# Efficiency and Forecasting of Dynamic Prices in UK Online Horse Racing Markets

MSc. Quantitative Finance
Masters Thesis

Cass Business School

Mark Best

July $31^{st}$ 2009

## Abstract

Many papers have been dedicated to studying the efficiency of horse racing markets. Initial studies observed participants in these markets from a purely psychological point of view; studying people's ability to make judgements while faced with uncertainty. Later studies look at categorisation of horses using factor models or late money flow in attempt to out perform the market. Although statistically significant anomalies such as long shot bias have been found, few studies found them sufficiently large as to overcome market frictions. With the advent of online betting markets the microstructure of gambling markets has changed. Lower commissions and the introduction of lay bets now allow for trading strategies not previously possible. This paper is unique as it studies the efficiency of horse betting markets both from the point of view of predicting race outcomes as well as testing the efficiency of price dynamics in the run up to races. Finally the start prices are forecast using Engle's autoregressive conditional multinomial (ACM) and autoregressive conditional duration models (ACD). The results of the models forecasting performance is very promising.

*Keywords: Horse Racing, Efficient Markets, Favourite and Long Shot Bias, Ultra High Frequency, Autoregressive Conditional Multinomial (ACM) model, Autoregressive Conditional Duration (ACD) model*

# Contents

# 1 Background

## 1.1 Introduction

The efficient markets hypothesis is an important keystone in financial theory. Investors are assumed to be rational such that market prices are a true reflection of fundamental value and incorporate all known information. The main implication of this hypothesis is prices should be unpredictable and it should not be possible to profit from private information due to it all ready being incorporated into the price. There have been many arguments against the existence of efficient markets on the basis of empirical and theoretical studies. Some argue the computational complexity required to achieve prefect rationality are unfeasible[Bei06]. Grossman and Stiglitz argue that if information is costly there is no incentive for agents to participate in markets unless the profits of doing so at least cover the outlay of information costs[GS80]. Shiller notes that the levels of volatility in the S&P 500 is higher than can be explained if traders are in agreement about price [Shi86]. Shiller notes excess trading and volatility seems more likely to be explained by a disagreement between heterogeneous market agents.

The abolishment of national gambling monopolies and extensive deregulation has lead to huge growth in volumes transacted on online betting exchanges thus prompting a great deal of research into the efficiency of these markets. Recent changes in online betting markets allow bettors to both bet a horse will win (back) and bet a horse will lose (lay). This has allowed punters a greater deal of flexibility in terms of trading strategies, leading to larger volumes and tighter spreads. Studies into the informational efficiency of football [VDM07], tennis [FM07] and cricket [EU06] markets have repeatedly found betting markets are efficient. Studies find some evidence of statistically significant mis-pricings such as favourite long shot bias but these rarely exceed market frictions i.e. bid-ask spread and commission rates.

This study is unique in its approach to testing horse racing market efficiency. Horse racing studies to date exclusively focus on prices as efficient estimators of race outcomes. Given recent changes in market microstructure is it possible to generate profits irrespective of race outcome. Thus for betting markets to be truly efficient the price discovery process must also be efficient at all times prior and during the race. This study will test the efficiency in the classical sense as well as testing the efficiency of the price discovery process.

This study is comprised of five parts. Section 2 is a literature review of papers regarding tests of market efficiency and those specific to horse race modelling and prediction. Section 3 discusses technical details of collecting and cleaning the data set used in the study. Section 4 tests the classical efficiency of market prices as efficient predictors of market outcome. Section 5 discusses the calibration and forecasting of an ultra-high frequency ACM-ACD model used to forecast start prices. Finally section 6 is left for results and conclusions.

The study was motivated by extensive work previously done by the author in market making and arbitrage strategies between the main UK online horse racing markets Betfair and Betdaq. Previous academic studies have found that market starting prices are efficient predictors of the outcome of a race. Although no study has been done into the dynamics of the price discovery process in the run up to the race it can be implied that it should remain within statistical boundaries implied by the handicapping. Price series examined up to thirty minute before a race starts seem to exhibit bias, trending and mean reversion behaviours.

The study's objectives are to identify stylised features of the data and use them to better forecast prices than with the price alone, to test whether classical statistical anomalies exist in dynamic prices and to test the forecasting performance of ACM-ACD models.

The intended audience are, exchange markets, market makers, gamblers. Efficiency of prices may seem innocuous but deviations from fundamental value suggest private information may be available which is not fully incorporated into the price. Betfair have agreements with horse racing authorities to report evidence of irregular betting activities. Deviations from common trends may be an indication of malicious activity.

Markets makers may include both exchange members who offer price to individuals as well as high street and on track book makers who hedge their risk in Betfair. There are well known mis-pricings in horse racing odds. These have implications in the levels that should be offered to punters.

Participation in betting markets is still comparatively low to those of financial markets. It is possible given reduced commission and non taxable winnings for investors to generate profits from relatively simple strategies in online betting markets. Another advantage to this form of investment is they are uncorrelated to other markets and may provide portfolio diversification benefits.

# 2 Literature Review

## 2.1 Market Efficiency

Market efficiency is an important keystone in financial theory due to the implication that market prices represent the true fundamental asset value. The studies in this section deal with the fundamental concepts of market and informational efficiency, how it may be tested and the pitfalls faced by empirical studies.

### Efficient Capital Markets

One of the first papers to comprehensively discuss empirical studies and theory of efficient markets is Fama's Efficient Capital Markets [Fam70]. It is initially important to discuss the definition, theory and empirical tests of market efficiency so as they can be applied to the study of betting markets. Rational investors are assumed to know the true fundamental value of assets such that market prices reflect all available information.

There are three standard classifications of market efficiency:

1. Weak Form: The information set only contains historical prices
2. Semi-Strong Form: Prices incorporate all publicly known information
3. Strong Form: Prices incorporate all privately and publicly known information.

It is difficult in many markets to test for semi-strong or strong forms of market efficiency because the set of public and private information which is significant to the prices is subjective (i.e. trainers experience, jockey etc.). Some semi strong tests of betting market efficiency have been completed however there is no consensus on the factors used in a pricing models.

The majority of empirical studies noted by Fama include tests of the weak form of market efficiency using fair games models. Before this can be explained further we need to add some meat to the bones of the underlying price process.

All expected returns theories are of the form,

$$E(\tilde{p}_{j,t+1}|\Phi_t) = [1 + E(\tilde{r}_{j,t+1}|\Phi_t)]p_{j,t+1} \tag{1}$$

By rearranging the equation in terms of actual and expected probability

$$x_{j,t+1} = p_{j,t+1} - E(\tilde{p}_{j,t+1}|\Phi_t) \tag{2}$$

It is possible to conclude from this that in a fair game the expected returns are zero

$$E(\tilde{x}_{j,t+1}|\Phi_t) = 0 \tag{3}$$

The fair game model has many useful implications as the null hypothesis of expected returns being zero is easily testable. These models have also been extended into random walk models where the tests assume price increments should be independent so as to ensure they are unpredictable.

### Anomalies and Market Efficiency

Since Fama's paper there has been a large amount of empirical analysis both for and against the efficient markets hypothesis. Many studies have investigated specific market anomalies that seem to violate the theory.

In Shwert's 2002 paper anomalies and market efficiency [Shw02] he discusses studies that find indicators better able to model the price assets in the market. Schwert notes studies such as Banz and Reiganum (1981) who showed small capitalisation firms over 1936-1975 show higher expected returns that expected by traditional asset pricing models. Keim (1983) and Reiganum (1983) observed the "turn of the year" effect in which returns of small firms were abnormally high in the first weeks of January. DeBondt and Thaler (1985) showed the "contrarian effect" of firms that under-performed the market in the previous three to five years are likely to outperform the returns of those who over-performed.

It seems from this collection of studies that there is one element in common. Each of the phenomena seems not to hold out of sample. The explanation given by Schwert is that once inefficiencies are highlighted they are removed by the very trading activity that exploits them.

In light of this research it is important that any evidence of inefficiency should hold out of sample while also accounting for transaction costs. This will be an important topic when testing any trading strategies suggested in this paper.

### The Impossibility of Informationaly Efficient Markets

From the papers discussed it seems that the majority of anomalies are either transitory or insignificant when market frictions are taken into account. One of the key assumptions of pure efficient markets is information is free and possessed equally by all agents. Grossman argues that if information is costly agents will only participate in the market if the benefits of doing so at least cover the cost of participation in the market. Trade results from difference in beliefs such that the market may reach a new equilibrium; Grossman argues trade takes place due to a difference in belief caused by heterogeneous expenditure on information.

Grossman introduces a series of conjectures for price systems (markets)

1. The more individuals who are informed the more informative the price.
2. The higher the cost of information the less likely the market is to reach equilibrium.
3. In the limit where there is no noise there is no incentive to purchase information.
4. The greater the magnitude of noise the less information the price.
5. Markets will be thinner when the ratio of informed individuals is either zero or one.

Grossman uses expected utility theory models similar to heterogeneous agent models to explain stylised features such as volatility clustering using models containing two types of agents, fundamentals traders and speculators to justify the existence of market inefficiencies[Hom06].

The implication of Grossman's paper is markets only exist because of the inefficiencies that allow market participants to be compensated for their expenditure on information. In horse racing there has been much less analysis into specific factors that affect prices such that it is difficult to know what information is necessary to build accurate pricing models. It may however be possible to gain insight into the ratio of informed vs. uniformed participants from the study of deviations of late money from the fundamental value taken as the start price.

## 2.2 Analysis of Betting Markets

Gambling markets are economically interesting as they are especially simple forms of financial markets in which the pricing problem is greatly reduced. Studies of the efficiency of betting markets existed well before the introduction of online betting exchanges. Various papers have highlighted anomalies such as long-shot bias however the comprehensive conclusion is similar to that of standard financial markets; betting markets are not sufficiently inefficient as to compensate for market frictions. It is important to reassess the findings of past studies in light of changes to micro market structure via the introduction of online betting markets. Market frictions have been greatly reduced and the set of possible betting strategies have significantly changed with the introduction of lay bets.

### Odds Adjustment by American Horse Racing Bettors

Richard Griffiths was the first to study the behaviour of market participants in betting markets [Gri49]. The study was initially aimed not at the economic implications of efficient markets, but at the psychological rationality in regards to uncertain outcomes.

Griffiths grouped horses from 1386 races via their expected payout from each race,

$$w_i = W_i/W \tag{4}$$

where $w_i$ is payout of horse i, $W_i$ total bet on horse i, W total betting pool.

Grifiths findings were "remarkably" that $w_i \approx p_i$ implying that payouts are an efficient estimator of the true underlying probability. Grifiths concludes market participants are rational and racing markets efficient.

### Stability of Choice Among Uncertain Alternatives

William McGlothlin [McG56] followed up Griffiths study examining 9248 races using an application of fair game tests of efficient markets,

$$E[R_i] = 0 \tag{5}$$

Grouping horses into buckets using implied probabilities from pool payouts he analyses expected returns from each bucket. Interestingly McGlothlin's study is the first to note the existence of favourite, long shot bias in which horses with low probabilities are over bet relative to the favourites in the market. The psychological reasoning given for the bias is people naturally over-estimate unlikely events and thus derive a higher utility from bets on horses with longer odds. This is the same theory as to why people play the lottery even though the expected utility is low.

Although there is evidence of bias in horse racing markets the conclusion from most academic research is that they do not exceed markets frictions. Strategies that exploit bias are not profitable when track take in included.

### Gambling and Rationality

Although many studies conclude that overall aggregate behaviour of market participants is rational Rosset pre-supposes that most gamblers are in fact probably not equipped to deal with the level of complexity required to calculate correct odds [Ros65]. This problem is made more difficult when bets are multiples based on the outcome of the results of more than one race. Rosset uses a data set originally collected by Weitzman to study the distribution between posted market odds and returns. Although Rosset's finds similar bias in horses prices with odds $\leq 0.03$ the overall findings are that market participants are rational. His reasoning is even if market participants are not able to mathematically estimate odds they are able to identify and follow the smart money. The explanation is seem to be more plausible given empirical evidence of betting activities in the market.

There are important implications to Rossets paper which it seems in gambling markets at least, participants are more likely to follow the smart money vs. have an in depth analytical assessment of the factor models are probability theory for pricing. This also seems evident due to the distinct lack of consensus in academic literature of factors determining probabilities. A herding effect may explain the existence of autocorrelation in prices and may lead to effective modelling strategies that can out think the market.

### Horse Racing: Testing Efficient Market Models

Many studies of market efficiency in horse racing market are fair game tests. Snyder [Sny78] gives an overview of past studies of Fabricant (1965), Griffiths (1949), McGlothlin(1956), Snyder (1978) and Weitzman(1965) .

| Author | Date Published | Racing Dates | No. of Races |
|---|---|---|---|
| Fabricant | 1965 | 1955-62 | 10,000 |
| Griffiths | 1965 | 1947 | 1,124 |
| McGlothlin | 1956 | 1947-56 | 9,248 |
| Snyder | 1978 | 1972-74 | 1,730 |
| Weitzman | 1965 | 1954-63 | 12,000 |

These studies find consistently returns are negative when track take is accounted for. Snyder notes that there are increasingly negative returns for betting on longer horses due to long shot bias. Although probabilities are a closed system and thus the returns on favourites should thus be greater than one no statistical anomalies found were sufficiently large so as to overcome market frictions.

These findings are interesting in light of online betting markets as commissions are much lower, there is no tax on gambling winnings and new markets structure allows new trading strategies using lay bets. In light of these changes it seems likely that if these statistical anomalies still exist they may finally exceed market frictions.

### The Economics of Wagering Markets

Prices observed in gambling markets are an amalgamation of various scarce data sources. Sauer provides a review of the studies on horse racing in light of utility theory to conclude whether market participants are rational [Sau98].

The study of efficiency of betting markets requires the introduction of some terminology.

$$R_i = QW/W_i = Q/w_i \tag{6}$$

$R_i$ is return on horse i, $W_i$ is total \$ bet, $\tau$ is take out rate and Q is total paid out i.e. $(1-\tau)$

Sauer shows that if agents are assumed rational and informed we can expect the returns from betting on all horses will be equal to one.

$$E[R_i] = p_i Q/w_i \tag{7}$$

Sauer gives a formal foundation of efficiency similar to that argued by Grossman that efficiency is a function of the number of informed market participants. William Hurley and Laurence McDonough derive a non competitive Nash equilibrium for informed agents [HM95].

$$E[R_i] = p\frac{N + \sum x}{N/3 + \sum x} x_m - x_m \tag{8}$$

where $\sum x$ is aggregated betting of all informed agents.

As the number of informed agents increase the expected returns quickly tend to one. Thus it does not take many informed agents to achieve efficient equilibrium.

Sauer points out that efficiency is not a standalone concept. The implications are efficient prices embody the same assumptions employed in the original pricing model. For this reason it is not always simple when rejecting the hypothesis of market efficiency as this may in fact be due to weaknesses in the original pricing model.

### Market Efficiency at the Derby: A Real Horse Race

Most literature investigating market efficiency of horse racing markets examines the relationship between posted odds and race outcomes. Dolvin and Pyles use data from the Kentucky Derby from 1920-2005 to examine the significant explanatory variable of posted odds[DP08]. They note their study is the only one that they know of using factor models to predict race odds.

Dolvin specifically examines how the experience of the breeder, trainer and jockey affect the prices observed in the market. A two stage model is used; the first to show significant factors in determining odds, the second is to compare the start prices predicted by the model with posted odds and race outcomes.

$$\begin{aligned} Prob = \quad & \alpha + \beta_1 JockeyEx + \beta_2 BreederEx + \beta_3 TrainerEx + \beta_4 Geilding + \\ & \beta_5 Filly + \beta_6 Inside + \beta_7 Mid + \beta_8 Outside + \beta_9 FieldSize \\ & \beta_1 0 GoodDum + \beta_1 1 HeavyDum + \beta_1 2 MuddyDum + \beta_1 3 SlowDum + \epsilon \end{aligned} \tag{9}$$

The factors investigated include experience, post position, race length and track conditions. Analysis of the various races found these factors to be significant in explaining the posted odds. The second stage tests the significance of factors from the first model in explaining the results of each race.

$$Results = \alpha + \beta_0 Probability + \beta_1 JockeyEx + \beta_2 BreederEx + \beta_3 TraninerEx + \epsilon \tag{10}$$

Although it seems significant relationships between factors and probabilities exist the same factors are insignificant in explaining the outcomes of races. The conclusion is in strong support of efficient market hypothesis, which is disappointing for bettors as it seems that these factors are fully incorporated into the price and hence it is not possible to use these factors alone to outperform the market.

The $R^2$ values for the first model above are in the range of 10-15 suggesting there may be significant factors omitted. It is not possible to suggest if this is because of omitted variables or noise in the model. During the course of the study it will be useful to analyse the level of noise in data so as to make a prediction about the reason as to why the model has such a low $R^2$.

## 2.3 Gap between literature and Practice

There have been a number of studies into the efficiency of horse racing markets however a substantial amount of research is based on tests between posted odds and race outcomes. Recently with the change to micro market structure it is important to reassess the findings of prior studies in light of tighter spreads, lower commission, zero taxation and lay bets.

There is a lack of in depth analysis of factor models similar to those of Dolvin and Pyles explaining offered probabilities. These findings are interesting given the number of studies that confidently state bettors are able to rationally deduce prices but give no insight as to how the process by which this is achieved.

Studies almost conclusively find horse racing markets are efficient or that anomalies are not sufficiently large as to form profitable trading strategies. If this is the case or if like Dolvin's pre supposes probabilities are explained by variables such as track length, trainer, jockey etc then there should be no reason for prices to change dramatically in the short run up to a race. This seems to indicate swings in market prices prior to a race are likely to be the result of market inefficiencies or variables omitted from static models.

This study will be unique in its analysis of the dynamic price discovery process shortly before the start of a race. The study will be the first to use ultra high frequency models to describe and forecast future odds.

# 3 Data

## 3.1 Definition

The data set for this study was collected directly from the Betfair online exchange. Horse racing markets are managed via a set of limit order books. The exchange offers a free SOAP API which allows programmatic querying of tiered back-lay prices, depths, volumes and last-traded price. Prices are decimal odds and quoted in the range [1.01 1000]. Implied probabilities or a runner winning are calculated as the reciprocal of the price.

The final data set was collected from March $28^{th}$ 2009 to July $12^{th}$ 2009, the final data set contains 947 UK horse races, 7872 horses and 15 million ticks. For each horse 1800 ticks, one tick per second for 30 minutes prior to the scheduled start of the rate were collected. The techniques used to clean the data are discussed in this section.

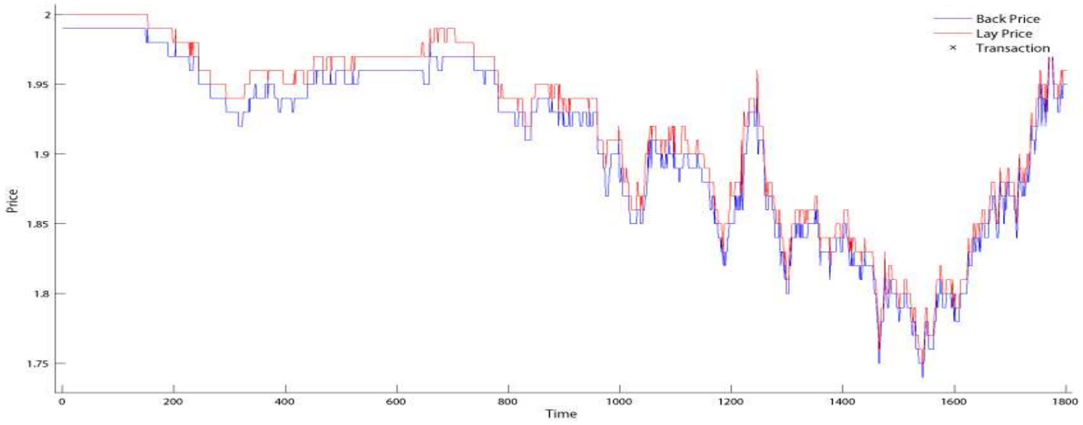

Figure 1: Price Series for Right Stuff at 19:20 Kempton Park $8^{th}$ April 2009

## 3.2 Collection And Cleaning

Data was collected via Apache AXIS2 on a Linux Virtual Private Server (VPS). The latency of the SOAP price requests are approximately 15-25ms. Although the connection is very stable over 15 million ticks were collected thus it is inevitable that some errors and missing ticks occurred in the data set. For this reason the data was cleaned in three stages.

### 3.2.1 Outlier Detection

Price tiers for horse races can be volatile due to a lack of volume trading a long time from the start of a race. Occasionally before a race significant changes may take place in the price on only one side of the market. These asymmetric changes are then corrected without any trade taking place at the new back or lay price. False ticks are identified as ”large” single sided moves without trade at the new price.

Let $\{p_i\}_{i=1}^{N}$ be an ordered series of ticks. Invalid ticks are identified via the model

$$|(p_i - \bar{p}_i(k)| > 3s_i(k) + \gamma \tag{11}$$

where $\bar{p}_i$ and $s_i(k)$ are the sample mean and standard deviation respectively. k is the number of lagged variables and $\gamma$ the minimum expected price variation which is important during periods of low tick volatility [BG06]. If ticks are identified as anomalous and no trade has taken place at that price it is replaced with the last valid price.

### 3.2.2 Interpolation And Time Mapping

For the analysis we would ideally like to have a continuous and regularly spaced series of observations. For technical reasons observation time stamps are unlikely to be perfectly interspaced and exactly at second intervals prior to a race. If we consider timestamps in milliseconds for five seconds up the start of a race we would like to map the observed values to the cleaned time stamps.

| | | | | | | |
|---|---|---|---|---|---|---|
| Observed: | 5014, | 3995, | 3032, | 2002, | 988, | 15. |
| Cleaned: | 5000, | 4000, | 3000, | 2000, | 1000, | 0. |

One technique to map observations to time points is to use the closest neighbour function,

$$f(x_t) = \begin{cases} x_i & \text{if } (x_t - x_i) < (x_{i+1} - x_t) \\ x_{i+1} & \text{if } (x_t - x_i) > (x_{i+1} - x_t). \end{cases} \tag{12}$$

Where $x_t$ cleaned timestamp $x_i$ is the $i^{th}$ observed timestamp and $x_i < x_t < x_{i+1}$. Alternately the most recent observation prior to the cleaned time point to be mapped via the function,

$$f(x_t) = max(x_i \mid x_i < x_t) \tag{13}$$

Although in some regards the latter method is more parsimonious as it does not use data not yet observed. The standard deviation between observations is order of a tenth the interval length so in this case using the nearest neighbour technique is a better approximation.

### 3.2.3 Noise Filters

The final stage of data cleansing involves removing noise associated with small, short term asymmetric adjustments of the back or lay prices. Model fitting and forecasting will use a synthetic mid price calculated as the average of the bid and ask prices. There is a large amount of noise associated with the observed back and lay prices. Back and lay price tend to oscillate around a stable mid with each price moving by one tick at a time. If this noise is not filtered out then it directly causes the mid price to also oscillate. This behaviour is however not a true reflection of the fair prices as it is unlikely that trade will take place around this adjusted mid.

Median filters and mean clip filters were considered to remove impulsive noise from the data set. Median filters consider the last k ticks and returns the median of them. This is a simple and effective method of removing noise from a stochastic process. The alternative mean clip filter takes the last k ticks, removing the largest and smallest values and averages the rest. Different specifications of filter were considered and the results are below,

| Filter | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| Median RSS | 0.53770 | 0.58210 | 0.58750 | 0.61530 | 0.64310 | 0.68070 |
| Avg Clip RSS | 0.53770 | 0.54350 | 0.56510 | 0.58380 | 0.60860 | 0.62120 |

Values are mean RSS of smoothing filter over entire data set

The RSS was calculated as the difference between the filtered mid price and the last transacted price. The best filter is consistently the median filter with k equal to three as it is best able to balance noise with quick adaption to true price adjustments.

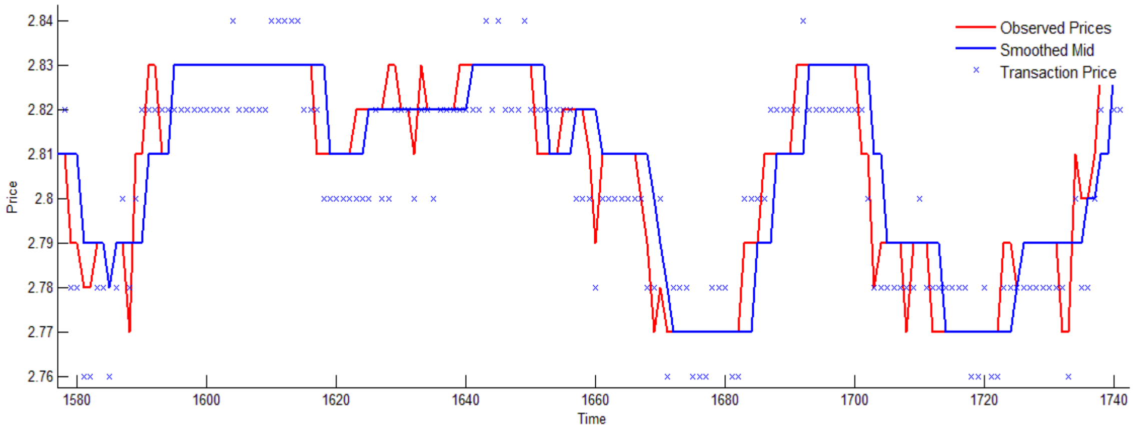

Figure 2: Comparison of Observed Mid Price and Transacted Price vs. Smoothed Mid

Further analysis requires the computation of implied probabilities for a horse winning a race. Filtered mid prices are the most accurate proxy to a bets true fair value. In the course of this paper implied probabilities will be calculated using mid prices smoothed using a median filter with k equal to three.

# 4 Classical Analysis

Previous studies analyse the efficiency of market prices as predictors of race outcomes. Studies use implied probabilities from market prices along with forms of categorisation to try to out predict the market. Studies have found consistent bias in prices such as favourite/long shot bias, also bias in markets where prices move significantly in a short space of time before the race starts. In this section market prices and information flow are used to predict the outcome of markets to test their efficiency of predicting race outcomes.

## 4.1 Comparison of Bucketed Winners

The implied probability of a horse winning a race is given by the reciprocal of its price.

$$pr_i = 1/p_i \tag{14}$$

In the case of the collected data set the filtered mid prices are used to calculate implied probabilities.

Horses are grouped into buckets by their implied probability using a similar technique to that of Griffiths paper[Gri49]. The number of expected wins per bucket can be calculated as number of horses divided by the average mid price. The random variable $W_i$ defined as the number of wins per bucket follows a binomial distribution with the expectation and variance given by,

$$E[W_i] = N_i * pr_i \tag{15}$$

$$Var[W_i] = N_i * pr_i * (1 - pr_i) \tag{16}$$

Given the data set it is possible to test the efficiency of prices as estimators of race outcome via confidence intervals from the binomial distribution. The null hypothesis is prices are efficient estimators of race outcome is tested using a two tailed test. An interesting extension to the classical analysis will be to test the efficiency of prices both immediately before a race starts as well as thirty minutes prior.

Table of Results for tests of efficiency of prices

| | 30 Mins From Race Start | | | | | | Race Start | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bucket | Horses | *Mid* | Wins | P(x) | E(x) | p | Horses | *Mid* | Wins | P(x) | E(x) | p |
| 0-1 | 408 | 410 | 0 | 0.24 | 1.0 | .63 | 403 | 416 | 2 | 0.24 | 1.0 | .45 |
| 1-2 | 673 | 111 | 8 | 0.90 | 6.1 | .43 | 755 | 107 | 6 | 0.93 | 7.0 | .72 |
| 2-3 | 514 | 51.2 | 10 | 1.95 | 10.0 | 1.0 | 629 | 51.0 | 13 | 1.96 | 12.3 | .78 |
| 3-4 | 554 | 34.0 | 15 | 2.95 | 16.3 | .80 | 480 | 34.2 | 16 | 2.93 | 14.1 | .60 |
| 4-5 | 520 | 25.3 | 33 | 3.96 | 20.6 | .005* | 560 | 25.2 | 29 | 3.98 | 22.3 | .13 |
| 5-6 | 549 | 19.9 | 25 | 5.01 | 27.5 | .62 | 501 | 19.9 | 39 | 5.02 | 25.2 | .004* |
| 6-7 | 453 | 16.6 | 33 | 6.01 | 27.2 | .23 | 397 | 16.7 | 25 | 5.98 | 23.8 | .75 |
| 7-8 | 408 | 14.4 | 32 | 6.96 | 28.4 | .43 | 357 | 14.5 | 20 | 6.92 | 24.7 | .35 |
| 8-9 | 383 | 12.7 | 31 | 7.89 | 30.2 | .85 | 291 | 12.7 | 20 | 7.85 | 22.8 | .59 |
| 9-10 | 279 | 11.3 | 22 | 8.88 | 24.8 | .60 | 312 | 11.2 | 22 | 8.90 | 27.8 | .28 |
| 10-11 | 280 | 10.0 | 20 | 10.0 | 28.0 | .13 | 273 | 10.0 | 26 | 10.0 | 27.2 | .84 |
| 11-12 | 278 | 9.05 | 30 | 11.0 | 30.7 | .92 | 332 | 9.10 | 34 | 11.0 | 36.5 | .73 |
| 12-13 | 200 | 8.34 | 22 | 12.0 | 24.0 | .74 | 201 | 8.32 | 23 | 12.0 | 24.2 | .83 |
| 13-14 | 240 | 7.73 | 31 | 12.9 | 31.0 | 1.0 | 185 | 7.69 | 22 | 13.0 | 24.0 | .67 |
| 14-15 | 217 | 7.14 | 44 | 14.0 | 30.4 | .006* | 202 | 7.10 | 43 | 14.1 | 28.4 | .002* |
| 15-20 | 572 | 5.95 | 87 | 16.8 | 96.2 | .31 | 649 | 5.90 | 112 | 17.0 | 110.0 | .83 |
| 20-25 | 384 | 4.65 | 94 | 21.5 | 82.6 | .15 | 372 | 4.59 | 81 | 21.8 | 81.0 | .95 |
| 25-30 | 243 | 3.75 | 67 | 26.6 | 64.7 | .72 | 245 | 3.73 | 59 | 26.8 | 65.6 | .35 |
| 30-35 | 146 | 3.16 | 38 | 31.7 | 46.2 | .15 | 142 | 3.14 | 44 | 31.8 | 45.2 | .85 |
| 35-40 | 91 | 2.72 | 28 | 36.7 | 33.4 | .28 | 86 | 2.74 | 27 | 36.5 | 31.4 | .37 |
| 40-45 | 53 | 2.39 | 24 | 41.9 | 22.2 | .58 | 63 | 2.38 | 24 | 42.0 | 26.5 | .61 |
| 45-50 | 25 | 2.15 | 11 | 46.6 | 11.6 | .84 | 27 | 2.15 | 13 | 46.5 | 12.6 | .85 |
| 50-55 | 36 | 1.94 | 20 | 51.5 | 18.5 | .62 | 38 | 1.93 | 19 | 51.7 | 19.7 | .87 |

$p^*$ significant at 1%

As can be seen from the table above there is little evidence that we can reject the null. More so prices are stable predictors at all points up to 30 minutes prior to the race start. This gives a preliminary indication that the bias in market prices is likely to be low.

## 4.2 Late Breakers

In Rossets paper he gives evidence of the profitability of strategies that bet on horses where their prices move by more than a threshold value late in the betting. The implication of this research is markets are unable to incorporate new information effectively into prices such that profits may be made from this. The Analysis in this section looks at subsets of horses in which prices change more than ± 20% in the last five minutes.

| Implied<br>Position | Category | Num<br>Horses | Back<br>Price | Wins | P(x) | E(x) | Lo | Hi | p |
|---|---|---|---|---|---|---|---|---|---|
| $1^{st}$ | | | | | | | | | |
| | No Late Move | 671 | 3.74 | 198 | 26.71% | 179.2 | 160.5 | 198.0 | .097 |
| | Late Mover | 84 | 5.12 | 22 | 19.55% | 16.4 | 10.8 | 22.0 | .097 |
| $2^{nd}$ | | | | | | | | | |
| | No Late Move | 595 | 5.67 | 117 | 17.63% | 104.9 | 92.8 | 117.0 | .179 |
| | Late Mover | 160 | 7.22 | 29 | 13.85% | 22.2 | 15.3 | 29.0 | .109 |
| $3^{rd}$ | | | | | | | | | |
| | No Late Move | 533 | 8.30 | 81 | 12.05% | 64.2 | 47.5 | 81.0 | .023* |
| | Late Mover | 222 | 9.11 | 29 | 10.98% | 24.4 | 19.8 | 29.0 | .283 |

$p^*$ significant at 5%

Apart from the 3rd favourites it is not possible to reject the null hypothesis that prices are efficient predictors of race outcomes. This is consistent across all favourites even if there is more that a 20% movement in prices less than five minutes before the race.

It seems market prices are efficient even in light of large changes in market prices late in betting. These findings are some what surprising as the level of accuracy market makers achieve is very high even when prices are extremely volatile.

### 4.3 Bias and the Information Curve

The vast majority of money bet on horse races is placed in the last five minutes before a race. In this torrent of trading is information as to the final market view on a horse's implied probability of winning. The price bias and information flow studied in this section give insight into the efficiency of prices as well as time at which the majority information enters the market.

Bias at time t is calculated as the error between the price at time $p_t$ and the price at the scheduled race start $p_0$. RMSE is the root mean squared error of this difference.

$$Bias_t = \frac{p_0 - p_t}{p_0} \tag{17}$$

$$RMSE_t = \sqrt{(\frac{p_0 - p_t}{p_0})^2} \tag{18}$$

Positive bias implies a horse is under priced before the start of a race such that the market has an overly optimistic view of the horse's chances of winning. A negative bias implies the inverse and the final price should actually be higher than that in the market. Under the null hypothesis of market efficiency the mean bias in price should be zero.

The table below give the mean bias and RMSE for $1^{st}$, $2^{nd}$ and $3^{rd}$ favourites. Favourites are horses with the highest implied probabilities of winning taken thirty minutes before the start of a race.

| Implied |  | Time in Minutes Before Race |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Position | Type | 30 | 25 | 20 | 15 | 10 | 5 | 4 | 3 | 2 | 1 | 0 |
| 1st | Bias | 1.91 | 1.95 | 1.43 | 1.19 | 1.06 | 0.07 | 0.00 | -0.19 | -0.41 | -0.55 | 0.00 |
|  | RMSE | 17.5 | 17.1 | 16.8 | 16.5 | 15.1 | 12.7 | 10.4 | 8.7 | 6.6 | 4.7 | 0.00 |
|  | Spread | 1.99 | 1.97 | 1.97 | 1.84 | 1.67 | 1.59 | 1.58 | 1.58 | 1.55 | 1.55 | 1.56 |
| 2nd | Bias | 1.60 | 1.58 | 1.68 | 1.75 | 1.35 | 1.21 | 1.24 | 0.53 | 0.24 | 0.15 | 0.00 |
|  | RMSE | 24.1 | 24.0 | 23.6 | 22.9 | 21.1 | 17.3 | 14.5 | 12.1 | 9.5 | 6.1 | 0.00 |
|  | Spread | 3.14 | 3.17 | 2.95 | 3.02 | 2.75 | 2.29 | 2.28 | 2.29 | 2.27 | 2.29 | 2.29 |
| 3rd | Bias | -2.40 | -1.93 | -1.46 | -1.27 | -0.04 | 0.50 | 0.42 | 0.16 | -0.08 | -0.03 | 0.00 |
|  | RMSE | 28.6 | 28.1 | 27.4 | 26.5 | 23.6 | 19.8 | 16.7 | 13.5 | 10.6 | 7.2 | 0.00 |
|  | Spread | 4.24 | 4.25 | 4.05 | 3.91 | 3.71 | 3.10 | 2.98 | 2.95 | 2.93 | 2.89 | 2.88 |

Values shown are mean % errors across all horses

The first derivative of the RMSE gives an indication as rate of information flow into the market. Interestingly the majority of price setting is done about five minutes before the start of the race. This non linear flow of information in time has implications to forecasting models as it implies that forecasts may only be accurate very shortly before a race starts as before this point little information is available in the market.

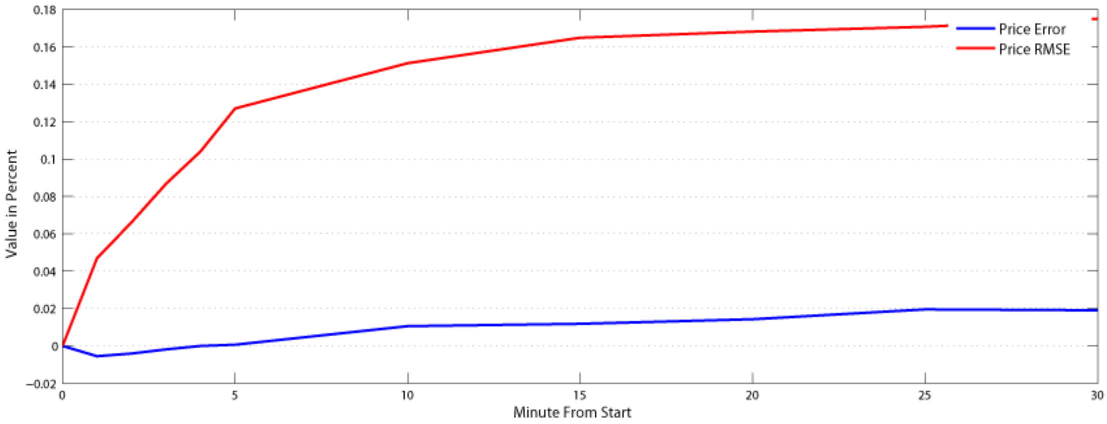

Figure 3: Graph of Mean Error and RMSE of Prices with Regard to Starting Price

The data implies about a 2% positive bias in prices of the first and second favourites and negative bias in the third. This is contrary to other studies findings of negative bias in the prices of favourites as they are under bet due to over betting of long shots. An explanation for this positive bias could in fact be market makers compensating for favourite bias. In order to get the volume required to balance their books it seems bettors require compensation for the natural bias against laying favourites. The dynamics of favourites are also interesting as bias is overcompensated for three minutes before the start of a race only to be re-corrected again.

It initially seems prices are efficient as at no point does the bias exceed market spreads. If this were the case then it would be possible to generate profits from either backing or laying all favourites thirty minutes before a race and then making the opposite trade when the race begins. The profits captured from such a strategy would be the difference between the average spread and average bias.

Spread offered on horses however are not uniform but in fact a function of price. If the $1^{st}$ favourite horses are group by price it is possible to see the bias for favourites priced between 2.75 and 4.75 are consistently larger than market spreads.

| Bucket | Races | Mean Lay @ 30 Min | Mean Back @ 1 Min | Average Spread @ 30 Min | Average Bias @ 30 Min | Profitable |
|---|---|---|---|---|---|---|
| 2.25 - 2.75 | 111 | 2.51 | 2.60 | 1.2% | 1.2% | False |
| 2.75 - 3.25 | 137 | 3.00 | 3.17 | 1.5% | 2.8% | True |
| 3.25 - 3.75 | 134 | 3.52 | 3.71 | 1.9% | 2.6% | True |
| 3.75 - 4.25 | 101 | 4.01 | 4.19 | 2.2% | 2.4% | True |
| 4.25 - 4.75 | 93 | 4.49 | 4.75 | 2.8% | 3.2% | True |
| 4.75 - 5.25 | 76 | 5.02 | 5.21 | 2.9% | 2.5% | False |

A profitable trading strategy can be achieved by laying horses in this group thirty minutes before the start of a race and backing them just before the race begins. Profits if this strategy were to be implemented on the collected data set are below.

| Month | Num Races | Profit w/o Comm | Profit w Comm |
|---|---|---|---|
| April | 139 | 249.03 | 191.79 |
| May | 150 | 79.21 | 24.55 |
| June | 112 | 349.12 | 300.40 |
| Total | 401 | 677.36 | 516.74 |

Profits based of £100 laid per race and commission rates of 5% on net winnings.

Profits of 500% in 3 months seem high except this is equivalent to 1.28% (bias - spread) per race over 400 races. It is also worth considering the volatility and maximum draw down of such a strategy to check the feasibility of such a strategy.

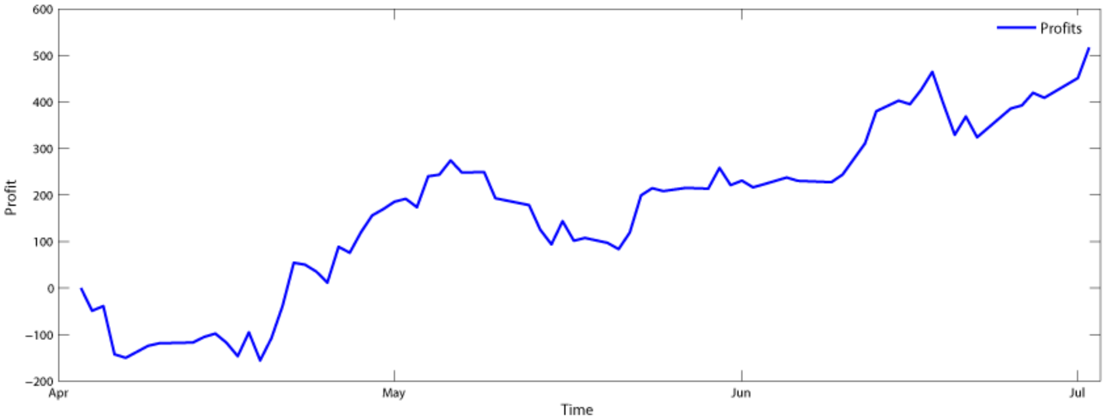

Figure 4: Profit of Strategy over April to July

The biggest draw down from the strategy is £155.63 which is quite large relative to the per bet outlay of £100. There are 174 races with negative profits in the set of 401 races. If we assume the drift to be a geometric brownian motion with zero bias then the probability of losing a race should be 0.5. The probability of having 174 losers over 401 races given by the binomial distribution is 0.812 %. This is significant at both a 5% and 1% level so the null hypothesis is rejected hypothesis that the market is efficient.

It can be concluded that market prices are efficient estimators of race outcome. This is somewhat unsurprising as it is consistent with the majority previous studies. Interestingly this is also the case for runners irrespective of time to race of if prior to the race beginning large price moves occur.

Prices before the start of a race show evidence of significant favourite long-shot bias. This bias is however opposite of that suggested by other studies. More importantly this bias is larger than market spreads; the implication of this is profits can be achieved without any specialist information. In light of this evidence it can be concluded that although prices are efficient predictors of race outcomes the dynamic price process itself is not efficient.

# 5 ACM-ACD model

Horse racing price discovery processes are collections of ultra high frequency discrete returns. The data exhibits stylised features such as asymmetric information flow, volatility clustering, and autocorrelation. It seems given the efficiency of prices thirty minutes from the start of a race the level of volatility in the market is unlikely to be purely from noise trading alone. If prices follow a random walk it should be impossible to forecast prices with any level of accuracy. The existence of bias and autocorrelation would imply that the price series is not efficient such that there may be room for forecasting models to outperform prices alone as indicators of race outcome.

Normal econometric models such as ARMA processes fail to deal with the irregularity of interspacing between price changes and the discreteness of returns[Sim07].

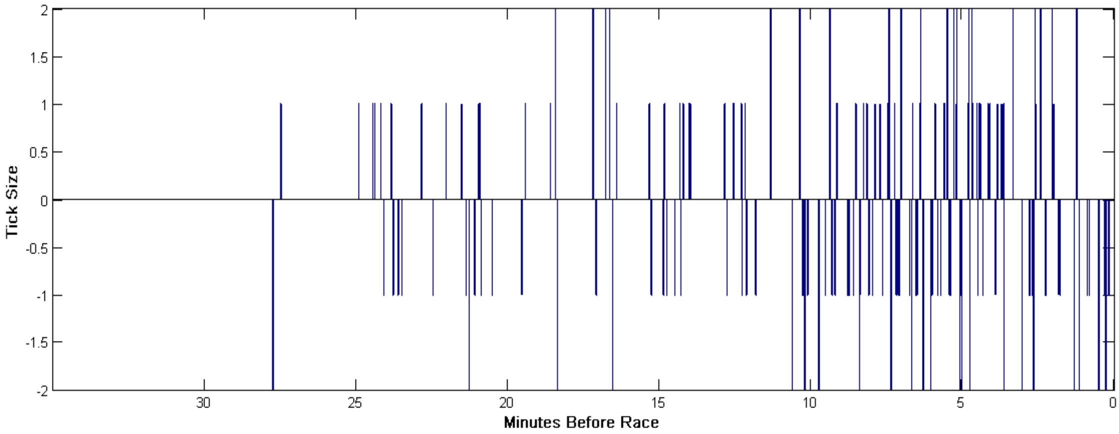

Figure 5: Example of Tick series for 3:15 at Southwell 31$^{st}$ March 2009

In this section Engle's autoregressive conditional multinomial (ACM) and autoregressive conditional duration models (ACD) are used to model these features. The ACD model is used to describe the duration between price ticks as an auto correlated random process, the ACM then models discrete price ticks using a multinomial logistic regression. Although the discrete tick series may be modelled as a first order markov chain, the ACM model allows the inclusion of explanatory variables such as duration, spreads and prices to achieve dynamics not possible using first order markov chains alone[Eng00].

It is important if we are to forecast prices at the start of a race then we can accurately forecast trading intensity as well as direction. The ACM-ACD model makes this possible in irregular and discrete time series.

## 5.1 ACD model

In this section a brief description of the ACD model is given and calibration of the model is discussed. The duration $x_i$ between price changes $t_i$ and $t_{i-1}$ is given by $x_i = t_i - t_{i-1}$. The conditional expectation of $x_i$ is given by,

$$x_i = \phi_i \, \psi_i \, \epsilon_i \tag{19}$$

The equation consists of three components, $\phi_i$ models seasonality of the data, in the case of horse racing data this will be the monotonically decreasing expectation of duration in time to the race start. $\psi_i$ is an autoregressive GARCH component that models the clustering of ticks. There is a strong dependence between ultra high frequency data series such that large price changes are often are followed by more trading activity which is captured by $\psi_i$ [BGGV04].

$$\psi_i = \omega + \sum_{j=1}^{p} \alpha_j d_{i-j} + \sum_{j=1}^{q} \beta_j \phi_{i-j} \tag{20}$$

$\psi_i$ captures volatility clustering and the long run expected duration in the ACD(1,1) model,

$$E[\psi_i] = \frac{\omega}{1 - \alpha - \beta} \tag{21}$$

The final component $\epsilon_i$ is an i.i.d. error term of standardised durations given by $\epsilon_i = \frac{x_i}{\phi_i \, \psi_i}$. Engle describes the use of any suitable hazard function and chooses an exponential distribution with $\lambda$ equal to one. In the case of this analysis a weibull distribution restricted to have scale one is calibrated. The pdf of the weibull distribution is given by,

$$F(x; \lambda, k) = \frac{k}{\lambda} \left( \frac{x}{\lambda} \right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} \tag{22}$$

Where $\lambda$ is the scale parameter and k the shape. The distribution is restricted by setting $\lambda$ equal to one.

### 5.1.1 Seasonality

To model the seasonality of price durations a six piece natural cubic spline was used. For each horse race the durations between ticks $x_i$ are calculated as $x_i = t_i - t_{i-1}$ and a sparse series created in which value at the midpoint between $t_i$ and $t_{i-1}$ is equal to $x_i$. This process is repeated and durations at all time points collected. Finally at each time point the mean duration is calculated and this series used to calibrate a six piece natural cubic spline. A natural cubic spine is a series of separate cubic splines with the value for each cubic function given by,

$$s(x) = \alpha + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 \tag{23}$$

At each of the seven knot points a natural spline is restricted such that each of the adjacent cubic function has the same level, slope and curvature and satisfies the conditions,

$$\begin{aligned}
s_i(x_i) &= s_{i-1}(x_i), \quad i = 1, \ldots, n-1. && (24)\\
s'_i(x_i) &= s'_{i-1}(x_i), \quad i = 1, \ldots, n-1 && (25)\\
s''_i(x_i) &= s''_{i-1}(x_i), \quad i = 1, \ldots, n-1. && (26)\\
s''_0 x_0 &= s''_{n-1}(x_n) = 0 && (27)
\end{aligned}$$

The spline knots were set at 30, 25, 20, 15, 10 5 and 0 minutes before a race. The final calibrated spline is shown below and will be used to standardise duration in the ACD model.

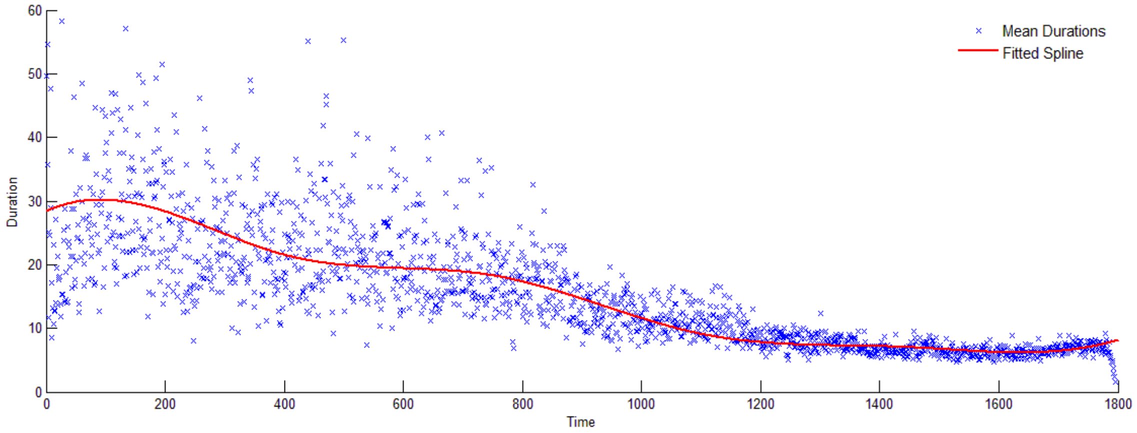

Figure 6: Mean Durations Between Ticks vs. Fitted Six Piece Natural Cubic Spline

### 5.1.2 Estimation

Parameters of the ACD model were estimated via maximising the likelihood of a suitably parameterized weibull distribution. The model uses durations standardised for the seasonality and volatility clustering given by $\psi$. The following likelihood function was used [Pan04],

$$
\begin{aligned}
L(\epsilon;\lambda,k) &= \prod_{i=1}^{n} L(\epsilon_i;\lambda,k) &\quad (28)\\
&= \prod_{i=1}^{n} \frac{k}{\lambda}\left(\frac{\epsilon_i}{\lambda}\right)^{k-1} e^{-(\frac{\epsilon_i}{\lambda})^k} &\quad (29)
\end{aligned}
$$

The error term is the standardised durations $\epsilon_i$ given by,

$$\epsilon_i = \frac{x_i}{\phi_i\ \psi_i} \tag{30}$$

Where $\psi_i$ given by equation 20 and $\lambda$ the scale parameter of the weibull distribution is restricted to be one.

It is important that the model is well specified. To ensure forecast accuracy before a GARCH process was fitted to the seasonally adjusted durations Engles ARCH test was used to test whether the series contain autocorrelated conditional heteroscedasticity to be modelled. Engles ARCH test uses a linear regression to test the dependence between lagged squared residuals. Engle's ARCH test

$$\hat{\epsilon}_t^2 = \hat{\alpha}_0 + \sum_{i=1}^{q} \hat{\beta}_i \hat{\epsilon}_{t-i}^2 \tag{31}$$

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0.$

The seasonally adjusted durations were tested with the ARCH test with 5 lags. If the null was accepted then the ACD model was fitted with $\psi = 1$ otherwise a GARCH(p,q) process was fitted.

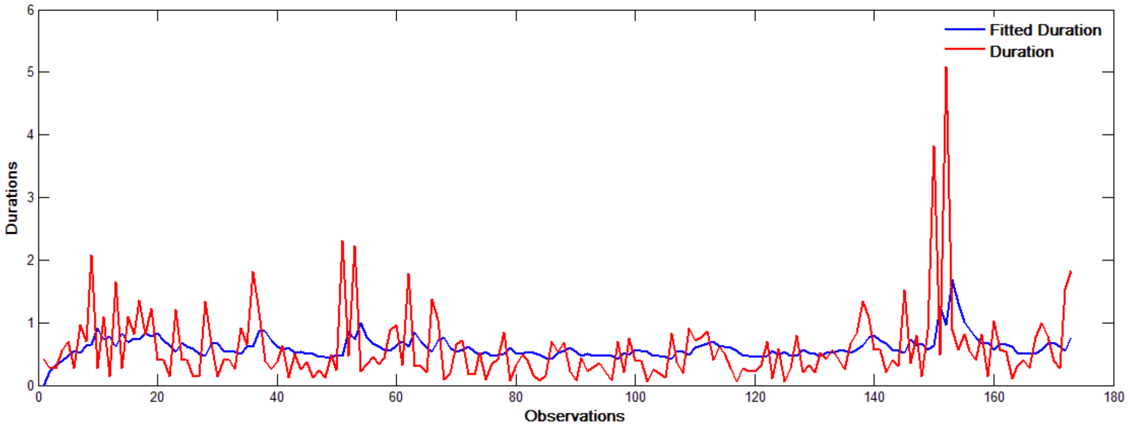

Figure 7: Seasonality Adjusted Durations and Fitted Model

The performance of the model fitted with a GARCH process in sample is compared with those without using the akaike information criterion and bayesian information criterion [MT06].

$$AIC = -2\frac{LogL}{n} + \frac{k}{n} \tag{32}$$
$$BIC = -2\ LogL + k\ log\ (n) \tag{33}$$

These two metrics give relative comparisons of the models adjusted for the number of observations and number of explanatory variables. In both cases smaller values of the information criterion are desired. In the case of this analysis they are used to determine which model should be used to forecast durations.

The quality of the in sample model fitting is reported via the $R^2$ statistic. The $R^2$ is a ratio bound between [0,1] of the explanatory capability of the model relative to a model with just a constant. $R^2$ of the model is given by,

$$R^2 = 1 - \left(\frac{L(0)}{L(\hat{\theta})}\right)^{2/n} \tag{34}$$

Where n is number of observations, L(0) is the logitistic transform of model log likely hood with only the intercept $\omega$ and $L(\hat{\theta})$ the fully parameterised model.

### 5.1.3 Forecasting

To test the accuracy of the ACD model of forecasting trading intensity a ACD(1,1) model was calibrated for each race from thirty to five minutes before the race starts. The average duration for the next 5 minutes was then estimated using 1000 Monte Carlo simulations using the calibrated model. The results of these simulations are below,

| Observed | 5% | Mean | 95% | | | |
|---|---|---|---|---|---|---|
| Duration | Forecast | Forecast | Forecast | $R^2$ | $A\bar{I}C$ | $B\bar{I}C$ |
| 6.363 | 4.358 | 6.032 | 9.833 | 0.712 | 198.614 | 206.811 |

The mean observed duration between trades is 6.363 seconds. The model forecast out of sample is 6.032. The mean $R^2$ is 0.712 in sample which is very promising.

The explanatory power of the ACD forecasts over the panel data was tested with the model.

$$x_i = \alpha + \beta \hat{x_i} \tag{35}$$

Where $x_i$ is the observed average duration and $\hat{x_i}$ is the forecast mean duration from 10,000 monte carlo simulations of the calibrated ACD model. The quality of the ACD model over the panel data is given by the $R^2$ from the estimation of this model.

| $\hat{\alpha}_{val}$ | $\hat{\alpha}_{s.e.}$ | $\hat{\alpha}_p$ | $\hat{\beta}_{val}$ | $\hat{\beta}_{s.e.}$ | $\hat{\beta}_p$ | $R^2$ |
|---|---|---|---|---|---|---|
| 11.595 | (3.593) | 0.002 | 0.758 | (0.069) | 0.000 | 0.608 |

The value of $\beta$ is significant and the $R^2$ of the model is 60.8%. The ACD(1,1) model is very good at explaining the duration between price changes in the data set even out of sample. In the next section the ACD model will be combined with the ACM model for price direction to generate price and directional forecasts for the final minutes before a race.

### 5.2 ACM model

The ACD model is useful for modelling and forecasting the intensity and timing of price changes in the market. To forecast the starting price the model needs to be extended to include the size and direction of price changes. Engles autoregressive conditional multinomial (ACM) model allows price movements to be modelled as a discrete series of ticks [ER05].

Given that time is modelled separately it is possible to modify the returns series so to remove the time between ticks and consider the return series as a set of sequential price changes. It should noted that the magnitude of these price change is rarely greater than ± 2 ticks. Under this assumption if a price change does occur then it is possible to model this using one of four states [-2, -1, +1, +2] ticks. At any time we can consider the price process to be in one of four states which allows the application of a first order markov chain.

The state of the price process is given by $x_i$ which represents the $i^{th}$ state of a markov chain.

$$x_i = \begin{cases} \text{[1 0 0 0]'} & \Delta p_i \geq +2\, ticks \\ \text{[0 1 0 0]'} & \Delta p_i = +1\, tick \\ \text{[0 0 1 0]'} & \Delta p_i = -1\, tick \\ \text{[0 0 0 1]'} & \Delta p_i \leq -2\, ticks \end{cases} \tag{36}$$

The probabilities of observing a tick of a size j are conditional upon the previous state i. Transition probabilities $\pi_i$ can be modelled as a first order markov chain given by,

$$\tilde{\pi_i} = P\tilde{x}_{i-1} \tag{37}$$

where P is a k$x$k matrix of transition probabilities satisfying the conditions that elements are positive and columns sum to unity. The probability of moving to state j is then given by the $j^{th}$ element of $\pi_i$ [RE98].

The ACM model is fitted to historical data via a maximum likelihood estimation technique. The estimation of parameters need satisfy constraints that transition probabilities must be positive and columns of P must sum to unity. In order to discuss the calibration of the ACM model first the technical details of logistic transforms need to be introduced. A logistic transform is used to ensure that parameters satisfy the first constraint. The transform maps numbers on the real axis on to the range [0,1] and is defined as,

$$F(x) \quad = \quad \frac{e^x}{1+e^x} \tag{38}$$

Estimation will take place in real numbers space representing the inverse logistic transform of transition probabilities. The likelihood will be calculated via probabilities using these transformed variables. This process ensures transition probabilities are positive and bounded between zero and one.

The second constraint is achieved by converting transition probabilities $\tilde{\pi_i}$ into set of log ratios. Representing the probabilities as ratios instead of absolute values ensures that the sum of each column of transformed variables is unity. Log ratios are calculated from the transition probabilities via the function,

$$\pi_{ij} = log\left(\frac{\tilde{\pi}_{ij}}{\tilde{\pi}_{ik}}\right) \tag{39}$$

Such that $j < k$ and k is the last element in the vector. The k dimensional vector $\tilde{\pi}_i$ has now been transformed into a k-1 dimensional vector $\pi_i$. The initial markov chain can now be represented via,

$$h(\pi_i) = P^* x_i + c \tag{40}$$

Where h is the inverse logistic transform, $P^*$ is (k-1) $x$ (k-1) matrix, $x_i$ is at k-1 element state vector, c is equal to $\pi_1$ and $P^* x_i = \pi_i - c \quad \forall i > 1$. From this representation the original transition probabilities are easily recovered by the logistic transform.

$$\pi_i = \frac{exp(P^* x_i + c)}{1 + \iota' exp(P^* x_i + c)} \tag{41}$$

Where $\iota'$ is a conforming vector of ones. Using these transforms it is possible to ensure probabilities are positive and columns sum to unity.

The ACM model allows a first order markov chain to be extended so as to include dynamics and explanatory variables. The generalised model form for an ACM(p,q) model is,

$$h(\pi_i) = \sum_{j=1}^{p} A_j (x_{i-j} - \pi_{i-j}) + \sum_{j=1}^{q} B_j h(\pi_{i-j}) + \chi z_i \tag{42}$$

A is a k-1 $x$ k matrix of coefficients with similar function to $\alpha$ in a univariate GARCH, B is k-1 $x$ k matrix of coefficients with similar function to $\beta$ in a univariate GARCH and $\chi$ is a k-1 $x$ r matrix of coefficients for r explanatory variables. z is a vector of explanatory variables that may include durations, spreads etc.

### 5.2.1 Estimation

The ACM model is estimated in a similar way to that of a standard first order markov chain. If $\tilde{\pi}_i$ is a vector of k transition probabilities given by the logistic transform of $h(\pi_i)$ and $x_i$ the current state the likelihood function for N observations is,

$$likelihood = \sum_{i=1}^{N} x_i' \, log(\pi_i) \tag{43}$$

The ACM model used for forecasting was restricted so as A was equal to zero and B equal to one similar to a first order markov chain. The vector $z_i$ contained the difference between the current price and a one minute exponentially weighted moving average. When $\chi$ is calibrated it represents the probability adjustment factors for mean reversion of prices through time. The model was calibrated using historical data from thirty minutes to five minutes to race start.

### 5.2.2 Forecasting

Once the ACD model is fitted two sets of parameters are known, $P^*$ and $\chi$. Prices at the start of the race are forecast using a two stage process. First a simulation of durations between ticks is produced for the next five minutes using a previously calibrated ACD model. Given the expected number of price changes the ACM model is used to simulate the aggregate price change. This process is repeated 10,000 times and the forecast taken as the mean of simulated prices.

### 5.3 Results

The reliability of forecasts from the ACM-ACD model are considered both from the point of view of the absolute error as well as the directional forecast error. The bias in prices five minutes from the start of the race is very low and spreads comparatively high so it is unlikely profits can be generated from a simple strategy such as laying every horse. Alternately in this section the ACM-ACD model will be used to make directional forecasts and trading positions taken accordingly.

| $Price_5$ | $Price_0$ | $Forecast_0$ | Price Error | Forecast Error | Directional Err |
|---|---|---|---|---|---|
| 3.8432 | 3.8912 | 3.9272 | 0.01233 | 0.00925 | 0.5433 |

As can be seen from the results the absolute error of the ACM-ACD forecast is less than that of the five minute prices alone. Although this looks promising the bias in prices is only 1.233% which is less than the spread at this time of 1.59%. The directional forecasting performance is also very promising at 54.5%. The direction of 514 horses from 946 races was correctly forecast. If the price series is a random walk the probability of correctly forecasting 514 races by chance alone is .5078 %. This is significant at a 1% level so it is concluded the ACM-ACD model has some superior forecasting performance over price alone.

Statistically the mid price forecasts are significant however more importantly is to test whether it is also economically significant. Profits from a trading strategy using ACM-ACD forecasts were calculated such that back and lay prices were used at the times the trades would have been entered. If bid offer spreads and commission are taken into account the strategy placing *£*100 per race would have lost £879.9 over the 3 month period. This is a large loss relative to the staked amount however it should be noted this is an average loss of only 1.1% per race.

The forecasting performance however is not uniform across all races. For some races, prices have low volatility and show insignificant signs of trending. For other races it seems prices show consistent trends that are more significant than any mean reversion characteristics. These races can be identified by the magnitude of the forecast compared with prices five minutes from the beginning of a race.

If horses in which the difference between the ACM-ACD forecast mid price and the mid price five minutes before race start is greater than 20% the models directional forecasting performance is 71.05%. Of the entire 946 races this gives a subset of 76 horses of which the direction of 56 were forecast correctly. Although this number is high however bid offer spreads and commission are also significant and need to be taken into account. If a strategy was to be implemented such that prices were bought and sold at the back and lay prices then 49 races of the 76 horses would still be profitable. The ratio of profitable bets is 64.5% even when bid offer spreads are taken into account. The overall strategy would have given profits of 264.6% over 3 months with a draw down of only .35% .

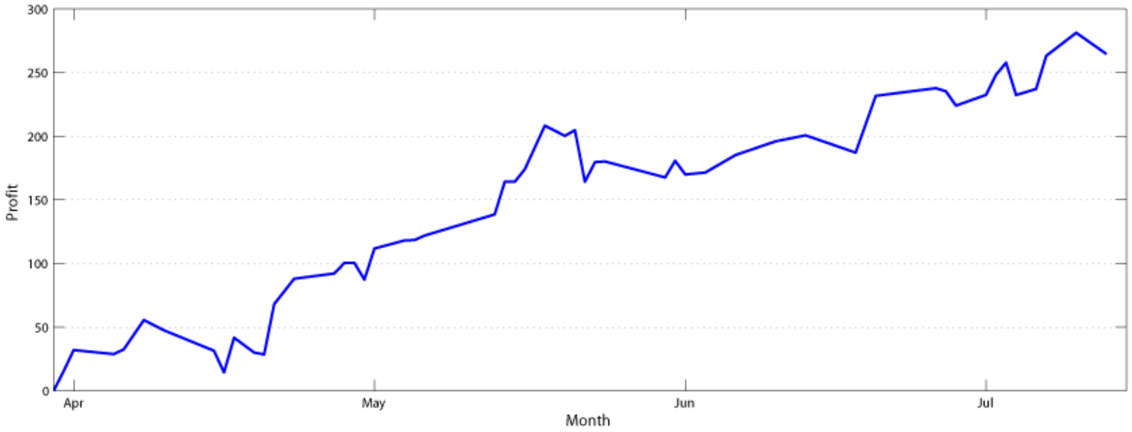

Figure 8: Profits from ACM-ACD Directional Strategy

# 6 Conclusion

In light of this paper's findings it can be concluded that online horse racing markets are efficient in the sense prices are efficient predictors of race outcomes. In none of the examples investigated can the null hypothesis of efficiency be rejected. These results are also robust even if horses are categorised by late movements or the implied positions as suggested in other papers.

Even though prices are efficient in regards to the classical definition the study of market efficiency should be expanded to also include that of dynamic price process such as those found in online betting exchanges. The introduction of lay bets allows complex strategies to be implemented that generate profits independently of race outcome. Interestingly prices before a race show significant levels of bias similar to favourite long-shot bias documented by other studies. Favourite bias in this study has opposite sign to that of previous studies and its most likely explanation seems to be market makers being required to offer compensation to bettors to offset their natural biases. More importantly is the fact that in some cases this bias is larger than the cost of trading leading to profitable trading strategies.

Results from the ACM-ACD model show positive results for the directional forecasts of prices in horse racing markets. The ACD duration model shows a high $R^2$ even out of sample and results from trading strategies exploiting these forecasts show economically significant profits.

Prices in Betfair are as if not more efficient predictors of outcomes, however the price process in which these prices are set seems far from efficient. Significant levels of bias and forecastability can only lead to a conclusion that prices for horses in online betting exchanges are not efficient.

# 7 Future Work

There is a great deal of future work that should be undertaken as a result of this paper. Firstly the models as presented in this paper should be implemented and their economic significance proved out of sample. In regards to the ACM-ACD model work should be undertaken to test the optimality of the assumption that the matrices A and B are constant. Other explanatory variables should be included in the ACM model to try to improve its forecasting performance. All in all it seems unlikely that economically significant profits can be made in betting markets by betting of the outcome of a race. The main focus of future work should be towards strategies which hold positions only up to the start of a race. In this paper two such statistically and economically significant strategies are proposed, however these I am sure are only just the beginning.

# References

[Bei06] Eric D. Beinhocker. *Origin of Wealth: Evolution, Complexity, and the Radical Remaking of Economics.* Harvard Buisness School, Press., 2006.

[BG06] Christian T. Brownlees and Giampiero M. Gallo. Financial econometric analysis as ultra high frequency: Data handling concerns. *Working Paper*, Feb 2006.

[BGGV04] Luc Bauwens, Pierre Giot, Joachim Grammig, and David Veredas. A comparison of finacial duration models via density forecasts. *International Journal of Forecasting*, 20:589–609, 2004.

[DP08] Steven D. Dolvin and Mark K. Pyle. Market efficiency at the derby: A real horse race. *Journal of Applied Economics and Policy*, 27, 2008.

[Eng00] Robert F. Engle. The econometrics of ultra high frequency data. *Econometrica*, 68:1:22, 2000.

[ER05] Robert Engle and Jeffery Russell. A discrete state continuous time model of financial transactions prices and times: The autoregressive conditional multinomial-autoregressive conditional duration model. *American Statistical Association*, 23(2):166:180, April 2005.

[EU06] Stephen Easton and Katherine Uylangco. An examination of in-play sports betting using one-day criket matches. *Working Paper Series*, Nov 2006.

[Fam70] Eugene F. Fama. Efficicent capital markets: A review of theory and empirical work. *Journal of Finance*, 25:384:417, 1970.

[FM07] David Forrest and Ian McHale. Anyone for tennis? (betting). *The European Journal of Finance*, 8:751:768, Dec 2007.

[GM09] Mashall Gramm and Nicolas C. McKinney. The effect of late money on betting market efficiency. *Applied Econometric Letters*, 16:369–372, Nov 2009.

[Gri49] Richard M. Griffiths. Odds adjustment by american horse race bettors. *American Journal of Psychology*, 1949.

[GS80] Sanford Grossman and Joseph Stiglitz. On the impossibility of informationally efficicent marktes. *The American Economic Review*, 70:393–408, Jun 1980.

[HM95] William Hurley and Lawrence McDonough. A note on the hyaek hypothesis and the favourite long shot bias in paramutual betting. *American Economics Review*, 1995.

[Hom06] Cars H. Hommes. *Heterogeneous Agent Models In Economics and Finance.* 2006.

[McG56] W. H. McGlothlin. Stability of choice among uncertain alternatives. *American Journal of Psycology*, 1956.

[MT06] Mika Meitz and Timo Terasvirta. Evaluation models of autoregressive conditional duration. *American Statistical Association*, 24(1), January 2006.

[Pan04] Wan-Kai Pang. Parameter estimation of the weibull distribution. *Department of Applied Mathematics, The Hong Kong Polytechnic University*, Nov 2004.

[RE98] Jeffery Russell and Robert F. Engle. Econometric analysis of discrete-value irregularly-spaced financial transaction data using a new autoregressive conditional multinomial model. *The Center for Research in Security Prices*, April 1998.

[Ros65] Richard N. Rosset. Gambling and rationality. *Journal of Political Economy*, 73, Dec 1965.

[Sau98] Raymond D. Sauer. The economics of wagering markets. *Journal of Economic Literature*, 36, Dec 1998.

[Shi86] Robert J. Shiller. Financial markets and macroeconomic fluctuations. *American Economic Review*, 76:499:503, 1986.

[Shw02] W. G. Shwert. Anomalies of market efficicency. *Handbook of Economic Finance*, page 939:960, 2002.

[Sim07] Ola Simonsen. An empirical model for duration in stocks. *Annals of Finance*, 3:241–255, 2007.

[Sny78] Wayne W. Snyder. Horse racing: Testing efficient market models. *Journal of Finance*, 1978.

[VDM07] Nikolaos Vlastakis, George Dotsis, and Raphael N. Markellos. How efficicent is the european football markets? *Journal of Forecasting*, page 1:22, Jan 2007.