be coded by 20700234.

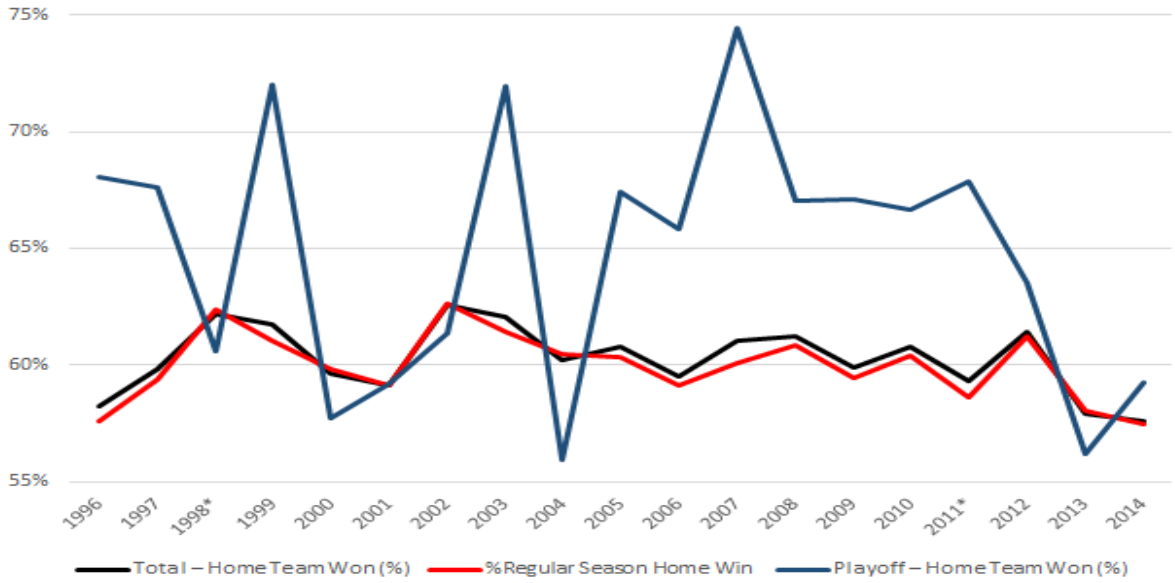
[45] Note that this represents the average of the number of events in each game and not the average of the variable *"eventnum" per se*.

[46] After 4, each number represents an overtime period. In that sense, observations registering the number 5 mean that the play occurred during the first overtime, entries registering the number 6 report to the second overtime and so forth up to the fourth overtime (represented by the number 8).

[47] *e.g.* Jones (2007); Entine and Small (2008); Gandar et al. (1988).

[48] In the last 3 NBA seasons, the average number of regular season games was 1230 whereas the average number of playoff games was 85.

Figure 5: Fluctuation of the home team winning frequency




cross checking the quality of the dataset fuelling our model, as if one estimates the winning probability of the home team with nothing but an intercept, it shall return precisely this home court advantage effect, thus expressing not a 50%-50% duality for each team, but an empirically observable advantage of the home team. This idea will be explored further ahead in this section, when we discuss the model's framework. Notwithstanding, as observable in figures 1, 2, 3 and 4, betting markets tend to absorb in the moneyline odds[49] not only the home-court advantage effect, but also the favourite/underdog effect of each team, as they vary way beyond the minimums and maximums presented in table 3. In that sense, it is essential that we incorporate these moneyline odds in our model to serve as a "reference start" to the estimation we want to undertake and thus avoid ignoring the important favourite/underdog duality. To that extent, in the next subsection we show the moneyline odds dataset we have gathered and describe it briefly.

[49]We refer to moneyline odds as the odds prevailing in the market at the event's start

### 3.2.2 Moneyline dataset

For the construction of our *moneyline* dataset, we adopted a similar approach as we did for the extraction of the *baseline* dataset – the inclusion of all possible games through a time-efficient process which renders data with the required quality –, in order to facilitate the inevitable merging process with the remaining datasets. While for the *baseline* dataset we had to design a complex Python oriented extraction approach, the process was fairly easier for the *moneyline* dataset, since there are a number of websites through which historical moneylines can be obtained for many sports.[50] To that extent, we opted to include data from *sportsbookreviewsonline.com*, which includes historical moneyline odds from the Las Vegas bookmakers[51] from the 2007-2008 to the 2015-2016 NBA season – regular season and playoffs. However, note that, for coherence reasons, we only included information spanning from the 2007-2008 to the 2014-2015 seasons, to match the seasons covered in the *baseline* dataset. Additionally, this provider offers the data in the user-friendly .csv format, from which we have mined the following relevant information:[52]

Table 4: Summary of the *moneyline* dataset

| **Variable** | **Description** | **Number of Observations** | **Min** | **Max** |
|---|---|---|---|---|
| Date | Indicates the month and day during which the game was played | 20 510 | - | - |
| VH | Indicates if the odd in the observation respects the home or the visiting team | 20 510 | - | - |
| Team | Indicates to which NBA team the odd respects | 20 510 | - | - |
| ML | Indicates, in american format, the moneyline odd of the team identified in variable "Team" | 20 510 | -13 000 | 3 000 |

Note that table 4 presents the data concerning the home and visiting teams. However, since we will estimate the winning probability of the home team on a play-by-play basis, we

[50] *E.g oddsportal.com* or *covers.com*.
[51] This odds are alson known as "Vegas lines" in the betting literature (Schnytzer and Weinberg, 2004).
[52] Again, note that we are only presenting the relevant information for this studies' purpose, since the raw dataset included additional non-relevant information (*E.g* 1st, 2nd, 3rd and 4th quarter score of each team)

removed from this dataset the information concerning the visiting team. Moreover, given that our intention is to use this dataset as a complement to the *baseline* dataset, choosing this source has two immediate implications: on the one hand, it is clear that the span of seasons covered will be much shorter, as the number of games collected is 10 255 NBA games, which represents a decrease of 13 602 games when compared to the total covered in the *baseline* dataset; on the other hand, this choice allows us to avoid having to consider including structural breaks in the data due to systemic changes occurring in the league between 1996 and 2007,[53] which is further justified since the league has consistently held 1.230 regular season games[54] since the 2007-2008 season, thus improving the stability of our data.

Choosing this provider also yielded another question needing to be addressed: the american format of the moneyline, which is not very easy to interpret and to incorporate in our model, given that it has a negative value (when the respective team is the favourite) and positive value (when the team is the underdog). To overcome this issue, we converted the moneyline odds to their implied probabilities such that:[55]

$$\frac{100}{moneyline_j + 100} \quad if moneyline_j > 0 \tag{6}$$

$$\frac{-moneyline_j}{-(moneyline_j + 100)} \; if moneyline_j < 0 \tag{7}$$

$$\forall j \; games \in moneyline \; dataset$$

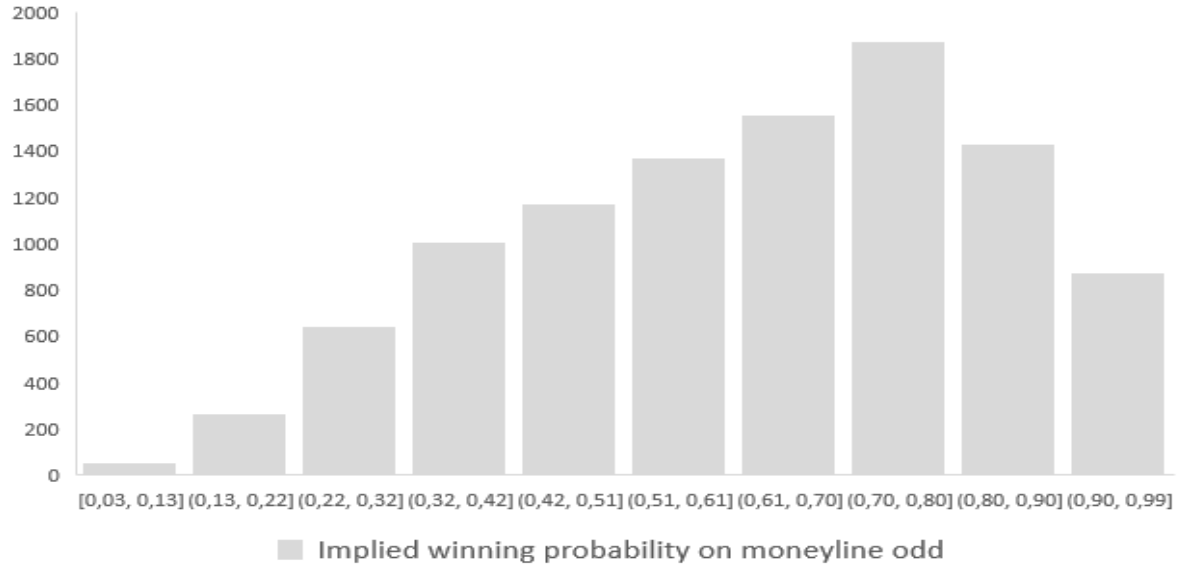
With this conversion duly processed, table 5 sjows the descriptive statistics of the resulting dataset. Table 5 and figure 6 allow us to retain two key ideas: *(i)* the data is slightly skewed to the right of the distribution, meaning that the majority of the observations occur when the home team is the moneyline favourite, which further translates not only the favourite/underdog duality each game entails but also the home court advantage we approached previously; and *(ii)* the average of the implied winning probability on the moneyline odds is effectively within the maximums and minimums we identified for the home team winning frequency for

[53] As, for example, the end of the hand-check rule implemented in the 2000-2001 season or the introduction of the three second defensive rule in the following season (NBA, 2008).
[54] With the exception of the lock-out shortened 2011-2012 regular season where only 990 games were played.
[55] In line with Cortis (2015).

Table 5: Summary of the implied winning probabilities on *moneyline* dataset

| | **Moneyline odd home team$_j$** |
|---|---|
| Number of observations | 10 255 |
| Min | 0,0322581 |
| Max | 0,9923664 |
| Mean | 0,6236185 |
| Variance | 0,0429001 |
| Standard Deviation | 0,2071233 |
| Skewness | -0,3497312 |
| Kurtosis | 2,2557570 |

Figure 6: Histogram of the implied winning probabilities on moneyline dataset




the seasons contemplated in the baseline data (see table 3) and slightly above its average for the seasons contemplated in the *moneyline* dataset (approximately 60% versus the 62% in the *moneyline* dataset), which further translates the alignment between the moneyline odds, the home team winning frequency and the concept of home-court advantage.

### 3.2.3 The model's framework

With our databases and supporting variables clearly defined in the subsections above, we are now able to decide the framework under which we will perform the estimation of the theoretical winning probabilities of the home team, on a play by play basis, which will serve as the crucial information element in the efficiency test *à la* Fama (1970).

The first step towards the construction of this model is, naturally, finalizing and specifying the dataset which will serve as input to the estimation process. This is done by merging the *baseline* dataset to the *moneyline* dataset using the game codes identified in table 2, thus creating a new dataset which we henceforth identify as the *play-by-play* dataset[56] and describe in table 6.[57]In this process, we also opted to mine the *play-by-play* dataset further, thus creating a set of additional variables describing with greater detail what is happening in each play, which will be essential for our model.

With the relevant variables and corresponding databases dully defined, we are finally able to start crafting the aforementioned model. As explored in the literature review, there is not a consensual method through which economists and statisticians approach the game of basketball for the purposes we strive for. In fact, the literature has shown that there are a multitude of ways through which one could model the winning probabilities inherent to a basketball game – *e.g.* Štrumbelj and Vračar (2012)'s Markov based approach, Kenter (2015)'s combinatorial sequences approach or even Beuoy (2015)'s weighted logistic regression method, among many possible others.For our model, taking into account that we have a binary dependent variable – which reduces the estimation methods and models applicable (Wooldridge (2009, p. 578))–, we have opted to anchor its construction on the richness of the *play-by-play* dataset which supports it, thus opting for a cross sectionally oriented maximum likelihood estimation (MLE) approach, which we explore further ahead.

[56]Note that the original *moneyline* dataset did not include a native identification code as the *baseline* one did. Notwithstanding, using the date and home/visiting teams information contained in the dataset, we were able to derive an identification code which enabled the merging process.

[57]Note that we are only presenting the variables that were added in relation to the *baseline* dataset. The original variables were retained and matched to the *moneyline* dataset.

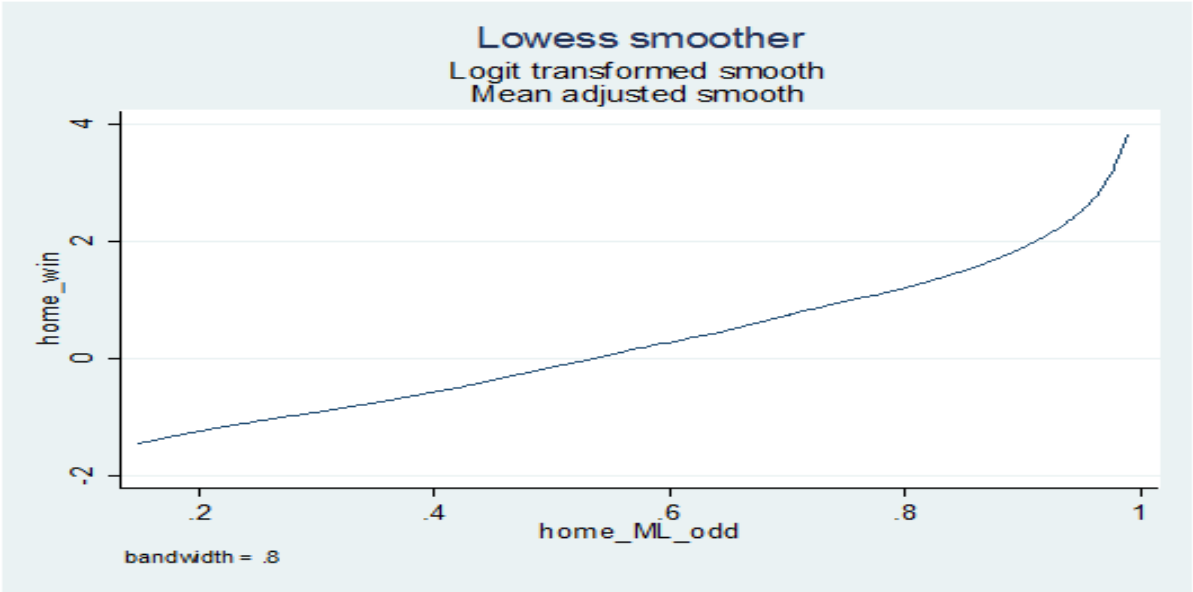
[58]Note that for the *"time_elapsed"* and the *"margin_home"* variables, the averages presented report to end of game averages and not in-game averages

Table 6: Summary of the *play by play* dataset

| Variable | Description | Number of Observations | Min | Max | Average[58] |
|---|---|---|---|---|---|
| home_win | Indicates if the home team has won for each game identification | 4 108 439 | 0 | 1 | - |
| home_ML_odd | Indicates the winning probability of the home team implied in the moneyline odd | 4 108 439 | 0,032258 | 0,992366 | 0,623619 |
| home_ML_odd2 | Square of the home_ML_odd variable | 4 108 439 | 0,001041 | 0,984791 | - |
| margin_home | Indicates the scoring difference between the home team and the away team, at the moment the play was recorded | 4 108 439 | -58 | 58 | 2,955303 |
| time_elapsed | Indicates how much seconds have been played since the start of the first period for the respective game identification code | 4 108 439 | 0 | 4142 | 2 899 |
| time_elapsed2 | Square of the time_elapsed variable | 4 108 439 | 0 | 17 156 164 | - |
| elapsed_margin | Represents the product of the margin_home variable and the time_elapsed variable | 4 108 439 | -167 040 | 164 430 | - |
| elapsed_margin2 | Square of the elapsed_margin variable | 4 108 439 | 0 | 27 037 224 900 | - |
| home_FGMiss | Binary variable indicating that the home team has missed a field goal attempt | 4 108 439 | 0 | 1 | - |
| home_FTMiss | Binary variable indicating that the home team has missed a free throw attempt | 4 108 439 | 0 | 1 | - |
| home_TO | Binary variable indicating that the home team has committed a turnover | 4 108 439 | 0 | 1 | - |
| home_foul | Binary variable indicating that the home team has committed a foul | 4 108 439 | 0 | 1 | - |
| home_steal | Binary variable indicating that the home team has stolen the ball | 4 108 439 | 0 | 1 | - |
| home_Oreb | Binary variable indicating that the home team has obtained an offensive rebound | 4 108 439 | 0 | 1 | - |
| home_Dreb | Binary variable indicating that the home team has obtained a defensive rebound | 4 108 439 | 0 | 1 | - |
| away_FGMiss | Binary variable indicating that the away team has missed a field goal attempt | 4 108 439 | 0 | 1 | - |
| away_FTMiss | Binary variable indicating that the away team has missed a free throw attempt | 4 108 439 | 0 | 1 | - |
| away_TO | Binary variable indicating that the away team has committed a turnover | 4 108 439 | 0 | 1 | - |
| away_foul | Binary variable indicating that the away team has committed a foul | 4 108 439 | 0 | 1 | - |
| away_steal | Binary variable indicating that the away team has stolen the ball | 4 108 439 | 0 | 1 | - |
| away_Oreb | Binary variable indicating that the away team has obtained an offensive rebound | 4 108 439 | 0 | 1 | - |
| away_Dreb | Binary variable indicating that the away team has obtained a defensive rebound | 4 108 439 | 0 | 1 | - |
| clutch | Indicates that there are less than 120 seconds to be played | 4 108 439 | 0 | 1 | - |
| clutch_margin | Represents the product of the clutch variable and the margin_home variable | 4 108 439 | -58 | 58 | - |
| clutch_margin2 | Square of the clutch_margin variable | 4 108 439 | 0 | 3 364 | - |
| clutch_elapsed | Represents the product of the clutch variable and the time_elapsed variable | 4 108 439 | 0 | 4 142 | - |
| clutch_elapsed2 | Square of the clutch_elapsed variable | 4 108 439 | 0 | 17 156 164 | - |

In this sense, the first natural step we took was evaluating how the dependent variable we are modelling – the home team victory – relates to a reduced set of key variables included in the *play-by-play* dataset, in order to detect any possible non-linear relation. In this set, we have considered the *"time_elapsed"*, *"margin_home"* and *"home_ML_odd"* variables as the remaining variables are essentially binary indications of the type of play which has occurred. Moreover, it is our understanding that these three variables capture many relevant factors we wish to embed in the model – the home-court advantage and the favourite/underdog duality (*"home_ML_odd"*), the run-time effect (*"time_elapsed"*) and the scoring of points by both teams (*"margin_home"*). To this extent, we have applied Cleveland (1979)'s locally weighted regression of our dependent variable on each of the independent variables enunciated, using StataCorp (2013)'s default bandwidth of 0.8, in order to smooth the binary dependent variable and understand how it relates to the aforementioned regressors. Note that since the calculation of these weighted regressions are extremely computationally intensive,[59] we have opted to run this procedure considering a random sample of 5% of the games in the *play-by-play* dataset, thus using a total of 512 games and 204 397 observations.[60]

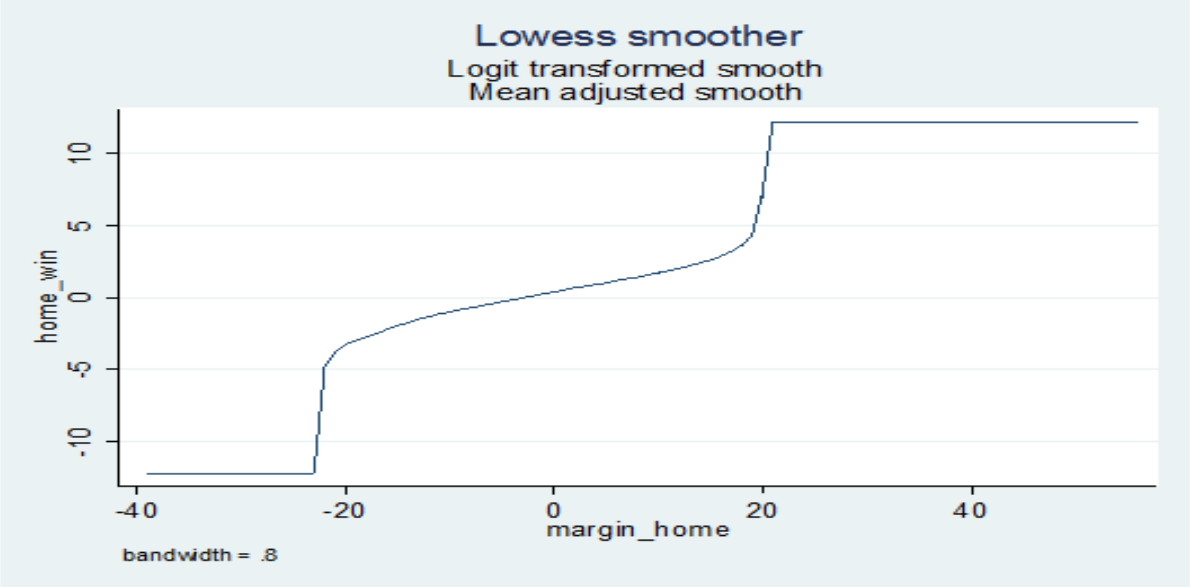
Figure 7: Locally weighted regression between "*home_win*" and "*home_ML_odd*"




[59] Since they impose the calculation of a regression *per* observation (StataCorp, 2013).
[60] In Annex 1, we present the code for this computation.

Figure 8: Locally weighted regression between "*home_win*" and "*time_elapsed*"


Figure 9: Locally weighted regression between "*home_win*" and "*margin_home*"




The results of this procedure shown in figures 7,8 and 9 portray the following relations between the dependent variable and each of the variables considered:

- *"home_ML_odd"*: Figure 7 shows that as this independent variable increases, there is a slight persistent non-constant increase in the slope of the smoothed series, partic-

ularly at the end of its domain. This denotes that for the same 1 unit increment in the home team's moneyline odd, the eventual winning probability of the home team is not adjusted homogeneously, *i.e.* it increases progressively more throughout the independent variable's domain. Therefore, our model shall include a squared term of the *"home_ML_odd"* variable in order to capture this relation;

- *"time_elapsed"* : Just like for the *"home_ML_odd"*, the locally weighted regression between the home team victory and elapsed time variables, portrayed in figure 8, also returned a smoothed series which also clearly depicts a non-linear relation between these two variables. This implies that the impact of each observation during the course of the game increases as the game flows, which justifies the addition of a squared term of the elapsed time to account for this effect. In a nutshell, one would easily comprehend this phenomena by acknowledging that, as the game approaches its end, each play is more important in determining the final outcome of the game given that the time available to reverse the current result is shortening, which further decreases the possibilities of such an event.

- *"margin_home"*: Contrary to the relation identified in the previous point, figure 9 elucidates how particular the relation between the home team winning margin and the eventual victory of the home team is, as there seems to exist a "two-sided" non-linear relation between these variables. In effect, the smoothed series suggests that there is an asymmetry around the point at which the home team's margin is zero, given the non-linear behaviour towards each side of its distribution. This behaviour is due to two factors: *(i)* the asymmetry portrayed is due to the effect of the home team being circumstantially winning or losing, which respectively increases or decreases the home teams' winning probability; and *(ii)* the non-linearity observed is a consequence of the intuitively understandable effect of winning/losing by different amounts – naturally, the 1-point increment between being up by 1 to 2 points has a very different effect on the home team winning probability than the 1-point increment between 11 to 12 points, as they progressively represent transitions to winning margins with different implications on the game outcome. Indeed, being up by 2 points is easily annulled by the opponent

in one possession, whereas being up by 12 points takes substantially more possessions to erase.

Following the relations identified in the previous points, we have opted to introduce in our model the "*home_ML_odd*" and the "*time_elapsed*" variables alongside a squared term, in order to capture the aforementioned non-linearities. Moreover, to capture the asymmetry effect induced by the "*margin_home*" variable, we have opted to include it as is, alongside a crossed term with the "*time_elapsed*" variable and their respective squared term. This was done in order to capture the intuitively understandable effect of the relation between the time elapsed and the home team winning margin – naturally, it is more important to be up by 2 points when 2000 seconds have been played versus when only 200 seconds were played.

Having defined how the key variables shall be introduced in our model, the final step before formulating it is clearly defining all the remaining key assumptions we adopted and reminding some of the underlying principles:

1. As defined in the *play-by-play* dataset description, it is worth stressing that we are not considering *all* of the games occurred from the 2007-2008 to the 2014-2015 NBA seasons, as punctually there were extraction errors and non-available information;

2. Simultaneously, we have disregarded all the observations which contained non-relevant information, especially substitutions, team rebounds and neutral observations[61];

3. We have opted not to control if either team is in possession of the ball, as this feature is already indirectly implied in some of the remaining variables – *e.g.* rebounds and turnovers;

4. Moreover, to avoid multicollinearity problems, we have opted to exclude from the model the variables signalling that either team has stolen the ball, since this information is already incorporated as a turnover of the opposite team. In this spirit, we have also removed the variables identifying the observations when either team missed a field goal or a free throw, as this information is covered by the rebounds variables;

[61] Observations not containing any information.

5. Furthermore, given that each season encompasses a very high number of games and since we are not interested in studying how the impact of each variable has changed over the course of the years, we also opted to neglect the seasons effect on the model, thus resorting to a cross sectional regression instead of applying a panel data approach;

6. In addition, we decided to not include an intercept term. In theory, this term would represent the value which the dependent variable would assume when all the remaining variables are zero (Wooldridge (2009, p. 32)), which, applied to our model, would be the winning probability of the home team right when the game is about to start. However, we do not wish to estimate this probability via an intercept, as it would only render a reproduction of the home court advantage effect. Despite this estimate is possibly useful as a quality control mechanism of our database formulation,[62] we wish to also incorporate the favourite/underdog duality each game encompasses, which is done by introducing the "*home_ML_odd*" variable, as it theoretically includes both effects we wish to capture. For this reason, the addition of this variable shall, in turn, make the inclusion of the intercept unnecessary, as it will itself perform as an intercept for each game, but with improved quality.

7. Finally, in order to adequately capture the dramatic effect on the game outcome of the last minutes, we have introduced the dummy variable "*clutch*", which signals that there are less than 120 seconds to be played in the 4th quarter or subsequent overtimes, in line with the end-game approach undertaken by Kenter (2015). To complement this feature, we have also included a crossed term between this "*clutch*" variable and the "*time_elapsed*" and the "*margin_home*", alongside their respective squared terms, to capture the understandably augmented effect of these variables in the final seconds of the game. Moreover, we also included the squared term of the "*clutch_elapsed*" variable, to grasp the intuitively understandable[63] non-linear increase in importance effect of the last seconds of the game.

[62] Comparing the predicted probabilities of the intercept only model with the sample home team winning frequency would, theoretically, show whether our database is capable of reproducing the home court advantage. We explore this idea in the next subsection.

[63] Naturally, an observation occurring with 120 seconds left is not as important as one occurring with 20 seconds left, specially in games where the home team winning/losing margin is slim.

Under these assumptions and the relations identified above, we have opted to perform the estimation of the theoretical winning probability of the home team, using the cross-sectionally oriented *play-by-play* dataset such that:

$$home\_win_{ij} = \beta_1 home\_ML\_odd_{ij} + \beta_2 home\_ML\_odd2_{ij} + \beta_3 margin\_home_{ij} + \tag{8}$$

$$+\beta_4 time\_elapsed_{ij} + \beta_5 time\_elapsed2_{ij} + \beta_6 elapsed\_margin_{ij} + \beta_7 elapsed\_margin2_{ij} + \beta_8 home\_Oreb_{ij} +$$
$$\beta_9 home\_Dreb_{ij} + \beta_{10} home\_TO_{ij} + \beta_{11} home\_foul_{ij} + \beta_{12} away\_Oreb_{ij} + \beta_{13} away\_Dreb_{ij} +$$
$$\beta_{14} away\_TO_{ij} + \beta_{15} away\_foul_{ij} + \beta_{16} clutch_{ij} + \beta_{17} clutch\_margin_{ij} + \beta_{18} clutch\_elapsed_{ij} +$$
$$\beta_{19} clutch\_elapsed2_{ij}$$

$$\forall\ j\ games \in play\text{-}by\text{-}play\ dataset \quad \wedge \quad \forall\ i\ EVENTNUM\ \in\ j\ games$$

With the model fully defined and dully justified, in the next section we proceed to test its quality through logical tests and through the standard statistical procedures applied to models with binary dependent variables.

### 3.2.4 The model's quality

The first step to assess the quality of the model we have formulated is, naturally, estimating it. As Wooldridge (2009, p. 578) puts it, in the context of binary dependent variables, "*For estimating limited dependent variable models, maximum likelihood methods are indispensable*". To this extent, we have estimated equation 8's parameters through maximum likelihood estimation using both a probit and a logistic (logit) regression of equation 8, in order to understand which one evidences the more plausible results.

The pertinence of the parameters estimated for our model are of quintessential importance for meeting the purposes we are striving for. In fact, the results shown in table 7 are very encouraging, with punctual caveats, and allow to draw some important conclusions:

- In both regressions, the sign of all parameters corresponds to the exact expected effects that each variable would exert on the dependent variable. Indeed, one would effectively anticipate that when the home team commits a turnover or a foul, or when the

Table 7: Estimation of equation 8 through Logit and Probit regressions

| Logit Regression | | |
|---|---|---|
| **Variable** | **Coefficient** | **Std.Error** |
| **home_ML_odd** | -2,86575*** | 0,01515 |
| **home_ML_odd2** | 5,65768*** | 0,01659 |
| **margin_home** | 0,02586*** | 0,00043 |
| **time_elapsed** | -0,00027*** | 0,00001 |
| **time_elapsed2** | 6,19E-08*** | 2,27E-09 |
| **elapsed_margin** | 0,00008*** | 2,85E-07 |
| **elapsed_margin2** | 7,83E-11*** | 5,77E-12 |
| **home_Oreb** | 0,03420*** | 0,00759 |
| **home_Dreb** | 0,04170*** | 0,00496 |
| **home_TO** | -0,13042*** | 0,00707 |
| **home_foul** | -0,14658*** | 0,00596 |
| **away_Oreb** | -0,09976*** | 0,00765 |
| **away_Dreb** | -0,12482*** | 0,00495 |
| **away_TO** | 0,04645*** | 0,00703 |
| **away_foul** | 0,09843*** | 0,00592 |
| **clutch** | -3,77171** | 1,6189 |
| **clutch_margin** | 0,42156*** | 0,00333 |
| **clutch_elapsed** | 0,00259** | 0,00104 |
| **clutch_elapsed2** | -4,53E-07*** | 1,67E-07 |

| Probit Regression | | |
|---|---|---|
| **Variable** | **Coefficient** | **Std.Error** |
| **home_ML_odd** | -1,59641*** | 0,00903 |
| **home_ML_odd2** | 3,23264*** | 0,00973 |
| **margin_home** | 0,01795*** | 0,00025 |
| **time_elapsed** | -0,00018*** | 3,60E-06 |
| **time_elapsed2** | 4,26E-08*** | 1,33E-09 |
| **elapsed_margin** | 0,00005*** | 1,62E-07 |
| **elapsed_margin2** | 3,30E-12 | 3,63E-12 |
| **home_Oreb** | 0,02019*** | 0,00444 |
| **home_Dreb** | 0,02489*** | 0,0029 |
| **home_TO** | -0,07738*** | 0,00414 |
| **home_foul** | -0,08756*** | 0,00349 |
| **away_Oreb** | -0,05925*** | 0,00448 |
| **away_Dreb** | -0,07404*** | 0,0029 |
| **away_TO** | 0,02765*** | 0,00411 |
| **away_foul** | 0,05868*** | 0,00345 |
| **clutch** | -2,13175*** | 0,92399 |
| **clutch_margin** | 0,22198*** | 0,00183 |
| **clutch_elapsed** | 0,00147*** | 0,00059 |
| **clutch_elapsed2** | -2,60E-07*** | 9,56E-08 |

| | |
|---|---|
| **N** | 4 108 439 |
| **Log-likelihood** | -1 806 245,46068 |

| | |
|---|---|
| **N** | 4 108 439 |
| **Log-likelihood** | -1 806 088,40 |

**Significance levels: * : 10% — ** : 5% — *** : 1%**

away team corrals an offensive/defensive rebound, the home team winning probability decreases, which is effectively signalled by the negative parameters of the respective variables. Conversely, the positive parameter associated to the variables when the home team obtains an offensive/defensive rebound, increases the scoring margin, or when the

away team commits a foul or a turnover, captures the positive effect that these events confer to the home team's winning probability.

- The non-linear relations identified in the previous subsection are effectively captured as the parameters associated to the "*home_ML_odd*" and the "*home_ML_odd2*" variables are different, which translates the effect we seek. This fact is also verified for the elapsed time variables,[64] for both regressions, which further warrants the incorporation of the intended effect.

- As for the statistical significance of the variables, in both regressions all the considered variables are statistically significant at the 5% significance level, with the exception of the squared term between the time elapsed and the home team's margin for the probit regression, as the p-value associated to the inherent t-test are all lower than 0.05, which allows the rejection of the null hypothesis of statistical non-significance inherent to this test (Wooldridge, 2009, p. 120-123)

Although the parametric results are satisfactory, it is also prudent to adjudge whether the variables included induce multicollinearity problems into the estimations, to further assure the quality of the estimated parameters. To do so, we computed the variance inflation factor (VIF), as in Wooldridge (2009, p. 99), for the variables presented in equation 8, excluding their squared and crossed terms.[65]

The results shown in table 8 are clear: there are no signs of multicollinearity in our model, as both the individual and the mean VIF are below 10, which, as Wooldridge (2009, p. 99) suggests, signals that there are no multicollinearity issues needing to be addressed.

The next important issue that could be explored, as in common econometric studies, would be the verification of the homoscedasticity assumption. However, we will disregard this step as it is unnecessary for the type of estimation methods we are using – "*Because maximum likelihood estimation is based on the distribution of $y$ given $x$, the heteroskedasticity in $Var(y|x)$ is automatically accounted for*" Wooldridge (2009, p. 578). The same could be ruled for the usual test of the error normality assumption, as this step is not necessary in the

[64] Both as a standalone and crossed with the "*clutch*" variable.
[65] Including these terms would, naturally, render the analysis unfruitful, as it would be heavily biased by these linear combinations of independent variables.

Table 8: Computation of the Variance Inflation Factors (VIF)

| **Variable** | **VIF** | **SQRT VIF** | **R-Squared** |
|---|---|---|---|
| home_ML_odd | 1,13 | 1,06 | 0,1177 |
| margin_home | 1,14 | 1,07 | 0,1250 |
| time_elapsed | 1,25 | 1,12 | 0,1995 |
| home_Oreb | 1,02 | 1,01 | 0,0175 |
| home_Dreb | 1,04 | 1,02 | 0,0378 |
| home_TO | 1,02 | 1,01 | 0,0200 |
| home_foul | 1,03 | 1,01 | 0,0285 |
| away_Oreb | 1,02 | 1,01 | 0,0173 |
| away_Dreb | 1,04 | 1,02 | 0,0368 |
| away_TO | 1,02 | 1,01 | 0,0206 |
| away_foul | 1,03 | 1,01 | 0,0292 |
| clutch | 1,24 | 1,11 | 0,1931 |
| **Mean VIF: 1,08** | | | |

framework of a MLE estimation under a probit or logit regression since, "*In either case, e is symmetrically distributed about zero*" Wooldridge (2009, p. 577).

In this sense, we resorted to Peng and So (2002) to explore additional possible quality measures of our model. The authors state that, for an overall evaluation of the model, "*a logistic regression model is said to provide a better fit to the data if it demonstrates an improvement over the intercept-only model (also called the null model, which has no predictors). Such an improvement is examined by inferential and descriptive statistics*" Peng and So (2002, p. 42). To that extent, we have estimated the intercept-only model to enable the aforementioned assessment. Note that we estimated this model only through a logistic regression, since the log likelihood and the predicted probabilities would be the same under a probit regression, but with a different intercept value.[66]

[66]Under a probit regression, the null model returns an estimate for the intercept of 0,25101, a log likelihood of -2 766 524,83857 and a predicted probability of 59,91%.

Table 10: Intercept-only model estimation

| **Variable** | **Coefficient** | **Std. Err.** |
|---|---|---|
| Intercept | 0,40170** | 0,00101 |
| N | 4 108 439 | |
| Log-likelihood | -2 766 524,83857 | |

Significance levels : † : 10% ∗ : 5% ∗∗ : 1%

As one would expect, the predicted probability of the home team *per* the null model equates to 59,91%, which is exactly in-line with the average of the home team winning frequency for the seasons considered (59,78%),[67] thus adequatly translating nothing but the home-court advantage effect, which attests the validity of our *play-by-play* database and the intuition of using the moneyline odds in our model. In this spirit, we have computed the log likelihoods, the Akaike and the Bayesian information criterias and classified the predicted probabilities[68] through our model, both using the logistic and probit regressions, and through the logistic null model, in line with Peng and So (2002)'s approach for the inferential and descriptive statistics assessment.

Table 11 demonstrates that the quality of our model for estimating the home team's winning probability, on a play by play basis, is evident when compared to the null model. Not only the log likelihood is significantly higher for both the logit and probit regressions, but also the information criterias presented are much smaller than those obtained in the null model, hence revealing the comparatively higher quality of our model (Peng and So,

[67] Again, note that the values are not *exactly* the same given that we do not incorporate *all* of the games occurred in the seasons considered due to punctual extraction errors and data availability.

[68] This procedure seeks to assess the validity of the probabilities predicted by dividing the observations according to their predicted probabilities and classifying them as over or under a certain threshold. Following this classification, the observations whose predicted probability are higher than the threshold are classified as the binary sucess and below it as the failure. Afterwards, these imputed successes and failures are compared against the sample sucesses and failures, thus rendering the relative percentage of sucesses predicted and not predicted (StataCorp, 2013). For this procedure, we decided to set this dividing threshold at 0,5 and classify as sucesses the predicted probabilites above that threshold and as failures those below it, as this value is the one which makes more sense for the type of phenomena we are modelling – any higher/lower threshold would render as insucesses/sucesses observations where, in average, the home team would have won/lost more frequently.

Table 11: Comparative statistics between the null model and equation 8

| | **Log likelihood** | **Akaike Information Criteria** | **Bayesian Information Criteria** | **Correctly classified observations** |
|---|---|---|---|---|
| **Null Model** | -2 766 524,8 | 5 533 052 | 5 533 065 | 59,91% |
| **Logit regression** | -1 806 245,5 | 3 612 527 | 3 612 765 | 78,18% |
| **Probit regression** | -1 806 088,4 | 3 612 211 | 3 612 436 | 78,16% |

2002). In addition, the percentage of correctly classified observations is significantly higher for our model, which further assures its quality and effectiveness in meeting its purpose. Furthermore, when one compares directly the results of the logit and the probit regression of equation 8, the results are somewhat mixed: on the one hand, the probit estimation returns a slightly higher log likelihood and relatively lower information criterias; on the other hand, the logit regression correctly classifies more 0,02 percentage points of the total number of observations, which equates to approximately more 821 observations correctly classified, thus counterbalancing the relatively better performance of the probit regression in the remaining statistics.

Complementary to the analysis above, we have also computed an additional statistic which renders a further ideia on the quality of the model: McFadden (1974)'s pseudo $R^2$. This statistic is a standard goodness of fit measure suitable for binary dependent variable regressions, as it overcomes the impossibility of calculating the standard $R^2$ and provides an idea about the explanatory power of the model (Wooldridge, 2009, p. 581-582).[69] The calculation of this measure incorporates the log likelihood value associated to the unrestricted model – the envisaged model (equation 8) – and to the restricted model – the null one. The effective workaround to the referred limitation is simply done by computing a ratio between these two values and correcting it as such:[70]

$$Pseudo\ R^2 = 1 - \frac{Log\ likelihood_{unrestricted}}{Log\ likelihood_{restricted}} \tag{9}$$

Using the statistics above computed, this procedure renders a pseudo $R^2$ of 0,347 to both

[69] However, this "*idea*" must be taken lightly, as it does not imply the same conclusions as the conventional $R^2$ of standard OLS linear regressions.

[70] In line with McFadden (1974).

the logistic and the probit regression of equation 8, which provides an additional theoretical indication of its explanatory power.

On a final note, and to provide a complement to the analysis of the suitability of our model, we have computed Pearson's goodness of fit test[71] for the logistic and probit regression of equation 8. This test, whose null hypothesis is that the model chosen fits accurately the data, follows a $\chi^2$ distribution with $(M - k)$ degrees of freedom, with $M$ number of covariate patterns, $m_j$ number of observations having covariate pattern $j$, $y_j$ number of positive responses among observations with covariate pattern $j$ and $p_j$ predicted probability of a positive outcome in covariate pattern $j$ (StataCorp, 2013, 500-501):

$$\chi^2 = \sum_{j=1}^{M} \frac{(y_j - m_j p_j)^2}{m_j p_j (1 - p_j)} \ , \quad \chi^2 \sim \chi^2_{M-k} \tag{10}$$

$$\forall j = 1, 2, 3, ..., M$$

Table 12: Pearson goodness of fit test

| **Pearson Goodness of Fit test** | | |
|---|---|---|
| | **Logit regression** | **Probit regression** |
| **Number of observations** | 4 108 439 | 4 108 439 |
| **Number of covariate patterns** | 3 588 974 | 3 588 974 |
| **Pearson $\chi^2$** | 3 499 418,28 | 6 039 303,74 |
| **Prob $>\chi^2$** | 0,99999 | 0,00000 |

The results shown in table 12 are, again, ambiguous and must be taken lightly. For the logistic regression, the p-value associated to the test is much higher than the conventional 5% significance level, which effectively allows us to not reject the null hypothesis and adjudge our model's fit as correct. Conversely, for the probit regression, the p-value associated to the test is much lower than the 5% significance, leading to the rejection of the aforementioned null hypothesis. However, note that in both cases the number of covariate patterns identified

[71] In line with Hosmer Jr et al. (2013).

is somewhat high and close to the total number of observations, but the underlying test statistics are very different in both regressions. Despite this, we take this test's result as a further indication of the quality of our model given that it does not make it "necessarily inappropriate" (StataCorp, 2013, p. 495).

Following all of the statistical and logical tests performed, we conclude that our model is indeed pertinent to be used for the purposes we have designed it. Therefore, and despite of the fact that the probit regression returned a higher log likelihood and lower information criterias, we decided to run equation 8 using a logistic regression, given that it has yielded all parameters as statistically significant but, more importantly, it has provided a relatively higher percentage of correctly classified observations, which is, in essence, the very purpose of these models.

In this sense, with the model and regression method duly specified and tested, we are now able to use it to compute the all necessary information element in our testing framework, thus completing the necessary tools for this study. However, before we proceed to this estimation, it is pertinent to adequately formalize how we actually test for efficiency and the underlying hypothesis. In the next subsection, we describe the testing framework, its theoretical basis and its elements' complex merging process.

## 3.3 The testing framework

The methodological and logical soundness of the framework under which we test for market efficiency of the *moneyline* dataset is a crucial element for the validity and applicability of this study. Taking this matter into careful consideration, we constructed our testing framework around the weak-form market efficiency tests that Schnytzer and Weinberg (2004) and Zuber et al. (1985) perform to NBA and NFL point spreads, respectively. In these studies, the authors regress the actual point spread of their samples' games on the respective point spreads predicted by the Las Vegas bookmakers – the Vegas lines – such that:[72]

$$Actual\ Point\ Spread_{ijt} = \beta_0 + \beta_1 Vegasline_{ijt} + \epsilon_{ijt} \tag{11}$$

$$\forall\ i, j\ teams \wedge\ time\ t,\ with\ \epsilon_{ijt} \sim i.i.d.$$

[72] As in Schnytzer and Weinberg (2004, p.5).

Under this setting, the markets are deemed to be efficient if jointly $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$ hold. The intuition for this testing is simple: if the markets are indeed efficient and the condition holds, then the Vegas lines are an unbiased complete predictor of the actual point spreads and shall render any additional element statistically non-significant (Schnytzer and Weinberg, 2004). Therefore, if the parameter associated to the Vegas line and the intercept prove to be statistically different from one and from zero, respectively, then the test demonstrates that there is additional information, not included in the Vegas lines, which explains a statistically important portion of the dependent variable, thus proving the inefficiency of the market. However, note that this test is designed to assess market efficiency through an *ex-post* approach, *i.e.* the Vegas lines predictions for the point spreads are compared against the actual point spreads, *after* they are observed.

In our study, we wish to perform a similar analysis as in Schnytzer and Weinberg (2004) and Zuber et al. (1985) but, instead of doing it through an *ex-post* scope, we strived to analyse market efficiency of the home team winning odds as the game progresses, *i.e.* on a near second by second *live* basis. In this sense, we have opted to adapt equation 11's testing framework and fit it to our purposes, as it effectively is able to incorporate both the information element we have crafted and the price element obtained. Hence, to introduce in the test the in-game second by second logic we seek, we run an OLS-based regression of the home team winning probabilities ($y_t$), which are estimated on a play by play basis through the model chosen previously, on the winning probabilities of the home team implied in the "*last price matched*" variable ($x_t$), derived from the *market* dataset:

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \tag{12}$$

$$\forall\ t\ \in Time\ stamp\ (market\ dataset),\ with \epsilon_t\ \sim\ i.i.d.$$

By estimating this regression through ordinary least squares (OLS), it will theoretically show how much of the estimated winning probability of the home team is explained by the betting prices prevailing on the market for that team, thus crystallizing Fama (1970)'s definition of efficiency. Following this rationale, if the markets were efficient while the game is in play, then, for all $t$'s, the joint condition of $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$ should hold.

Moreover, to facilitate the testing framework, we manipulated a bit further equation 12 to fit the test to a simpler joint F-test. This was done by simply subtracting $x_t$ to each side of equation 12, such that:

$$y_t - x_t = \beta_0 + (\beta_1 - 1)x_t + \epsilon_t \tag{13}$$

By substituting $y_t - x_t$ and $(\beta_1 - 1)$ by $z_t$ and $\theta_t$, respectively, the equation is now simplified as :

$$z_t = \beta_0 + \theta_1 x_t + \epsilon_t \tag{14}$$

Note that under equation 14's setting, the efficiency test performed by Schnytzer and Weinberg (2004) and Zuber et al. (1985) is effectively modified to a much easier joint test. Indeed, as mathematically expressed above, testing for $\hat{\beta_0} = \hat{\theta_1} = 0$ is the same as jointly testing for $\hat{\beta_0} = 0$ and $\hat{\beta_1} = 1$, as $\beta_1 - 1 = \theta_1$. In this sense, the modified testing framework we designed is simply performed through a standard joint F-test, which follows an F distribution with $(q, n-k-1)$ degrees of freedom, such that:[73]

$$F = \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{n-k-1}}, \; F \sim F_{(q,n-k-1)} \tag{15}$$

Although the testing framework expressed in equation 14 is quite simple, it has a major caveat associated to its setting: it demands that the *play by play* and the *market* datasets are merged into the same dataset and aligned by the time stamp $t$, to enable the near second by second analysis we seek. However, this time stamp $t$, which encompasses the indication of the hour, minute and second respective to each play, is only included in the *market* dataset. Indeed, the only information available in the *play by play* dataset that identifies the moment in time correspondent to each play is the "*WCTIMESTRING*" variable, which includes both the relevant hour and minute, but not the second. This mismatch between the time identifiers in both datasets effectively creates an additional challenge for meeting our end goal, as it

[73] With $SSR_r$ as the sum of the squared residuals of the restricted model, $SSR_{ur}$ as the sum of the squared residuals of the unrestricted model, $q$ restrictions imposed (in this case, 2), $n$ observations and $k$ independent variables, in line with Wooldridge (2009, p. 145-147).

does not, in any way, allow the information to be matched by the second on a 1 to 1 basis, as this merge is only possible on a minute basis – something we are not interested in.[74]

To workaround this important shortcoming, we decided to manually match the plays identified in the *play by play* dataset to the relevant time stamp included in the *market* dataset. This was done by reviewing every second of the games considered in the *market* dataset and, taking as reference the hour and minute expressed both in the "*time stamp*" and the "*WCTIMESTRING*" variables, manually accounting for the real time elapsed between each play expressed in the *play by play* dataset. With this real-time elapsed duly calculated, we appended the value included in the "*EVENTNUM*" variable to the corresponding relevant time stamp of the *market* dataset. This unorthodox method immediately showed its virtues and deficiencies: on the hand, it effectively enabled the merging of the databases we sought, thus warranting the applicability of equation 14; on the other hand, it proved to be a very cumbersome task,[75] with the error margin associated to a manual imputation process.

To identify if this manual process has led to systematic errors on a minute basis, we compared the minute inscribed in the "*WCTIMESTRING*" variable to the minute contained in the time stamp of the matched observation in the *market* dataset, for all "*EVENTNUM*" observations of the games considered in the *market* dataset, having obtained a perfect 100% correspondence between the minutes inscribed in both datasets. Despite this, the same 1 to 1 guarantee cannot be made on a second basis, due to the identified and unavoidable mismatch between the *market* and the *play by play* datasets.

With the testing framework duly specified, justified and theoretically framed, we are now able to put it into practice. In the next section, we show the outcome of the merging process between the *play by play* and the *market* datasets, the results of the test we designed and comment on its findings.

[74] In practical terms, this 1 to 1, second by second, merging process is virtually impossible, as the NBA API does not disseminate play by play data with a time stamp detailed up to the second.

[75] This procedure impused the thorough revision and computation of the time elapsed between over 1700 plays, dispersed through almost 8 hours of game footage.

# 4 Results

Before delving into the actual application of the testing framework we have designed, the visualization of the results of our model's estimation and its matching results are imposed, in order to understand its suitability in meeting its purpose. To that extent, figures 10, 11, 12 and 13 show the fluctuations of our estimation of the home team winning probabilities, *per* a logistic regression of equation 8, against the winning probability implied in the home team winning odds for the selected games.

The analysis of these four figures allow to extract important notes about the pertinence of our approach and of the model's design:

1. Firstly, all figures suggest that the results obtained through our model ( *"y"*) follow, in a general way, the winning probability of the home team implied by the corresponding market odds ( *"prob_market"*), as both curves roughly evolve towards the same path. This provides an early indication of the validity and plausibility of the approach endured and of the model we have constructed;

2. Secondly, the combined role of the home team winning margin and of the time stamp in determining the outcome of each game seems to be included in both curves and appears to evidence the type of non-linear relations that were identified previously. Indeed, when one considers, as an example, game 1 (figure 10), the 5 point increment on the home team's winning margin occurred in the early first quarter (between 2:24 PM and 2:29 PM) resulted in an increase of the winning probability of approximately 4 percentage points, which was much smaller than the 11 percentage point increase registered in the fourth quarter (between 4:25 PM and 4:28 PM) due to the very same increase in the winning margin. This translates the non linear effect of this combination identified in the previous section and demonstrates the ability of our model in capturing it;

3. Thirdly, despite both series generally follow the same path, they do not do so in a completely synchronized and homogeneous fashion. This might be due to several factors:

   (a) On the one hand, the discrete nature of the predictions of our model might lead

Figure 10: Game 1 – *"implied winning probability of the home team"* (*prob_market*) and estimation of home team winning probability ($y$)

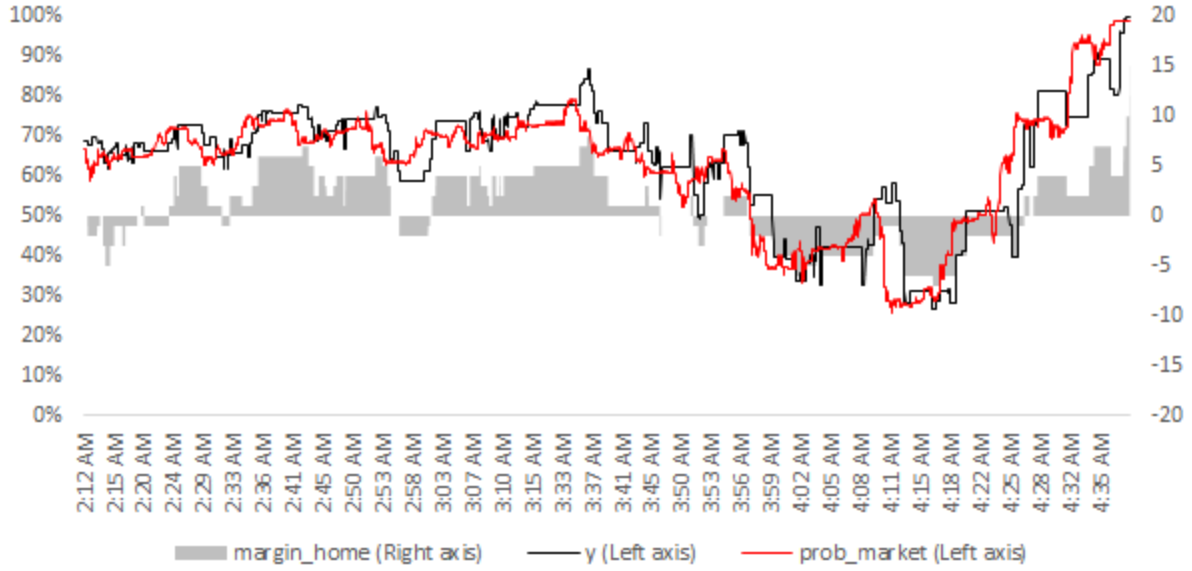



Figure 11: Game 2 – *"implied winning probability of the home team"* (*prob_market*) and estimation of home team winning probability ($y$)

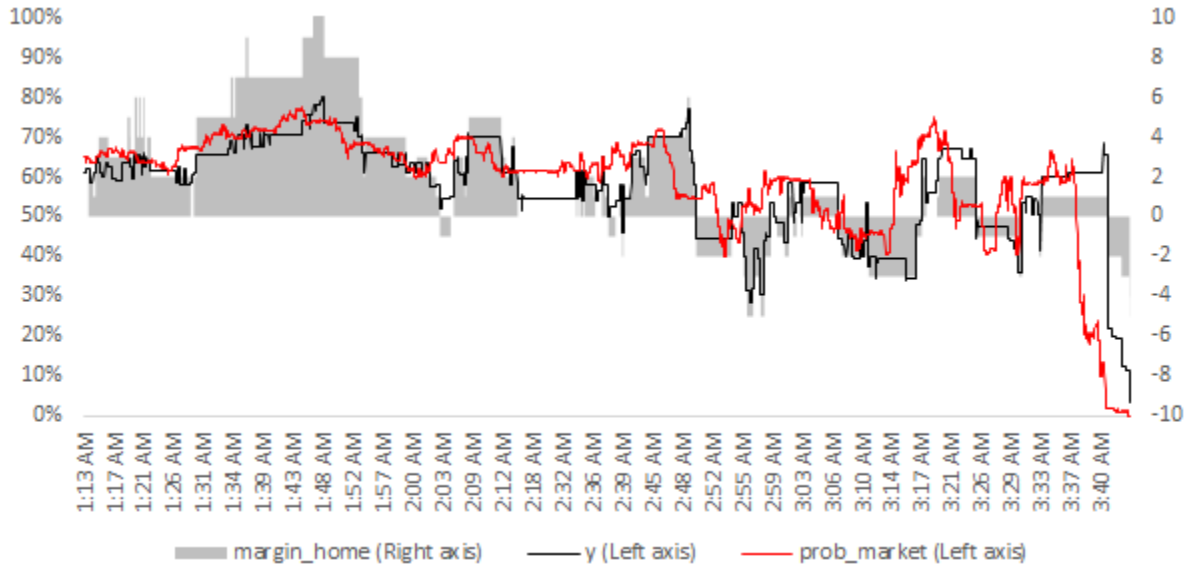


Figure 12: Game 3 – *"implied winning probability of the home team"* (*prob_market*) and estimation of home team winning probability ($y$)

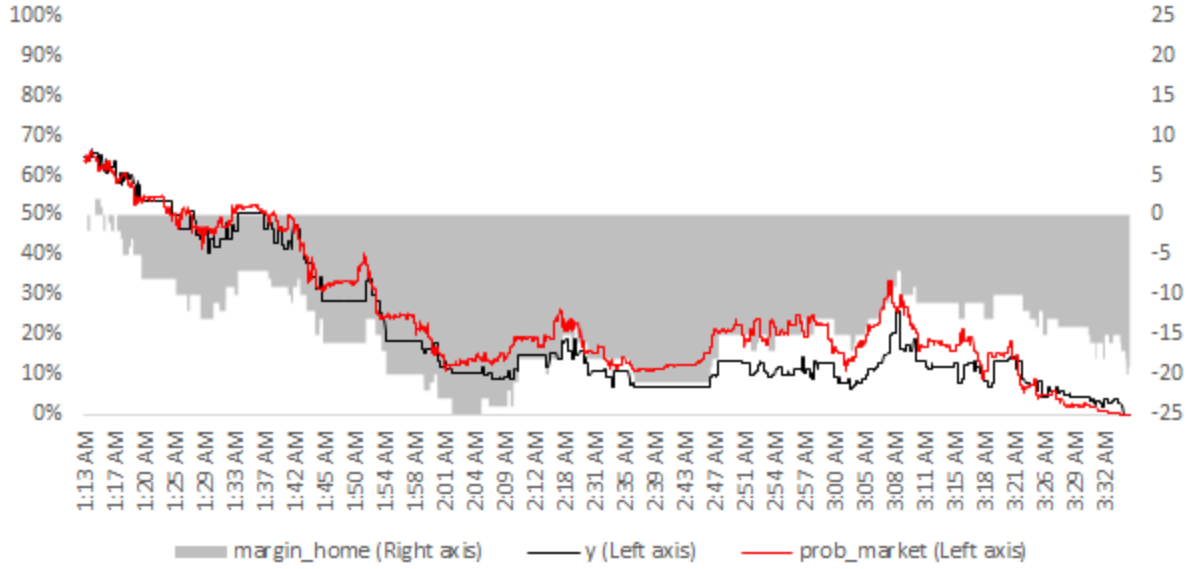
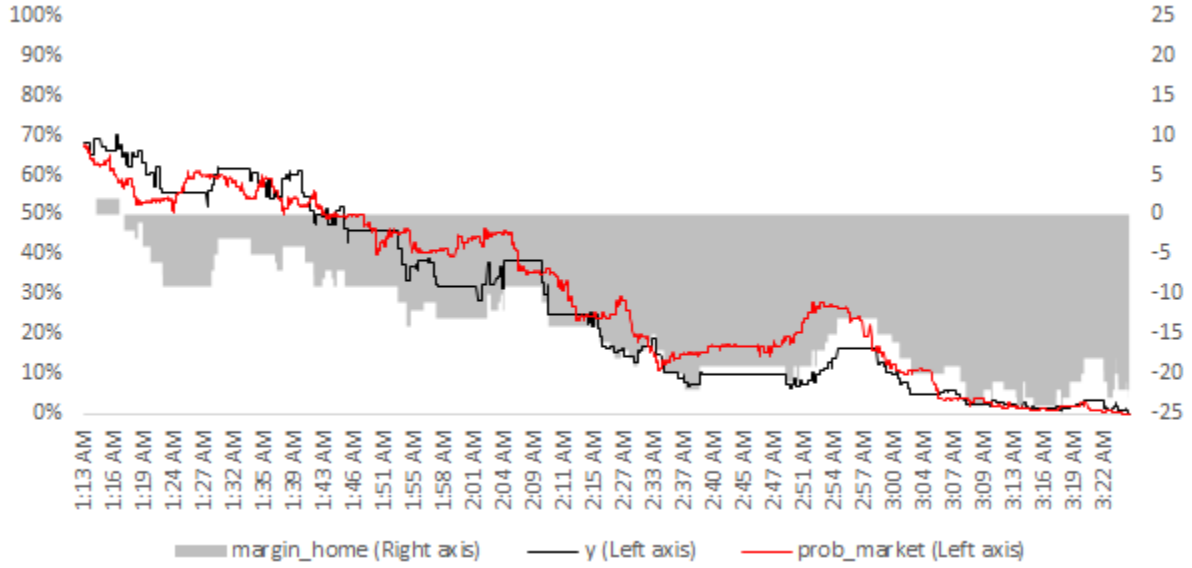



Figure 13: Game 4 – *"implied winning probability of the home team"* (*prob_market*) and estimation of home team winning probability ($y$)



it, in very specific cases,[76] to react to important changes with a certain delay. For example, when one considers the end of game 2, the period between 3:33 AM and 3:37 AM stands out due to the difference between both curves. However, note that during this period, a technical time-out was called and the possession was reviewed, both of which escape the scope of our model, hence explaining the flat shape of the estimated winning probabilities. Conversely, the market immediately reacted to the evolution of events and quickly adjusted to the odds that our model eventually reached afterwards, when the following relevant events occurred, thus anticipating this estimate. This *delay* of the model's output is naturally due to the choice of not including these very specific situations in our model, as their effect on the game outcome are subject to many factors – not only on the time elapsed and the home team winning margin but also, *inter alia*, the substituted players, the momentum of each team when the time-out is called or the outcome of the possession review;

(b) On the other hand, one of the effects we are most interested in is the potential heterogeneity of the fluctuation of both curves, for the same game related incentive, given that if it occurs consistently, then there would be evidence of market inefficiency – the market would react to changes in the information set that would differ from their real implications on the probability of the home team winning. On that note, when one focuses on game 3, it is noteworthy that between the end of the second quarter and the early stages of the third quarter, the home team quickly increased the scoring margin by 6 points, which is translated, *per* our model, in an increase of 6 percentage points of its estimated winning probability. Conversely, for the very same period, the winning probability of the home team implied by the market rose by 10 percentage points, thus reacting completely differently to the stimulus identified. In this sense, our testing framework is exactly designed to capture this type of phenomena and reject, or not reject, the hypothesis that the betting markets are efficient for the selected games, on a near second by second

[76] When any event outside of those captured in our model (*e.g.* Substitutions, jump balls and video-reviews of possession) occurs, the model only changes the probability of the home team for the ensuing relevant play.

basis.

Although these figures provide valuable input into the fit of our model to the underlying datasets and of its output against the market odds, they do not allow any definitive conclusion on the efficiency behaviour of the markets under analysis. Indeed, to meet the ambitious goal we have set, it is imposed that we run the testing framework designed in the previous section – equation 14 – and thoroughly interpret the estimates it renders. To this extent, in the next subsections, we explore the properties of the OLS estimator for that regression and perform the designed test.

## 4.1 The properties of the OLS estimate of the testing equation

Equation 14 presents a cross sectionally oriented simple linear regression, whose parameters can be estimated through ordinal least squares (OLS). As Wooldridge (2009, p. 102) puts it, in the context of cross-sectionally oriented regressions, the verification of the Gauss-Markov theorem hypothesis "*justifies the use of the OLS method rather than using a variety of competing estimators*". This is warranted since, under this theorem, the OLS estimator is "*the best linear unbiased estimator (BLUE)*" (Wooldridge, 2009, p. 109) for the simple/-multiple regression parameters, thus assuring they are unbiased and the most efficient.[77] To that extent, it is important to specify what are the hypothesis underlying the Gauss Markov theorem, in order to adequately test them when equation 14 is estimated. According to Wooldridge (2009, p. 84-105), these hypothesis are:

1. Linearity in the parameters, which implies that the model is formulated as in equation 16, with $\beta_j$ as an unknown constant parameter and $u_t$ as an unobservable random error;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u_t \tag{16}$$

2. Random sampling, which implies that the model incorporates $n$ observations following the population model assumed in the previous assumption;

[77] Note that we are assuming that the model's specification is correct and that there are no measurement errors in the dependent/independent variables. If some of these hypothesis are not verified, one might be calculating biased parameters (see, for example, Wooldridge (2009, p. 89-93)).

3. No multicollinearity among regressors, which implies that none of the independent variables evidences a perfect linear relation with any of the remaining explanatory variables;

4. Zero conditional mean, which implies that the expected value of the error term $u$ is zero, for any values of the independent variables, such that:

$$E(u|x_1, x_2, ..., x_k) = 0 \tag{17}$$

5. Homoskedasticity, which implies that the error term $u$ has the same variance for any values of the regressors, such that:

$$Var(u|x_1, x_2, ..., x_k) = \sigma^2 \tag{18}$$

Under this setting, Wooldridge (2009, p. 274-275) notes that hypothesis 5 can be somewhat relaxed to a less strict premise. Indeed, the author claims that it "*can be replaced with the weaker assumption that the squared error, $u^2$, is uncorrelated with all the independent variables ($x_j$), the squares of the independent variables ($x_j^2$) and all the cross products ($x_j, x_h,\ for\ j \neq h$)*" (Wooldridge, 2009, p. 274-275), which effectively fosters the possibilities to test for this assumption.

Additionally, a stronger hypothesis is also commonly referenced when one is dealing with cross-sectional data: the normality of errors. This implies that the "*population error $u$ is independent of the explanatory variables $x_1$, $x_2$, ..., $x_k$ and is normally distributed with zero mean and variance $\sigma^2$, such that $u \sim Normal(0, \sigma^2)$*" (Wooldridge, 2009, p. 118-119). According to the same author, this hypothesis is somewhat stronger than the previously appointed hypothesis 4 and 5, as it already implies that a constant variance ($\sigma^2$) and a zero expected value for $u_t$ are verified. This hypothesis, together with the 5 hypothesis enumerated previously, comprise the so called "*classical linear model (CLM) assumptions*" (Wooldridge, 2009, p. 118) which, when jointly validated, assure the estimated parameters are the BLUE estimator.

In this sense, the verification of these hypothesis is of quintessential importance, in order to validate the parametric estimates obtained through the OLS estimation of equation 14,

hence assuring that the efficiency test is being performed to unbiased and consistent estimates of the parameters. At first glance, when one compares the CLM hypothesis against the formulation of the test equation (equation 14), it is clear that both hypothesis 1 and 2 are already verified, as our testing equation does not encompass any non-linearity in the parameters and incorporates observations in line with the population of the supporting databases. Moreover, the validity of hypothesis 3 is also necessarily implied in our testing framework, as we are dealing with a simple linear regression with 1 regressor, thus the possibility of multicollinearity among independent variables is not applicable. Despite this, note that the remaining 2 hypothesis are not necessarily verified and need to be tested.

To test for the validity of the remaining hypothesis, we have considered three tests: the Breusch and Pagan (1979) test, the White (1980) test and the Jarque and Bera (1980) test. The first two are tests which seek to find linear and non-linear forms of heteroskedastic residuals,[78] in the spirit of the less strict premise identified for hypothesis 5. The Breusch and Pagan (1979) test is simply a Lagrange multiplier test as in equation 19, where $\hat{u}^2$ corresponds to the square of the residuals of the OLS regression of the testing equation and $R^2_{\hat{u^2}}$ is the standard $R^2$ of an auxiliary regression of these residuals on the original model's independent variables:

$$LM = n * R^2_{\hat{u}^2} \tag{19}$$

The resulting test statistic follows a $\chi^2$ distribution, with $k$ (number of independent variables) degrees of freedom, under the null hypothesis of homoskedastic residuals (Wooldridge, 2009, p. 274).

A somewhat different test for the presence of heteroskedastic residuals is the White (1980) test, which complements the former test by considering squared and crossed products of the independent variables in the aforementioned auxiliary regression. In this sense, for a model with 1 independent variable as equation 14, the auxiliary regression is performed such that:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_1^2 + \epsilon \tag{20}$$

[78] The residuals are simply the difference between the predicted values for the dependent variable and its original values : $\hat{u}_i = \hat{y}_i - y_i$, with $\hat{u}_i$ as the regression residuals, $\hat{y}_i$ as the predicted values and $y$ as the original dependent variable (Wooldridge, 2009, p. 38).

Under this setting, this test, which is capable of capturing non-linear forms of heteroskedasticity, is also simply computed through a Lagrange multiplier statistic, which follows a $\chi^2_k$ distribution, under the null hypothesis that all $\delta_j$ are equal to zero, with the exception of the intercept (Wooldridge, 2009, p. 275).

Contrary to the previous two tests, the Jarque and Bera (1980) test seeks to validate the somewhat stronger hypothesis we have identified: the normality of errors. To that extent, the test seeks to assess if the skewness and kurtosis of the residuals generated by the OLS regression are compliant with those verified for a standard Gaussian distribution, hence assuring that they are normally distributed. In that sense, the test statistic is computed as in equation 21, with $S^2$ as the sample skewness and $K^2$ as the sample kurtosis, and follows a $\chi^2$ distribution, with 2 degrees of freedom, under the null hypothesis that the testing residuals are normally distributed (Jarque and Bera, 1980):

$$JB = n(\frac{S^2}{6} + \frac{k-3}{24}),\ with\ JB \sim \chi^2_2 \tag{21}$$

***

With the classical linear regression model hypothesis duly specified and their necessary tests described, we estimated equation 14 through ordinal least squares, in order to assess the applicability of its parameters to the efficiency testing framework:

Table 13: OLS regression of equation 14

| **Variable** | **Coefficient** | **(Std. Err.)** |
|---|---|---|
| home_ML_odd | 0,00449* | (0,00202) |
| Intercept | -0,01931** | (0,00102) |

*Continued on next page...*

*... table 13 continued*

| Variable | Coefficient | (Std. Err.) |
|---|---|---|
| N | 27 639 | |
| $R^2$ | 0,00018 | |
| $F_{(1,27637)}$ | 4,95237 | |

Significance levels : † : 10% ∗ : 5% ∗∗ : 1%

As discussed previously, with some of the aforementioned classical linear model assumptions already verified, we proceed to test if the residuals are heteroskedastic and if they follow a normal distribution. To that extent, we computed the Breusch and Pagan (1979), White (1980) and the (Jarque and Bera, 1980) tests, for which we have obtained the following results:

Table 14: Breusch and Pagan (1979), White (1980) and Jarque and Bera (1980) tests

| | **Test Statistic** | **P-value** |
|---|---|---|
| **Breusch and Pagan (1979) test** | 448,05 | 0,0000 |
| **White (1980) test** | 78,25 | 0,0000 |
| **Jarque and Bera (1980) test** | 1,1E05 | 0,0000 |

The p-value associated to the Jarque-Bera test is below the 5% significance level, which leads us to reject its null hypothesis of normally distributed residuals. Although this rejection might apparently be troublesome, Wooldridge (2009, p. 758-759) demonstrates that, under the central limit theorem, for large samples, the residuals are considered to approximately follow a normal distribution, hence being considered as asymptotically normally distributed and retaining the desired properties. On that note, when one takes into consideration that that our regression is run on 27 639 observations, it is evidently the case that this theorem is applicable and that the residuals can be considered as asymptotically normally distributed going forward.

In addition, note that both the White and the Breusch-Pagan test are rejecting that the regression's residuals are homoskedastic, as both p-values are lower than the 5% significance level. This comes as no surprise since, as studied before, our model's estimates of the home team winning probability and the implied winning probabilities on the market odds tend to fluctuate more widely as the game approaches its end – thus meeting the definition of heteroskedasticity –, due to the augmented effect of the last moments of the game. Notwithstanding, note that, as Wooldridge (2009, p. 264-265) shows, the violation of this hypothesis does not directly imply that the estimated parameters are either inconsistent or biased. Indeed, this violation indicates that the variance of the estimated parameters is biased and, consequently, so are the standard errors derived, which is an important insight since they are no longer suitable to perform statistical inference. In that sense, to workaround this shortcoming, we have used White (1980)'s heteroskedasticity robust standard errors, which enable the calculation of the standard errors suitable for statistical inference.

Table 15: OLS estimate of equation 14 with heteroskedasticity robust standard errors

| **Variable** | **Coefficient** | **(Std. Err.)** |
|---|---|---|
| home_ML_odd | 0,00449* | (0,00208) |
| Intercept | -0,01931** | (0,00112) |
| | | |
| N | 27 639 | |
| $R^2$ | 0,00018 | |
| $F_{(1,27637)}$ | 4,67326 | |

Significance levels : † : 10% ∗ : 5% ∗∗ : 1%

Under this procedure, note that the parameters shown in table 15 are exactly the same as those calculated without robust standard errors. However, their standard errors are now somewhat higher, but accurate to perform statistical inference, as the parametric estimates

are still unbiased and consistent, but more importantly, their standard errors are no longer biased. In that sense, the parameters for both variables are statistically different from zero, which, although at the individual level, might already preclude the conclusions of our efficiency test.

On that note, now that the accurate parameters for statistical inference are dully calculated and to fully comply with the essence of Schnytzer and Weinberg (2004)'s and Zuber et al. (1985)'s efficiency tests, it is imposed that we test if both the intercept and the variable representing the market driven implied winning probability of the home team are jointly statistically equal to zero, thus meeting our designed testing framework. In the next subsection we show the results of this test and comment on its results.

## 4.2 The efficiency test results

Having estimated the parameters for equation 14 and thoroughly discussed their properties under an OLS estimation with heteroskedasticity robust standard errors, it is now finally pertinent to perform the efficiency testing we have designed. Recall that, under our test, we seek to demonstrate whether the estimated winning probabilities of the home team calculated through our model are fully explained by the market odds for the same event, thus meeting Fama (1970)'s preposition of prices *fully reflecting* all available information. In that sense, when one considers the testing equation (14), the markets are deemed efficient, on a second by second basis, if both $\hat{\beta}_0 = 0$ and $\hat{\theta}_1 = 0$ jointly hold. To that extent, we have computed this test for all games and for each game individually, for which we have obtained the following results:

The results shown in table 16 are clear and allow for the major conclusion of this study: they prove that the betting markets are not efficient for the games considered, on a near second by second basis, as the p-values associated to each of the tests are lower than the conventional 5% significance level, which leads us to reject that the markets are efficient, either for all games combined or for each game individually. Taking into account these results, what the test appears to be concluding is that the information element is not fully explained by the price element, thus providing an evidence on the inefficiency of the markets under analysis.

Table 16: Market efficiency test for equation 14

| | **Testing hypothesis:** $\hat{\beta}_0 = 0 \wedge \hat{\theta}_1 = 0$ | | | | |
|---|---|---|---|---|---|
| | All games | Game 1 | Game 2 | Game 3 | Game 4 |
| **Test Statistic** | 655,42 | 224,41 | 1 125,17 | 3 013,98 | 975,30 |
| **F (2, 27 637)** | F (2, 27 637) | F (2, 6 817) | F (2, 7 437) | F (2, 6 964) | F (2, 6 413) |
| **P-value** | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |

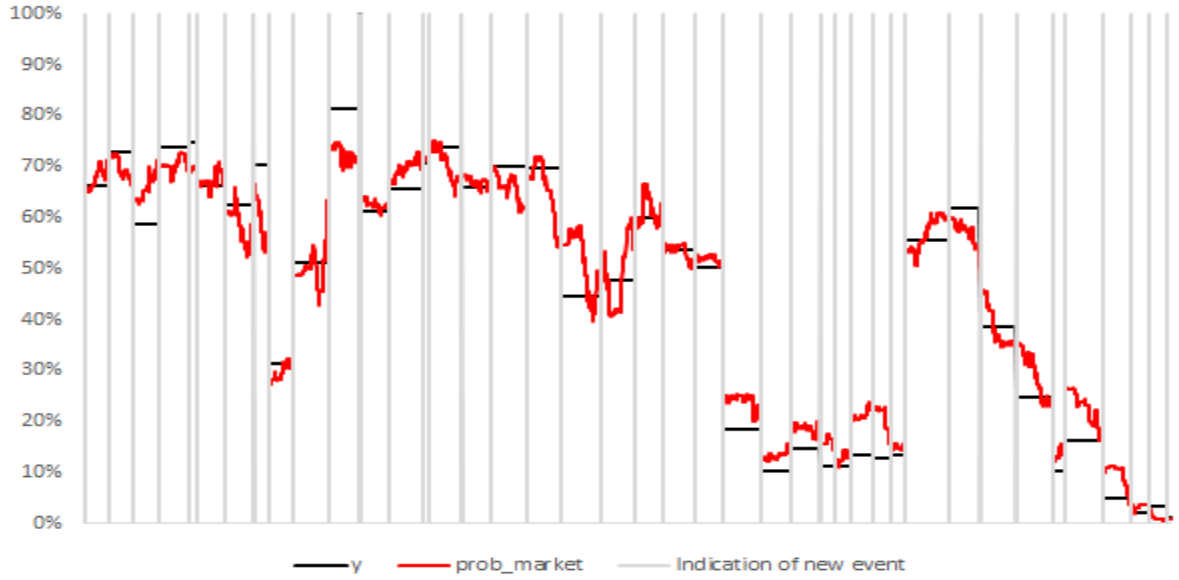
This conclusion is interestingly contrary to the studies we have surveyed on the pre-game market efficiency of NBA games, namely Paul et al. (2004)'s non-rejection of the hypothesis that the pre-game totals market is economically efficient, Baryla Jr et al. (2007)'s conclusion that after the first four games of the season the markets tend to correct any early season bias and swiftly evidence an efficient behavior, and Schnytzer and Weinberg (2004)'s demonstration of the weak form efficiency of the pre-game point spreads predicted by the Las Vegas bookmakers. In fact, although we must underline that the results obtained are only valid for the games surveyed, this contradiction seems to illustrate the interesting results that the analysis of the odd fluctuations during the game might render, as opposed to the conventional pre-game studies.

In this sense, in spite of the fact that this testing formulation provides a very intuitive and direct interpretation, the major shortcoming of adapting Schnytzer and Weinberg (2004)'s and Zuber et al. (1985)'s testing frameworks is exactly that we cannot identify and/or isolate the factor/s that is/are inducing the empirically observed behaviour. Nevertheless, recall that this study's main objective is simply providing empirical proof on whether the NBA betting markets are efficient, on a near second by second basis, for the set of games considered, hence leaving outside of its scope the reasons leading to these results. On that note, the conclusions laid in this study can be the starting point of future studies on the causes of the inefficiency verified for the in-play NBA betting markets.

Notwithstanding, from our point of view, two avenues of study could be explored in order to justify this inefficiency.The first could be grounded on the Kahneman and Tversky (1979)

idea that the individuals value gains and losses asymmetrically and on the role of heurestics argued by Shiller (2000) to justify the non-efficient behaviour of individuals when betting for basketball games as they are being played. Under this stream of thought, it could be arguable that the market processes information asymmetrically and tends to deviate from the equilibrium value[79] obtained from the information element. As an example, we have compiled the fluctuations of our price and information elements in a reduced set of two key situations where no information is being incorporated in the information set[80] – during time-outs and during the intermission between one period and the other –, to assess whether or not the market fluctuates around the estimated winning probability:

Figure 14: Fluctuation of market odds *vs* estimated winning probability of the home team during time-outs

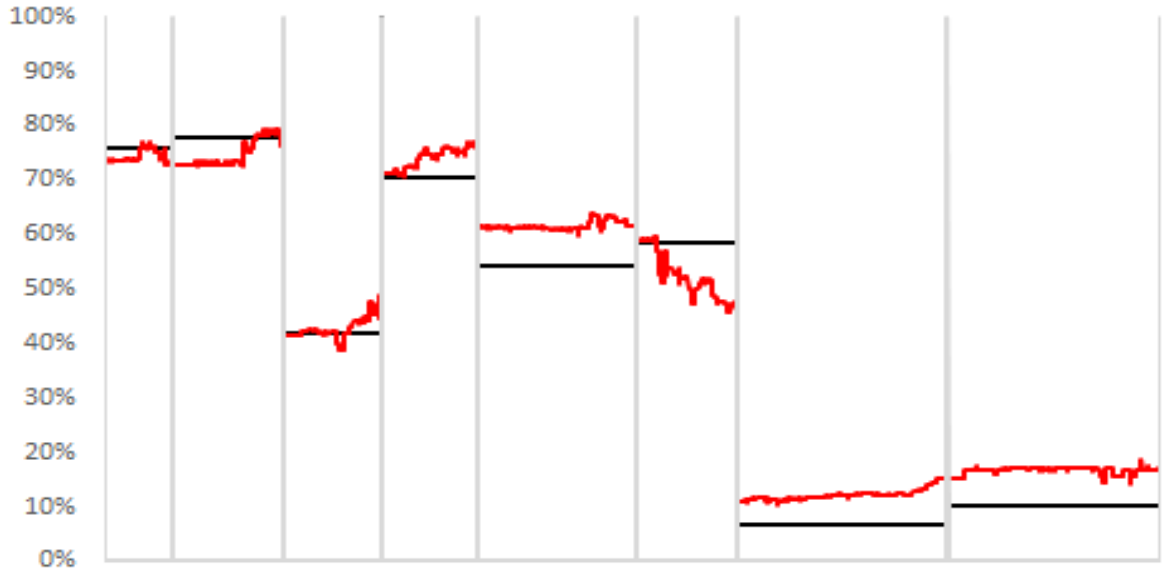



Under the efficient markets hypothesis discussed previously, it would be expectable that the market fluctuated, in average, in line with the estimated winning probability. However, as both figure 14 and 15 suggest, there are many cases in which both curves diverge, despite that no *new* informations is being generated. Although this is not necessarily the only cause behind the conclusion for our games, it certainly is one to take into account when future studies are conducted on the causes of in-play betting markets inefficiency.

[79] See, for examples, Shiller (2000)'s work on the rationality of market bubbles for a further exploration of the concept of deviation from equilibrium level.

[80] Or, at least, no information is being reported on the NBA's API as a relevant play event.

Figure 15: Fluctuation of market odds *vs* estimated winning probability of the home team during reduced set of intermissions between quarters




Another possibly pertinent avenue of study that could provide additional input onto the empirically observed market inefficiency is the market infrastructure in which the betting market is inserted, which, among other rigidities, might prevent bettors from instantly placing their bets on the betting exchange market. This phenomena is potentially troublesome since it may prevent the bettors from acting swiftly, thus possibly delaying their reaction to the introduction of a new element in the information set and arguably introducing an inefficiency in the market, as the aforementioned preconditions for market efficiency are somewhat weakened. As an example of this rigidity, consider that Betfair's betting exchange market is currently forcing English premier league (football) bettors to wait for a period of 5 seconds between the moment the bettor places the bet and the moment it is placed on the market to be corresponded,[81] which might be seen and studied as a rigidity measure potentially triggering market inefficiency.

Apart from the aforementioned future research hypothesis on the causes of the market inefficiency uncovered for the games considered in this study, one other very interesting analysis that can be derived from it is the exploration of the existence of profitable long-run

[81] In line with www.betangel.com/forum/viewtopic.php?f=6t=11148, consulted on the 10th of September of 2016.

betting strategies. In fact, the very definition of market efficiency in Fama (1970) would warrant that if the market is deemed inefficient then it could hold some particular trading strategies as profitable on the long-run. To that extent, one very simple strategy that could be designed and explored could revolve around the average deviation between the market implied winning probabilities of the home team and the theoretical model's estimates for the same reality. The computation of this deviation would, in theory, render the curve towards which the market curve would converge in relation to the estimated winning probabilities, thus precluding its movement throughout the game and possibly uncovering trading moments where a surplus can be extracted from the market. Taking as an example the four games considered in this study, we have computed the aforementioned average deviation between the market curve and the theoretical estimation curve, which has equated to a net positive difference of 1,73 percentage points. Therefore, by adding this difference to the results of our model, we obtain an adjusted curve which reflects the sample's average deviation between the price and the information element, which would translate the point towards which the market curve will, in average, expectedly converge.

In this sense, one can use this adjusted curve as a reference for the study of the existence of long term profitable trading strategies. For example, when the market curve is above the adjusted curve, there could be an opportunity to bet against the market – *lay* the odd on the market – and then bet in favour of the market odd – *back* the odd on the market – once it has converged to the adjusted curve.[82] This would effectively allow for a profit to be made, as the expected variation of the market odd towards the adjusted curve would render the trade profitable. Take as an example the *lay-back* strategy area highlighted in yellow in figure 16. It represents a period of time in which it seems that the market odd is unadjusted in relation to its average deviation against our estimation of the home team's winning probability. In this sense, by laying this market odd, the bettors are seemingly able to extract profit from the market with limited risk, thus meeting the concept of market inefficiency.

Conversely, for moments where the market odd is below the adjusted curve, the inverse strategy (*back-lay*) might also possibly yield long term net profits. This would imply that, for these specific moments, the bettors would bet in favour (*back*) of the market odd and bet

[82]For a broad explanation on the *back-lay* and *lay-back* trading strategies, see Rebelo (2012).

against it once it has converged to the adjusted curve, thus capitalizing on the odd fluctuation with limited risk. As an example, consider figure 17 which represents the possible moments where this *back-lay* strategy could be applied on game 1 of the considered games. As expected, in moments where the market odd is below the adjusted curve, it swiftly converges to its average deviation against the estimated winning probabilities, thus enabling the identified strategy.

Although we could identify in these figures particular moments where these strategies could prove profitable, it would be naive to assume that this judgement would hold for every imbalance between the aforementioned curves in any NBA game. In fact, to rule whether these strategies can prove to be consistently profitable and to assess their potential returns, one would need a broader set of testing games, throughout several NBA seasons, in order to prove the consistency of both the market inefficiencies and of the extracted long-term net positive returns. In this sense, this assessment can be explored in future studies alongside the possible causes of the identified in-play NBA betting markets inefficiency, in line with the two hypothesized causes and the strategies underlined in this study.

Figure 16: Game 3: Possible *lay-back* strategy

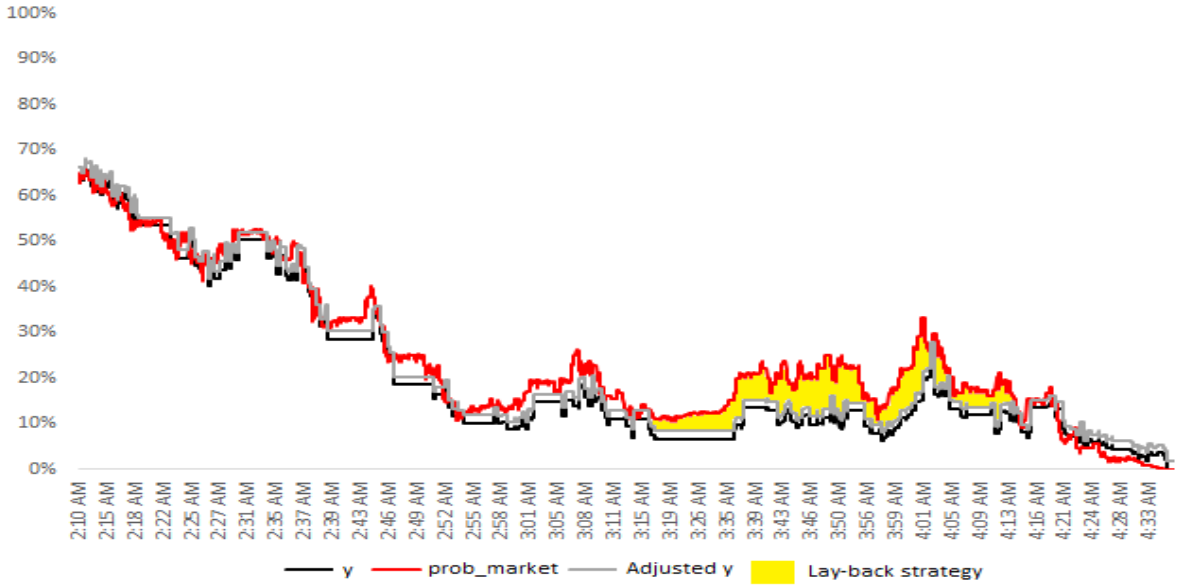
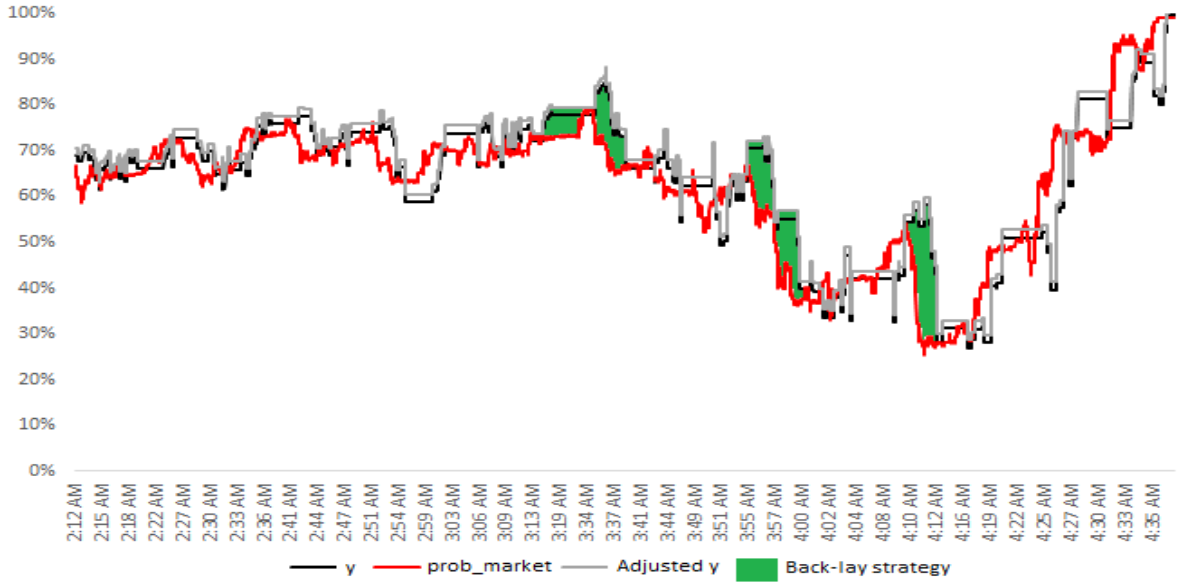



Figure 17: Game 1: Possible *back-lay* strategy



# 5 Conclusions

The purpose of this study is setting forth a first approach to the testing of the efficiency of sports betting markets, as the inherent game is being played. In this sense, the novelty of this endeavour was precisely shifting the pre-game odds *versus* actual outcomes scope that many of the surveyed studies employed, in order to understand if the markets resemble the same efficiency behaviour as the game is being played and, if possible, evaluate the existence of profitable long-term betting strategies.

To attain this goal, the first step we took was thoroughly assessing the relevant literature on market efficiency and on empirical studies performed to different sports, including basketball. In this process, Fama (1970)'s definition of market efficiency surfaced as the most consensual theoretical framing to the efficiency problem. Although it is not exempt of the identified criticisms, we took Fama (1970)'s preposition of prices *fully reflecting* all available information and its expected return model as cornerstone concepts for our empirical testing, in line with the vast majority of the literature we have surveyed. Concurrently, the empirical studies on sports betting literature we considered, for the game of basketball and other sports, have provided several key inputs that allowed the construction of our testing framework. On the one hand, we were able to identify some particular forms of widely documented market inefficiencies (*e.g.* the favourite/longshot bias in horse racing (Gramm and Owens, 2005; Asch and Quandt, 1990; Ali, 1977; Thaller and Ziemba, 1988)) and some factors potentially inducing market efficiency, namely, *inter alia*, the betting volume, the number of participants and the notoriety of the event (Busche and Walls, 2000; Gramm and Owens, 2005; Johnson et al., 2010). On the other hand, we uncovered several possible ways of modeling the game of basketball, which were particularly helpful when deciding how to construct the information element of our testing framework.

Bearing in mind the considerations extracted from this preliminary step, we opted to perform an efficiency test to the betting markets of 4 games of the NBA Finals, as the games occurred, by checking whether Fama (1970)'s preposition of prices *fully reflecting* all available information was verified. In this framework, the choice of these particular four games was grounded on two important factors, which, according to the literature surveyed, helped to

minimize the potential bias that the game choice could introduce. Firstly, these games are widely regarded as the ones with more attention throughout the NBA season (Karp, 2016; Pallota, 2016) and involve two relatively balanced teams playing repeatedly 2 times on each team's home floor, hence taking advantage of the benefits of the repetitiveness of basketball games (Schnytzer and Weinberg, 2004) and of the increment in the above identified efficiency factors for this restrict set of games *versus* the standard regular season ones. Secondly, taking into account the severe constraints we faced to access market data, these games comprised a very limited set of games we were allowed to access, hence further justifying the choice for the aforementioned 4 NBA Finals games, in accordance with the previous argument.

In the spirit of Fama (1970)'s preposition, we gathered a price and an information elements to be included in the testing framework. While the latter simply considered the market fluctuations of the home team winning odds throughout the selected four games, the former was much more complicated to derive, as the *actual* winning probabilities of the home team are not directly observable. Therefore, we constructed a model that allowed us to estimate the theoretical winning probability of the home team on a play by play basis, by using a dataset containing the moneyline odds and the plays of nearly all games between the 2007-2008 and the 2014-2015 NBA season. By regressing this model through a probit and a logistic (logit) regression, we concluded that the most suitable method for our endeavour is the logistic regression of the binary home team victory on a set of game-related key variables (elapsed time, moneyline odd and margin of the home) and some additional binary variables indicating which play is occurring, while taking into account relevant non-linear relations between the regressors and the dependent variable.

After the information and the price element were duly defined and tested for, the natural step we took was matching them into the same time frame, in order to enable the near second by second *live* analysis we strived for. In this process, we faced a big hurdle: the unavailability of the second at which the plays registered in the NBA's API occurred, which we overcame by manually accounting for the real time elapsed between each play, taking as reference the hour and minute inscribed in both elements of the testing framework. Under this setting, the matching of both datasets quickly portrayed the fact that both elements appeared to follow the same path, despite that they did it with some punctual differences possibly arising

due to market inefficiencies or due to very specific punctual cases not contemplated in the information element.

Although the former caveat was not that impactful on our conclusions, the latter is of quintessential importance, since if we managed to prove that the market tends to react to changes in the information set in a way that it deviates from its estimated implications, then we would be able to conclude for market inefficiency. In this spirit, we opted to adapt the efficiency testing framework employed in Schnytzer and Weinberg (2004) and Zuber et al. (1985) to our purposes, thus regressing the difference between the information element and the price element on the price element itself. Under this setting, we concluded for market inefficiency of the home team winning odds for all games combined and for each game individually, as we rejected that the intercept and the parameter associated to the price element are jointly different from zero.

This preposition is the key conclusion of our study: we provide empirical evidence on the inefficiency of in-play betting markets, thus bridging the identified gap in the literature and producing findings which are interestingly different in relation to the pre-game analysis paradigm where, although some particular forms of inefficiency were found, many of the surveyed studies held the markets under analysis as efficient.

Even though these conclusions are quite strong, it is of the essence to underline that we can only assure their validity for the surveyed games alone. Indeed, the testing of these results in a broader set of games would constitute a first avenue of research that can follow this study, in order to assess the consistency of the inefficiency verified and its possible patterns throughout the games and the season/s. In this sense, a critical measure that would not only enable the assessment of a wider set of games but that would also deeply foster the quality of future studies, is the introduction, by the system owner, of the real-time second at which the plays, registered in the NBA's API, take place, thus enabling the ideal 1 to 1 match that is, for now, impossible.

In light of the major conclusions identified, we must underline the major drawback of the efficiency testing framework we have chosen: the inability to identify which factor is inducing the empirically observed behaviour. Notwithstanding, from our point of view, this behaviour might possibly be happening due to the bounded rationality of individuals and

their asymmetric valuation of gains and losses, in line with Kahneman and Tversky (1979) and Shiller (2000), or due to betting exchange markets infrastructures, both of which might prevent the verification of the preconditions for efficiency, as laid down by Fama (1970). However, the ruling on these specific reasons – or on any other – for the empirically observed in-play betting markets' inefficiency demands a broader set of games (and respective in-play bets) and a more complex efficiency testing framework – both of which can be much easily tackled in future studies with the implementation of the measure set forth in the previous paragraph.

Finally, having concluded for market inefficiency, it would be expectable, under Fama (1970)'s expected return model, that profitable long-term betting strategies could be identified. In this sense, we have computed the average difference between the price element and the information element, thus obtaining the curve towards which the market would, in average, converge towards. Taking this adjusted curve as reference, we identified particular moments in game 1 and game 3 of the considered NBA Finals where trading strategies – *back-lay* and *lay-back*, respectively – could potentially yield profits, hence taking advantage of the progressive convergence of the price element towards the adjusted curve. Nevertheless, the consequences of inefficiency would, in light of the expected return model, imply that these strategies should prove profitable over the long run, which is something that we are not able to adjudge, due the limited span of games tested for. In this sense, although the identification of these possibly profitable mechanisms is of quintessential importance and duly spotted in this study, future studies may tackle the assessment of the consistency of these strategies over the long-run, by using a broader set of games, alongside the exploration and validation of the aforementioned possible causes of in-play betting market inefficiency.

# Bibliography

Ali, M. M. (1977). Probability and utility estimates for racetrack bettors. *Journal of Political Economy*, 85:803–815.

Asch, P. and Quandt, R. E. (1990). Risk love. *Journal of Economic Education*, 21:422–426.

Baryla Jr, E. A., Borghesi, R. A., Dare, W. H., and Dennis, S. A. (2007). Learning, price formation and the early season bias in the nba. *Finance Research Letters*, 4:155–164.

Beaver, W. H. (1981). Market efficiency. *The accounting review*, 56(1):23–37.

Beuoy, M. (2015). Updated nba win probability calculator. www.inpredictable.com/2015/02/updated-nba-win-probability-calculator.html.

Boettke, P. J. (2010). What happened to "efficient markets". *The independent review*, 14(2):363–375.

Breusch, T. and Pagan, A. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294.

Bruce, A. and Marginson, D. (2013). Power, not fear: A collusion-based account of betting market inefficiency. *International Journal of the Economics of Business*, 21(1):77–97.

Busche, K. and Walls, W. D. (2000). Decision costs and betting market efficiency. *Rationality and Society*, 12(4):477–492.

Busche, K. and Walls, W. D. (2001). Breakage and betting market efficiency: evidence from the horse track. *Applied Economics Letters*, 8:601–604.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

Cortis, D. (2015). Expected balues and variances in bookmaker payouts: a theoretical approach towards setting limits on odds. *The Journal of Prediction Markets*, 9(1):1–14.

Dare, W. H. and Holland, S. (2004). Efficiency in the nfl betting market: modifying and consolidating research methods. *Applied Economics*, 36:9–15.

Entine, O. A. and Small, D. S. (2008). The role of rest in the nba home-court advantage. *Journal of Quantitative Analysis in Sports*, 4(2):1–10.

Fama, E. F. (1970). Efficient capital markets: a review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Gandar, J., Zuber, R., O'Brien, T., and Russo, B. (1988). Testing rationality in the point spread betting market. *Journal of Finance*, 43:995–1007.

Gramm, M. and McKinney, N. (2009). The effect of late money on betting market efficiency. *Applied Economics Letters*, 16:369–372.

Gramm, M. and Owens, D. H. (2005). Determinants of betting market efficiency. *Applied Econmics Letters*, 12:181–185.

Hosmer Jr, D., Lemeshow, S., and Sturdivant, R. (2013). *Applied Logistic Regression*. Wiley, New Jersey, 3rd edition.

Jarque, C. and Bera, A. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(8):255–259.

Johnson, J., Bruce, A., and Yu, J. (2010). The ordinal efficiency of betting markets: an exploded logit approach. *Applied Economics*, 42:3703–3709.

Jones, M. B. (2007). Home advantage in the nba as a game-long process. *Journal of Quantitative Analysis in Sports*, 3(4):1–16.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–91.

Karp, A. (2016). Nba media partners rebound with gains across the board for '15-16 game broadcasts. sportsbusinessdaily.com.

Kenter, F. H. J. (2015). An analysis of the basketball endgame: When to foul when trailing and leading. In *MIT Sloan Sports Analytics Conference*.

Levitt, S. (2004). Why are gambling markets organised so differently from financial markets? *Economic Journal*, 114:223–246.

Malkiel, B. (1973). *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*. W. W. Norton.

Marshall, A. (1890). *The Principles of Economics*. McMaster University Archive for the History of Economic Thought.

McFadden, D. (1974). *Frontiers of Econometrics*, chapter 3: Conditional logit analysis of qualitative choice behavior, pages 105–142. Academic. in P. Zarembka(Ed.).

NBA (2008). Nba rules history. nba.com.

Pallota, F. (2016). Nba finals game 7 audience tops 30 million, biggest in 18 years. money.cnn.com.

Parker, R. J. (2010). Modeling basketball's points per possession with application to predicting the outcome of college basketball games.

Paul, R. J., Weinbach, A. P., and Wilson, M. (2004). Efficient markets, fair bets, and profitability in nba totals 1995–96 to 2001–02". *The Quarterly Review of Economics and Finance*, 44:624–632.

Peng, C.-Y. J. and So, T.-S. H. (2002). Logistic regression analysis and reporting: A primer. *Understanding statistics*, 1(1):31–70.

Rebelo, P. (2012). *Ganhar com as Apostas Desportivas*. Marcador.

Sauer, R., Brajer, V., Ferris, S., and Marr, M. (1988). Hold your bets: Another look at the efficiency of the betting market for nfl games. *Journal of Political Economy*, 96(1):206–213.

Schnytzer, A. and Weinberg, G. (2004). Is the nba betting market efficient. In *The economics and management of mega athletic events: Olympic games, professional sports, and other essays*.

Shiller, R. J. (2000). *Irrational exuberance*. Princeton University Press.

Shin, S. (1993). Measuring the incidence of insider trading in a market for state-contingent claims. *Economic Journal*, 103:1141–1153.

Smith, M. A., Paton, D., and Vaughan Williams, L. (2009). Do bookmakers possess superior skills to bettors in predicting outcomes? *Journal of Economic Behavior and Organization*, 71:539–549.

StataCorp (2013). *Stata 13 Base Reference Manual*. Stata Press, College Station, Texas.

Thaller, R. H. and Ziemba, W. T. (1988). Parimutuel betting markets: Racetracks and lotteries. *Journal of Economic Perspectives*, 2:161–174.

Štrumbelj, E. and Vračar, P. (2012). Simulating a basketball match with a homogeneous markov model and forecasting the outcome. *Simulating a Basketball Match with a Homogeneous Markov Model and Forecasting the Outcome*, 28(2):532–542.

Vale, S. (2013). Classification of types of big data. unece.org.

Vaughan Williams, L. and Patton, D. (1988). Do betting costs explain betting biases? *Applied Economics Letters*, 5:333–335.

Vorhies, W. (2014). How many "v's" in big data? the characteristics that define big data. datasciencecentral.com.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(5):817–838.

Wooldridge, J. M. (2009). *Introductory econometric – a modern approach*. South-Western Cengage Learning, 4th edition.

Zuber, R., Gandar, J., and Bowers, B. (1985). Beating the spread: Testing the efficiency of the gambling market for national football league games. *Journal of Political Economy*, 93(4):800–806.

# 6 Appendix 1

## 6.1 Python routine for the extraction of the *baseline* dataset

To extract the raw data used in our *baseline* dataset, we have firstly used Grant Fiddyment's repository of NBA python scripts[83] to compile a list of the games we wanted to consider in our dataset and their respective identification code. Having compiled this list – to which we have named "games.csv" – through a routine stored in this repository, we have crafted, with important help,[84] a Python script which queries the NBA's API for all the plays of the games considered in the aforementioned list and converts the .json input into a user-friendly .csv output. The resulting code is as follows:

```
from __future__ import print_function
import json
import csv
import requests
u_a = "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/48.0.2564.82 Safari/537.36"
url_pattern = "http://stats.nba.com/stats/playbyplayv2?GameID=%(GameID)s&Start
Period=%(StartPeriod)s&EndPeriod=%(EndPeriod)s&tabView=%(tabView)s"

def write_csv(game_id, resultSet):
    fn = resultSet['name'] + '_' + str(game_id) + '.csv'
    if resultSet['name'] not in ['PlayByPlay', 'PlayBlahBlah']:
        return
    with open(fn, 'w') as fout:
        csv_file = csv.writer(fout, quotechar='"')
        csv_file.writerow(resultSet['headers'])
        for rowSet in resultSet['rowSet']:
            csv_file.writerow(rowSet)
def process_game_id(game_id, tabView='playbyplay',
                    start_period='0', end_period='0'):
    url_parms = {
        'GameID': game_id,
        'StartPeriod': start_period,
        'EndPeriod': end_period,
        'tabView': tabView,
    }
    r = requests.get((url_pattern % url_parms), headers={"USER-AGENT":u_a})
    if r.status_code == requests.codes.ok:
        data = json.loads(r.text)
        for rset in data['resultSets']:
            write_csv(url_parms['GameID'], rset)
    else:
        r.raise_for_status()
if __name__ == '__main__':
    #
    # assuming that the 'games.csv' file contains all Game_IDs ...
    #
    with open('games.csv', 'r') as f:
        csv_reader = csv.reader(f, delimiter=',')
```

[83] See github.com/gmf05/nba.

[84] See the important contributes of user "MaxU" in stackoverflow.com/questions/35444430/converting-nba-play-by-play-specific-json-to-csv/35444833?noredirect=1comment58616761_35444833

```
        for row in csv_reader:
            process_game_id(row[0])
```

## 6.2 Stata code

Following the construction of the *baseline* dataset, we have constructed the following Stata code to: *(i)* construct the *play by play* dataset, *(ii)* estimate the information element of the testing framework, *(iii)* assess the quality of the aforementioned element, *(iv)* merge the price element of the testing framework to its information element, *(v)* estimate the designed testing equation, and *(vi)* test for in-play betting markets efficiency. This code is as follows:

```
clear
set excelxlsxlargefile on
cap log close _all
cd "G:\Stata\output\"
*1st STEP: Gathering all play by play games. This is done by importing each
season individually
* Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\14\" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\14\/`file'"
 , clear
 save 14, replace
 restore
 append using 14, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 14.dta
save 14, replace
clear
*-------------------------------------------------------------------------------
*2013 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\13\" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\13\/`file'"
 , clear
 save 13, replace
 restore
 append using 13, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 13.dta
save 13, replace
clear
*-------------------------------------------------------------------------------
*2012 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\12\" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\12\/`file'"
 , clear
 save 12, replace
 restore
 append using 12, force
```

```
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 12.dta
save 12, replace
clear
*-------------------------------------------------------------------------------
*2011 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\11\" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\11\/`file'"
 , clear
 save 11, replace
 restore
 append using 11, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 11.dta
save 11, replace
clear
*-------------------------------------------------------------------------------
*2010 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\10\" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\10\/`file'"
 , clear
 save 10, replace
 restore
 append using 10, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 10.dta
save 10, replace
clear
*-------------------------------------------------------------------------------
*2009 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\09\" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\09\/`file'"
 , clear
 save 09, replace
 restore
 append using 09, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 09.dta
save 09, replace
clear
*-------------------------------------------------------------------------------
*2008 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\08\" files
"*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\08\/`file'"
 , clear
 save 08, replace
 restore
 append using 08, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
```

```
homedescription visitordescription score scoremargin
rm 08.dta
save 08, replace
clear
*-------------------------------------------------------------------------------
*2007 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\07\" files
  "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\07\/`file'"
 , clear
 save 07, replace
 restore
 append using 07, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 07.dta
save 07, replace
clear
*-------------------------------------------------------------------------------
*2006 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\06\" files
  "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\06\/`file'"
 , clear
 save 06, replace
 restore
 append using 06, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 06.dta
save 06, replace
clear
*-------------------------------------------------------------------------------
*2005 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\05\" files
  "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\05\/`file'"
 , clear
 save 05, replace
 restore
 append using 05, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 05.dta
save 05, replace
clear
*-------------------------------------------------------------------------------
*2004 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\04\" files
  "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\04\/`file'"
 , clear
 save 04, replace
 restore
 append using 04, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 04.dta
```

```
save 04, replace
clear
*-------------------------------------------------------------------------------
*2003 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\03\" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\03\/`file'"
 , clear
 save 03, replace
 restore
 append using 03, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 03.dta
save 03, replace
clear
*-------------------------------------------------------------------------------
*2002 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\02\" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\02\/`file'"
 , clear
 save 02, replace
 restore
 append using 02, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 02.dta
save 02, replace
clear
*-------------------------------------------------------------------------------
*2001 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\01\" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\01\/`file'"
 , clear
 save 01, replace
 restore
 append using 01, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 01.dta
save 01, replace
clear
*-------------------------------------------------------------------------------
*2000 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\00" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\00\/`file'"
 , clear
 save 00, replace
 restore
 append using 00, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 00.dta
save 00, replace
clear
```

```
*-------------------------------------------------------------------------------
*1999 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\99" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\99\/`file'"
 , clear
 save 99, replace
 restore
 append using 99, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 99.dta
save 99, replace
clear
*-------------------------------------------------------------------------------
*1998 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\98" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\98\/`file'"
 , clear
 save 98, replace
 restore
 append using 98, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 98.dta
save 98, replace
clear
*-------------------------------------------------------------------------------
*1997 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\97" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\97\/`file'"
 , clear
 save 97, replace
 restore
 append using 97, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 97.dta
save 97, replace
clear
*-------------------------------------------------------------------------------
*1996 Season
local csvfiles: dir "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\96" files
 "*.csv"
foreach file of local csvfiles {
 preserve
 import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\Json_PBP\96\/`file'"
 , clear
 save 96, replace
 restore
 append using 96, force
}
keep game_id eventnum eventmsgtype period wctimestring pctimestring
homedescription visitordescription score scoremargin
rm 96.dta
save 96, replace
clear
append using 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14, force
keep game_id eventnum eventmsgtype period wctimestring pctimestring
```

```
 homedescription visitordescription score scoremargin
save PBP_FINAL, replace
clear
cd "G:\Stata\output\"
import delimited "G:\Drive\Faculdade\Mestrado\TESE\Data\moneyline\moneyline
 .csv"
rename v1 game_id
rename v2 game_id_date
rename v3 home_ML
generate home_ML_odd = cond(home_ML<0, (home_ML*-1)/     (home_ML*-1+100),
 100/(home_ML+100))
generate home_ML_decimal = home_ML_odd^-1
save moneyline.dta, replace
merge m:m game_id using moneyline.dta
recast int _merge
keep if _merge ==3
drop _merge
drop if home_ML_odd==.
*2nd step: Draw descriptive statistics of complete database
*2.1 Clear collection errors
drop if homedescription=="" & visitordescription=="" & eventmsgtype ==18
*2.2 Generate the clock
sort game_id eventnum
generate period_clock = substr(pctimestring,1,5)
generate double period2=clock(period_clock,"ms")
generate time_remaining = cond(period<2 & eventmsgtype>11 & eventmsgtype<13,
 clock("48:00","ms"),cond(period<3 & eventmsgtype>11 & eventmsgtype<13,
 clock("36:00","ms"),cond(period<4 & period>2 & eventmsgtype>11 &
 eventmsgtype<13, clock("24:00","ms"),cond(period>3 & eventmsgtype>11 &
 eventmsgtype<13, period2,0))))
replace time_remaining = time_remaining[_n-1] - period2[_n-1] + period2[_n]
if time_remaining<=0
format period2 %tcMM:SS
replace time_remaining = time_remaining/1000
generate time_elapsed =0
replace time_elapsed = abs(time_remaining-2880)
replace time_elapsed =2880 if period ==5 & eventmsgtype==12
replace time_elapsed =3180 if period ==6 & eventmsgtype==12
replace time_elapsed =3480 if period ==7 & eventmsgtype==12
replace time_elapsed =3780 if period ==8 & eventmsgtype==12
replace time_elapsed =4080 if period ==9 & eventmsgtype==12
replace time_elapsed = time_elapsed[_n-1]+abs(time_remaining[_n-1]-
 time_remaining[_n]) if period>=5 & game_id[_n-1]==game_id[_n] &
 (eventmsgtype<11 | eventmsgtype==13)
drop period2
*2.3 Generate margin of home team using scoremargin collumn
generate margin_home = cond(eventmsgtype==12 & period ==1, 0,
 cond(scoremargin=="TIE", 0, real(scoremargin)))
split score, p("-")
destring score1 score2, replace
generate z = score2-score1
generate zz = cond(missing(z),0,1)
replace margin_home = z if zz==1
replace margin_home = margin_home[_n-1] if missing(margin_home)
drop zz z
*2.4 Debugging strange observations (eventnum after game ends)
destring game_id, replace
sort game_id eventnum
generate a = cond(eventmsgtype==13 & time_remaining==0, 1,0)
replace a =0 if period[_n+1]>period | period[_n+20]>period
generate b = cond(a[_n-1]==1 & game_id==game_id[_n-1],1,0)
replace b = 1 if b[_n-1]==1 & game_id==game_id[_n-1]
drop if b==1
drop b a
```

```
save PBP_Final, replace
*2.5 Get the result of each game and generate descriptive dataset
generate temp = cond(game_id[_n+1] > game_id,1,0)
egen max = max(game_id)
keep if temp ==1
generate home_win = cond(margin_home>0,1,0)
generate final_score = margin_home
save descriptive, replace
*This will yield relative win% of home team across years
tostring game_id, replace
generate season = substr( game_id, 2,2)
generate phase = substr(game_id,1,1)
replace phase ="Regular" if phase=="2"
replace phase ="Playoff" if phase=="4"
egen season_full = concat( season phase)
log on
tabulate home_win season_full
log off
*3rd Step: Mining the raw information gathered through the web scrapper
 and merging to moneyline odds
clear
use PBP_Final
merge m:m game_id using descriptive
sort game_id eventnum
drop temp _merge max
save PBP_Final, replace
*3.1 Start mining the data - eliminating non relevant observations
 and blank observations
drop if eventmsgtype ==8
drop if eventmsgtype ==9 & homedescription =="" & visitordescription==""
*3.2 Describe the plays through the event message type
*Identify home team blank observations
generate home_play= cond(visitordescription =="", eventmsgtype,0)
replace home_play =0 if homedescription==""
*introduct home block
replace home_play = 15 if strpos(homedescription, "BLOCK")>0
*introduct home steal
replace home_play = 16 if strpos(homedescription, "STEAL")>0
*Identify away team blank observations
generate away_play= cond(homedescription =="", eventmsgtype,0)
replace away_play =0 if visitordescription==""
*introduct away block
replace away_play = 15 if strpos(visitordescription, "BLOCK")>0
*introduct away steal
replace away_play = 16 if strpos(visitordescription, "STEAL")>0
*FGMiss in case the opponent blocks
replace away_play=2 if home_play==15
replace home_play=2 if away_play==15
*TO in case the opponent steals the ball
replace away_play=5 if home_play==16
replace home_play=5 if away_play==16
*drop team rebounds
generate home_team_REB = cond(strpos(homedescription, "Rebound")>0 &
 strpos(homedescription[_n-1], "Free Throw")>0 & strpos(homedescription[_n+1],
"Free Throw")>0, 1,0)
generate away_team_REB = cond(strpos(visitordescription, "Rebound")>0 &
 strpos(visitordescription[_n-1], "Free Throw")>0 &
 strpos(visitordescription[_n+1], "Free Throw")>0, 1,0)
drop if home_team_REB ==1 | away_team_REB ==1
drop home_team_REB away_team_REB
*Describe detailed plays
gen home_3PT = cond(strpos(homedescription, "3PT")>0 & home_play==1, 1,0)
gen home_2PT = cond(home_play==1 & home_3PT ==0,1,0)
gen home_FGMiss = cond(home_play==2,1,0)
```

```
gen home_FTM = cond(strpos(homedescription, "MISS")==0 & home_play==3, 1,0)
gen home_FTMiss = cond(strpos(homedescription, "MISS")>0 & home_play==3, 1,0)
gen home_Dreb = cond(home_play==4 & away_play[_n-1]==2,1,0)
gen home_Oreb = cond(home_play==4 & home_play[_n-1]==2,1,0)
gen home_TO = cond(home_play == 5,1,0)
gen home_foul = cond(home_play ==6,1,0)
gen home_block = cond(home_play==15,1,0)
gen home_steal = cond(home_play==16,1,0)
gen away_3PT = cond(strpos(visitordescription, "3PT")>0 & away_play==1, 1,0)
gen away_2PT = cond(away_play==1 & away_3PT ==0,1,0)
gen away_FGMiss = cond(away_play==2,1,0)
gen away_FTM = cond(strpos(visitordescription, "MISS")==0 & away_play==3, 1,0)
gen away_FTMiss=cond(strpos(visitordescription,"MISS")>0 & away_play==3, 1,0)
gen away_Dreb = cond(away_play==4 & home_play[_n-1]==2,1,0)
gen away_Oreb = cond(away_play==4 & away_play[_n-1]==2,1,0)
gen away_TO = cond(away_play == 5,1,0)
gen away_foul = cond(away_play ==6,1,0)
gen away_block = cond(away_play==15,1,0)
gen away_steal = cond(away_play==16,1,0)
*introduct crossed terms
gen time_margin = margin_home*time_remaining
gen time_margin2 = (time_margin)^2
gen ML_margin= home_ML_odd* margin_home
gen ML_time = home_ML_odd * time_remaining
gen time_elapsed2 = (time_elapsed)^2
gen time_elapsed3 = (time_elapsed)^3
gen elapsed_margin = margin_home*time_elapsed
gen elapsed_margin2 = (elapsed_margin)^2
gen elapsed_margin3 = (elapsed_margin)^3
gen ML_elapsed = home_ML_odd*time_elapsed
gen home_ML_odd2=home_ML_odd^2
*Introduct clutch factor: last 2 minutes of gameplay
gen clutch = cond(1,time_elapsed >2760,0)
gen clutch_margin = clutch*margin_home
gen clutch_margin2 = clutch*margin_home*margin_home
gen clutch_elapsed = clutch*time_elapsed
gen clutch_elapsed2 = clutch*time_elapsed2
save PBP_Final, replace
*4. Specify and estimate the models
use descriptive.dta
generate random = runiform()
sort random
*selecionar 5% dos jogos aliatoriamente
generate insample = _n <= _N*0.05
drop if insample ==0
drop random
save random.dta, replace
use PBP_Final.dta
merge m:m game_id using random.dta
keep if insample==1
log on
lowess home_win home_ML_odd, logit bwidth(.8) adjust gen(y_home_ML_odd)
clear
use PBP_final.dta
drop score scoremargin home_ML_decimal final_score home_play away_play
home_3PT home_2PT home_FTM away_3PT away_2PT period_clock
sort game_id eventnum
logit home_win home_ML_odd home_ML_odd2 margin_home time_elapsed time_elapsed2
 elapsed_margin elapsed_margin2 home_Oreb home_Dreb home_TO home_foul away_Oreb
 away_Dreb away_TO away_foul clutch clutch_margin clutch_elapsed
 clutch_elapsed2, noconst vce(robust)
probit home_win home_ML_odd home_ML_odd2 margin_home time_elapsed time_elapsed2
 elapsed_margin elapsed_margin2 home_Oreb home_Dreb home_TO home_foul away_Oreb
```

```
 away_Dreb away_TO away_foul clutch clutch_margin clutch_elapsed
 clutch_elapsed2, noconst vce(robust)
*5. Assessing the quality of the models and mergint the testing elements
collin home_ML_odd margin_home time_elapsed home_Oreb home_Dreb home_TO
 home_foul away_Oreb away_Dreb away_TO away_foul clutch
estat class
estat gof
tostring eventnum, replace
save PBP_modelo.dta, replace
* NBA Finals game 1
keep if game_id==41300401
twoway (scatter y time_elapsed if game_id==41300401, msymbol(p) connect(l)
 yscale(range(0 1)) ylabel(#10) xline(720 1440 2160 2880 3600 4320))
cd "G:\Drive\Faculdade\Mestrado\TESE\Data\output Stata\"
graph export Game1_logit.png, replace
cd "G:\Stata\output\"
save game1.dta, replace
clear
import excel "G:\Drive\Faculdade\Mestrado\TESE\Data\Betfair data\bets_0041300401
 .xlsx", firstrow
rename eventid eventnum
merge m:m eventnum using game1.dta
sort Timestamp
drop if _merge==2
replace y= y[_n-1] if _merge==1
generate prob_market=1/lastpricematched
replace margin_home=margin_home[_n-1] if margin_home==.
save game1.dta, replace
clear
use PBP_modelo.dta
keep if game_id==41300402
* NBA Finals game 2
twoway scatter y time_elapsed if game_id==41300402, msymbol(p) connect(l)
 yscale(range(0 1)) ylabel(#10) xline(720 1440 2160 2880 3600 4320)
cd "G:\Drive\Faculdade\Mestrado\TESE\Data\output Stata\"
graph export Game2_logit.png, replace
cd "G:\Stata\output\"
save game2.dta, replace
clear
import excel "G:\Drive\Faculdade\Mestrado\TESE\Data\Betfair data\bets_0041300402
 .xlsx", firstrow
rename eventid eventnum
tostring eventnum, replace
merge m:m eventnum using game2.dta
sort Timestamp
drop if _merge==2
replace y= y[_n-1] if _merge==1
generate prob_market=1/lastpricematched
replace margin_home=margin_home[_n-1] if margin_home==.
save game2.dta, replace
clear
cd "G:\Stata\output\"
use PBP_modelo.dta
keep if game_id==41300403
* NBA Finals game 3
twoway scatter y time_elapsed if game_id==41300403, msymbol(p) connect(l)
 yscale(range(0 1)) ylabel(#10) xline(720 1440 2160 2880 3600 4320)
cd "G:\Drive\Faculdade\Mestrado\TESE\Data\output Stata\"
graph export Game3_logit.png, replace
cd "G:\Stata\output\"
save game3.dta, replace
clear
```

```
import excel "G:\Drive\Faculdade\Mestrado\TESE\Data\Betfair data\bets_0041300403
  .xlsx", firstrow
rename eventid eventnum
tostring eventnum, replace
merge m:m eventnum using game3.dta
sort Timestamp
drop if _merge==2
replace y= y[_n-1] if _merge==1
generate prob_market=1/lastpricematched
replace margin_home=margin_home[_n-1] if margin_home==.
save game3.dta, replace
clear
cd "G:\Stata\output\"
use PBP_modelo.dta
keep if game_id==41300404
* NBA Finals game 4
twoway scatter y time_elapsed if game_id==41300404, msymbol(p) connect(l)
  yscale(range(0 1)) ylabel(#10) xline(720 1440 2160 2880 3600 4320)
cd "G:\Drive\Faculdade\Mestrado\TESE\Data\output Stata\"
graph export Game4_logit.png, replace
cd "G:\Stata\output\"
save game4.dta, replace
clear
import excel "G:\Drive\Faculdade\Mestrado\TESE\Data\Betfair data\bets_0041300404
  .xlsx", firstrow
rename eventid eventnum
tostring eventnum, replace
merge m:m eventnum using game4.dta
sort Timestamp
drop if _merge==2
replace y= y[_n-1] if _merge==1
generate prob_market=1/lastpricematched
replace margin_home=margin_home[_n-1] if margin_home==.
save game4.dta, replace
clear
cd "G:\Stata\output\"
append using game1 game2 game3 game4
drop if y==. | temp==.
save PBP_modelo.dta, replace
*6. Assessing the properties of the estimation of the testing equation
 and performing the test
use PBP_modelo.dta
rename prob_market home_ML_odd
gen z = y - home_ML_odd
reg z home_ML_odd
predict z2, residuals
test (home_ML_odd=0)(_cons=0)
estat hettest
estat imtest, white
jb6 z2
reg z home_ML_odd, vce(robust)
test (home_ML_odd=0)(_cons=0)
log close _all
```