Statistical Models of Horse Racing Outcomes Using R

Dr Alun Owen, Coventry University, UK

aa5845@coventry.ac.uk

Royal Ascot 20th June 2019
The Britannia Stakes (1 mile = 8 furlongs)

https://www.youtube.com/watch?v=sZsF3Q3IJEE

Frankie Dettori riding Turgenev

Frankie Dettori had won 4/4 races so far

He was the strong favourite @ 7/2 to win this 5th race of the day

Harry Bentley riding Biometric @ 28/1

Data: Flat Turf Handicaps in the UK
16,685 horses taking part in 1,693 races.

`race.id` - unique reference number for each race;

`horse.ref` - reference number (or name) for each horse in each race (must be unique within a race);

`age` - age of the horse (years);

`sireSR` - win percentage by offspring of the horse's sire (father) prior to this race;

`trainerSR` - win percentage achieved by the horse's trainer prior to this race;

`daysLTO` - days since last race (days since Last Time Out);

`position1` - finishing position in the previous race (1, 2, 3 or 4, 0 = anywhere else);

`position2` - finishing position two races ago (1, 2, 3 or 4, 0 = anywhere else);

`position3` - finishing position three races ago (1, 2, 3 or 4, 0 = anywhere else);

`finpos` - finishing position in the current race;

`entire` - male horse that has not been castrated (1=yes, 0=no) ;

`gelding` - male horse that has been castrated (1=yes, 0=no) ;
note that a horse that is neither a gelding nor an entire was female;

`blinkers`, `visor`, `cheekpieces` or `tonguetie` (each 1=yes if they were wearing these, 0=no).

`win` **- indicator of whether each horse won (yes) or not (no);**

`sp` **- starting price obtained from Betfair (adjusted for commission);**

race.id	horse.ref	age	sireSR	trainerSR	daysLTO	position1	position2	position3	finpos	win	sp	entire	gelding	blinkers	visor	cheekpieces	tonguetie
1	1	7	6.2	5.4	96	0	0	0	5	no	18	0	1	0	0	0	0
1	2	7	10	9.7	4	3	1	2	9	no	3.5	0	1	0	0	0	0
1	3	4	8	11.1	23	0	4	1	6	no	8	0	0	0	0	0	0
1	4	6	8.8	11.4	40	4	1	0	3	no	3.5	0	0	0	1	0	0
1	5	8	4.7	11.9	14	0	1	3	4	no	11	0	1	0	0	1	0
1	6	9	2.5	2.8	16	3	0	0	1	yes	6	0	1	1	0	0	0
1	7	5	9.5	8.7	16	0	0	0	2	no	4.5	0	1	0	0	0	0
1	8	6	8.1	9	2	0	2	0	7	no	9	0	1	0	1	0	0
1	9	7	8.3	9	23	0	0	0	8	no	20	0	0	0	1	0	0
2	1	9	8.1	5.2	16	3	0	3	3	no	4	0	1	0	1	0	0
2	2	6	7.4	8.8	159	0	0	2	7	no	8	0	1	0	0	0	0
2	3	10	0	0	5	0	0	0	8	no	16	0	1	0	0	0	0
2	4	6	8.8	14	5	0	0	1	5	no	9	0	1	0	0	0	0
2	5	5	9	13.6	23	4	0	1	2	no	2.25	0	0	0	0	0	0
2	6	9	8.3	8.7	19	4	1	2	1	yes	7	0	1	0	0	0	0
2	7	8	7.3	11.4	31	0	0	0	6	no	12	0	1	0	0	0	0
2	8	7	7.1	10	14	2	0	0	4	no	5	0	1	0	0	0	0

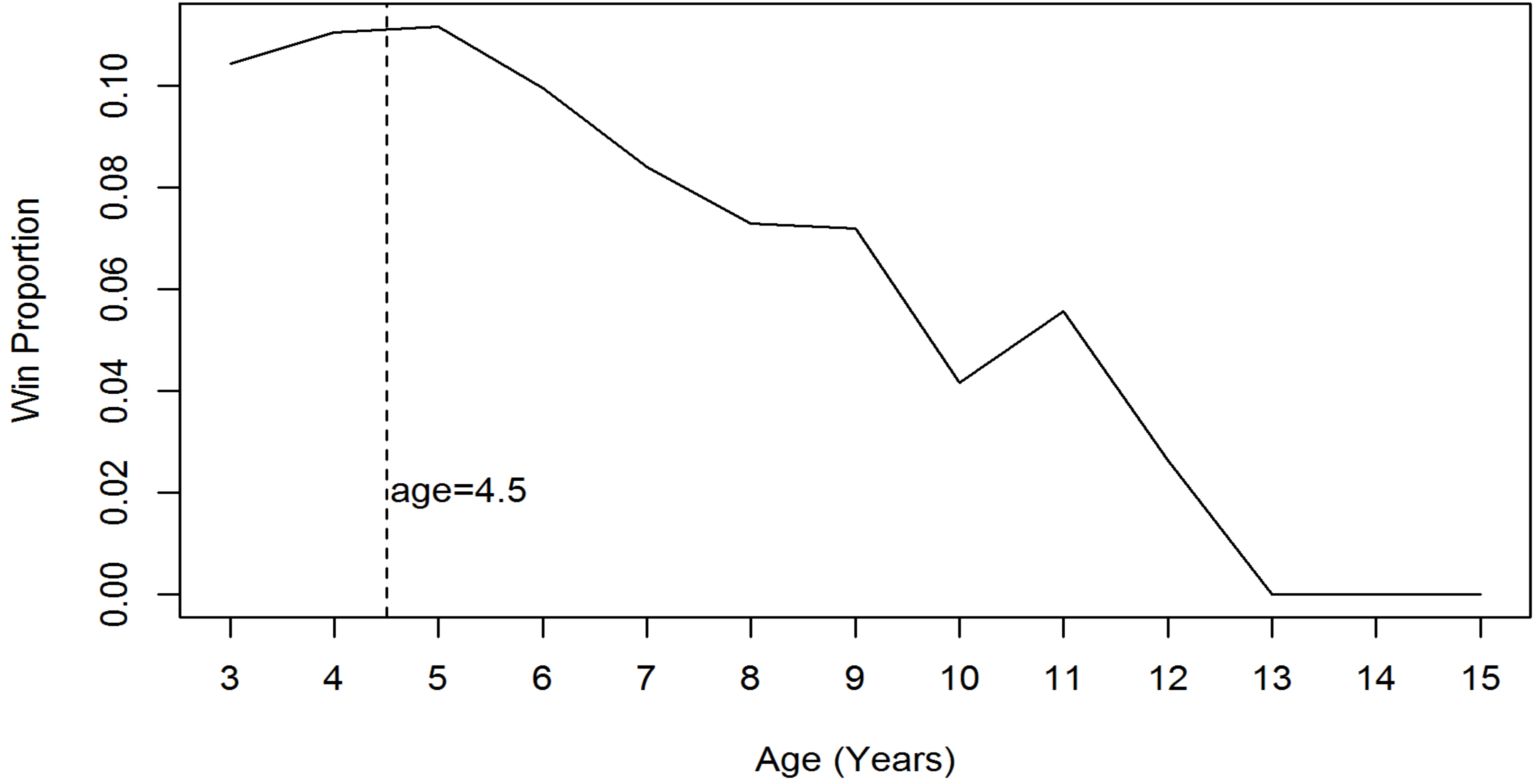
Data Management

- Sire SR and Trainer SR both capped at 20%
- daysLTO capped at 60 days
- SP adjusted for Betfair Commission assumed to be 5%
- Training set 70% of races (11,710 horses taking part in1,181) to develop a model and possible betting strategy;
- Test set 30% of races 4,975 horses from 512 races for out-of-sample assessments.

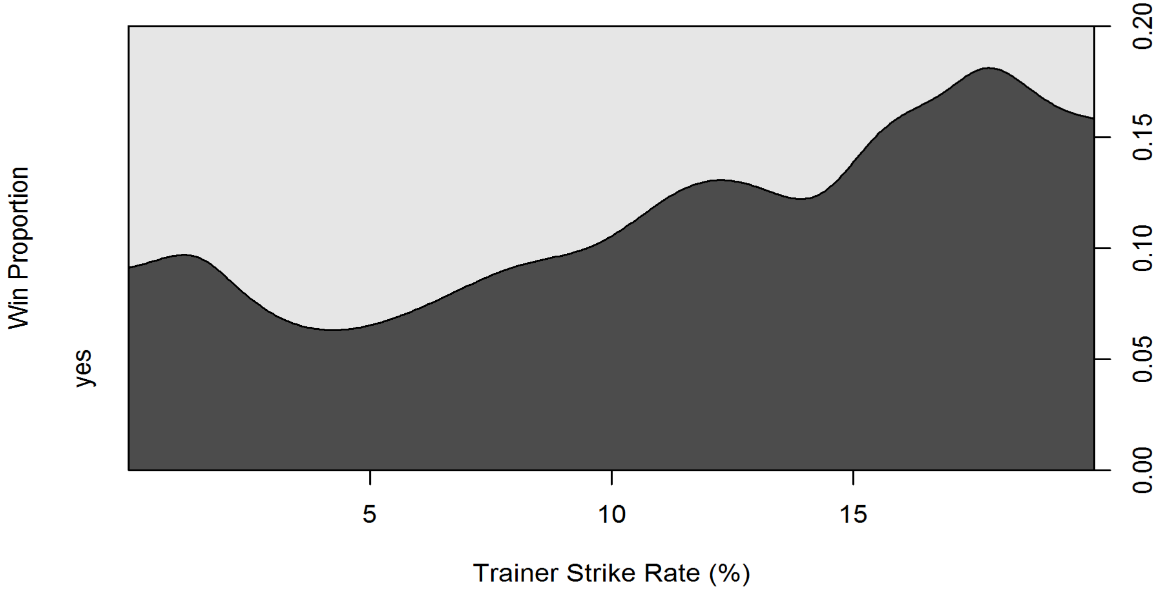
Win Proportion versus Age (Training Set)

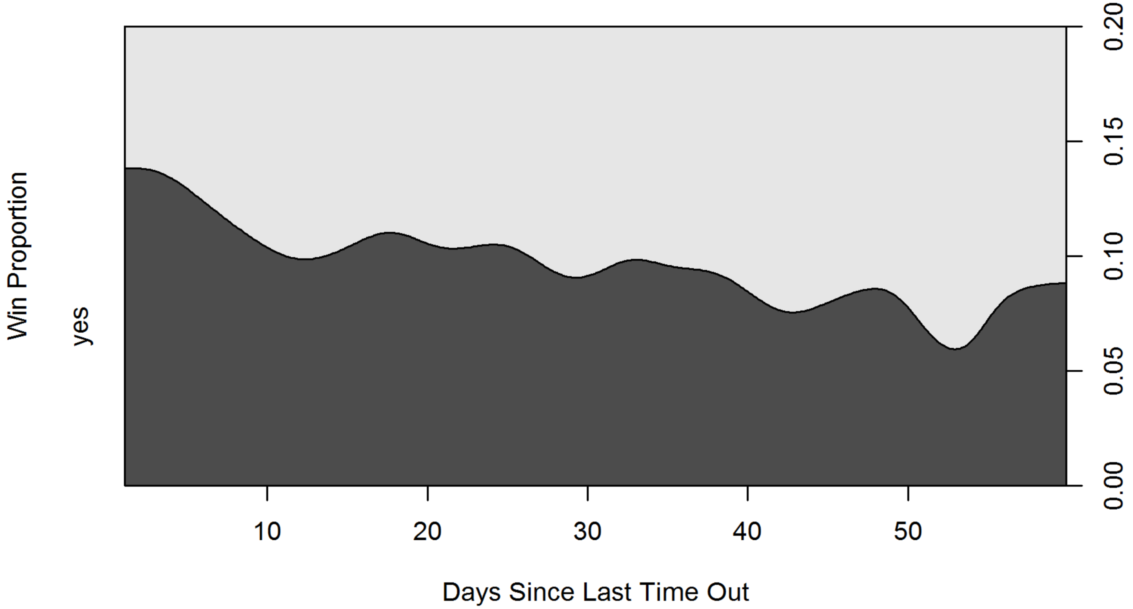
Hence define new variable: `age.diff=abs(age-4.5)`

Supports evidence in Gramm and Marksteiner (2011)

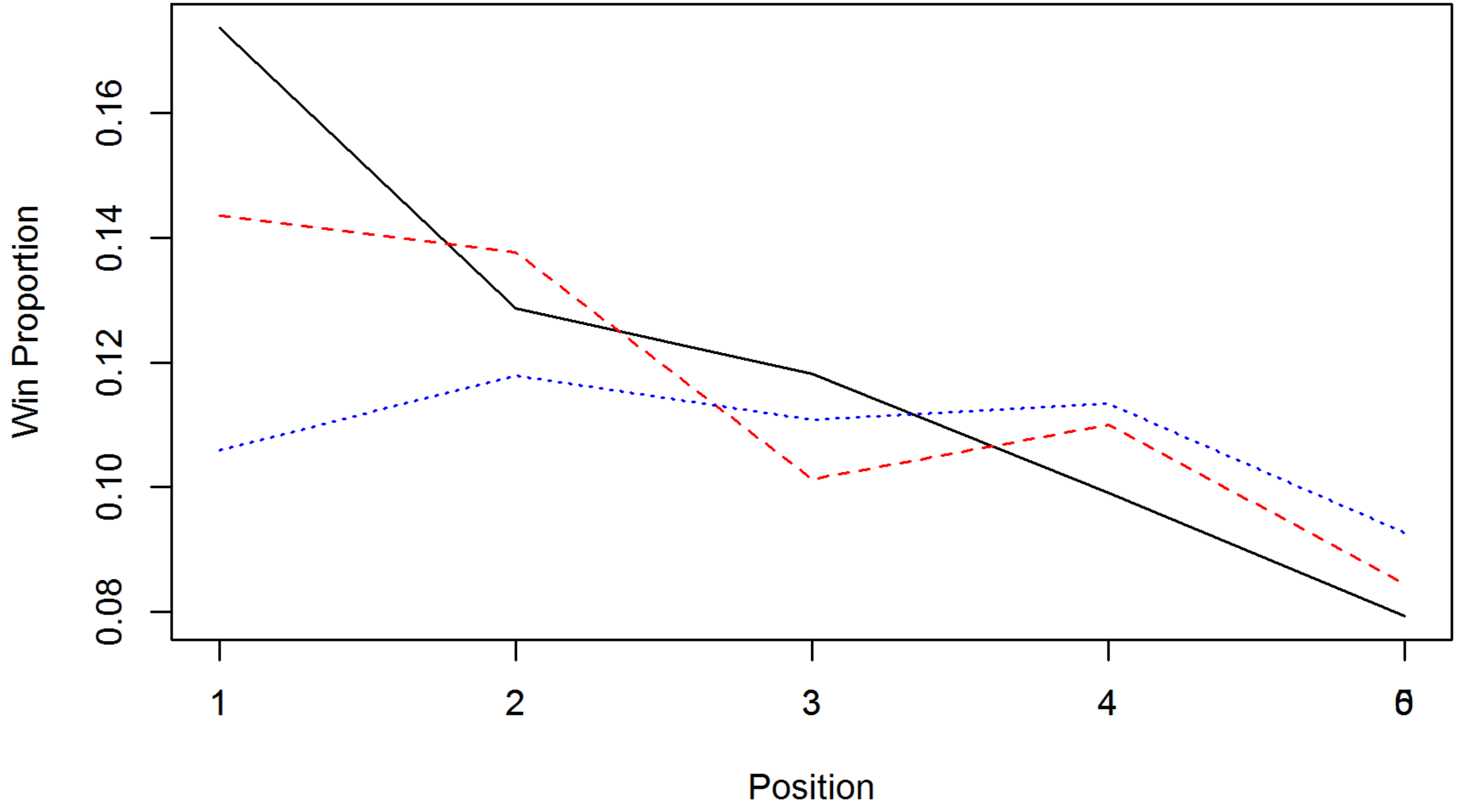


Win Proportion versus SireSR, TrainerSR, daysLTO (Training Set)





(Training Set) Win Proportion v Position in the horse's:
previous race (——)
two races ago (- - - - -)
three races ago (··········)



(Training Set) Win Proportion versus blinkers, visor, cheekpieces or tongue-tie

	Entire	Gelding	Blinkers	Visor	Cheek Pieces	Tongue Tie
Yes	0.115	0.106	0.111	0.103	0.069	0.084
No	0.099	0.091	0.100	0.101	0.103	0.102

Multinomial logistic regression model (Discrete choice models)

Consider "estimated" relative ratings or utilities, V_i, for horses i =1,…,n in a race

And "true" (unknown) ratings/utilities U_i, then:

$$U_i = V_i + \varepsilon_i,$$

ε_i is the (random) difference between the estimated and true ratings/utilities

Probability that horse i will win the race is:

$$P_i = \mathrm{Prob}\left(U_i > U_j \;\forall j \neq i\right)$$
$$= \mathrm{Prob}\left(V_i + \varepsilon_i > V_j + \varepsilon_j, \forall j \neq i\right).$$
$$= \mathrm{Prob}\left(\varepsilon_j < \varepsilon_i + V_i - V_j, \forall j \neq i\right).$$

This is the cumulative distribution of ε_j over all $j \neq i$.

The logistic model derived by assuming that ε_i follows an extreme value distribution (Gumbel distribution):

$$\mathrm{F}\left(\varepsilon_j\right) = \exp\left\{-\exp\left(-\varepsilon_j\right)\right\}$$

Multinomial logistic regression model
(Discrete choice model)

By making the assumption above, it can then be shown that the probability P_i that horse i will win a race involving n horses is given by:

$$P_i = \frac{\exp(V_i)}{\sum_{i=1}^{n} \exp(V_i)}.$$

We relate the rating/utility, V_i, for horse i to horse-specific variables (age, sireSR etc.) using

$$V_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip},$$

where $x_{i1,}\, x_{i2},\ldots, x_{ip}$ are the p horse-specific variables (age, sireSR etc.) for horse i

and $\beta_1, \beta_2, \ldots, \beta_p$ are model parameters to be estimated.

Specification in R Using `mlogit` package

```
mlogit(win~
  age.diff+sireSR+trainerSR+daysLTO+
  position1+position2+position3+entire
  +gelding+blinkers+visor+cheekpieces+tonguetie
|0|0,data=h.dat)
```

Alternative-specific variables are the **horse**-specific variables.

Individual-specific variables are the **race**-specific variables.

Often this is the source of confusion that prevents many implementing the multinomial logistic model for horse racing.

Specification in R

```
h.dat<- mlogit.data(data=model.data,
choice="win",chid.var="race.id",
alt.var="horse.ref",shape="long")
```

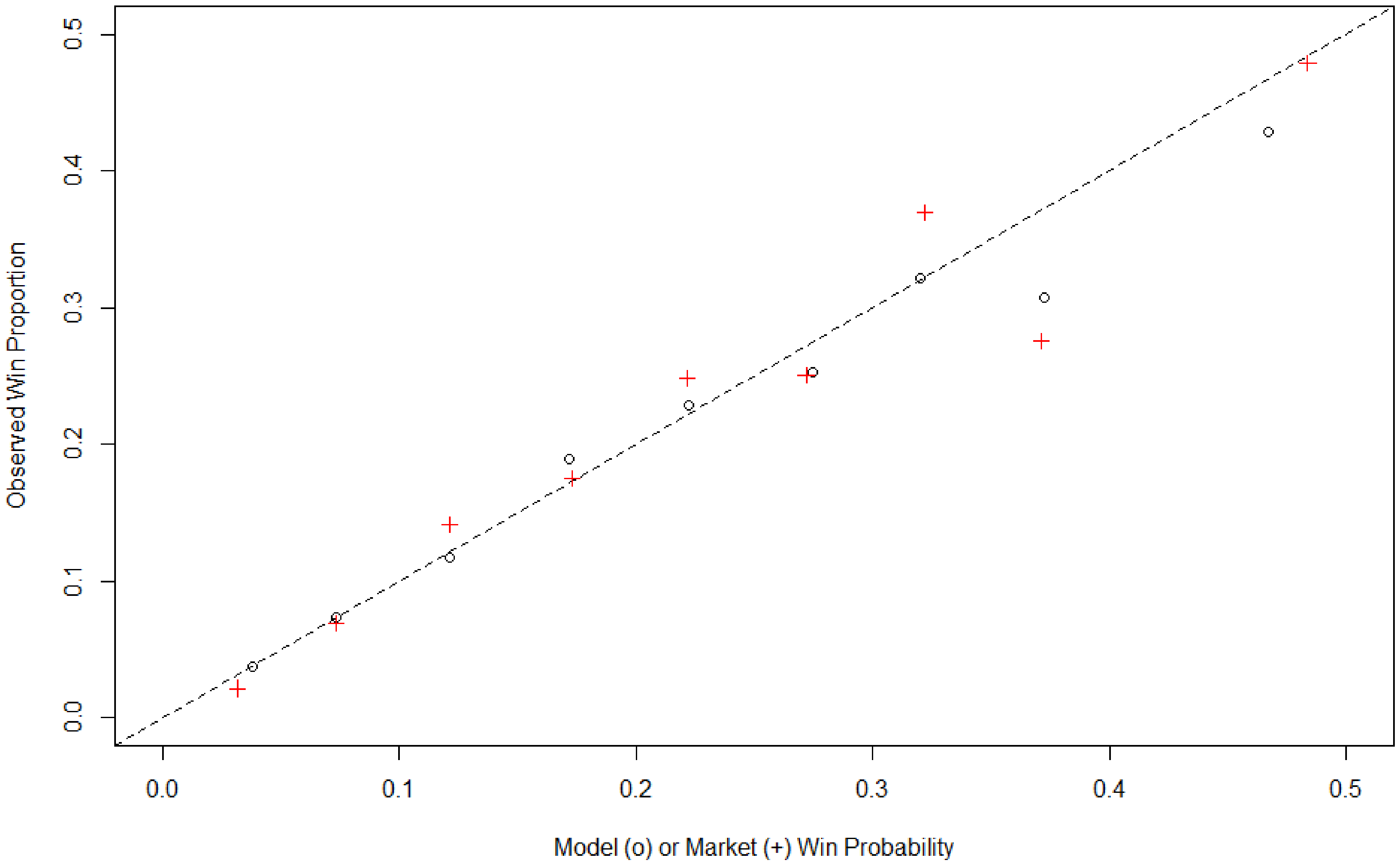
`choice` indicator of which horse won each race (in our data set this is the variable called `win`);

`chid.var` defines the choice sets (races) from which winner is chosen (in our data set this is `race.id`);

`alt.var` defines the choice alternatives (horses) in each set (race) (in our data set this is `horse.ref`)

Parameter		Estimate	Std. Error	p
age.diff		-0.153	0.0314	<0.001
sireSR		0.048	0.0093	<0.001
trainerSR		0.051	0.0093	<0.001
daysLTO		-0.004	0.0018	0.020
Position1	1	0.602	0.0919	<0.001
	2	0.324	0.1006	
	3	0.312	0.1027	
	4	0.159	0.1082	
Position2	1	0.368	0.0974	<0.001
	2	0.363	0.0982	
	3	0.066	0.1074	
	4	0.213	0.1050	
~~Position3~~	1	-0.046	0.1061	~~0.43~~
	2	0.109	0.1000	
	3	0.117	0.1036	
	4	0.130	0.1013	
entire		0.499	0.1297	<0.001
gelding		0.557	0.0948	<0.001
~~blinkers~~		~~0.016~~	~~0.1125~~	~~0.89~~
~~visor~~		~~0.027~~	~~0.1443~~	~~0.85~~
cheekpieces		-0.504	0.1470	0.001
~~tonguetie~~		~~-0.297~~	~~0.1632~~	~~0.069~~

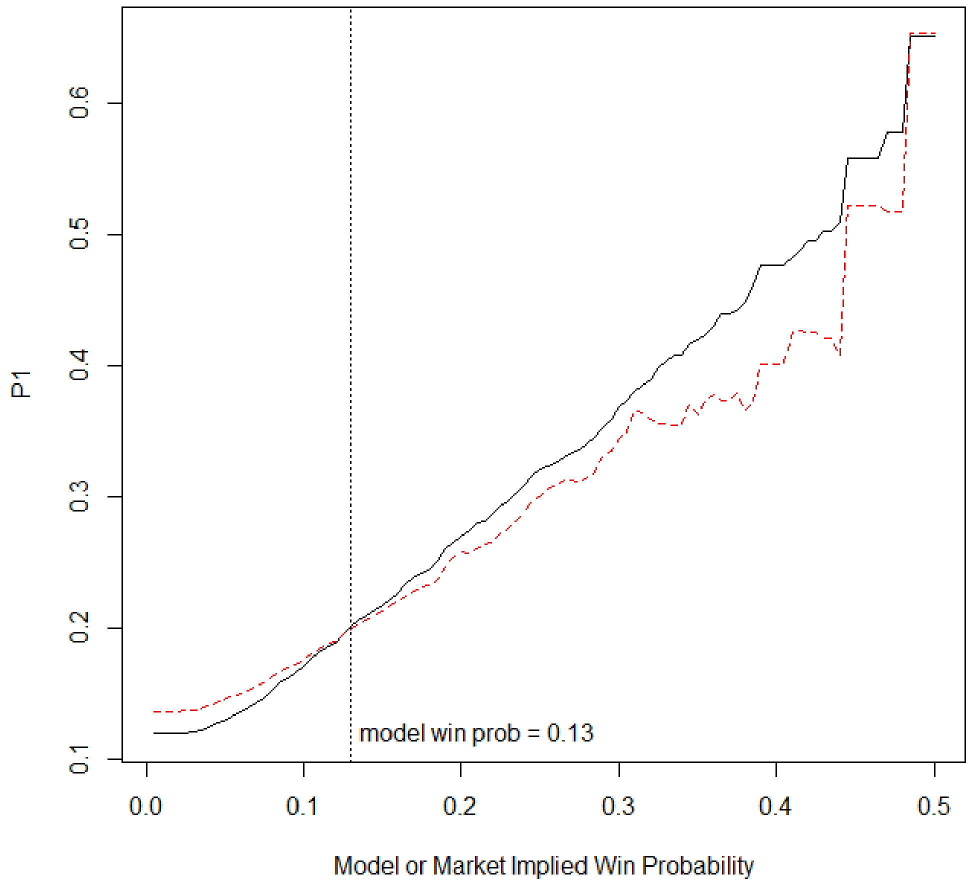
Calibration for the model (o) and market implied win probabilities (+)
Here we adjust market probabilities to account for Betfair Commission

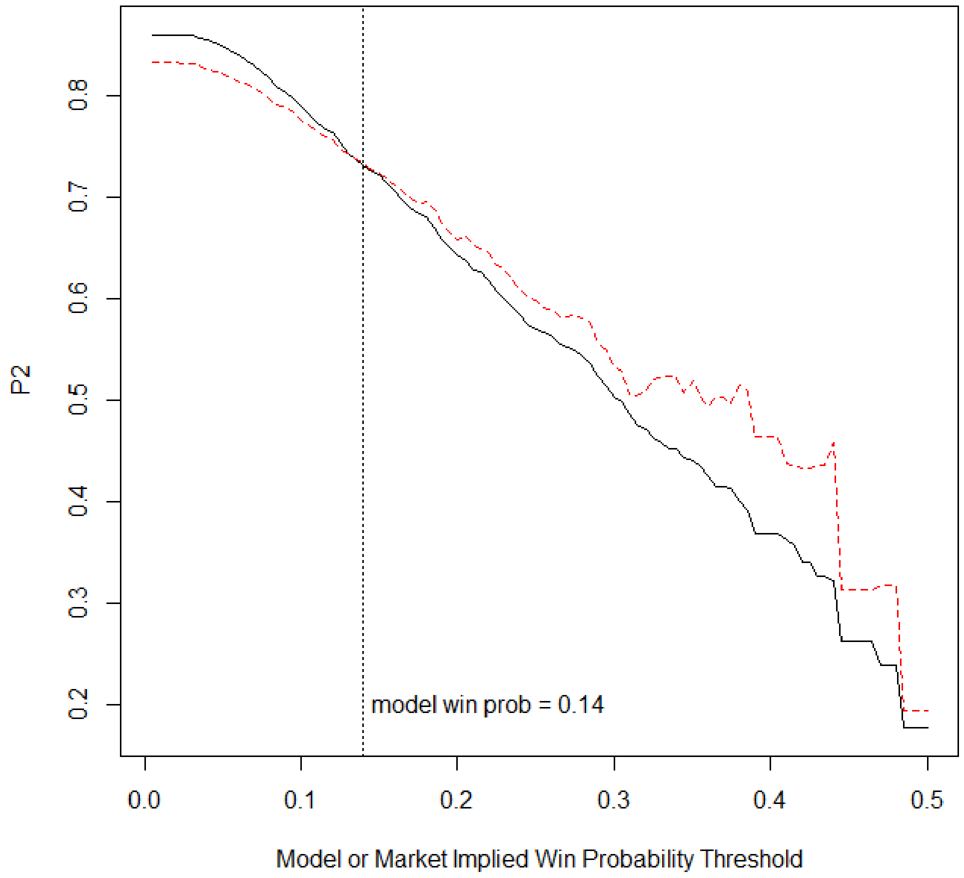


$P1$ and $P\,2$ v Model (——) and Market (------) Win Probabilities

$$P1 = \exp\left\{\frac{1}{N}\sum_{k=1}^{N}\log(P_{jk})\right\}$$

$$P2 = \exp\left\{\frac{1}{N}\sum_{k=1}^{N}\left[\log(1-P_{jk})^{2} + \sum_{i\neq j}^{n_k}\log(P_{ik})^{2}\right]\right\}$$



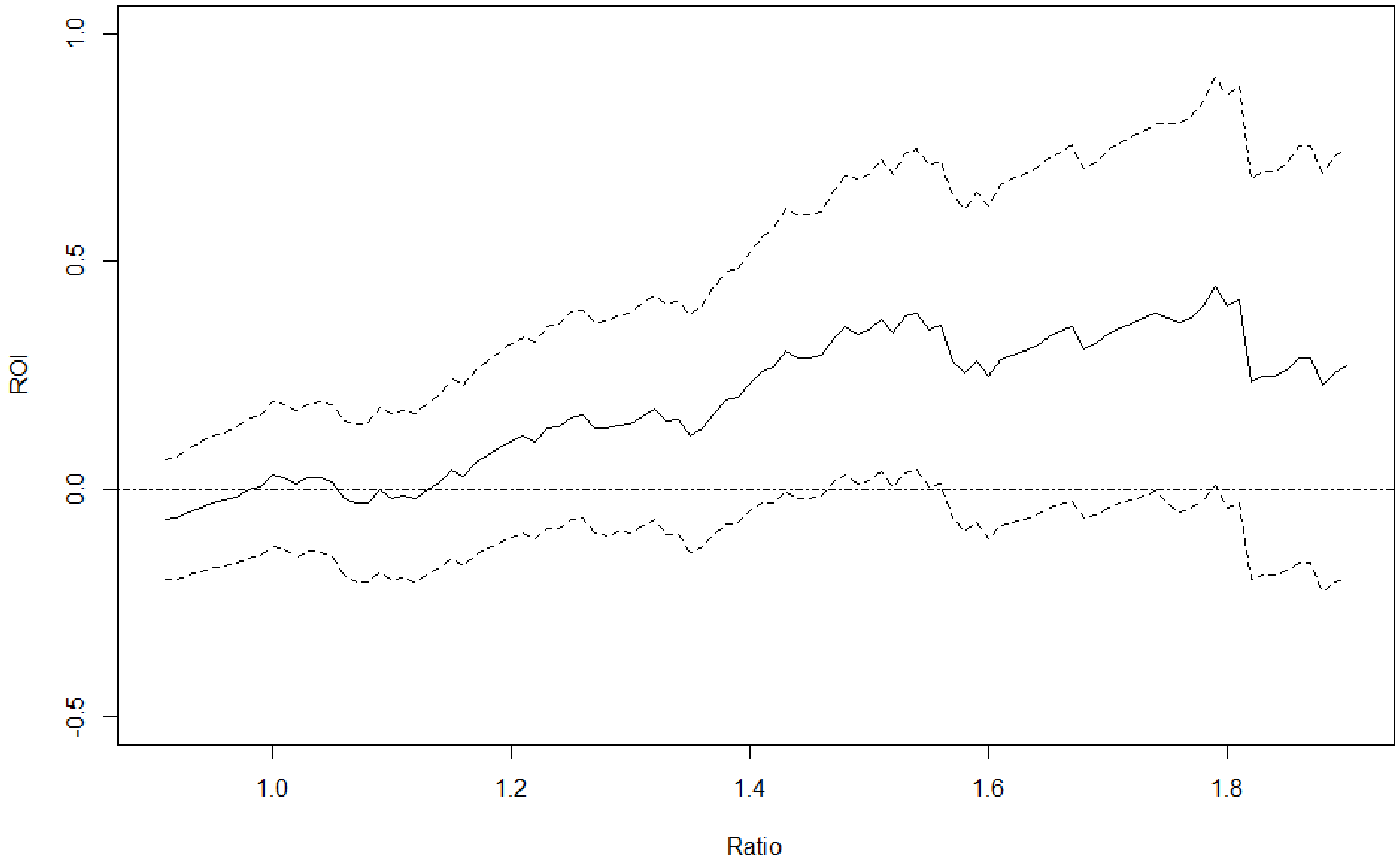


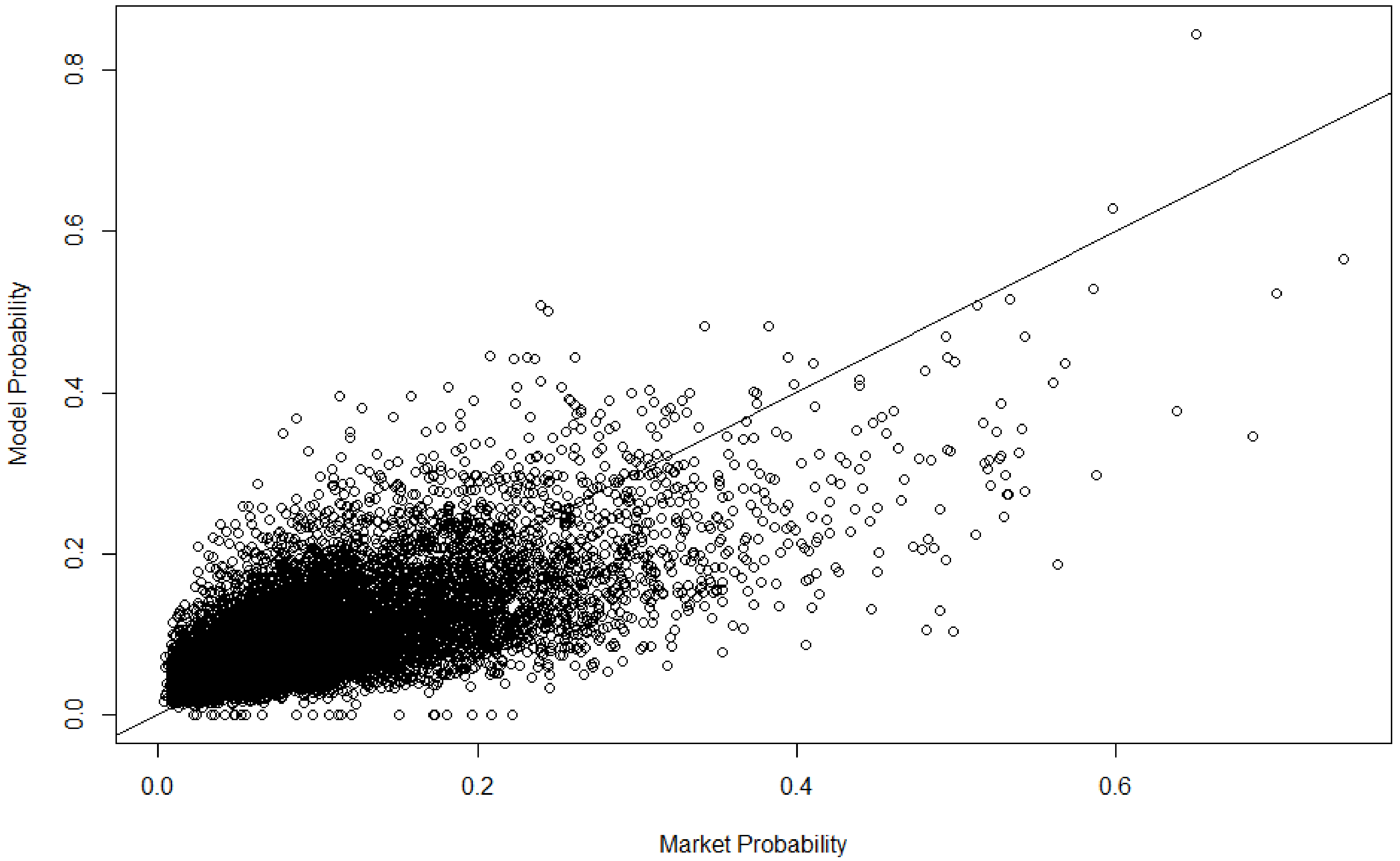
Betting on unseen data (Test Set)

Unit bets placed "virtually" on horses where:

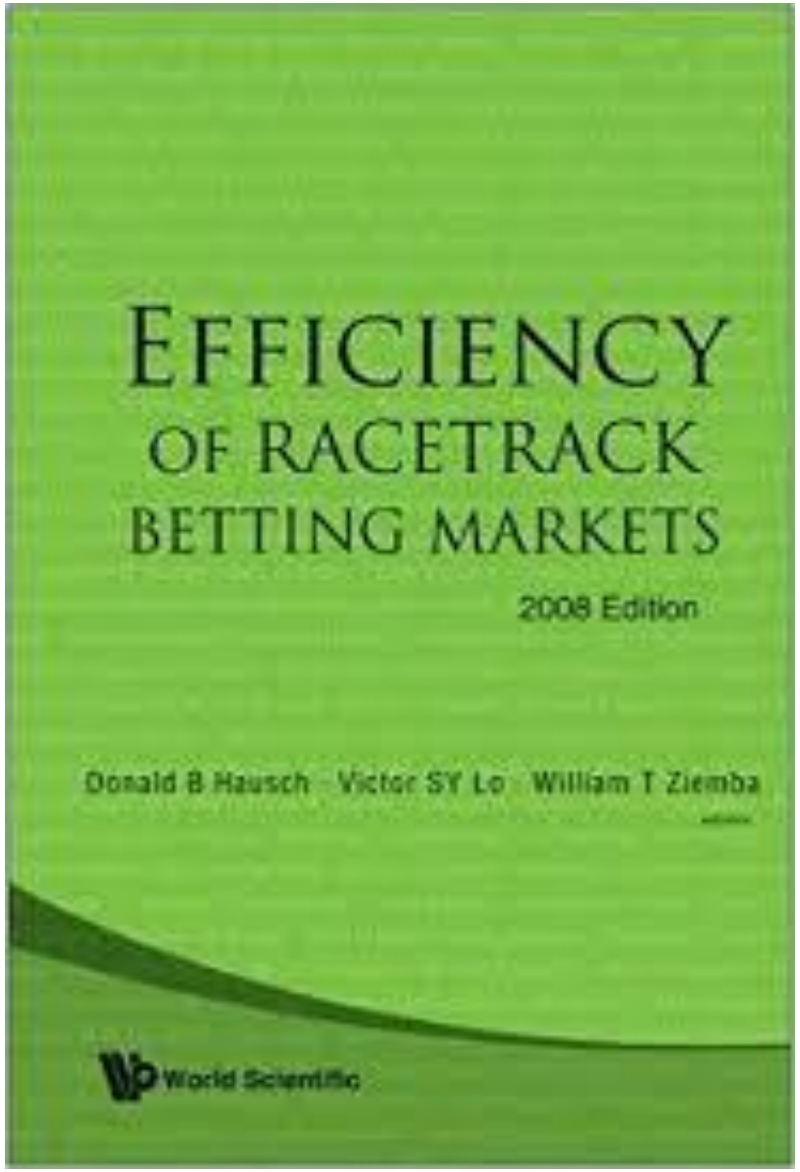
1. model win probability was greater than 0.15

2. ratio of win probability of model/market (adjusted for commission) > 1.3

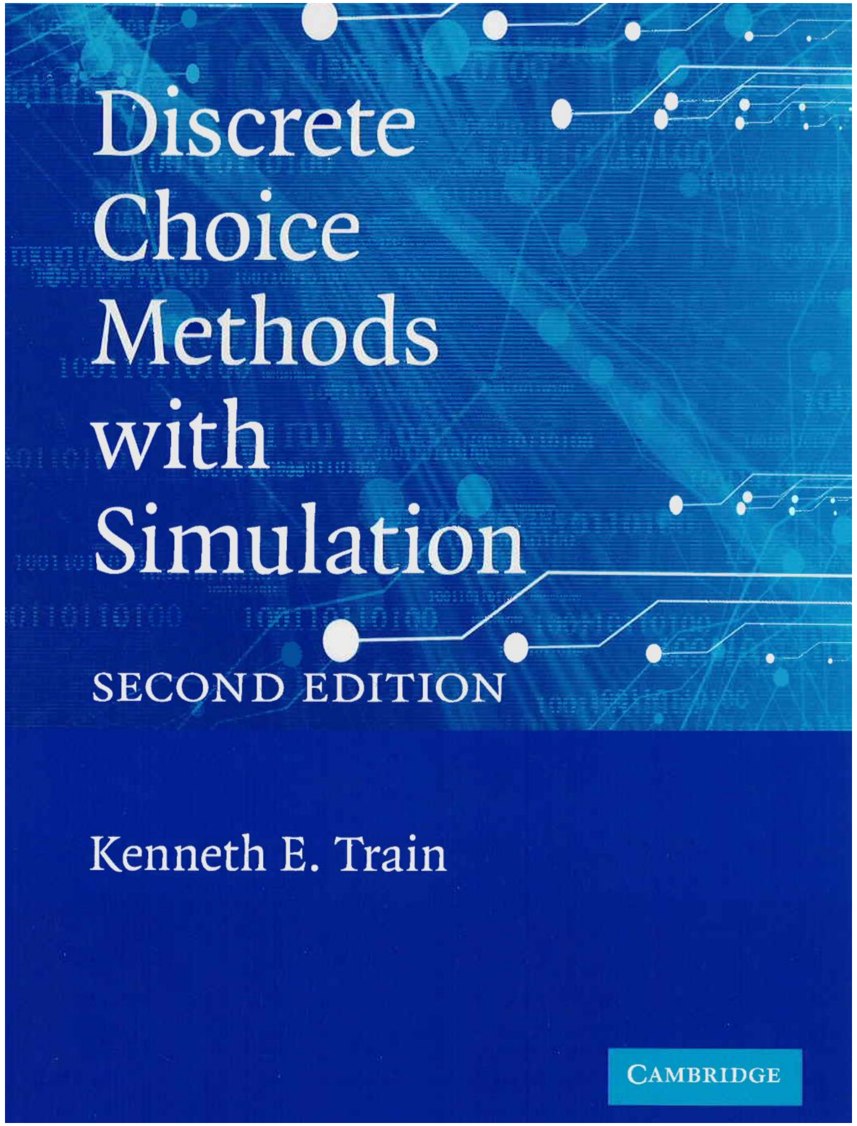



Model Probability
0.0
0.2
0.4
0.6
0.8
Market Probability
0.0
0.2
0.4
0.6

Efficiency of Race Track Betting Markets, eds. Haush, Lo and Ziemba.



Discrete Choice Methods and Simulation, Kenneth Train.



Independence from irrelevant alternatives (IIA)

$$\frac{P_i}{P_j} = \frac{\exp(V_i)}{\sum_{i=1}^{n} \exp(V_i)} \Bigg/ \frac{\exp(V_j)}{\sum_{i=1}^{n} \exp(V_i)} = \frac{\exp(V_i)}{\exp(V_j)}$$

Depends only on horses i and j

Suppose have three horses A, B and C with model win probabilities 0.4, 0.4, 0.2 and hence model implied (decimal) odds 2.5, 2.5, 5.0

If horse A becomes a non-runner the probabilities will change to 0.4/0.6=0.67 for B and 0.2/0.6=0.33 for C and hence odds of 1.5 and 3.0.

Need to be happy this is sensible??

Ευχαριστώ


THANK YOU FOR YOUR ATTENTION
PLEASE CLAP AND DO NOT MAKE TOUGH QUESTIONS
...emegenerator.net