Estimations for Statistical Arbitrage in Horse Racing Markets

XIONG, Liying

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Master of Philosophy
in
Risk Management Science

The Chinese University of Hong Kong
September 2010

Thesis/Assessment Committee

Professor Ben CHAN (Chair)

Professor Minggao GU (Supervisor)

Professor Samuel Po-Shing WONG (Committee Member)

Professor Bing-Yi JING (External Examiner)

The Chinese University of Hong Kong
Graduate School

The undersigned certify that we have read a thesis, entitled "Estimations for Statistical Arbitrage in Horse Racing Market" submitted to the Graduate School by XIONG, Liying (   ) in partial fulfillment of the requirements for the degree of Master of Philosophy in Statistics. We recommend that it be accepted.

Prof. Minggao GU
Supervisor

Prof. Samuel Po-Shing WONG

Prof. Ben CHAN

Prof. Bing-Yi JING
External Examiner

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Acknowledgements

I would like to express my gratitude to all those who have helped me during the writing of this thesis. I gratefully acknowledge the help of my supervisor Professor Minggao Gu. I do appreciate his patience, encouragement, and professional instructions during my thesis writing. Without his patient instruction, insightful criticism and expert guidance, the completion of this thesis would not have been possible.

Also, I deeply appreciate the devoted time, effort and support from the thesis committee Professor Samuel Po-Shing WONG, Professor Ben CHAN and external examiner Professor Bing-Yi JING. Their valuable suggestions benifited me a lot and helped my thesis reach its present form.

Abstract

Arbitrage opportunities exsit in every inefficient markets. Hong Kong horse racing and its related betting market have very long histories since 1884 and have trained a large number of sophisticated gamblers, who make the betting market nearly efficient. However arbitrage opportunities still exist. Before we can take advantage of these chances, an accurate model for the probabilities of horses to win should be developed. This article aims at providing such a computer based model by firstly constructing regression model and then combining estimate from this model and public estimate. We show that the combined estimate works better, under some very general conditions. Real data from Hong Kong horse racing are used to justify the claims.

摘　要

任何非有效市場都存在套利機會。香港賽馬運動以及與此緊密聯繫的賽馬博彩市場有著漫長的歷史，從 1884 賽馬會建立至今，它吸引了大量經驗豐富的賭博者，而這些參與者讓這個市場變得很接近有效市場。儘管如此，套利機會仍然存在。想要把握這些套利機會，我們需要一個準確的模型來估計每一匹馬贏得比賽的概率。本文旨在介紹一種以電腦程序為基礎的回歸模型用以估計馬屁贏得比賽的概率，並且嘗試將模型得出的估計量與公眾反映的估計量結合起來定義一個新的估計量。我們將證明新的估計量在實際中表現得更好，尤其是當某些特定情況發生的時候。我們將會用實際的賽馬數據來證實我們的論點。

Contents

Chapter 1

Introduction

Horse Racing is a very famous entertainment in Hong Kong and also popular in many countries around the world. Long history and large size of participants make horse racing markets nearly efficient. The implied win probabilities from odds shown on the tote board, which is essentially driven by many individual gamblers, is extremly accurate. Thus, the public implied win probabilities play a very important role in any regression model which attempts to accurately estimate the horse win probabilities. Some researches has shown certain ways of using odds as a regression factor or combining model estimate and public estimate. People did achieve a great improvement as they add odds as an predictor in regression model[1]. In his article[2], Benter has shown his achievement in combining model and public estimates. However, as people take public odds as an predictor or combine two estimates as a new one, they care little or nothing about the potential relationship between model estimate and public estimate. People seldomly ask the possible consequences of the scenario 1. if model estimate is larger than

[1] Bolton, Ruch N. and Randall G. Chapman (1986). Searching For Postitive Returns at the Track: A Multinomial Logit Model For Handicapping Horse Races.

White, E.M. , Dattero, R. , Flores, B. (1992). Combining vector forecasts to predict thoroughbred horse race outcomes.

[2] William Benter (1994). Computer Based Horse Race Handicapping and Wagering Systems: A Report.

public estimate or 2. if model estimate is smaller than the public counter part. Ignoring these differences would have fatal consequences for an investor. In this article,we are going to discuss how should the estimate be adjusted when one of the two above scenarios happens.

The rest of the thesis is organized as follows. Chapter 2 introduces the horse racing gambling machanism and some existed regression models on horse racing. Chapter 3 provides a possible methods to improve the estimation and especially produce new estimators under particular conditions. Chapter 4 shows performance of improved estimations. Chapter 5 concludes.

Chapter 2

Hong Kong Horse Racing Market and Models in Horse Racing

2.1 Hong Kong Horse Racing Market

Horse racing is an equestrian sport that has been practiced over the centuries. The British tradition of horse racing left its mark as one of the most important entertainment and gambling instituion in Hong Kong. Established as the Royal Hong Kong Jockey Club(HKJC) in 1884, currently it conducts over 700 races every season at two race tracks in Happy Valley and Sha Tin. Also, off-track betting is available from oversea bookmakers. Having attracted so many participants, Hong Kong horse racing wagering market is the largest per race in the world. As a non-profit organization, HKJC is the largest tax income contributor to the Hong Kong Government, and it is behind many social programs in Hong Kong.

The style of racing, the distances and the type of events varies very much by the country in which the races are occuring, and many countries offer different types of horse races. In Hong Kong, there is a great variety of betting pools, e.g. win(betting a horse winning the first), place(betting a horse winning the first, second or third), quinella(betting two horses winning the first two, ignoring the order), tierce(betting three horses winning the first, second and third in exact order) and some combinations of these, etc.

In Hong Kong horse racing, most pools are in a pari-mutual machanism(Jockey Challenge is of fixed-odds betting type). In these wagering markets, returns from investments are uncertain, which is essentially dependent on winning probabilities and odds. The number of participants is huge and there is a variety of information concerning investments and participants. Thus, efficiency of the wagering market is of interest. An efficient market, in economic sense, is a market in which prices(odds particularly in horse racing) reflect all relevant information and no arbitrage opportunities exist. Therefore, if a profitable betting strategy depending on odds and other publicly available information can be found, the market is inefficient.

As mentioned before, horse race wagering market adopt the pari-mutuel system: Let $B_i, i = 1, 2, 3, ..., I$ be the amount of money wagered on the horse i in the win pool, then the payoff or odds if horse i won the race is

$$Odd_i = (1-\rho)\frac{B_1 + ... + B_I}{B_i}, \tag{2.1}$$

where ρ is the track take percentage. ($\rho = 17.5\%$) The odds, which can reflect the public's attitudes toward horses. We can summerize them into the "public win probability estimate", which can be written as:

$$\begin{aligned} P_i^{pub} &= \frac{B_i}{B_1 + ... + B_I} \\ &= \frac{1/Odd_i}{1/Odd_1 + ... + 1/Odd_I} \\ &= \frac{1-\rho}{Odd_i} \end{aligned} \tag{2.2}$$

Historical data has shown that there is no obvious long or short price bias by the public estimate in Hong Kong[1]. The public implied win probabilities are extremely accurate. The wagering market is close to efficient. However, a little inefficiency can bring golden arbitrage opportunities. A prerequisite for developing a profitable betting strategy is to construct an accurate model for predicting the outcomes in races.

2.2 Models in Horse Racing

Several models have been already constructed to estimate each horse's current performance potential."Current performance potential" is a single overall summary index of a horse's expected performance in a particular race. To construct a

[1] K.Busche (1994). Efficient Market Results in an Asian setting.

model to estimate current performance potential, one must investigate the available data to find those variables or *factors* which have predictive significance. The profitability of the resulting betting system will be largely determined by the predictive power of the factors chosen. Various types of factors can be classified into groups:

Current condition:

- performance in recent races
- time since last race
- age of horse

Past performance:

- finishing position in past races
- lengths behind winner in past races
- normalized times of past races

Adjustments to past performance:

- strength of competition in past races
- weight carried in past races
- jockey's contribution to past performances

Present race situational factors:

- weight to be carried
- today's jockey's ability
- advantages or disadvantages of the assighed post position

Performances which could influence the horse's performance in today's race:

- distance preference
- surface preference(turf vs dirt)
- condition of surface preference(wet vs dry)

It is important to define factors which extracts as much imformation as possible out of the data in each of the relevant areas. And also, the general thrust of model development is to continually experiment with refinements of various factors. Although time-comsuming, the gains are worthwhile.

It can be presumed that valid fundamental information exists which can not be systematically or parctically incorporated into a statistical model. Therefore, any statistical model, however well developed, will always be incomplete. Since the sophisticated public's implied probability estimates generally correspond well with the actual frequencies of winning. An extremely important step in model development is the estimation of the relation of the model's probability estimates to the public's estimates, and the adjustment of the model's estimates to incorporate whatever information can be gleaned from the public's estimates. In a sense, what is needed is a way to combine the judgements of two experts. Some practical techniques for accomplishing this have been provided. Here is one from White, Dattero and Flores(1992):

Estimate a second logit model using two probability estimates as independent variables. For a race with entrants$(1, 2, ..., N)$ the win probability of horse i is given by:

$$c_i = \frac{\exp(\alpha f_i + \beta\pi_i)}{\Sigma\exp(\alpha f_j + \beta\pi_j)}, \tag{2.3}$$

where

f_i = log of fundamental model probability estimate

π_i = log of public's implied probability estimate

c_i = combined probability estimate

(Natural log of probability is used rather than probability as this transformation provides a better fit)

Given a set of past races$(1, 2, ..., R)$ for which both public probability estimates and fundamental model estimates are available, the parameters α and β can be estimated by maximizing the log likelihood function of the given set of races with respect to α and β:

$$\exp(L) = \Pi c_{ji_j^*} \qquad (j = 1 \mathrm{to} R) \tag{2.4}$$

where $c_{ji_j^*}$ denotes the probability as given by equation (2.3) for the horse i_j^* observed to win race j.

Chapter 3

Probit Regression Model Incorporating with Public Estimates

Though a variety of techniques provided in literature have shown a great improvement to fundamental models' estimates. There is still one important and interesting question, which the author believes has been generally overlooked in the literature. Should the estimate be adjusted when particular condition is detected? Specifically speaking, should the estimate adopt different forms when model estimate is detected to be larger than public estimate and what happens if it is less? In this chapter, we try to answer this question. Firstly, we investigate into our fundamental model, and then try to combine two estimates of the *strength* of each horse rather than the winning probability so as to estimate the winning probabilities based on strength argument.

3.1 Estimation under No Particular Conditions

The fundamental model we choose is probit regression model, and we use the *strength* of a horse as the reponse variable, rather than directly using the win probability of this horse, where the strength of horse i(denoted as S_i) represents the horse's performance potential. Horses with relatively higher strength value are more likely to win over others with relatively lower strength value. How to obtain the strength data will be further discussed in Chapter 4. In this chapter, we assume the strength data is available. By probit regression model, we assume the strength S_i of horse i follows:

$$\begin{aligned} S_i &= \beta_0 + \beta_1 F_{i1} + \ldots + \beta_p F_{ip} + \epsilon_i, \quad \epsilon_i \sim N(0,1) \\ &= s_i + \epsilon_i \end{aligned} \tag{3.1}$$

or

$$\boldsymbol{S} = \boldsymbol{F\beta} + \boldsymbol{\epsilon} \ , \tag{3.2}$$

where F_{i1},...,F_{ip} are factors whose information is available to public.

This probit regression model can be fitted from historical data. We leave the detail procedures to chapter 4.

We assume our model estimate to be unbiased:

$$E\hat{S}_i = ES_i \tag{3.3}$$

Denote $X = \hat{S}_i$ [1]. Since current factor values are available, when the predicted value X is observed to be x_i for horse i in real situation, if no extenal information

[1] the subscript i of X is omited here for convenience. Unless otherwise noted, X;Y,Z and W refer to horse i.

is added, the resulting win probability of horse i is estimated by:

$$\hat{P}'_i = P(x_i + \epsilon_i > x_k + \epsilon_k, \quad \text{for all } k \neq i) \tag{3.4}$$

where ϵ_i and ϵ_k are independent errors, both following $N(0, 1)$.

Denote Y as the implied public estimate of the strength of horse i from odds. Since odds are available on HKJC website, Y is always observable. Now if Y and X have both been observed to be y_i and x_i for horse i, we intend to find a appropriate combination, commonly linear combination $\alpha x_i + \beta y_i$, such that the new win probability estimate is:

$$\hat{P}_i = P(\alpha x_i + \beta y_i + \epsilon_i > \alpha x_k + \beta y_k + \epsilon_k, \quad \text{for all } k \neq i) \tag{3.5}$$

We assume that X and Y follow joint normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}\right),$$

where we assume X and Y are both unbiased estimator of μ, which is the *true* expected current strength of Horse i. This assumption is reasonable since: $EX = E\hat{S}_i = ES_i$ and public estimate is precise in win probability. Thus it is also believed to be precise in estimating the strength of horse since the strength estimate is actually implied by the win probability estimate. Σ can be estimated from historical data and assumed to be known. We aim at constructing a better estimator by combining X and Y, i.e. $\lambda_1 X + \lambda_2 Y$, such that this new estimator would have a minimum"error".

Among all these linear combinations, we constrain them to be unbiased, which gives the expression $\lambda X + (1 - \lambda)Y$. Further, we need it to have the minimum variance. Simply solve the following problem:

$$\min_{\lambda} \operatorname{var}[\lambda X + (1-\lambda)Y] \tag{3.6}$$

The minimum variance is achieved at

$$\lambda^* = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}} \tag{3.7}$$

Thus $\lambda^* X + (1-\lambda^*)Y$ is the estimator with minimum variance among all unbiased linear combinations.

However, unbiased estimator does not necessarily perform better than biased estimators. Some other factors, such as variance, should be taken into account simultaneously. It is sometimes the case that a trade-off occurs between variance and bias in such a way that a small increase in bias can be traded for a larger decrease in variance, resulting in an improvement in mean square error(MSE)[2].

Here, we use MSE as our primary criteria to measure the goodness of estimators. Under this criteria, our problem becomes how to decide appropriate λ_1 and λ_2 in $\lambda_1 X + \lambda_2 Y$ such that it can achieve a minimum MSE. We want to solve the following problem:

$$\min_{\lambda_1, \lambda_2} E[(\lambda_1 X + \lambda_2 Y - \mu)^2]. \tag{3.8}$$

[2] MSE is defined as MSE $= E(est - m)^2$, where est stands for estimator and m stands for the true value estimated.

Let $G(\lambda_1, \lambda_2) = E[(\lambda_1 X + \lambda_2 Y - \mu)^2]$, we want to minimize G.

Firstly, rewrite G by following:

$$\begin{aligned} G(\lambda_1, \lambda_2) &= E[(\lambda_1 X + \lambda_2 Y - \mu)^2] \\ &= E(\lambda_1^2 X^2 + \lambda_1^2 Y^2 + \mu^2 2\lambda_1\lambda_2 XY - 2\lambda_1 \mu X - 2\lambda_2 \mu Y) \\ &= \lambda_1^2 EX^2 + \lambda_2^2 EY^2 + 2\lambda_1\lambda_2 E(XY) - 2\lambda_1 \mu EX - 2\lambda_2 \mu EY + \mu^2 \\ &= \lambda_1^2(\sigma_X^2 + \mu^2) + \lambda_2^2(\sigma_Y^2 + \mu^2) + 2\lambda_1\lambda_2(\sigma_{XY} + \mu^2) - 2\lambda_1\mu^2 - 2\lambda_2\mu^2 \\ &\quad +\mu^2. \end{aligned} \tag{3.9}$$

Then we minimize it by setting partial derivatives to be zero:

$$\begin{aligned} \frac{\partial G}{\partial \lambda_1} &= 2\lambda_1(\sigma_X^2 + \mu^2) + 2\lambda_2(\sigma_{XY} + \mu^2) - 2\mu^2 = 0 \\ \frac{\partial G}{\partial \lambda_2} &= 2\lambda_2(\sigma_Y^2 + \mu^2) + 2\lambda_1(\sigma_{XY} + \mu^2) - 2\mu^2 = 0, \end{aligned} \tag{3.10}$$

which gives unique pair

$$(\lambda_1^*, \lambda_2^*) = \left(\frac{(\sigma_Y^2 - \sigma_{XY})\mu^2}{\sigma_X^2\sigma_Y^2 + (\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\mu^2}, \frac{(\sigma_X^2 - \sigma_{XY})\mu^2}{\sigma_X^2\sigma_Y^2 + (\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\mu^2}\right) \tag{3.11}$$

which minimizes G.

To verify the result above, we check the *Jacobian* $J(\lambda_1, \lambda_2)$:

$$\begin{aligned}
\frac{\partial^2 G}{\partial \lambda_1^2} &= 2(\sigma_X^2 + \mu^2) \geq 0 && (3.12)\\
\frac{\partial^2 G}{\partial \lambda_2^2} &= 2(\sigma_Y^2 + \mu^2) \geq 0 && (3.13)\\
\frac{\partial^2 G}{\partial \lambda_1 \partial \lambda_2} &= 2(\sigma_{XY} + \mu^2) \\
\frac{\partial^2 G}{\partial \lambda_1^2}\frac{\partial^2 G}{\partial \lambda_2^2} - \left(\frac{\partial^2 G}{\partial \lambda_1 \partial \lambda_2}\right)^2 &= 4(\sigma_X^2 + \mu^2)(\sigma_Y^2 + \mu^2) - 4(\sigma_{XY} + \mu^2)^2 \\
&= 4[\sigma_X^2\sigma_Y^2 + (\sigma_X^2 + \sigma_Y^2)\mu^2 + \mu^4 - \sigma_{XY}^2 - 2\sigma_{XY}\mu^2 - \mu^4] \\
&= 4[\sigma_X^2\sigma_Y^2 + (\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\mu^2 - \sigma_{XY}^2] \\
&= 4[(\sigma_X^2\sigma_Y^2 - \rho^2\sigma_X^2\sigma_Y^2) \\
&\quad +(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y)\mu^2], && (3.14)
\end{aligned}$$

where ρ is the correlation coefficient of X and Y.

Since $0 \leq \rho \leq 1$, then $-1 \leq \rho^2 \leq 0$. Therefore,

$$\begin{aligned}
\frac{\partial^2 G}{\partial \lambda_1^2}\frac{\partial^2 G}{\partial \lambda_2^2} - \left(\frac{\partial^2 G}{\partial \lambda_1 \partial \lambda_2}\right)^2 &= 4[(\sigma_X^2\sigma_Y^2 - \rho^2\sigma_X^2\sigma_Y^2) + (\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y)\mu^2] \\
&\geq 4[(\sigma_X^2\sigma_Y^2 - \sigma_X^2\sigma_Y^2) + (\sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y)\mu^2] \\
&= 4(\sigma_X - \sigma_Y)^2\mu^2 \\
&\geq 0. && (3.15)
\end{aligned}$$

Therefore, from (3.12), (3.13) and (3.14), we know that $J(\lambda_1, \lambda_2)$ is positively definite and thus $J(\lambda_1^*, \lambda_2^*)$ is postitively definite, which implies G reaches its minumum at $(\lambda_1^*. \lambda_2^*)$.

Unfortunately, since MSE is a function of the parameter, there will not be one overall "best" estimator in general. Here we notice that

$$(\lambda_1^*, \lambda_2^*) = \left(\frac{(\sigma_Y^2 - \sigma_{XY})\mu^2}{\sigma_X^2\sigma_Y^2 + (\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\mu^2}, \frac{(\sigma_X^2 - \sigma_{XY})\mu^2}{\sigma_X^2\sigma_Y^2 + (\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\mu^2}\right)$$

are both dependent on μ, which is unknown. Therefore, $\lambda_1^*X + \lambda_2^*Y$ is not an estimator. Problem (3.8) has no uniformly minimum solution as we expected. An intuitional idea to solve this problem is replacing μ in $(\lambda_1^*, \lambda_2^*)$ by $\lambda^*X + (1-\lambda^*)Y$, which we derived previously. Since $\lambda^*X + (1-\lambda^*)Y$ is unbiased and has minimum variance, it may approximate μ well. Thus, the best estimator under MSE is approximated by:

$$\begin{aligned} & \lambda_{1,appr}^*X + \lambda_{2,appr}^*Y \\ = \; & \frac{(\sigma_Y^2 - \sigma_{XY})\left(\lambda^*X + (1-\lambda^*)Y\right)^2}{\sigma_X^2\sigma_Y^2 + (\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\left(\lambda^*X + (1-\lambda^*)Y\right)^2}X \\ & + \frac{(\sigma_X^2 - \sigma_{XY})\left(\lambda^*X + (1-\lambda^*)Y\right)^2}{\sigma_X^2\sigma_Y^2 + (\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\left(\lambda^*X + (1-\lambda^*)Y\right)^2}Y \end{aligned} \qquad (3.16)$$

where $\lambda^* = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$.

Remark: *appr* here stands for *approximate*.

3.2 Estimators under Particular Condition

In previous section, we have already constructed a new estimator, approximated by $\lambda_{1,appr}^*X + \lambda_{2,appr}^*Y$, which has a minimum mean square error among all linear combination of X and Y. In real situations, we do not use this result directly, especially when horse racing is concerned here. Before continuing discussion, we first do some investigation into the relationship between X and Y.

Case 1: $X \leq Y$

In this situation, the resulting public's estimated win probability is higher than ours. Thus, in our point of view, the public has over-estimated. As a result, public will over bet on this horse, which will drive the odd lower than "fair". Hence the public *advantage*[3]($Adv_{pub} = P_{pub} * odd$) of this horse will possibly drop dramtically. As we pointed out before, because of the tax, public is generally always losing. So the public advantage will not exceed 1 on average. Not to mention our advantage($Adv_{model} = P_{model} * odd < P_{pub} * odd = Adv_{pub}$). Since our advantage is less than 1, we won't bet.

Case 2: $X > Y$

Inversely, in this situation, our resulting estimated win probability is higher than public's, which means our model tells us that the public is under-estimating. As shown previously, this time, the public will bet less than "fair" on this horse, which drives the odd relatively higher. As a result, our advantage increases. Particularly, when $Adv > 1$, we bet.

The best chance comes when the condition $X > Y$ is fulfilled, and we can ourperform the public. Then our problem now becomes how to estimate μ more precisely and accurately using both X and Y under this particular condition. Again, we use "mean square error" to measure. Thus, we intend to minimize the "mean square error" of the estimator $\alpha'X + \beta'Y$ with respect to α' and β', under condition $X > Y$. However, we prefer to provide a more general result under a more general condition $X > aY + b$. Then the problem is as follow:

$$\min_{\alpha',\beta'} E[(\alpha'X + \beta'Y - \mu)^2 \mid X > aY + b]. \tag{3.17}$$

where a and b are given known constants set according to our preference.

Like all other conditional expection problems of normal variables, we can't find a

[3] *advantage* is defined as the expected payoff on a 1-dollar bet: $Adv = P * odd$. Thus if $Adv > 1$, it's a profitable game to play.

solutuion of closed form generally. To solve (3.17), we need numerical motheds. But before that, in next steps, we firstly try to simplify it by using independent properties.

Now, we define two new variables $Z = X - aY - b$ and $W = AX + BY$, following normal distributions, since X and Y follow joint normal distribution. Further, we choose appropriate A and B to make Z and W to be independent. Here we can set:

$$A = a\sigma_Y^2 - \sigma_{XY} \text{ and } B = \sigma_X^2 - a\sigma_{XY}.^4 \tag{3.18}$$

New random variables Z and W are used here instead of X and Y. Problem (3.18) is then equivalent to solve for α^* and β^*, such that:

$$F(\alpha^*, \beta^*) = \min_{\alpha,\beta} E[(\alpha Z + \beta W + \alpha b - \mu)^2 \mid Z > 0].^5 \tag{3.19}$$

where Z and W are independent nomal random variables, which follow:

$$\begin{pmatrix} Z \\ W \end{pmatrix} \sim N\left(\begin{pmatrix} (1-a)\mu - b \\ (A+B)\mu \end{pmatrix}, \begin{pmatrix} \sigma_Z^2 & 0 \\ 0 & \sigma_W^2 \end{pmatrix}\right),$$

where $\sigma_Z^2 = \sigma_X^2 + a^2\sigma_Y^2 - 2a\sigma_{XY}$ and $\sigma_W^2 = A^2\sigma_X^2 + B^2\sigma_Y^2 + 2AB\sigma_{XY}$.

Denote:

$$F[\alpha, \beta] = E[(\alpha Z + \beta W + \alpha b - \mu)^2 \mid Z > 0]. \tag{3.20}$$

[4] A and B are not unique, but only need to satisfy: $A(\sigma_X^2 - a\sigma_{XY}) = B(a\sigma_Y^2 - \sigma_{XY})$.

[5] αb is added here to eliminate μ inside Z, such that $\alpha Z + \beta W + \alpha b$ is an estimator equivalent to the linear combination of X and Y.

Then

$$\begin{aligned} F &= E[\alpha^2(Z+b)^2 + \beta^2 W^2 + 2\alpha\beta(Z+b)W - 2\alpha\mu(Z+b) - 2\beta\mu W + \mu^2 \mid Z > 0] \\ &= \alpha^2 E[(Z+b)^2 \mid Z>0] + \beta^2 E[W^2 \mid Z>0] + 2\alpha\beta E[(Z+b)W \mid Z>0] \\ &\quad -2\alpha\mu E[(Z+b) \mid Z>0] - 2\beta\mu E[W \mid Z>0] + \mu^2 \qquad (3.21) \end{aligned}$$

Since W is independent of Z, and $W \sim N((A+B)\mu, \sigma_W^2)$,

$$\begin{aligned} F &= \alpha^2 E[(Z+b)^2 \mid Z>0] + \beta^2[\sigma_W^2 + (A+B)^2\mu^2] + 2\alpha\beta(A+B)\mu E[(Z+b) \mid Z>0] \\ &\quad -2\alpha\mu E[(Z+b) \mid Z>0] - 2\beta(A+B)\mu^2 + \mu^2 \end{aligned} \qquad (3.22)$$

Denote the first and second conditional moments of $Z+b$ as

$$M_1 = E[(Z+b) \mid Z>0] \text{ and } M_2 = E[(Z+b)^2 \mid Z>0] \qquad (3.23)$$

Then,

$$\begin{aligned} F &= \alpha^2 M_2 + \beta^2[\sigma_W^2 + (A+B)^2\mu^2] + 2\alpha\beta(A+B)\mu M_1 \\ &\quad -2\alpha\mu M_1 - 2\beta(A+B)\mu^2 + \mu^2 \\ &= C_{\alpha^2}\alpha^2 + C_{\beta^2}\beta^2 + C_{\alpha\beta}\alpha\beta + C_\alpha\alpha + C_\beta\beta + \mu^2 \qquad (3.24) \end{aligned}$$

where

$$\begin{aligned}
C_{\alpha^2} &= M_2 \\
C_{\beta^2} &= \sigma_W^2 + (A+B)^2\mu^2 \\
C_{\alpha\beta} &= 2(A+B)\mu M_1 \\
C_{\alpha} &= -2\mu M_1 \\
C_{\beta} &= -2(A+B)\mu^2. \qquad (3.25)
\end{aligned}$$

To minimize $F(\alpha, \beta)$, we set:

$$\begin{aligned}
\frac{\partial F}{\partial \alpha} &= 2C_{\alpha^2}\alpha + C_{\alpha\beta}\beta + C_{\alpha} = 0 \\
\frac{\partial F}{\partial \beta} &= 2C_{\beta^2}\beta + C_{\alpha\beta}\alpha + C_{\beta} = 0, \qquad (3.26)
\end{aligned}$$

which gives unique pair

$$\begin{aligned}
(\alpha^*, \beta^*) &= \left(\frac{2C_{\alpha}C_{\beta^2} - C_{\beta}C_{\alpha\beta}}{C_{\alpha\beta}^2 - 4C_{\alpha^2}C_{\beta^2}}, \frac{2C_{\beta}C_{\alpha^2} - C_{\alpha}C_{\alpha\beta}}{C_{\alpha\beta}^2 - 4C_{\alpha^2}C_{\beta^2}} \right), \\
\alpha^* &= \frac{\mu\sigma_W^2 M_1}{M_2\sigma_W^2 + (A+B)^2\mu^2(M_2 - M_1^2)} \qquad (3.27) \\
\beta^* &= \frac{(A+B)\mu^2(M_2 - M_1^2)}{M_2\sigma_W^2 + (A+B)^2\mu^2(M_2 - M_1^2)} \qquad (3.28)
\end{aligned}$$

which minimizes $F(\alpha, \beta)$, where the C's are specified previously. To verify this result,

$$\begin{aligned}
\frac{\partial^2 F}{\partial \alpha^2} &= 2C_{\alpha^2} = 2M_2 \geq 0 \qquad (3.29) \\
\frac{\partial^2 F}{\partial \beta^2} &= 2C_{\beta^2} = 2[\sigma_W^2 + (A+B)^2\mu^2] \geq 0 \qquad (3.30) \\
\frac{\partial^2 F}{\partial \alpha \partial \beta} &= C_{\alpha\beta} = 2(A+B)\mu M_1
\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 F}{\partial \alpha^2}\frac{\partial^2 F}{\partial \beta^2}-\left(\frac{\partial^2 F}{\partial \alpha \partial \beta}\right)^2 &= 4C_{\alpha^2}C_{\beta^2}-C_{\alpha\beta}^2 \\ &= 4M_2[\sigma_W^2+(A+B)^2\mu^2]-4(A+B)^2\mu^2 M_1^2 \\ &= 4M_2\sigma_W^2+4(A+B)^2\mu^2(M_2-M_1^2)\geq 0. \quad (3.31)\end{aligned}$$

The above nonequality is essentially because

$$\begin{aligned}0 &\leq E[(Z+b-M_1)^2 \mid Z>0] \\ &= M_2-2M_1E[Z+b \mid Z>0]+M_1^2 \\ &= M_2-M_1^2\end{aligned}$$

Thus, from (3.29), (3.30) and (3.31),

$$J=\begin{pmatrix} \frac{\partial^2 F}{\partial \alpha^2} & \frac{\partial^2 F}{\partial \alpha \partial \beta} \\ \frac{\partial^2 F}{\partial \alpha \partial \beta} & \frac{\partial^2 F}{\partial \beta^2} \end{pmatrix}$$

is positively definite, which means $F(\alpha,\beta)$ achieves its minimum at (α^*,β^*). The "best" estimator under condition $X>aY+b$ is:

$$U^*=\alpha^*(Z+b)+\beta^* W \quad (3.32)$$

Now we encounter the same problem as in Section 3.1. Note that (α^*,β^*) are all dependent on μ, the unknown parameter we are trying to estimate. Again, we try to use unbiased linear combination of X and Y, or equivalently linear combination of $(Z+b)$ and W, to approximate μ in α^* and β^*, i.e. we want:

$$E[\lambda_1(Z+b)+\lambda_2 W \mid Z>0]=\mu \quad (3.33)$$

where

$$\begin{aligned} LHS &= \lambda_1 \left[\frac{\sigma_Z e^{-[(1-a)\mu-b]^2/2\sigma_Z^2}}{\sqrt{2\pi}P(Z>0)} + (1-a)\mu \right] + \lambda_2(A+B)\mu \\ &= \lambda_1 \frac{\sigma^2}{\int_0^{+\infty} e^{-\frac{z^2-2z[(1-a)\mu-b]}{2\sigma^2}} dz} + \lambda_1(1-a)\mu + \lambda_2(A+B)\mu \end{aligned} \quad (3.34)$$

Thus, (3.34) is essentially:

$$\lambda_1 \frac{\sigma^2}{\int_0^{+\infty} e^{-\frac{z^2-2z[(1-a)\mu-b]}{2\sigma^2}} dz} = \mu - \lambda_1(1-a)\mu - \lambda_2(A+B)\mu \quad (3.35)$$

Note that the LHS of equation (3.35) is a non-linear function of μ, while the RHS is linear on μ. In order to keep this equaion valid for all μ, both sides of the equation have to be constant, thus zero. Hence,

$$\lambda_1 = 0, \text{ and } \lambda_2 = \frac{1}{A+B} \quad (3.36)$$

As a result, we approximate μ in (3.27) and (3.28) by $\frac{A}{A+B}X + \frac{B}{A+B}Y$.[6] Noting that M_1 and M_2 are also dependent on μ, we should also approximate μ inside M_1 and M_2 by $\frac{A}{A+B}X + \frac{B}{A+B}Y$. Thus, the new estimator is approximated by $U = \alpha^*_{appr}(Z+b) + \beta^*_{appr}W$.

Similarly, for the opposite condition $X \leq aY + b$ ($Z \leq 0$), the "best" estimator

[6] Though the value of A and B are not unique(we only require them to satisfy: $A(\sigma_X^2 - a\sigma_{XY}) = B(a\sigma_Y^2 - \sigma_{XY})$), the ratio of A or B to $A+B$ is constant, and thus the variance of $\frac{A}{A+B}X + \frac{B}{A+B}Y$ is constant.

under mean square error criteron is $u^* = \gamma^*(Z+b) + \delta^* W$, where:

$$\begin{aligned}
\gamma^* &= \frac{\mu\sigma_W^2 m_1}{m_2\sigma_W^2 + (A+B)^2\mu^2(m_2 - m_1^2)} \\
\delta^* &= \frac{(A+B)\mu^2(m_2 - m_1^2)}{m_2\sigma_W^2 + (A+B)^2\mu^2(m_2 - m_1^2)} \\
m_1 &= E[Z+b \mid Z \le 0] \\
m_2 &= E[(Z+b)^2 \mid Z \le 0]
\end{aligned}$$

and u^* is approximated by $u = \gamma^*_{appr}(Z+b) + \delta^*_{appr} W$ by replacing μ in γ^* and δ^* by $\frac{A}{A+B}X + \frac{B}{A+B}Y$.

More generally, denote $\mathcal{A}$ as an event concerning only Z(e.g. $Z \le 1$). Then under the condition $\mathcal{A}$, the best estimator for the strength of a horse is: $U^{\mathcal{A}} = \alpha^{\mathcal{A}}(Z+b) + \beta^{\mathcal{A}} W$, where

$$\begin{aligned}
\alpha^{\mathcal{A}} &= \frac{\mu\sigma_W^2 M_1^{\mathcal{A}}}{M_2^{\mathcal{A}}\sigma_W^2 + (A+B)^2\mu^2(M_2^{\mathcal{A}} - (M_1^{\mathcal{A}})^2)} \\
\beta^{\mathcal{A}} &= \frac{(A+B)\mu^2(M_2^{\mathcal{A}} - (M_1^{\mathcal{A}})^2)}{M_2^{\mathcal{A}}\sigma_W^2 + (A+B)^2\mu^2(M_2^{\mathcal{A}} - (M_1^{\mathcal{A}})^2)}
\end{aligned} \tag{3.37}$$

where,

$$\begin{aligned}
M_1^{\mathcal{A}} &= E[(Z+b) \mid \mathcal{A}] \\
M_2^{\mathcal{A}} &= E[(Z+b)^2 \mid \mathcal{A}]
\end{aligned} \tag{3.38}$$

The proof is similar to previous disccusion.

Chapter 4

Prediction and Testing

In chapter 3, by minimizing conditional mean square error, we have derived a new estimator $U' = \alpha^*(Z+b) + \beta^* W$ for the strength of a particular horse under a specific condition $X > aY + b$ (and $u' = \gamma^*(Z+b) + \delta^* W$ under $X < aY + b$). Since unknown μ is included in both α^* and β^*, we approximate μ inside U' by $\frac{A}{A+B}X + \frac{B}{A+B}$ and then obtain the approximated estimator, denoted as U (and u for the case $X < aY + b$). In this chapter we are going to test the prediction accuracy of the estimators we have obtained in chapter 3.

4.1 Prediction of Win Probability

First, we try to fit the model parameters. As mentioned in chapter 3, the probit regression model can be summerized as:

$$\begin{aligned} S_i &= \beta_0 + \beta_1 F_{i1} + ... + \beta_p F_{ip} + \epsilon_i, \quad \epsilon_i \sim N(0,1) \\ &= s_i + \epsilon_i \end{aligned} \tag{4.1}$$

The resulting win probability of horse i can be calculted by:

$$P_i = P(s_i + \epsilon_i > s_k + \epsilon_k, k \neq i) \tag{4.2}$$

where ϵ_i and ϵ_k are independent. Denoting $v_i = s_i + \epsilon_i$, the equation above can be written as:

$$P_i = \int_{-\infty}^{\infty} \prod_{k \neq i} \Phi(v_i - s_k)\phi(v_i - s_i)dv_i \tag{4.3}$$

$\Phi()$ is the cumulative distribution function of ϵ_i and $\phi()$ is the probability density function of ϵ_i.

Further, if we make some rearrangement to our data, we can set the horse 1 to be the one who wins the first, horse 2 to be the one who wins the second and so on. Then the tierce probability of hrose 1, 2 and 3 (the probability of horse 1 winning exactly first, horse 2 winning exactly second and horse 3 winning exactly the third place)is calculated as:

$$\begin{aligned} P_{1,2,3} &= \iiint\limits_{v_1 > v_2 > v_3 > v_j, j \neq 1,2,3} \prod_{i=1}^{N} [\phi(v_i - s_i)] dv_i \\ &= \iiint\limits_{v_1 > v_2 > v_3 > v_j} \prod_{i=4}^{N} [\Phi(v_3 - s_i)] \prod_{i=1}^{3} [\phi(v_i - s_i) dv_i] \end{aligned} \tag{4.4}$$

The tierce probability shown above can be considered as a likelihood function. By maximizing this likelihood function, we can fit the β's, and then the strengths of horses can be predicted once current factor information is known:

$$\tilde{s}_i = \hat{\beta}_0 + g_{i1}\hat{\beta}_{i1} + g_{i2}\hat{\beta}_{i2} + ... + g_{ip}\hat{\beta}_{ip} \tag{4.5}$$

Here, the data used for fitting this model is from season 2001-2004 while data from season 2005-2006 is used for prediction testing. For distinguishing, we denote $G(g_{ij})$ for the factors data of season 2005-2006 and $F(f_{ij})$ for 2001-2004. Since the model estimated strength for each horse is obtained, the resulting predicted probability of each horse, particularly horse i, to win is:

$$\tilde{P}_i = Pr(\tilde{s}_i + \epsilon_i > \tilde{s}_j + \epsilon_j,\ j = 1, 2, ..., l, j \neq i) \tag{4.6}$$

Table 4.1: Accuracy of Regression Model

Prob. Range	No. of Horses	Exp. No. Win	Act. No. Win	Std. Diff.
0-0.02	4367	43.11	54	1.6592
0.02-0.03	1653	41.04	45	0.6180
0.03-0.04	1426	49.80	48	-0.2556
0.04-0.05	1135	50.75	51	0.0351
0.05-0.06	1061	58.24	63	0.6236
0.06-0.09	2474	183.33	160	-1.7229
0.09-0.12	1736	181.16	174	-0.5323
0.12-0.15	1194	160.28	151	-0.7326
0.15-0.18	838	137.49	128	-0.8094
0.18-0.23	872	176.77	185	0.6188
0.23-0.29	589	150.92	157	0.4950
0.29-0.35	262	82.96	91	0.8829
0.35-1	207	84.05	101	1.8493
			Chi-Sq Stat.	12.8612

Based on model predicted probabilities, the accuracy table(table 4.1) can be constructed. From the last column, the absolute values of standard differences (Z-values) are all less than 2, which shows that the difference is not significant. The Chi-square statistics is only 12.86. While the public estimated win probabilities implied by the final odds has a Chi-square statistics of 11.78(accuracy table is omitted here), which is extremely accurate. The model estimate seems to perform nearly as good as the public estimate. However, there is problem hiding behind the table.

Before we look into the problem, how to find model estimated strength should be introduced. As mentioned previously, the public implied probability of horse i to win can be calculated as follows, since *odds* is available to download from Jockey Club website:

$$\begin{aligned} P_i^{pub} &= \frac{B_i}{B_1 + ... + B_N} \\ &= \frac{1/Odd_i}{1/Odd_1 + ... + 1/Odd_N} \\ &= \frac{1-\rho}{Odd_i} \end{aligned} \tag{4.7}$$

Please carefully note that since final odds is not available when people really bet in practice. Thus even though the implied win probabilities are not so accurate as implied by final odds, the odds obtained two minutes before each race starts is used. From now on, when we mention *odds*, we refer to the odds obtained two minutes before each race starts.

On the other hand, based on the strength argument, the public implied probability of horse i to win should also be calculated in terms of public estimated strength Y_i by:

$$P_i^{pub} = P(Y_i + \epsilon_i > Y_k + \epsilon_k, k \neq i), \quad \epsilon_i, \epsilon_k \sim iid\ N(0,1) \tag{4.8}$$

Inversely solve the equation above for Y_i's,[1] we can get the public estimated strength.

Now, we classify the result into two subsets: we pick out the results where model

[1] Minggao Gu, Yueqin WU. (2007). Rank Based Marginal Likelihood Estimation for General Transformation Models.

Table 4.2: Accuracy of Regression Model including Public Factors under $X_i > Y_i$

Prob. Range	No. of Horses	Exp. No. Win	Act. No. Win	Std. Diff.
0-0.02	1699	14.72	22	1.8959
0.02-0.03	603	14.94	16	0.2731
0.03-0.04	568	19.94	19	-0.2101
0.04-0.05	482	21.50	13	-1.8337
0.05-0.06	517	28.47	30	0.2866
0.06-0.09	1349	100.45	82	-1.8406
0.09-0.12	1053	110.03	94	-1.5285
0.12-0.15	763	102.53	87	-1.5334
0.15-0.18	559	91.86	77	-1.5503
0.18-0.23	593	120.47	110	-0.9537
0.23-0.29	413	106.15	108	0.1796
0.29-0.35	183	58.00	59	0.1320
0.35-1	147	59.58	71	1.4798
			Chi-Sq Stat.	20.7860

Table 4.3: Accuracy of Regression Model including Public Factors under $X_i < Y_i$

Prob. Range	No. of Horses	Exp. No. Win	Act. No. Win	Std. Diff.
0-0.02	2668	28.38	32	0.6791
0.02-0.03	1050	26.10	29	0.5683
0.03-0.04	858	29.87	29	-0.1584
0.04-0.05	653	29.25	38	1.6186
0.05-0.06	544	29.77	33	0.5920
0.06-0.09	1125	82.88	78	-0.5361
0.09-0.12	683	71.13	80	1.0516
0.12-0.15	431	57.75	64	0.8227
0.15-0.18	279	45.63	51	0.7946
0.18-0.23	279	56.30	75	2.4915
0.23-0.29	176	44.77	49	0.6322
0.29-0.35	79	24.96	32	1.4083
0.35-1	60	24.47	30	1.1183
			Chi-Sq Stat.	16.3221

estimated strength($\tilde{s}_i$ or X_i) is larger than the public estimated strength(i.e. $X_i > aY_i + b$, when $a = 1, b = 0$) and put them into Table 4.2 and the rest (when $X_i < Y_i$) into Table 4.3. Outstanding bias is revealed. Table 4.3 shows when model estimate is smaller than public estimate, the actual win numbers are generally larger than expected win numbers, which means our model is on average under-estimating. On the other hand, when model estimate is larger than public, the "one-side" bias also exists. Both of the subsets show larger Chi-square statistics than overall.

This is a generally ignored problem. When people construct an estimate which seems to be overall accurate, there may exist oustanding biases in the subsets. Possible case is that bias towards negative side exists in subset 1 while bias towards positive side exists in subset 2, and they cancel out with each other when considered as a whole and thus produce a overall "good" result. However, it's fatal that the payoff does not cancel out each other in two subsets to produce a nice payoff as people expect. Actually, unfortunately sometimes they worse the bad situation and reduce the good sitution. The reason is simple, when over-estiamted, people over-bet and they under-bet when they under-estimate. Over-bet drives people losing more while under-bet wins less. For this very sake, new estimator derived in chapter 3 becomes extremly important.

As model estimate and public estimate can be observed as x_i and y_i previously, the new estimate of horse i's strength is calculated as specified in chapter 3:

$$U_i = \alpha^*_{i,appr}(z_i + b) + \beta^*_{i,appr} w_i, \text{ if } x_i > ay_i + b. \tag{4.9}$$

and

$$u_i = \gamma^*_{i,appr}(z_i + b) + \delta^*_{i,appr} w_i, \text{ if } x_i < ay_i + b. \tag{4.10}$$

Thus the new estimate of horse i's strength is essentially:

$$newstr_i = \begin{cases} U_i, \text{ if } x_i > ay_i + b \\ u_i, \text{ if } x_i < ay_i + b \end{cases} \tag{4.11}$$

The resulting win probability of horse i within one race is given by:

$$P_i = Pr(newstr_i + \epsilon_i > newstr_j + \epsilon_j,\ j \neq i) \tag{4.12}$$

Based on the probability prediction, the overall accuracy is summerized in Table 4.4. Table 4.5 and Table 4.6 are also shown here for situation $X > Y$ and $X < Y$ respectively. Neither of Table 4.5 and Table 4.6 shows obvious single-side bias. Both of each has reduced its own Chi-square statistics and in result cooperates with each other to further enhance the overall accuracy. This result assures us a lot not only because it has a overall smaller Chi-square statistics, but also because the Chi-square statistics of Table 4.5 and 4.6 is quite close to that of Table 4.4, the new estimation performs well in either subset.

Table 4.4: Accuracy of New Estimator

Prob. Range	No. of Horses	Exp. No. Win	Act. No. Win	Std. Diff.
0-0.02	4073	38.76	46	1.1629
0.02-0.03	1520	37.85	39	0.1874
0.03-0.04	1432	50.22	45	-0.7370
0.04-0.05	1436	64.55	58	-0.8150
0.05-0.06	1271	69.85	69	-0.1012
0.06-0.09	2752	201.26	188	-0.9347
0.09-0.12	1560	162.41	171	0.6737
0.12-0.15	1057	142.16	130	-1.0201
0.15-0.18	779	127.66	119	-0.7665
0.18-0.23	843	171.19	177	0.4440
0.23-0.29	564	144.57	155	0.8671
0.29-0.35	284	89.45	93	0.3750
0.35-1	243	99.72	118	1.8310
			Chi-Sq Stat.	10.0032

Table 4.5: Accuracy of New Estimator Under $X_i > Y_i$

Prob. Range	No. of Horses	Exp. No. Win	Act. No. Win	Std. Diff.
0-0.02	470	6.67	4	-1.0353
0.02-0.03	529	13.34	12	-0.3667
0.03-0.04	588	20.69	17	-0.8106
0.04-0.05	736	33.13	31	-0.3694
0.05-0.06	691	37.96	37	-0.1560
0.06-0.09	1675	123.36	120	-0.3028
0.09-0.12	1138	119.08	132	1.1840
0.12-0.15	859	115.71	109	-0.6242
0.15-0.18	666	109.19	104	-0.4970
0.18-0.23	735	149.59	161	0.9332
0.23-0.29	528	135.49	145	0.8167
0.29-0.35	269	84.80	92	0.7816
0.35-1	236	97.14	115	1.8122
			Chi-Sq Stat.	9.5874

Table 4.6: Accuracy of New Estimator Under $X_i < Y_i$

Prob. Range	No. of Horses	Exp. No. Win	Act. No. Win	Std. Diff.
0-0.02	3603	32.09	42	1.7504
0.02-0.03	991	24.51	27	0.5034
0.03-0.04	844	29.54	28	-0.2827
0.04-0.05	700	31.42	27	-0.7888
0.05-0.06	580	31.88	32	0.0204
0.06-0.09	1077	77.90	68	-1.1213
0.09-0.12	422	43.33	39	-0.6585
0.12-0.15	198	26.45	21	-1.0594
0.15-0.18	113	18.47	15	-0.8067
0.18-0.23	108	21.60	16	-1.2057
0.23-0.29	36	9.08	10	0.3053
0.29-0.35	15	4.65	1	-1.6929
0.35-1	7	2.58	3	0.2634
			Chi-Sq Stat.	11.9662

Table 4.7: Accuracy of other Estimators

Estimators	Chi-sq. Stat.	Chi-sq Under $X > Y$	Chi-sq Under $X > Y$
MSE	11.1730	14.8041	15.3225
Unbiased	15.1988	11.2050	19.6168
Public	22.3323	84.8406	53.2622

For comparison, some other estimation results are summerized in Table 4.7. "MSE" is the estimator $\lambda^*_{1,appr}X + \lambda^*_{2,appr}Y$ with minimum mean square error under no particular conditions derived in section 1 of chapter 3. "Uubiased" is the estimator $\lambda^*X + (1 - \lambda^*Y)$ with minimum variance among all linear unbiased estimators derived in section 1 of chapter 3. "Public" is the estimator derived from public odds two minutes before the race begeins.

Even though public estimate is not so accurate as final odds implied estimate performs. The information provided by the public estimate has supported enough to improve our estimation.

Chapter 5

Conclusion

This thesis provides an estimation of horses' win probabilities based on probit regression model in horse racing market, and then provides several methods to improve the estimation by cooperating with public estimation. Even though the "best" new estimator is an approximate result, not a perfect one in a ideal close form, it still performs very well and better than the fundamental regression prediction. What's more, we find out an extemely important problem, which has been generally ignored. That may be why some people lose money in real markets while their models look pretty good theoretically. The situation - mistakes hide themselves by canceling out each other as if they have disappeared, always happen in daily lifes. The result of this situation could bring fatal damage to our investment. Over-investment aggressively is dangerous while under-investment negatively is foolish. By deep investigation, we can detect the problems and solve them. This thesis does not only provide a method to enhance the overall estimation performance, it also provides an estimation which can perform well in each case specified and thus enhance the overall accuracy in the root. The result is also very useful for people who constrain their portfolios in certain conditions. A straightforward example is an investor, whose budget is limited for instance, would only invest in assets whose potential growth he believes is greater than others' prediction.

As far as horse racing is concerned here, our consideration is not simply limited within whether model estimate is larger than public estimate or not. The condition $X > aY + b$ varies with different a's and b's, and some other conditions can be set. Variety of requirements can be met. Our condition shown in this article is just the tip of an iceberg to remind people that their beliefs may have to be adjusted when particular case happens.

The new estimator provided in this thesis can be further developed if odds closer to final odds can be used practically. Otherwise, odds prediction may be required to reduce or cancel the public estimation quality reduction.

Bibliography

[1] William Benter (1994). Computer Based Horse Race Handicapping and Wagering Systems: A Report. *Efficiency of Racetrack Betting Markets by Academic Press, Inc.* 183-198.

[2] Bolton, Ruch N. and Randall G. Chapman (1986). Searching For Postitive Returns at the Track: A Multinomial Logit Model For Handicapping Horse Races. *Management Science*, 32, 8(August), 60-82.

[3] White, E.M. , Dattero, R. , Flores, B. (1992). Combining vector forecasts to predict thoroughbred horse race outcomes. *International Institute of Forecasters*, Vol. 8, Issue 4, 595-611.

[4] K.Busche (1994). Efficient Market Results in an Asian setting. *Efficiency of Racetrack Betting Markets by Academic Press, Inc.* 615-617.

[5] Minggao Gu, Yueqin WU. (2007). Rank Based Marginal Likelihood Estimation for General Transformation Models.

[6] P.Deuflhard and A. Hohmann. *Numerical Analysis. A First Course in Scientific Computing.* de Gruyter, Berlin, 1995.

[7] Randall G. Chapman (1987) Still Searching for Positive Returns at the Track: Empirical Results from 2000 Hong Kong Races. *Efficiency of Racetrack Betting Markets by Academic Press, Inc.* 173-182.

[8] Donald B. Hausch, Victor S. Y. Lo and William T. Ziemba. Introduction to the Efficiency of Win Markets and the Favorite-Longshot Bias. *Efficiency of Racetrack Betting Markets by Academic Press, Inc.* 251-255.

[9] Donald B. Hausch and William T. Ziemba (1990). Arbitrage Strategies for Cross-Track Betting on Major Horse Races. *The Journal of Business by The University of Chicago Press.* Vol. 63, No. 1, Part 1, 61-78.

[10] Peter Asch, Burton G. Malkiel and Richard E. Quandt (1986). Market Efficiency in Racetrack Betting: Further Evidence and a Correction. *The Journal of Business by The University of Chicago Press.* Vol. 59, No. 1, 157-160.

[11] Lo, V.S.Y., Bacon-Shone, J. and Busche, K. (1995) The Application of Ranking Probability Models to Racetrack Betting. *Management Science* 41, 1048-1059.